Received: January 2, 2024. Revised: March 6, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Optimizing Feature Selection Method in Intrusion Detection System Using Thresholding

Muhammad Arif Faizin[1] **Dias Tri Kurniasari**[1] **Nazhifah Elqolby**[1]
Muhammad Aidiel Rachman Putra[1] **Tohari Ahmad**[1]*

[1]*Department of Informatics Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia*
* Corresponding author's Email: tohari@its.ac.id

Abstract: Information and communication technology is growing rapidly, making it the target of various attacks. The attacks can be in the form of data theft, phishing, and Denial of Service (DoS). There are many ways to handle attacks on communication networks, including developing an Intrusion Detection System (IDS) model. Research on IDS has developed a lot and focuses on certain things such as feature selection, dealing with data imbalance problems. Feature selection is essential to the IDS model because of the dataset's characteristics, which have many features. Besides, the number of features included in the classification can affect the detection performance of the IDS model. This research proposes an IDS combining mutual information with thresholding feature selection and XGBoost classification algorithm. Mutual information is used to measure the dependency between every input feature and the target features. After the amount of information is obtained with mutual information, thresholding is used to decide the best number of features in the classification process. Then, the data are classified using XGBoost selected features. The proposed method was tested using four metrics: accuracy, precision, recall, and f1-score. This study used UNSW-NB15 as the primary dataset to analyze the best combinations of feature selection method and thresholding value. In addition, the proposed method has also been tested using NSL-KDD and CIC-IDS2017 datasets to evaluate the performance compared with previous research. The proposed method performs best using the CIC-IDS2017 dataset with 99.89 % accuracy and 99.68 % F1 score. Furthermore, it can reduce computational training time compared with other IDS methods that only use feature selection or tree-model-based algorithms without thresholds.

Keywords: Feature selection, Information security, Intrusion detection system, National security, Thresholding.

1. Introduction

The growth of the information and communication technology is considered massive. However, various threats to user security, information and communication infrastructure also grow [1], which include data theft, fraud, ransomware, and things that threaten communication and information infrastructure, such as Denial of Service (DoS). Thus, dealing with information and communication security threats requires appropriate handling mechanisms. An Intrusion Detection System (IDS) is one way to handle network security issues by monitoring and finding suspicious activity in network traffic [2]. Users and network administrators commonly use IDS to predict incidents that may occur, analyze logs, and identify attempted attacks [3], as well as classify them.

Common techniques can be used to develop IDS: signature-based and anomaly-based [4]. Each IDS classification method has its advantages and disadvantages, including anomaly-based. Signatured-based detects intrusions based on predefined patterns, whereas anomaly-based detects intrusions based on current user activity [5]. Anomaly-based intrusion has a broader scope of detecting new attacks for better accuracy [6].

One of the challenges of the IDS model is the large number of features that need to be analyzed [7]. On the other hand, the number of features included in the classification process can affect the model performance [8]. Besides, not all features significantly contribute to detecting the attack [9]. Therefore, some studies use various feature selection methods to obtain the most significant features in detecting attacks.

Feature selection is carried out by measuring the correlation and influence of each feature on detection performance [10]. Previous research used various feature selection methods in the IDS detection model: mutual information [11], chi-square test [12], ANOVA-f test [13], and variance influence factor [14]. Besides, this study used mutual information to quantify the information obtained about all input features through the label. Then, threshold analysis determines the number of features allowed for the classification process. This research used XGBoost for classification because this algorithm performs better than other tree-model-based algorithms [15].

This paper is structured as follows. Section 2 provides related work about feature selection in IDS. The methodology of the research is described in Section 3. Section 4 shows and explains the result and discussion. Finally, conclusions are explained in Section 5.

2. Related work

Intrusion detection is a popular method that used by many network security professionals to address network intrusion issues [16–18]. Many algorithms and methods have been developed to overcome this issue, such as using anomaly-based detection. One of the research projects is done by Liu et al. [19] that implements machine learning methods to detect anomalies in IOT network intrusion detection systems (IDS). It results achieving high accuracy on detection and raising the effectiveness. Anomaly-based intrusion has a broader scope for detecting new attacks. Thus, anomaly-based gets better accuracy than signature-based [6].

In the IDS field, network datasets show common challenges due to the complexity and varied features. Feature selection became popular for decreasing irrelevant features [20, 21]. Some feature selection methods include mutual information, chi-square test, ANOVA-f test, and variance influence factor [22, 23]. Mutual information is one of the feature selection methods that applies a tree-model-based algorithm. Sulaiman and Labadin [24] used a typical greedy feedforward feature selection method using mutual information. Mutual information can select features that retain relevant information, improving prediction accuracy and reducing training time. However, the experiment only tested the proposed method with an ANN classifier.

There are various efficient and high-performance methods available to classify intrusion. One of these methods is using tree-model-based algorithms, such as XGBoost, Decision Tree, and Random Forest. These classifiers are known for their ability to provide good explanations of the overall model structure using high-quality local explanations, regardless of the data domain [25]. Some studies show that XGBoost is better than others tree-model-based due to its ensemble classifier. Putrada et al. [26] proposed and successfully implemented XGBoost for IDS on the WSN cyberattack dataset with imbalanced data. Deepak et al. [27] had the best-generalized model to detect real-time cyber-attacks using XGBoost. Out of all algorithms, XGBoost gave high accuracy on the validation set. Sood et al. [28] show that XGBoost can produce faster comparison, provide visualization of results, and provide excellent accuracy and precision. However, the study requires further testing on the dataset to improve the efficiency of the security measure.

Other research uses optimization for feature selection methods to get the optimal limit in determining important features, such as thresholding [29]. Megantara and Ahmad [30] successfully improved the classification model using a hybrid machine-learning approach that combines feature selection with thresholding. The method removes features with zero values to distinguish between high and low importance features and divides the remaining by the median data. Sulaiman and Labadin [24] use feature selection by choosing thresholds based on statistical criteria with percentile and average. The experiment result indicates that feature selection can reduce overfitting and decrease training time against the full set of features. The research requires further exploration to expand the proposed method.

Therefore, this research proposed an IDS using mutual information to measure the correlation and influence of each feature on detection performance. Besides, thresholding feature selection was conducted to optimize the feature selection method. XGBoost is used as a classification model to accompany the best feature with an efficient method.

3. Methodology

Many experiments and developments have been carried out in the intrusion detection system on various datasets. However, each dataset has different characteristics based on the existing features. The

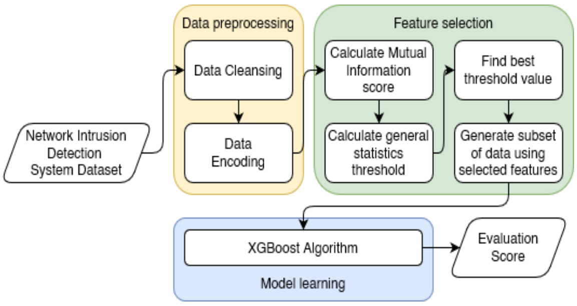
Data preprocessing
Feature selection
Network Intrusion Detection System Dataset
Data Cleansing
Data Encoding
Calculate Mutual Information score
Calculate general statistics threshold
Find best threshold value
Generate subset of data using selected features
XGBoost Algorithm
Model learning
Evaluation Score

Figure. 1 Proposed method

number of existing features can affect the results of the intrusion detection system model, such as decreasing accuracy. One way to prevent this problem is to only employ certain features using feature selection method. There are some feature selection methods can be used, while this study aims to get a high accuracy value with subset of features by implementing feature selection methods with thresholding analysis. The overall proposed method is provided in Fig 1.

Based on Fig 1, the network intrusion detection system dataset is being preprocessed, which includes data cleansing and encoding. Furthermore, feature selection will be carried out on the processed dataset. This research proposed a method using optimization of mutual information feature selection using thresholding with XGBoost classification.

For comparison, this research also implements other feature selection: chi-square test, ANOVA-f test, and variance inflation factor. Also, the machine learning classification will be compared with several methods, namely Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Artificial Neural Network (ANN), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM). Finally, these results can be evaluated and compared based on performance and efficiency.

3.1 Data preprocessing

This phase focused on processing data before entering the training stage. The preprocessing phase consists of two sub-phases: data cleansing and data encoding. Data cleansing is carried out to prevent unused data from participating in detection. Data cleansing consists of several processes, which are removing unused features, removing duplicate features and data, removing one-value column, filling the NaN data, and remove any infinite data. Thus, data encoding is also needed to convert categorical data into numerical data. This study uses a label encoder in this preprocessing stage. It was chosen because it did not change the dimension of the data. Consequently, the result of the encoding does not imply a relationship or degree in each category.

3.2 Feature selection

The IDS dataset has many features that can affect accuracy in the IDS model. Therefore, a selection feature is required. Feature selection aims to determine the most relevant and informative features to improve the model's performance and decrease the model’s complexity without eliminating important features. To compare with mutual information, several feature selection methods are used in this study: chi-square test, ANOVA-f test, and variance inflation factor. Some methods use statistical definitions from the data, which can be explained as follows.

Definition 1. (Contingency table) Contingency table or also known as cross-tabulation table or two-way table, is a table that represents the frequencies between categorical features. For the two random categorical variables X and Y, each cell in the table represents the count of the frequency of specific categories $x \in X$ for $y \in Y$. This table can be used for two or more categorical features by using the combination between the features.

Definition 2. (Groups Variability) To measure differences between the means of two or more groups, there are two kind of groups variability in the data, between groups and within groups. Between group variability focuses on differences in means between the groups, when within group variability focuses on variations within each group. For N observations, there are k groups in categorical features, and n_i observations in i-th group. Each of the group have $\overline{X}_i$ mean and combined into grand mean $\overline{X}_{grand}$.

Mutual information. Mutual information or commonly called information gain is a concept that measures the amount of the information gained from the target variable (usually categorical) by knowing the value of a potential feature (usually numerical) [31]. High mutual information indicates that the feature is good to differentiate between classes, making more valuable information in feature. Mutual information score is defined in Eq (1).

$$I(X,Y) = \sum_{x=1}^{X} \sum_{y=1}^{Y} p(x,y) \cdot \log\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right) \quad (1)$$

From the contingency table, the mutual information between $x \in X$ and $y \in Y$ variables are calculated using $p(x)$ and $p(y)$ as the probability mass or density function of X and Y, and $p(x, y)$ is the joint probability mass of X and Y.

Chi-square test. Chi-square is a method used to test the dependency between the data observed and the targeted class. This process determines whether a relationship exists between existing features and the observed data [12], while the higher chi-square score and lower degree of freedom value indicates that the observed data and the existing feature are dependent, shows that it has more important features. The chi-square test is good for finding dependency between categorical data due to its non-parametric nature characteristics [32]. The chi-square score χ^2 is defined in Eq (2).

$$\chi^2 = \sum_{x=1}^{X}\sum_{y=1}^{Y} \frac{\left(O_{xy} - E_{xy}\right)^2}{E_{xy}} \tag{2}$$

From the contingency table, the chi-square score calculated for each category $x \in X$ and $y \in Y$. It calculated the difference squares between O_{xy} observed frequency and E_{xy} expected frequency. In addition, $E_{xy} = p(x) \times p(y)$ which calculate using probability of independence between the features. The result score called independent if the score is located on rejected hypothesis range adjusted by the degree of freedom df formula in Eq (3).

$$df = (X - 1) \times (Y - 1) \tag{3}$$

ANOVA-f test. Analysis of variance, or ANOVA, is a method used to determine any statistically significant differences among the means of two or more groups into the target variable [33]. The high value of F-statistic and low p-value suggests that there are significant differences of the groups, suggesting that the features are relevant to the targeted variable [34]. This score is calculated with Eq (4-6).

$$F = \frac{MSB}{MSW} \tag{4}$$

$$MSB = \frac{\sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{X}_{grand}\right)^2}{k - 1} \tag{5}$$

$$MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2}{N - k} \tag{6}$$

The F-statistic F is described as division between MSB mean square between groups variability and MSW mean square within groups variability. MSB is defined as multiplication between number of observations in i-th group and the difference squares between the mean $\bar{X}_i$ and combined mean $\bar{X}_{grand}$. Therefore, MSW is defined as summation between difference squares of each X_{ij} data of the j-th observation in the i-th group.

The p-value is a value of evidence that the probability of obtaining test results is rejected the null hypothesis or not. The threshold of the p-value or called as significance level α on implementation level usually set to 0.005.

Variance inflation factor. Variance Inflation Factor (VIF) is a method used to measure and evaluate the level of multicollinearity between the independent variables with other independent variables in a regression [35]. The higher the VIF value, the higher the multicollinearity or correlated, indicate that it may not provide unique information of the model and needs to be removed. The VIF score can be calculated using Eq (7).

$$VIF_i = \frac{1}{1 - R^2{}_i} \tag{7}$$

The variance inflation factor VIF of X_i feature is calculated using coefficient of determination R_i^2 that obtained after regressing X_i against another features.

The sklearn machine learning library provides the implementation of chi-square test, ANOVA-f test, and mutual information using chi2(), f_classif(), and mutual_info_classif() function respectively. The implementation can be used in feature selection strategy, such as selecting k best features using SelectKBest class. Then the variance inflation factor is implemented by statsmodels library using variance_inflation_factor class.

In this study, thresholding is implemented in each method to get the most efficient results from each selected feature. Thresholding aims to obtain an optimal boundary in determining which features are used and the most important. With the optimal value limit, the worst value can be eliminated.

3.3 Model learning

Classification is an important part of IDS implementation. The primary purpose of classification in IDS is to identify whether an

Table 1. UNSW-NB15 Dataset Summary

Class	Training Set	Testing Set
Attack	119341	45332
Normal	56000	37000
Total	175341	82332

Received: January 2, 2024. Revised: March 6, 2024.

Table 2. UNSW-NB15 list of features

Feature Type	Features	Count
Integer	dpkts, spkts, dbytes, sbytes, sttl, dttl, sloss, dloss, swin, stpcb, dtpcb, dwin, smean, dmean, trans_depth, response_body_len, ct_srv_src, ct_state_rttl, ct_dst_ltm, ct_src_dport_ltm, ct_dst_sport_ltm, ct_dst_src_ltm, ct_ftp_cmd, ct_ftw_http_mthd, ct_src_ltm, ct_srv_dst	26
Float	dur, rate, sload, dload, sinpkt, dinpkt, sjit, djit, tcprtt, synack, ackdat	11
Categorical	proto, service, state	3
Binary	is_ftp_login, is_sm_ips-ports	2
Total		42

activity or network is running normally or exposed to an attack. This study proposed XGBoost classification method to evaluate the model. XGBoost is one of popular machine learning algorithm which efficient for classification and regression [36, 37]. XGBoost is a model that is built incrementally through a combination of weak learners, such as shallow decision trees. In this method, each decision obtained is described in a tree [38], where each tree always tries to correct the error obtained in the previous prediction model.

For validating the proposed method, this study compares the performance with several machine learning and deep learning methods. The machine learning method used are Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, k-Nearest Neighbours (k-NN), and Support Vector Machine (SVM). Meanwhile, the deep learning method used are Artificial Neural Network (ANN).

4. Result and discussion

This chapter will focus on the proposed method for feature selection and provide a comprehensive analysis of the results obtained from this method.

4.1 Dataset

The experiment is conducted on Google Collab using the Python programming language. This study uses general libraries for data processing, such as pandas, numpy, and matplotlib. Then, the experiment used the sklearn library for the classification and regression. This research uses three different datasets: UNSW-NB15, NSL-KDD and CIC-IDS2017. The UNSW-NB15 is used for analyzing the proposed feature selection and thresholding method. On the other hand, NSL-KDD, and CIC-IDS2017 are used for evaluating the proposed method’s performance compared with previous research.

The UNSW-NB15 dataset was published in 2015 by Moustofa and Slay [39]. This dataset has a high-dimensional feature of 42, with 3 non-numerical features (categorical) and 39 numerical features. This dataset consists of 257,673 rows of data, split into 175,341 training data and 82,332 testing data. The distribution of training and testing data with its classes are shown in Table 1. The NSL-KDD dataset has 42 features with 3 categorical features and 39 numerical features. The NSL-KDD dataset consists of 125,973 data for the training process and 22,544 data used for testing. Besides, the CIC-IDS2017 dataset has a high dimensional feature of 78 numerical features consisting of 2,830,743 rows of data. In this research, the CIC-IDS2017 dataset was split into 70%:30% as training and testing data.

The features in the UNSW-NB15 dataset are categorized into several types, including general features, content features, period features, additional features (connection features and status features), and labeled features arranged by categorical labels. The category consists of 9 types of attack: Reconnaissance, DoS, Analysis, Fuzzer, Backdoors, Exploits, Shellcode, Generic, and Worms and binary label features that indicate whether the data is normal (0) or an attack (1). Table 2 contains a list of features for UNSW-NB15 with their feature types.

4.2 Feature selection

The UNSW-NB15 dataset contains large number of features. Some features may be irrelevant, redundant, or noisy, which can negatively impact the performance of the models. Feature selection aims to improve the model's efficiency, interpretability, and generalization by focusing on more relevant features. The objective of this method can be achieved by determining the thresholding of the feature score to eliminate the unneeded feature.

To get the best thresholding analysis, this research uses four methods of feature selection: mutual information, chi-square test, ANOVA-f test, and variance inflation factor. Each method results in different selected features, which means that it depends on their unique score value, and no method has the same selected features. The statistical details of the resulted were shown in Table 3.

Received: January 2, 2024. Revised: March 6, 2024.

Table 3. Statistical summary of feature selection method score of UNSW-NB15 dataset

Index	Chi-square test	ANOVA-f test	Mutual information	Variance inflation factor
mean	1082784644026.63	13867.00	0.18	1984046497112.34
std	4574660562293.67	28887.63	0.13	8944976807159.86
min	22.77	0.07	0.00	1.12
Q1	5980.15	122.85	0.07	2.96
Q2	209770.98	2089.60	0.16	20.18
Q3	9497717.00	16425.29	0.29	171.06
max	21606970517356.50	161780.53	0.47	41507830667009.10

Table 4. Comparation of selected features with previous research of UNSW-NB15 dataset

No	Source	Method	Selected features	Total
1	Proposed	Chi-square test	stcpb, dtcpb, sload, dload, rate, dbytes, sinpkt, sbytes, response_body_len, djit, dmean, sttl, sjit, swin, dwin, dpkts, dinpkt, dloss, spkts, dttl	20
		ANOVA-f test	sttl, ct_state_ttl, state, dload, ct_dst_sport_ltm, dmean, rate, swin, dwin, ct_src_dport_ltm, ct_dst_src_ltm, stcpb, dtcpb, ct_src_ltm, ct_dst_ltm, ct_srv_src, ct_srv_dst, is_sm_ips_ports, sload, sinpkt	20
		Mutual information	sbytes, sttl, dbytes, ct_state_ttl, dttl, rate, sload, dur, smean, dmean, dinpkt, dload, dpkts, sinpkt, tcprtt, synack, ackdat, sjit, state, djit	20
		Variance inflation factor	trans_depth, ct_flw_http_mthd, response_body_len, djit, dur, dinpkt, smean, dload, sjit, sload, service, rate, dmean, stcpb, dtcpb, ct_src_ltm, sinpkt, is_sm_ips_ports, ct_dst_sport_ltm, proto	20
2	Kasongo and Sun [40]	XGBoost	sttl, ct_srv_dst, sbytes, smean, proto, ct_state_ttl, sloss, synack, ct_dst_src_ltm, dmean, ct_srv_src, service, ct_dst_sport_ltm, dbytes, dloss, state, tcprtt, ct_src_dport_ltm, rate	19
3	Nururrahmah and Ahmad [22]	CHI2CV	state, sbytes, dbytes, dttl, service, dtcpb, response_body_len, sinpkt, synack, ct_flw_http_mthd, is_ftp_login, ct_ftp_cmd, ct_srv_dst, ct_dst_src_ltm	14

4.2.1. Feature score

The sample of the top 20 selected features of each method and its comparison with previous research is shown in Table 4. Based on the results, the chi-square of selected feature more focused on network metrics (such as sload and dload), packet characteristics (such as spkts and dpkts), and timing features. Meanwhile, the ANOVA-f test result encompasses a wide range of network and connection-related attributes, including state information, packet counts, and source/destination ports. In mutual information, the features are more focused on data size (such as sbytes and dbytes), duration, and various TCP-related features. The variance inflation factors result in covering HTTP-related features, duration, and various packet and connection features.

Notably, several features consistently have good score across methods, such as rate, state, sttl, sload, dload, sbytes, dbytes, sinpkt, dmean, dtcpb. Those features are dominantly belonged to content features (state, sbytes, dbytes, sttl, sload, dload), time features (dtpcb, dmean), and connection features (sinpkt). The convergence of these feature categories indicates their importance in intrusion detection.

4.2.2. Threshold analysis on feature selection

From the Table 3, the chi-square and variance inflation factor score are scattered in a wide range, indicated by the large mean and standard deviation.

Those methods are more sensitive to data changes than the ANOVA-f test and mutual information method.

The chi-square test score is scattered in a wide range due to its sensitivity to the sample size and the nature of categorical data. The big sample of UNSW-NB15 data makes the score more significant even in insignificant relationship of feature. The other factor is there are more categorical features in this dataset, there are three category features (proto, state, and service) that can lead to high number of possible combinations and high range of score.

The variance inflation factors score is scattered in wide range due to its severity of the multicollinearity. The strong correlation between the independent variables in regression model, leading to inflated variances of regression coefficients, means that the UNSW-NB15 dataset was complex. The number of features may affect the complexity of the model to classify the data.

On the other hand, the ANOVA-f test score is more distributed, because the variability between the group means relative to the variability within groups. Because for this research the binary regression was conducted, and only attack and normal groups are tested.

The mutual information score result is evenly distributed, due to its limitation of equation which only limits between 0 and 1. This is affected by the reduction of uncertainty of the feature, makes it lower variability score among the methods.

In the other side, the data distribution across the methods are poorly distributed. The range of each quartile (Q1-25%, Q2-50%, and Q3-75%) across methods have big differences. The chi-square test and variance inflation factor mean are closer to the maximum value, indicating that the average value is affected by extreme outliers in the data. This suggests that the distribution of the data is skewed, with a few very large values influencing the mean. This condition can be happened because of the chi-square test are applied on highly categorical data, and the variance inflation factors value are applied in highly dimensional data.

Then, this research chose to tune the threshold value of the feature selection method using the variance of quartiles (Q1, Q2, and Q3) to analyse and get optimized score. This paper compared the result of each feature selection method against the complete set of features.

4.3 Result analysis

During the evaluation stage, the main objective is to assess the proposed method's performance thoroughly. This stage encompasses several key aspects, such as comparing the chosen feature selection technique with other available options, comparing the proposed machine learning model with other models, and comparing the results obtained with previous research on the same topic.

4.3.1. Performance score

In order to evaluate the proposed method, the evaluation results are compared by analysing the score of accuracy, precision, recall, F1 score, and computation of the training time, as provided in Eq (8-11), respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1 Score} = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

The accuracy, precision, recall, and F1 score are calculated using the help of the confusion matrix, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) refers to an attack activity that is correctly predicted as an attack, True Negative (TN) refers to a normal activity that is correctly predicted as normal, False Positive (FP) refers to a normal activity that is wrongly predicted as an attack, and False Negative (FN) refers to an attack activity that is wrongly predicted as a normal. Computation training time is calculated between the initial process with training process.

4.3.2. Comparative analysis with other feature selection

This scenario is to compare the mutual information feature selection with other feature selection methods, such as: chi-square test, ANOVA-f test, and variance inflation factor. This research is using XGBoost algorithm to create the model and apply thresholding across the method to get optimal analysis.

The result of the performance scores for various feature selection methods that applied with different thresholding criteria are shown in Table 5. The chi-

Table 5. Performance result comparation of feature selection with thresholding of UNSW-NB15 dataset

Feature Selection	Thresholding	Num. of Features	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Time (s)
No	-	42	80.29	99.28	73.90	84.73	6.07
Chi-square test	$\geq Q1$	31	87.13	97.49	82.37	89.29	5.46
	$\geq Q2$	21	87.38	98.48	82.14	89.57	3.40
	$\geq Q3$	11	86.09	97.27	81.19	88.51	2.43
	$\geq mean$	3	75.46	95.63	70.40	81.10	1.63
ANOVA-f test	$\geq Q1$	31	86.48	98.28	81.15	88.90	5.35
	$\geq Q2$	21	85.77	97.66	80.60	88.31	2.40
	$\geq Q3$	11	85.61	98.21	80.14	88.26	1.88
	$\geq mean$	11	85.61	98.21	80.14	88.26	1.69
Mutual information	$\geq Q1$	31	79.78	99.31	73.37	84.39	2.43
	$\geq Q2$	21	87.47	96.48	83.37	89.45	3.23
	$\geq Q3$	11	87.09	97.52	82.31	89.27	3.61
	$\geq mean$	19	87.63	96.35	83.66	89.56	2.11
Variance inflation factor	$\leq Q1$	11	86.18	97.23	81.32	88.56	2.10
	$\leq Q2$	21	78.36	99.70	71.88	83.54	2.72
	$\leq Q3$	31	79.20	99.57	72.73	84.06	3.22
	$\leq mean$	40	80.29	99.40	73.85	84.74	4.03

square test with a threshold based on mean shows lower accuracy and F1 score compared to other methods. The ANOVA-f test results relatively consistent performance across thresholding levels. Besides, the variance inflation factor with thresholds based on Q2 and Q3 accuracy and F1 score are need to be improved. Mutual information with a threshold based on mean outperforms another threshold with the highest accuracy of 87.63%, F1 score of 89,56%, and training time of 2.10 seconds. Although there is a slight difference in the F1 score compared to the chi-square, the proposed method is more efficient regarding the number of features and training time. Almost all feature selection methods outperformed the complete features, improving the accuracy and F1 score and reducing the training time.

The mutual information achieves high accuracy because the method is not affected by data type and is usable for categorical and numerical data, which is suitable for the UNSW-NB15 dataset. It measures the dependency between variables, capturing non-linear relationships and interactions between features [24].

The results indicate that the feature selection methods and thresholding criteria impact the classification performance compared with the full set of features, especially improving accuracy from 80.29% to 87.63%, improving the F1 score from 84.73% to 89.56%, and reducing the training time from 6.07 second into 2.10 second. This combination of feature selection and thresholding is used in the next scenario.

4.3.3. Comparative analysis with other machine learning models

This scenario is to compare the proposed method with other machine learning method performance, such as: Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, ANN, k-NN, and SVM. The comparison result shown in Table 6.

Based on the comparison result, XGBoost achieves the highest accuracy, F1 score, and training time, making it a strong performer across multiple metrics. Decision Tree and Random Forest also show good performance, particularly in precision. Logistic Regression, Naïve Bayes, and ANN still need to be increased, especially in ANN. k-NN performs well in accuracy, precision, recall, and F1 score, but the performance is still below the XGBoost. Despite its high precision value, SVM has low accuracy and recall values and has a long training time.

Table 6. Performance result comparation with other machine learning methods of UNSW-NB15 dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Time (s)
XGBoost	87.63	96.35	83.66	89.56	2.1
Logistic Regression	71.19	89.38	68.18	77.36	2.3
Naïve Bayes	69.94	96.31	65.42	77.91	0.1
Decision Tree	85.83	97.69	80.65	88.36	1.1
Random Forest	83.36	99.44	77.02	86.81	20.0
ANN	61.65	99.83	58.96	74.14	81.2
k-NN	85.87	94.62	82.35	88.06	0.4
SVM	74.15	63.24	86.12	72.93	2166.2

The XGBoost reach high accuracy because the XGBoost model uses the concept of mutual information to calculate feature importance. It measures the mutual information provided by each feature when splitting the data. It matched due to their shared ability to capture non-linear relationships and dependencies in the data, making well-suited for feature selection in complex, high dimensional datasets.

4.3.3. Comparative analysis with previous research

This research tested the proposed method with three different IDS datasets, UNSW-NB15, NSL-KDD, and CIC-IDS2017. Thus, the performance of the proposed method is compared with previous research that used the same dataset. Comparative analysis was carried out with two previous studies using a tree-model-based classification. Tables 7, 8, and 9 compare the performance of the proposed method with previous studies using UNSW-NB15, NSL-KDD, and CIC-IDS2017, respectively.

The proposed method performs better than previous studies in the experiment using the UNSW-NB15 dataset with an 89.56% F1-score. The proposed methods also have more efficient time than previous methods. For the experiment using the NSL-KDD dataset, the proposed methods show a good recall score of 96.73%. Even though it performs optimally

Table 7. Comparation of proposed method with previous research on UNSW-NB15 Dataset

Paper	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Time (s)
Das et al. [41]	Ensemble ML + Ensemble Feature Selection	88.10	80.80	93.50	86.70	6.02
Kasongo and Sun [40]	XGBoost Feature Selection + Decision Tree	90.85	80.33	98.38	88.45	-
Proposed Method	XGBoost + Mutual Information Thresholding	87.63	96.35	83.66	89.56	2.10

Table 8. Comparation of proposed method with previous research on NSL-KDD Dataset

Paper	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Time (s)
Tavallaee et al. [42]	NB Tree	66.16	-	-	-	-
Andalib and Vakili [43]	Random Forest	79.95	-	-	-	16.32
Das et al. [41]	Ensemble ML + Ensemble Feature Selection	88.10	95.90	82.60	88.70	0.23
Proposed Method	XGBoost + Mutual Information Thresholding	80.51	68.06	96.73	79.90	2.60

Table 9. Comparation of proposed method with previous research on CIC-IDS2017 Dataset

Paper	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Time (s)
A. Bansal and S. Kaur [44]	XGBoost	91.36	97.45	82.02	89.06	-
Das et al. [41]	Ensemble ML + Ensemble Feature Selection	99.50	99.50	99.60	99.50	0.29
Proposed Method	XGBoost + Mutual Information Thresholding	99.89	99.75	99.60	99.68	40.75

in the recall score, the proposed method is still optimal for reducing the training time. Last, the proposed methods outperformed the previous method while tested using the CIC-IDS2017 dataset. Almost all of the metrics show better scores compared to previous research, with an accuracy of 99.89%, precision of 99.75%, and F1 score of 99.68%. Thus, the proposed method is suitable for the UNSW-NB15 and CIC-IDS2017 datasets. However, it is not suitable for the NSL-KDD dataset due to the amount of data. The NSL-KDD dataset has less data than UNSW-NB15 and CIC-IDS2017, so the model built by the tree-based algorithm is not complex enough to cover every existing intrusion scenario.

5. Conclusion

The growth of information and communication technology causes various threats. One of them is information and communication security threats. Thus, dealing with information and communication security threats requires appropriate handling mechanisms, such as IDS. However, Various datasets used in IDS implementation have a large number of features, including UNSW-NB15, NSL-KDD, and CIC-IDS2017 datasets. The number of features included in the classification process can affect the model's performance. Thus, feature selection is needed to overcome this challenge.

This research aims to optimize the number of features used and improve the performance of the model in intrusion detection systems. This research proposed an approach to detect intrusion in network traffic with XGBoost and optimize the feature selection process. The feature analysis results using mutual information are used as a basis for feature selection using threshold analysis.

Based on the experiments, the results show that the combination of mutual information, thresholding feature selection, and XGBoost classification successfully improved the methods to 87.63%. The proposed method showed the stability of increasing the IDS performance and relatively efficient computation time compared with previous methods. However, there is still a need for improvement in another score.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology, MAF, DTK, and NE; formal analysis, MAF, DTK, and NE; writing—original draft preparation, DTK; writing—review and editing, DTK and MARP; supervision, funding acquisition, TA.

Acknowledgments

This work was supported by Pendidikan Magister menuju Doktor untuk Sarjana Unggul (PMDSU), The Ministry of Education, Culture, Research, and Technology, The Republic of Indonesia, and the Institut Teknologi Sepuluh Nopember (ITS).

References

[1] M. R. Aziz and A. S. Alfoudi, “Feature Selection of The Anomaly Network Intrusion Detection Based on Restoration Particle Swarm Optimization”, *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 5, pp. 592–600, 2022, doi: 10.22266/ijies2022.1031.51.

[2] M. Mulyanto, J.-S. Leu, M. Faisal, and W. Yunanto, “Weight embedding autoencoder as feature representation learning in an intrusion detection systems”, *Computers and Electrical Engineering*, Vol. 111, pp. 108949, 2023, doi: https://doi.org/10.1016/j.compeleceng.2023.108949.

[3] J. Jabez and B. Muthukumar, “Intrusion Detection System (IDS): Anomaly Detection Using Outlier Detection Approach”, *Procedia Computer Science*, Vol. 48, pp. 338–346, 2015,

doi: https://doi.org/10.1016/j.procs.2015.04.191.

[4] H. Huang, T. Li, Y. Ding, B. Li, and A. Liu, "An artificial immunity based intrusion detection system for unknown cyberattacks", *Appl Soft Comput*, Vol. 148, pp. 110875, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110875.

[5] Ashima, G. Shaheamlung, and K. Rote, "A comprehensive review for test case prioritization in Software Engineering", In: *Proc. of International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 331–336, 2020, doi: 10.1109/ICIEM48762.2020.9160217.

[6] V. Jyothsna and K. M. Prasad, "Anomaly-Based Intrusion Detection System", *Computer and Network Security*, 2019.

[7] C. Khammassi and S. Krichen, "A NSGA2-LR wrapper approach for feature selection in network intrusion detection", *Computer Networks*, Vol. 172, pp. 107183, 2020, doi: https://doi.org/10.1016/j.comnet.2020.107183.

[8] W. L. Al-Yaseen, A. K. Idrees, and F. H. Almasoudy, "Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system," *Pattern Recognit*, Vol. 132, pp. 108912, 2022, doi: https://doi.org/10.1016/j.patcog.2022.108912.

[9] I. Hidayat, M. Z. Ali, and A. Arshad, "Machine Learning-Based Intrusion Detection System: An Experimental Comparison", *Journal of Computational and Cognitive Engineering*, Vol. 2, No. 2, pp. 88–97, 2022, doi: 10.47852/bonviewJCCE2202270.

[10] H. Mamdouh Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification", *Soft comput*, Vol. 27, No. 16, pp. 11259–11274, 2023, doi: 10.1007/s00500-023-08587-x.

[11] F. Macedo, R. Valadas, E. Carrasquinha, M. R. Oliveira, and A. Pacheco, "Feature selection using Decomposed Mutual Information Maximization", *Neurocomputing*, Vol. 513, pp. 215–232, 2022, doi: https://doi.org/10.1016/j.neucom.2022.09.101.

[12] S. Rosidin, Muljono, G. F. Shidik, A. Z. Fanani, F. Al Zami, and Purwanto, "Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data", In: *Proc. of International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 32–36, 2021, doi: 10.1109/iSemantic52711.2021.9573196.

[13] N. Elssied, A. Prof. Dr. O. Ibrahim, and A. Hamza Osman, "A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification", *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 7, pp. 625–638, 2014, doi: 10.19026/rjaset.7.299.

[14] J. Cheng, J. Sun, K. Yao, M. Xu, and Y. Cao, "A variable selection method based on mutual information and variance inflation factor", *Spectrochim Acta A Mol Biomol Spectrosc*, Vol. 268, pp. 120652, 2022, doi: https://doi.org/10.1016/j.saa.2021.120652.

[15] N. Manju, B. S. Harish, and V. Prajwal, "Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier", *International Journal of Computer Network and Information Security*, Vol. 11, pp. 37–44, 2019, doi: 10.5815/ijcnis.2019.07.06.

[16] M. Vishwakarma and N. Kesswani, "A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection", *Decision Analytics Journal*, Vol. 7, pp. 100233, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100233.

[17] A. Verma and V. Ranga, "ELNIDS: Ensemble Learning based Network Intrusion Detection System for RPL based Internet of Things", In: *Proc. of 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, pp. 1–6, 2019, doi: 10.1109/IoT-SIU.2019.8777504.

[18] A. Sunyoto and Hanafi, "Enhance Intrusion Detection (IDS) System Using Deep SDAE to Increase Effectiveness of Dimensional Reduction in Machine Learning and Deep Learning", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 4, pp. 125–141, 2022, doi: 10.22266/ijies2022.0831.13.

[19] Z. Liu, N. Thapa, A. Shaver, K. Roy, X. Yuan, and S. Khorsandroo, "Anomaly Detection on IoT Network Intrusion Using Machine Learning", In: *Proc. of International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1–5, 2020, doi: 10.1109/icABCD49160.2020.9183842.

[20] A. N. Iman and T. Ahmad, "Improving Intrusion Detection System by Estimating Parameters of Random Forest in Boruta", In: *Proc. of International Conference on Smart Technology and Applications (ICoSTA)*, pp. 1–

6, 2020, doi: 10.1109/ICoSTA48221.2020.1570609975.

[21] W. A. Safitri and T. Ahmad, "Rank-Based Univariate Selection for Intrusion Detection System", In: *Proc. of 4th International Conference on Information and Communications Technology (ICOIACT)*, pp. 164–168, 2021, doi: 10.1109/ICOIACT53268.2021.9563981.

[22] A. T. Nururrahmah and T. Ahmad, "CHI2CV : Feature Selection using Chi-Square with Cross-Validation for Intrusion Detection System", In: *Proc. of 11th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–6, 2023, doi: 10.1109/ISDFS58141.2023.10131731.

[23] D. J. Perangin-Angin and F. A. Bachtiar, "Classification of Stress in Office Work Activities Using Extreme Learning Machine Algorithm and One-way ANOVA F-Test Feature Selection", In: *Proc. of 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 503–508, 2021, doi: 10.1109/ISRITI54043.2021.9702802.

[24] M. A. Sulaiman and J. Labadin, "Feature selection based on mutual information", In: *Proc. of 9th International Conference on IT in Asia (CITA)*, pp. 1–6, 2015, doi: 10.1109/CITA.2015.7349827.

[25] M. Dong, L. Yao, X. Wang, B. Benatallah, S. Zhang, and Q. Z. Sheng, "Gradient Boosted Neural Decision Forest", *IEEE Trans Serv Comput*, Vol. 16, No. 1, pp. 330–342, 2023, doi: 10.1109/TSC.2021.3133673.

[26] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, "XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data", In: *Proc. of International Symposium on Electronics and Smart Devices (ISESD)*, pp. 1–7, 2022, doi: 10.1109/ISESD56103.2022.9980630.

[27] T. Deepak, "XGBoost Classification based Network Intrusion Detection System for Big Data using PySparkling Water", *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, pp. 377–382, 2020, doi: 10.30534/ijatcse/2020/55912020.

[28] T. Sood, S. Prakash, S. Sharma, A. Singh, and H. Choubey, "Intrusion Detection System in Wireless Sensor Network Using Conditional Generative Adversarial Network", *Wirel Pers Commun*, Vol. 126, pp. 21, 2022, doi: 10.1007/s11277-022-09776-x.

[29] D. Donoho and J. Jin, "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak", In: *Proc. of the National Academy of Sciences*, Vol. 105, No. 39, pp. 14790–14795, 2008, doi: 10.1073/pnas.0807471105.

[30] A. A. Megantara and T. Ahmad, "A hybrid machine learning method for increasing the performance of network intrusion detection systems", *J Big Data*, Vol. 8, No. 1, pp. 142, 2021, doi: 10.1186/s40537-021-00531-w.

[31] M. Malikhah, R. Sarno, and S. Sabila, "Ensemble Learning for Optimizing Classification of Pork Adulteration in Beef Based on Electronic Nose Dataset", *International Journal of Intelligent Engineering and Systems*, Vol. 14, pp. 44–55, 2021, doi: 10.22266/ijies2021.0831.05.

[32] M. McHugh, "The Chi-square test of independence", *Biochem Med (Zagreb)*, Vol. 23, pp. 143–149, 2013, doi: 10.11613/BM.2013.018.

[33] A. Megantara and T. Ahmad, "ANOVA-SVM for Selecting Subset Features in Encrypted Internet Traffic Classification", *International Journal of Intelligent Engineering and Systems*, Vol. 14, pp. 536–546, 2021, doi: 10.22266/ijies2021.0430.48.

[34] A. F. Siegel, "Chapter 15 - ANOVA: Testing for Differences Among Many Samples and Much More", *Practical Business Statistics (Seventh Edition)*, A. F. Siegel, Ed. Academic Press, pp. 469–492, 2016, doi: https://doi.org/10.1016/B978-0-12-804250-2.00015-8.

[35] R. Salmerón, C. Garcia, and J. Pérez, "Variance Inflation Factor and Condition Number in multiple linear regression", *J Stat Comput Simul*, Vol. 88, pp. 1–20, 2018, doi: 10.1080/00949655.2018.1463376.

[36] S. D. Permai and K. Herdianto, "Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods", *Procedia Computer Science*, Vol. 227, pp. 1012–1019, 2023, doi: https://doi.org/10.1016/j.procs.2023.10.610.

[37] D. A. Anggoro and S. S. Mukti, "Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 6, pp. 198–207,2021, doi: 10.22266/ijies2021.1231.19.

[38] L. Yang, L. Guo, W. Zhang, and X. Yang, "Classification of Multiple Power Quality Disturbances by Tunable-Q Wavelet Transform

with Parameter Selection", *Energies (Basel)*, Vol. 15, pp. 3428, 2022, doi: 10.3390/en15093428.

[39] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)", *Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, 2015, doi: 10.1109/MilCIS.2015.7348942.

[40] S. M. Kasongo and Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset", *J Big Data*, Vol. 7, No. 1, pp. 105, 2020, doi: 10.1186/s40537-020-00379-6.

[41] S. Das, S. Saha, A. T. Priyoti, E. K. Roy, F. T. Sheldon, A. Haque, and S. Shiva, "Network Intrusion Detection and Comparative Analysis Using Ensemble Machine Learning and Feature Selection", *IEEE Transactions on Network and Service Management*, Vol. 19, No. 4, pp. 4821–4833, 2022, doi: 10.1109/TNSM.2021.3138457.

[42] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", In: *Proc. of IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, 2009, doi: 10.1109/CISDA.2009.5356528.

[43] A. Andalib and V. T. Vakili, "An Autonomous Intrusion Detection System Using an Ensemble of Advanced Learners", In: *Proc. of 28th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1–5, 2020, doi: 10.1109/ICEE50131.2020.9260808.

[44] A. Bansal and S. Kaur, "Extreme Gradient Boosting Based Tuning for Classification in Intrusion Detection Systems", *Advances in Computing and Data Sciences*, pp. 372–380, 2018.