Analysis of NSL-KDD Dataset for Classification of Attacks Based on Intrusion Detection System Using Binary Logistics and Multinomial Logistics

(Analisis Dataset NSL KDD untuk Klasifikasi Serangan Berdasarkan Sistem Deteksi Gangguan Menggunakan Logistir Biner dan Logistik Multinomial)

Novia Amilatus Solekha

Statistics Department, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology (ITS) Jalan Arif Rahman Hakim, Surabaya 6011 Indonesia
E-mail: noviasolekhah51@gmail.com

ABSTRACT

At present, the intrusion detection system is the most developed trend in society. The intrusion detection system acts as a defense tool to detect security attacks which has been increasing steadily in recent years. Therefore, global information security is a very serious problem. As the types of attacks that emerge are constantly changing, there is a need to develop adaptive and flexible security features. Selection feature is one of the focuses of research on data mining for datasets that have relatively many attributes. In this study, the author tries to analyze the NSL-KDD data set with the selected attributes classified in two ways, namely binary classification (attack or not attack) and five classification classes using multinomial logistics, namely Dos, R2L, U2R, Probe and Normal. The results showed that the NSL-KDD dataset for the classification of attacks on the Intrusion Detection System (IDS) using binary logistics can increase the classification accuracy to 92.3% and 91.7% for datasets with multinomial logistics.

Keywords: Intrusion detection system (IDS), NSL-KDD dataset, Binary logistics and Multinomial logistics

PRELIMINARY

During the last two decades, the global network and the internet have been become a necessity for most of the world's population [1]. Communication systems play an important role in everyday life. Computer networks are effectively used for business data processing, education, collaboration, broad data acquisition and entertainment [2][3]. The development of computer network technology makes the security system very important so that a system is needed that is able to detect and identify any unusual activity, especially unexpected malicious network traffic. Actually, network traffic can be categorized into two labels (normal traffic and malicious traffic). Not only that network traffic can also be divided into five categories: Normal, DoS (Attack Denial of Service), R2L (Root to Local attacks), U2R (User to Root attacks) and Probe (Probing attacks). [4][5].

The emergence of Intrusion Detection System (IDS) technology can give solution security, system this could applied to both host-based IDS (HIDS) and network-based IDS (NIDS) technique this have similarity however only on device soft or device hard the place from installed IDS system [4]. Intrusion Detection System (IDS) is used for monitor then cross network and activity suspicious as well as warn system or network administrator . IDS can also respond then dangerous traffic by taking actions such as blocking user or source IP address for access network [6] [7].

In general, the way IDS works is developed in 2 ways, namely using signature-based detection, namely matching the attack behavior pattern that has been defined in the database. This technique requires a relatively short execution time for the pattern matching process, but has the disadvantage of not being able to detect the type of attack that can modify itself. The next method is Anomaly-based IDS, this technique will detect any unusual activity on the network. This technique can detect new types of attacks, but the definition of normal conditions must first be ascertained. Because this type gives a lot of false positive messages [8].

In writing this recommended use of the NSL-KDD dataset is a network dataset from the enhanced version of its predecessor KDD CUP 99. NSL-KDD appears to solve the problems inherent in KDD PIALA 99. The NSL-KDD dataset is a dataset that can be used as a comparison for various classification method for intrusion detection. This dataset has a fairly high dimension with 41 features. Feature selection is one of the important preprocesses to reduce the dataset by removing unimportant features from the NSL-KDD dataset [9] [10].

Destination study this is for knowing characteristics and modeling of the developed NSL-KDD dataset use

The purpose of this study was to determine the characteristics and modeling of the NSL-KDD dataset which was developed using binary logistic regression and multinomial logistic regression, where classification accuracy would be obtained for each method. The formation of the model is carried out in two ways, namely the classification of attacks on the Intrusion Detection System using binary logistics (attack or non-attack) and five polynomial classification classes, namely Dos, R2L, U2R, Probe and Normal [10].

Research that is relevant to this is research conducted by [11] researcher uses different for the detection of different types of attacks. In his paper he differentiates classifiers in minority and majority base which concludes that the false detection of minority class will lead to many discrepancies in IDES (Intrusion Detection Expert System).

In [12], author proposed hybrid approaches which combine some techniques like J48 Decision Tree, Support Vector machine and Naïve Bayesian for detection of different types of attacks and also contains different types of accuracy according to algorithms. These all tests took place on NSL-KDD Dataset.

RESEARCH REVIEW

Description of KDD Trend+ Data Data Set

Various statistical analyzes revealed that the weaknesses inherent in the KDD cup 99 data set that affect the accuracy of IDS detection have been modeled by many researchers [9]. The NSL-KDD data set [3] is an enhanced version of its predecessor, it contains important records of the complete KDD data set. The number of records selected from each difficult level group is inversely proportional to the percentage of records in the original KDD data set. Attribute details, namely attribute names, descriptions and sample data are listed in the table. There is a collection of downloadable files available to researchers [14].

There are 41 attributes available in the NSL-KDD data set. The 42nd attribute contains data about various 5 classes of network connection vectors and they are categorized as one normal class and four attack classes. The 4 attack classes are further grouped as DOS, Probe, R2L and U2R [3].

Table 1. Attributes Independent NSL-KDD Dataset .

Type	Features
Nominal	Protocol_type (2), Service(3), Flags (4)
Binary	land(7), logged_in (12), root_shell (14), su_attempted (15), is_host_login (21),, is_guest_login (22)
Numeric	Duration (1), src_bytes (5), dst_bytes (6), wrong_fragment (8), urgent (9), hot(10), num_failed_logins (11), num_compromised (13), num_root (16), num_file_creations (17), num_shells (18), num_access_files (19), num_outbound_cmds (20), count (23) srv_count (24), serror_rate (25), srv_serror_rate (26), rerror_rate (27), srv_rerror_rate (28), same_srv_rate (29) diff_srv_rate (30), srv_diff_host_rate (31), dst_host_count (32), dst_host_srv_count (33), dst_host_same_srv_rate (34), dst_host_diff_srv_rate (35), dst_host_same_src_port_rate (36), dst_host_srv_diff_host_rate (37), dst_host_serror_rate (38), dst_host_srv_serror_rate (39), dst_host_rerror_rate (40), dst_host_srv_rerror_rate (41)

The specific types of attacks are classified into four major categories . The Table 2 . shows this details [15]:

Table 2. Classified Into Four Major Categories.

Attack Class	Attack Type
DoS	Back , Land , Neptune , Pod , Smurf , Teardrop , Apache2 , Udpstorm , Processtable , Worm (10)
Probes	Satan , Ipsweep , Nmap , Portsweep , Mscan , Saint (6)
R2L	Guess_Password , Ftp_write , Imap , Phf , Multihop , Warezmaster , Warezclient , Spy , Xlock , Xsnoop , Snmpguess , Snmpgetattack , Httptunnel , Sendmail , Named (16)
U2R	Buffer_overflow , Loadmodule , Rootkit , Perl , Sqlattack , Xterm , Ps (7)

Multicollinearity

one conditions that must fulfilled in regression model formation with a number of variable predictor is no there is case multicollinearity or no there is correlation between one predictor variable with other predictor variables . Detecting multicollinearity can be seen by looking at the value of *Variance inflation Factor* (VIF). The VIF value in the to- k regression coefficient is formulated as follows:

One of the conditions that must be met in the formation of a regression model with several predictor variables is that there is no case of multicollinearity or there is no correlation between one predictor variable and another predictor variable. Detecting multicollinearity can be seen by looking at the value of Variance Inflation Factor (VIF). The VIF value in the to- k regression coefficient is formulated as follows:

$$VIF_k = \frac{1}{1-R_k^2} \qquad (1)$$

where R_k^2 is the coefficient of determination obtained by regressing the variable X_k with the explanatory variable $x_{\neq k}$ [16].

Binary Logistics Regression

Binary logistic regression is a statistical analysis method used to find the relationship between response variables (y) that have a nominal data scale (two categories or binary) and one or more predictor variables (x) either categorical or continuou. In general, the logistic regression model involving several predictor variables can be written as follows [17].

$$\pi(x_i) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1+\exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \qquad (2)$$

where $\boldsymbol{\beta}^T : [\beta_0, \beta_1, \cdots, \beta_k]$ and $\mathbf{x}_i : [1, x_1, x_2, \cdots, x_k]^T$ represents the vector of the predictor variables The logistic model is a *nonlinear* model, which requires transformation to become linear function , the transformation used is logit transformation from $\pi(x_i)$ so based on equality (2) obtained

$$1-\pi(x_i) = \frac{1}{1+\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})} \qquad (3)$$

The ratio between $\pi(x_i)$ and $1-\pi(x_i)$ is written as follows:

$$\begin{aligned}\frac{\pi(x_i)}{1-\pi(x_i)} &= \frac{\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})}{1+\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})} \Big/ \frac{1}{1+\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})} \\ &= \exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})\end{aligned} \qquad (4)$$

Furthermore, in equation (4) / transformation is carried out $\ln$ on both sides

$$\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \ln\left[\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki})\right] \qquad (5)$$

So that we get the global binary logistic regression logit model is:

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \boldsymbol{\beta}^T \mathbf{x}_i \qquad (6)$$

Multinomial Logistic Regression

Multinomial logistic regression is a data analysis method used to look for connection Among variable response (y) which is polychotomous or multinomial [17]. Hosmer and Lemeshow explain that the model used in the regression multinomial logistics is :

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi})}{1+\exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi})} \qquad (7)$$

With use logit transformation will obtained two log function

$$\begin{aligned} P_1(x) &= \ln\left(\frac{P(Y=1)1|x}{P(Y=1)0|x}\right) \\ &= \beta_{10}+\beta_{11}x_1+\ldots+\beta_{1p}x_p \\ &= X^T\beta_1 \end{aligned} \qquad (8)$$

$$\begin{aligned} P_2(x) &= \ln\left(\frac{P(Y=1)1|x}{P(Y=1)0|x}\right) \\ &= \beta_{20}+\beta_{21}x_2+\ldots+\beta_{2p}x_p \\ &= X^T\beta_2 \end{aligned} \qquad (9)$$

Based on second the logit function so obtained a regression model logistics as following :

$$\pi_0(x)=\frac{1}{1+\exp P_1(x)+\exp P_2(x)} \qquad (10)$$

$$\pi_1(x)=\frac{\exp P_1(x)}{1+\exp P_1(x)+\exp P_2(x)} \qquad (11)$$

$$\pi_2(x)=\frac{\exp P_2(x)}{1+\exp P_1(x)+\exp P_2(x)} \qquad (12)$$

RESEARCH METHODOLOGY

Data source

The type of data used in this study is secondary data, namely data obtained indirectly to obtain information (information) from the object under study. Secondary data used is an NSL KDD Network Intrusion Detection Dataset that can be accessed at https://www.kaggle.com/hassan06/nslkdd .

In this research database NSL-KDD is used to create intrusions detection models. Where the NSL-KDD dataset is a predictive model that can determine is traffic normal or attack tissue [21]. Dataset it has two sections as NSL-KDD train , and NSL-KD D test sets. All type attacks on NSL-KDD datasets are categorized to in four class namely DoS, Probe, R2L, and U2R [4]. Where will the dataset be used in study this is the KDD Train + dataset.

Research variable

The variables used in this study are 32 independent attributes which can be seen in Table 3. And the dependent variable is classified in two ways, namely binary classification (attack or not attack) in Table 3. and five classification classes using multinomial logistics, namely Dos, R2L, U2R, Probe and Normal in Table 4.

Table 3. Attributes Variable dependent Binary Regression .

Attributes Dependent	Category	Scale	Description
y	0= No attack 1= Attack	Nominal	Dependent attributes for analysis regression binary logistics

Table 4. Attributes Variable dependent Regression Multinomial Logistics .

Dependent Attribute	Category	Scale	Description
y	1= Normal 2= Dos 3= Probes 4= U2R 5=R2L	Nominal	Dependent attributes for analysis regression logistics multivariate

On attribute variable independent data features used in study this is a type of numeric

Table 5. Attributes Variable Independent .

Feature	Attributes Independent	Name	Feature	Attributes Independent	Name	Feature	Attributes Independent	Name
1.	x_1	Duration	19.	x_{12}	Num_access s_file	32.	x_{23}	Dst_host_count
5.	x_2	Src_bytes	20.	x_{13}	Num_outbound_cmds	33.	x_{24}	Dst_host_srv_coun
6.	x_3	Dst_bytes	23.	x_{14}	Count	34.	x_{25}	Dst_host_same _ srv_rate
8.	x_4	Wrong_fragment	24.	x_{15}	Srv_count	35.	x_{26}	Dst_host_diff _ srv_rate
9.	x_5	Urgent	25.	x_{16}	Serror_rate	36.	x_{27}	Dst_host_same _ src_port_rate
10.	x_6	Hot	26.	x_{17}	Srv_serror_rate	37.	x_{28}	Dst_host_srv _ diff_host_rate
11.	x_7	Num_failed_logins	27.	x_{18}	error_rate	38.	x_{29}	Dst_host_serror_rate
13.	x_8	Num_compromised	28.	x_{19}	Srv_rerror_rate	39.	x_{30}	Dst_host_srv_serror_rate
16.	x_9	Num_root	29.	x_{20}	Same_srv_rate	40.	x_{31}	Dst_host_rerror_rate
17.	x_{10}	Num_file_creactions	30.	x_{21}	Diff_srv_rate	41.	x_{32}	Dst_host_srv_rerror_rate
18.	x_{11}	Num_shells	31.	x_{22}	Srv_diff_host _ rate	32.	x_{33}	Dst_host_count

Analysis Step

Stages research for modeling the NSL-KDD dataset using regression binary logistic and multinomial regression

1. Load NSL-KDD dataset
2. Describe NSL-KDD data characteristics
3. Test multicollinearity variable response with each variable predictor
4. To do analysis regression binary logistics and regression multinomial logistics
5. Conducting model suitability test (*goodness of fit*)
6. Count accuracy model classification

RESULTS AND DATA ANALYSIS

Data Description NSL-KDD Train+

The NSL KDD dataset is a dataset that is used as a comparison for research in the field of intrusion detection. Where in the NSL-KDD dataset there are 125,973 records in detail shown in Table 6. below:

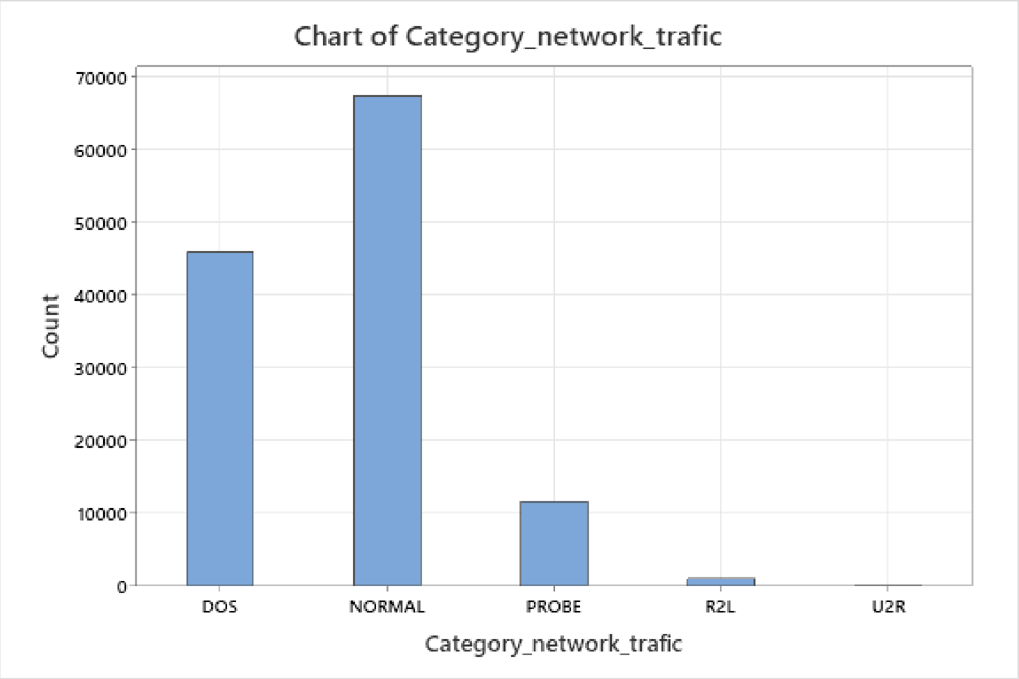
Table 6. Distribution of the NSL-KDD Train+ dataset.

Attack	Number of Records	Class	Number of records per class	Attack	Number of Records	Class	Number of records per class
Back	956	DOS	45927	Guess_Passwd	53	R2L	995
Land	18			ftp_write	8		
Neptune	41214			Imap	11		
Pod	201			Phf	4		
Smurfs	2646			Multihop	7		
teardrop	892			Warezmaster	20		
Satan	3633	PROBE	11656	Warezclient	890		
Lpsweeo	3599			Spy	2		
Nmap	1493			Buffer_overflow	30	U29	52
Postsweep	2931			Loadmodule	9		
				Perl	3		
Normal	67343	NORMAL	67343	Rootkit	10		

Next, we will analyze the attacks in the NSL-KDD dataset with binary logistic modeling (attack or non-attack) as shown in Table 7. and Picture 1. And multinomial regression modeling which consists of five classes of network traffic, namely Dos, R2L, U2R, Probe and Normal are shown in Table 8. and Picture 2.

Table 7. **Intrusion Category Network Traffic in NSL-KDD Training Dataset.**

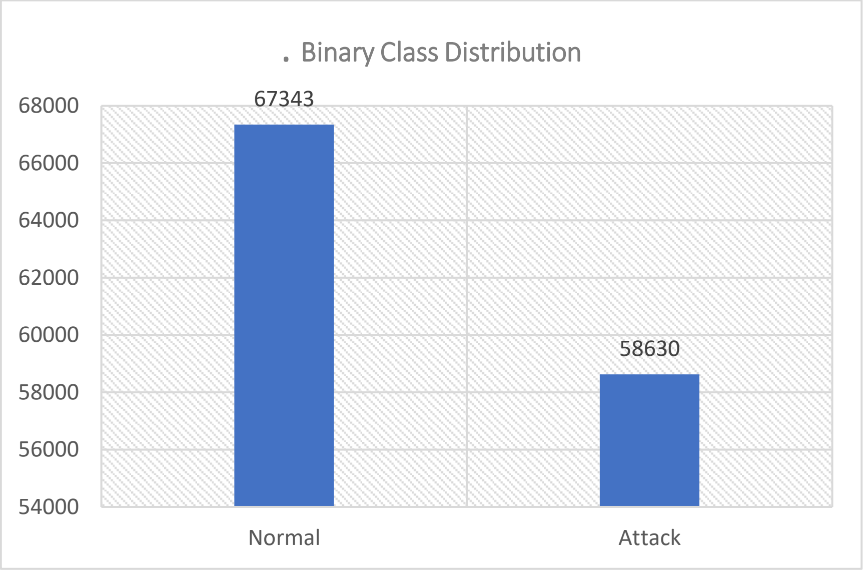
	Normal	DOS	Probes	R2L	U2R
Frequency	67343	45927	11656	995	52



Pictures 1. Intrusion Category Network Traffic in NSL-KDD Training Dataset

Table 8. Binary Class Distribution in NSL-KDD Training Dataset.

	Normal	Attack
Frequency	67343	58630



Picture 2. Binary Class Distribution in NSL-KDD Training Dataset

Multicollinearity Test

Multicollinearity calculation results can be determined using the VIF value, which can be seen in Table 9. as follows:

Table 9. VIF Attribute Value Independent .

Attributes Independent	Score VIF	Attributes Independent	Score VIF	Attributes Independent	Score VIF	Attributes Independent	Score VIF
X_1	1,289	X_9	703.979	X_{18}	66,233	X_{26}	1,860
X_2	1.007	X_{10}	1.038	X_{19}	11.026	X_{27}	1,579
X_3	1,001	X_{11}	1.004	X_{20}	2.058	X_{28}	42,876
X_4	1.091	X_{12}	1,819	X_{21}	1.316	X_{29}	62,044
X_5	1.012	X_{14}	4,670	X_{22}	2,136	X_{30}	9,667
X_6	1.063	X_{15}	3,178	X_{23}	8,047	X_{31}	20,163
X_7	1.018	X_{16}	104.675	X_{24}	11,445	X_{32}	66,233
X_8	690,182	X_{17}	123.845	X_{25}	2,554		

In Table 9. it can be seen that there are 11 attributes that have a VIF value > 10. So there are only 20 independent attributes that will be used in the analysis X_1, namely X_{30}, X_2, X_3, X_4, X_5, X_6, X_7, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{20}, X_{21}, X_{22}, X_{23}, X_{25}, X_{26} . X_{27}

Binary Logistics Regression Analysis

Next is to do Binary logistic regression modeling, the number of independent attributes used are 20 attributes and the dependent attribute is in two categories, namely attack and no attack.

1. Simultaneous Testing

Simultaneous parameter testing to determine the effect of parameters simultaneously with the enter method until the best model is obtained only from variables that have a significant effect with hypothesis testing as follows:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : least there is one $\beta_k \neq 0$

Table 10. Omnibus Tests of Model Coefficients.

		Chi- square	Df	Sig .
Step 1	Step	124855,719	23	0.000
	Block	124855,719	23	0.000
	Model	124855,719	23	0.000

Table 10. above show that χ^2 worth $124855{,}719 > \chi^2_{22,0,05}$ that is worth 33,924 so that decided reject H_0. This means that at least there is one independent attribute of the NSL-KDD dataset that affects binary class distribution network traffic.

2. Partial Test

Table 11. Independent Attribute Testing by Partial

Attributes	B	Wald	df	Sig .	Attributes	B	Wald	df	Sig .
x_1	0,000	20,746	1	0,000	x_{21}	-0,015	305,511	1	0,000
x_6	0,088	945,366	1	0,000	x_{22}	0,010	684,221	1	0,000
x_8	0,695	254,927	1	0,000	x_{23}	0,006	1000,287	1	0,000
x_9	-0,697	256,317	1	0,000	x_{24}	-0,010	1752,875	1	0,000
x_{10}	-0,372	34,469	1	0,000	x_{25}	0,010	259,215	1	0,000
x_{14}	0,015	641,567	1	0,000	x_{27}	0,025	5314,936	1	0,000
x_{15}	-0,008	163,105	1	0,000	x_{28}	0,039	1470,844	1	0,000
x_{16}	-0,007	11,944	1	0,001	x_{29}	0,006	28,634	1	0,000
x_{17}	0,032	197,627	1	0,000	x_{30}	0,019	168,305	1	0,000
x_{18}	-0,018	60,119	1	0,000	x_{31}	0,006	74,788	1	0,000
x_{19}	0,032	196,076	1	0,000	x_{32}	0,003	6,743	1	0,009
x_{20}	-0,018	524,763	1	0,000	Constant	-1,638	332,818	1	0,000

In Table 11. above are the results of the partial test using the enter method which has been iterated 11 times, so that the results of the Wald test of all attributes are significant. So that the best model can be formed from binary logistic regression analysis is

$$\begin{aligned} g_1(x) = {} & -1{,}638 + 0{,}088X_6 + 0{,}695X_8 - 0{,}697X_9 + {,}015X_{14} - 0{,}008\mathrm{X}_{15} - 0{,}007\mathrm{X}_{16} + 0{,}032\mathrm{X}_{17} - {,}018\mathrm{X}_{18} + \\ & 0{,}032\mathrm{X}_{19} - {,}018\mathrm{X}_{21} + 0{,}010\mathrm{X}_{22} + 0{,}006\mathrm{X}_{23} - 0{,}010\mathrm{X}_{24} + 0{,}010\mathrm{X}_{26} + 0{,}025\mathrm{X}_{27} + 0{,}039\mathrm{X}_{28} + 0{,}006\mathrm{X}_{29} + \\ & 0{,}019\mathrm{X}_{30} + 0{,}006\mathrm{X}_{31} + 0{,}003\mathrm{X}_{32} \end{aligned} \tag{13}$$

3. Model Fit Test

For test the suitability of the model using the Hosmer and Lemeshow test

Table 12. Hosmer and Lemeshow test .

Step	Chi- square	df	Sig .
1	611,236	8	,000

From Table 12 above, it is found that the χ^2 value is (61 1236) > 2.706 which means that it failed to reject so it can be decided that the model is appropriate.

4. Classification Accuracy

Table 13. Accuracy Classification Regression Binary Logistics .

Observed	Predicted		
	Y		
	No attack	Attack	Percentage Correct
No attack	61651	3718	94.3
Attack	5988	54616	90.1
Overall Percentage			92.3

Based on Table 13. It can be seen that the percentage of accuracy in the classification of Binary Class Distribution network traffic in the NSL-KDD Training Dataset as a whole by 92.3%. Showing that no attack is classified with Correct by 94.3% and classified attack with Correct by 90.1%. This thing means with using regression model binary logistics there are 116,267 NSL-KDD data sets from 125,973 classified NSL-KDD data sets with Correct in accordance with proportion of binary class distribution network traffic in NSL-KDD training dataset.

Multinomial Logistics Regression Analysis

1. Simultaneous Testing

This test is carried out to check the coefficients β simultaneously or simultaneously on the response variable, the best model is only from the variables that have a significant effect. with hypothesis testing as follows:

Table 14. Simultaneous Test Regression Multinomial Logistics .

	Fitting Model Criteria	Likelihood Ratio Tests		
Model	-2 Likelihood Logs	Chi- Square	df	Sig .
Intercept Only	240152,730			
Final	58674,504	181478,226	44	0.000

Based on Table 14, the results of the simultaneous test show the value in the final row with a chi-square value of 58674,504 > 65,410 (table) or a sig < alpha value so that H_0 it rejects so that it can be concluded that at least there is an independent attribute that affects the intrusion category network traffic in NSL- KDD training dataset

2. Partial Test

Table 15. Partial Test Regression Multinomial Logistics .

Effect	-2 Likelihood Logs	Chi-Square	df	Sig	Effect	-2 Likelihood Logs	Chi-Square	df	Sig
Intercept	68391,535	9717.032	4	0.000	x_{24}	61975,265	3300,762	4	0.000
x_4	64237,157	5562,653	4	0.000	x_{26}	63524,141	4849,638	4	0.000

x_{14}	63670,912	4996,408	4	0.000	x_{27}	66867,406	8192,902	4	0.000
x_{15}	60459,270	1784,766	4	0.000	x_{28}	63307.309	4632,805	4	0.000
x_{21}	59583,585	909.081	4	0 ,000	x_{30}	68558,957	9884.453	4	0.000
x_{23}	59000,379	325,876	4	0 ,000	x_{32}	62533,371	3858,867	4	0.000

Based on Table 15 above, it was found that the value of sig<alpha so reject H_0 showing eleven independent attributes significant affect intrusion category network traffic in NSL-KDD training dataset on attributes response .

3. Model Fit Test

Table 16. Multinomial Logistics Model Suitability Test

	Chi- Square	**df**	**Sig .**
Pearson	13609567,255	390640	0.000
Deviance	57032,855	390640	1,000

Based on the results in Table 16, the p-value (sig.) in the line of the Deviance method feasibility test results is 1,000. This value states that the feasibility test of the Deviance method is greater than alpha (5% = 0.05), then it fails to reject which explains that the fit model or model is appropriate. This means that the model can explain the data.

4. Parameter Estimation of Multinomial Logistics Regression Model

Table 17. Estimation of Regression Parameters Multinomial Logistics .

Type	**Normal**		**Dos**		**Probes**		**R2L**	
	β	Sig.	β	Sig.	β	Sig.	β	Sig.
Intercept	2,724	0,000	0,520	0,268	-1,955	0,000	0,211	0,662
x_4	-4,374	1,000	26,986	0,999	-7,034	1,000	-0,655	0,000
x_{14}	-0,025	0,000	0,003	0,604	-0,009	0,112	-0,444	0,000
x_{15}	0,655	0,004	0,636	0,005	0,641	0,005	0,577	0,012
x_{21}	0,004	0,630	-0,017	0,026	0,018	0,022	-0,025	0,004
x_{23}	0,009	0,001	0,011	0,000	0,010	0,000	0,015	0,000
x_{24}	0,065	0,000	0,059	0,000	0,058	0,000	0,044	0,000
x_{26}	0,074	0,000	-0,008	0,673	0,082	0,000	-0,056	0,006
x_{27}	-0,022	0,000	-0,009	0,023	0,020	0,000	0,017	0,000
x_{28}	-0,012	0,172	-0,073	0,000	0,053	0,000	-0,010	0,291
x_{30}	0,029	0,501	0,086	0,046	0,053	0,213	0,051	0,232
x_{32}	0,016	0,164	0,039	0,001	0,043	0,000	0,023	0,053

Based on the Table 17 , obtained parameter estimation for each variable and each categorical with the average odd ratio value of 1 which indicates that each estimation of significance parameter The logit function obtained is as following .

$$\begin{aligned} g_1\left(x\right) = {} & 2{,}724 - 4{,}374\mathrm{X}_4 - 0{,}025X_{14} + 0{,}655\mathrm{X}_{15} + 0{,}004\mathrm{X}_{21} + 0{,}009\mathrm{X}_{23} + 0{,}065\mathrm{X}_{24} + 0{,}074\mathrm{X}_{26} \\ & - 0{,}022\mathrm{X}_{27} - 0{,}012\mathrm{X}_{28} + 0{,}029\mathrm{X}_{30} + 0{,}016\mathrm{X}_{32} \end{aligned} \qquad (14)$$

$$\begin{array}{l} g_2\left(x\right)=0{,}520+26{,}986\mathrm{X}_4+0{,}003X_{14}+0{,}636\mathrm{X}_{15}-0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23}+0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26} \\ \qquad -0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28}+0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32} \end{array} \qquad (15)$$

$$\begin{array}{l} g_3\left(x\right)=-1{,}955-7{,}034\mathrm{X}_4-0{,}009X_{14}+0{,}641\mathrm{X}_{15}+0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23}+0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26} \\ \qquad +0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28}+0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32} \end{array} \qquad (16)$$

$$\begin{array}{l} g_4\left(x\right)=0{,}211-0{,}655\mathrm{X}_4-0{,}025X_{14}+0{,}577\mathrm{X}_{15}-0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23}+0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26} \\ \qquad +0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28}+0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32} \end{array} \qquad (17)$$

Based on the logit function , can Regression model is formed multinomial logistics as following

$$\pi_0\left(x\right)=\frac{1}{\begin{array}{l} 1+\exp\left(\begin{array}{l} 2{,}724-4{,}374\mathrm{X}_4- \\ 0{,}025X_{14}+0{,}655\mathrm{X}_{15} \\ +0{,}004\mathrm{X}_{21}+0{,}009\mathrm{X}_{23} \\ +0{,}065\mathrm{X}_{24}+0{,}074\mathrm{X}_{26} \\ -0{,}022\mathrm{X}_{27}-0{,}012\mathrm{X}_{28} \\ +0{,}029\mathrm{X}_{30}+0{,}016\mathrm{X}_{32} \end{array}\right)+\exp\left(\begin{array}{l} 0{,}520+26{,}986\mathrm{X}_4+ \\ 0{,}003X_{14}+0{,}636\mathrm{X}_{15} \\ -0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23} \\ +0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26} \\ -0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28} \\ +0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32} \end{array}\right)+ \\ \exp\left(\begin{array}{l} -1{,}955-7{,}034\mathrm{X}_4- \\ 0{,}009X_{14}+0{,}641\mathrm{X}_{15} \\ +0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23} \\ +0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26} \\ +0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28} \\ +0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32} \end{array}\right)+\exp\left(\begin{array}{l} 0{,}211-0{,}655\mathrm{X}_4- \\ 0{,}025X_{14}+0{,}577\mathrm{X}_{15} \\ -0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23} \\ +0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26} \\ +0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28} \\ +0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32} \end{array}\right) \end{array}} \qquad (18)$$

$$\pi_1\left(x\right)=\frac{\exp\left(\begin{array}{l} 2{,}724-4{,}374\mathrm{X}_4-0{,}025X_{14}+0{,}655\mathrm{X}_{15}+0{,}004\mathrm{X}_{21}+0{,}009\mathrm{X}_{23}+ \\ 0{,}065\mathrm{X}_{24}+0{,}074\mathrm{X}_{26}-0{,}022\mathrm{X}_{27}-0{,}012\mathrm{X}_{28}+0{,}029\mathrm{X}_{30}+0{,}016\mathrm{X}_{32} \end{array}\right)}{\begin{array}{l} 1+\exp\left(\begin{array}{l} 2{,}724-4{,}374\mathrm{X}_4- \\ 0{,}025X_{14}+0{,}655\mathrm{X}_{15} \\ +0{,}004\mathrm{X}_{21}+0{,}009\mathrm{X}_{23} \\ +0{,}065\mathrm{X}_{24}+0{,}074\mathrm{X}_{26} \\ -0{,}022\mathrm{X}_{27}-0{,}012\mathrm{X}_{28} \\ +0{,}029\mathrm{X}_{30}+0{,}016\mathrm{X}_{32} \end{array}\right)+\exp\left(\begin{array}{l} 0{,}520+26{,}986\mathrm{X}_4+ \\ 0{,}003X_{14}+0{,}636\mathrm{X}_{15} \\ -0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23} \\ +0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26} \\ -0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28} \\ +0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32} \end{array}\right)+ \\ \exp\left(\begin{array}{l} -1{,}955-7{,}034\mathrm{X}_4- \\ 0{,}009X_{14}+0{,}641\mathrm{X}_{15} \\ +0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23} \\ +0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26} \\ +0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28} \\ +0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32} \end{array}\right)+\exp\left(\begin{array}{l} 0{,}211-0{,}655\mathrm{X}_4- \\ 0{,}025X_{14}+0{,}577\mathrm{X}_{15} \\ -0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23} \\ +0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26} \\ +0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28} \\ +0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32} \end{array}\right) \end{array}} \qquad (19)$$

$$\pi_2\left(x\right)=\frac{\exp\left(\begin{array}{l}0{,}520+26{,}986\mathrm{X}_4+0{,}003X_{14}+0{,}636\mathrm{X}_{15}-0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23}+\\ 0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26}-0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28}+0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32}\\ \end{array}\right)}{\begin{array}{c}1+\exp\left(\begin{array}{l}2{,}724-4{,}374\mathrm{X}_4-\\ 0{,}025X_{14}+0{,}655\mathrm{X}_{15}\\ +0{,}004\mathrm{X}_{21}+0{,}009\mathrm{X}_{23}\\ +0{,}065\mathrm{X}_{24}+0{,}074\mathrm{X}_{26}\\ -0{,}022\mathrm{X}_{27}-0{,}012\mathrm{X}_{28}\\ +0{,}029\mathrm{X}_{30}+0{,}016\mathrm{X}_{32}\\ \end{array}\right)+\exp\left(\begin{array}{l}0{,}520+26{,}986\mathrm{X}_4+\\ 0{,}003X_{14}+0{,}636\mathrm{X}_{15}\\ -0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23}\\ +0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26}\\ -0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28}\\ +0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32}\\ \end{array}\right)+\\ \exp\left(\begin{array}{l}-1{,}955-7{,}034\mathrm{X}_4-\\ 0{,}009X_{14}+0{,}641\mathrm{X}_{15}\\ +0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23}\\ +0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26}\\ +0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28}\\ +0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32}\\ \end{array}\right)+\exp\left(\begin{array}{l}0{,}211-0{,}655\mathrm{X}_4-\\ 0{,}025X_{14}+0{,}577\mathrm{X}_{15}\\ -0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23}\\ +0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26}\\ +0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28}\\ +0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32}\\ \end{array}\right)\end{array}} \tag{20}$$

$$\pi_3\left(x\right)=\frac{\exp\left(\begin{array}{l}-1{,}955-7{,}034\mathrm{X}_4-0{,}009X_{14}+0{,}641\mathrm{X}_{15}+0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23}+\\ 0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26}+0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28}+0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32}\\ \end{array}\right)}{\begin{array}{c}1+\exp\left(\begin{array}{l}2{,}724-4{,}374\mathrm{X}_4-\\ 0{,}025X_{14}+0{,}655\mathrm{X}_{15}\\ +0{,}004\mathrm{X}_{21}+0{,}009\mathrm{X}_{23}\\ +0{,}065\mathrm{X}_{24}+0{,}074\mathrm{X}_{26}\\ -0{,}022\mathrm{X}_{27}-0{,}012\mathrm{X}_{28}\\ +0{,}029\mathrm{X}_{30}+0{,}016\mathrm{X}_{32}\\ \end{array}\right)+\exp\left(\begin{array}{l}0{,}520+26{,}986\mathrm{X}_4+\\ 0{,}003X_{14}+0{,}636\mathrm{X}_{15}\\ -0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23}\\ +0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26}\\ -0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28}\\ +0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32}\\ \end{array}\right)+\\ \exp\left(\begin{array}{l}-1{,}955-7{,}034\mathrm{X}_4-\\ 0{,}009X_{14}+0{,}641\mathrm{X}_{15}\\ +0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23}\\ +0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26}\\ +0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28}\\ +0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32}\\ \end{array}\right)+\exp\left(\begin{array}{l}0{,}211-0{,}655\mathrm{X}_4-\\ 0{,}025X_{14}+0{,}577\mathrm{X}_{15}\\ -0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23}\\ +0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26}\\ +0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28}\\ +0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32}\\ \end{array}\right)\end{array}} \tag{21}$$

$$\pi_4(x)=\frac{\exp\begin{pmatrix}0{,}211-0{,}655\mathrm{X}_4-0{,}025X_{14}+0{,}577\mathrm{X}_{15}-0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23}+\\ 0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26}+0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28}+0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32}\\ \end{pmatrix}}{1+\exp\begin{pmatrix}2{,}724-4{,}374\mathrm{X}_4-\\ 0{,}025X_{14}+0{,}655\mathrm{X}_{15}\\ +0{,}004\mathrm{X}_{21}+0{,}009\mathrm{X}_{23}\\ +0{,}065\mathrm{X}_{24}+0{,}074\mathrm{X}_{26}\\ -0{,}022\mathrm{X}_{27}-0{,}012\mathrm{X}_{28}\\ +0{,}029\mathrm{X}_{30}+0{,}016\mathrm{X}_{32}\\ \end{pmatrix}+\exp\begin{pmatrix}0{,}520+26{,}986\mathrm{X}_4+\\ 0{,}003X_{14}+0{,}636\mathrm{X}_{15}\\ -0{,}017\mathrm{X}_{21}+0{,}011\mathrm{X}_{23}\\ +0{,}059\mathrm{X}_{24}-0{,}008\mathrm{X}_{26}\\ -0{,}009\mathrm{X}_{27}-0{,}073\mathrm{X}_{28}\\ +0{,}086\mathrm{X}_{30}+0{,}039\mathrm{X}_{32}\\ \end{pmatrix}+\exp\begin{pmatrix}-1{,}955-7{,}034\mathrm{X}_4-\\ 0{,}009X_{14}+0{,}641\mathrm{X}_{15}\\ +0{,}018\mathrm{X}_{21}+0{,}010\mathrm{X}_{23}\\ +0{,}058\mathrm{X}_{24}+0{,}082\mathrm{X}_{26}\\ +0{,}020\mathrm{X}_{27}+0{,}053\mathrm{X}_{28}\\ +0{,}053\mathrm{X}_{30}+0{,}043\mathrm{X}_{32}\\ \end{pmatrix}+\exp\begin{pmatrix}0{,}211-0{,}655\mathrm{X}_4-\\ 0{,}025X_{14}+0{,}577\mathrm{X}_{15}\\ -0{,}025\mathrm{X}_{21}+0{,}015\mathrm{X}_{23}\\ +0{,}044\mathrm{X}_{24}-0{,}056\mathrm{X}_{26}\\ +0{,}017\mathrm{X}_{27}-0{,}010\mathrm{X}_{28}\\ +0{,}051\mathrm{X}_{30}+0{,}023\mathrm{X}_{32}\\ \end{pmatrix}} \qquad (22)$$

5. Classification Accuracy

Table 18. Accuracy Multinomial Regression Model Classification .

	Predicted					
Observed	Normal	DOS	Probes	R2L	U2R	Percent Correct
Normal	62294	785	2115	175	0	95.3%
DOS	3592	44190	106	13	0	92.3%
Probes	2508	180	8892	76	0	76.3%
R2L	484	56	287	168	0	16.9%
U2R	30	3	7	12	0	0.0%
Overall Percentage	54.7%	35.9%	9.1%	,4%	0.0%	91.7%

Based on Table 18. It can be seen that the percentage of accuracy in classifying the Intrusion Category Network Traffic in NSL-KDD Training Dataset as a whole by 91.7%. Showing that the normal class is classified with Correct by 95.3%, classified DOS class with Correct of 92.3%, Class Probe classified with Correct by 76.3%, classified R2L class with Correct of 16.9%, and the U2R class is not classified the truth . This thing means with using regression model logistics Multinomial there are 115,544 NSL-KDD data sets from 125,973 classified NSL-KDD data sets with Correct in accordance with Classification of Intrusion Category Network Traffic in NSL-KDD Training Dataset.

CONCLUSIONS AND RECOMMENDATIONS

Based on the estimation test of the binary logistic regression model with the multinomial logistic regression model, it was found that by using the binary logistic regression model there were 116,267 NSL-KDD data sets from 125,973 NSL-KDD data sets that were classified correctly according to the proportion of Binary Class Distribution network traffic in NSL-KDD Training Datasets. Meanwhile, in Multinomial logistic regression, there are 115,544 NSL-KDD data sets from 125,973 NSL-KDD data sets which are classified correctly according to the classification of Intrusion Category Network Traffic in NSL-KDD Training Dataset. The results showed that the NSL-KDD dataset for the classification of attacks on the Intrusion Detection

System (IDS) using binary logistics can increase the classification accuracy to 92.3% and 91.7% for datasets with multinomial logistics. Further research can add a variety of attributes in order to obtain a more refined model.

BIBLIOGRAPHY

[1] Kumar. V, Chauhan. H, and D. Panwar. D. "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset," *International Journal of Soft Computing and Engineering (IJSCE)*, ISSN: 2231-2307, Volume-3, Issue-4, September 2013

[2] Khorram. T, and Baykan. N. A, "Network Intrusion Detection using Optimized Machine Learning Algorithms," *European Journal of Science and Technology*, No. 25, pp. 463-474, August 2021.

[3] Dhanabal. L, and Shantharajah. S, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms". *International Journal of Advanced Research in Computer and Communication Engineering*, (2015), 446-451.

[4] Effendy. D. A, Kusrini, and Sundarmawan, "Algoritma K-Means untuk Diskretisasi Numerik Kontinyu Pada Klasifikasi Intrusion Detection System Menggunakan Naive Bayes," *Konferensi Nasional Sistem & Informatika 2017*, STMIK STIKOM Bali, 10 Agustus 2017, pp.61-66.

[5] Su. T, Sun. H, Zhu. J, Wang. S, and Li. A. Y, "BAT: Deep Learning Metthods on Network Intrusion Detection Using NSL-KDD Dataset", *Digital Object Identifier IEE Access*, 10 Februari 2020, pp.29575-29585.

[6] Deepa. M, and Dr. Sumitra. P, "Intrusion Detection System Using K-Means Based on Cuckoo Search Optimization*", IOP Conf. Series: Materials Science and Engineerin,g* 993 (2020) 012049, pp. 1-10.

[7] Ojugo. A. A, Eboka. A. O, konta. O. E, Yoro. R. E and Aghware. F. O, "Genetic Algorithm Rule-Based Intrusion Detection System", 2012, CIS, Vol. 3, No. 8. ISSN 2079- 8407.

[8] Neethu. B, "Classification of Intrusion Detection Dataset Using Machine Learning Approaches", *International Journal of Electronics and Computer Science Engineering*, 2012, pp.1044-51.

[9] Solanki. S, Gupta. C, and Rai. K, "A Survey on Machine Learning based Intrusion Detection System on NSL-KDD Dataset", International Journal of Computer Applications, Vol 176, No. 30 (2020), pp. 0975 – 8887.

[10] Jupriyadi, "Implementasi Seleksi Fitur Menggunakan Algoritma Fvbrm Untuk Klasifikasi Serangan Pada Intrusion Detection System (IDS)", Seminar Nasional Sains dan Teknologi 2018, e-ISSN: 2460 – 8416, pp. 1-6.

[11] Uikey. R, Cyanchandani. M, "Survey on Classification Techniques Applied to Intrusion Detection System and its Comparative Analysis" at 4th International Conference on Communication $ Electronics System (ICCES 2019) IEEE Conference Record #45898; IEEE Xplore ISBN; 978-1-7281-1261-9 in 2019.

[12] Pitale. A. B. A, "Detection of Network Intrusions Using Hybrid Intelligent System" at International Conferences on Advances in Information Technology in 2019.

[13] Fu. Y, Du. Y, Cao. Z, Li. Q, anad Xiang. W, " A Deep Learning Model for Network Intrusion Detection with Imbalanced Data", Electronics 2022, 11, 898. Pp. 1-13.

[14] Bhattacharjee. P. S, Fujail. A. K. M, and Begum. S. A, "Intrusion Detection System for NSL-KDD Data Set using Vectorised Fitness Function in Genetic Algorithm", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 2 (2017) pp. 235-246.

[15] Sapre. S, Ahmadi. P, and Islam. K, "A Robust Comparison of the KDDCup99 and NSL-KDD IoT Network Intrusion Detection Datasets Through Various Machine Learning Algorithms", Department of Information Sciences and Technology George Mason University Fairfax, USA, 31 Dec 2019.

[16] Hocking. R. R, *Method and Applications of Linear Models* (2nd Edition ed), New York: John Wiley and Sons, Inc, 1996.

[17] Hosmer. D. W, and Lemeshow. S, *Applied Logistic Regression*. USA: John Wliey & Sons, 2000