SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Impact of Number of Features Selected and Size of Training Data on the Accuracy of Machine Learning Based Cloud Security Algorithms – An Empirical Analysis

Tanko Yahaya Mohammed[1], Abdulrazaq Abdulrahim[2], Muhammad Aminu Umar[3], Aliyu Muhammad Kufena[4], and Hadiza Isa Abdullahi[5]
[1,2,3,4,5]Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria
tymohammed45@gmail.com[1], aabdulraahim@abu.edu.ng[2], maumar@abu.edu.ng[3], kufenah@gmail.com[4], hadeezah18@gmail.com[5]

Abstract

Background: *Machine learning (ML) techniques have proven to be very effective in providing security in a cloud environment considering the continuous evolving nature of threats. Some of the factors that influence the accuracies of ML models include the specific ML algorithm used, sample size, the number of features selected and portion of dataset used for training. Many studies have conducted empirical analyses of the effects of one or more combination of these factors on predicted accuracies of ML models. However, the effect of the portion of the entire dataset that is used for training the ML model as well as the number of features extracted from the dataset in predicting the accuracy of an ML model is yet to be investigated.* ***Aim*** *This study uses Ordinary Least Square (OLS) regression to investigate if the number of features selected and the size of training data are useful in predicting the accuracies obtained in ML based approaches to cloud security.* ***Method***: *For this research, we have two independent variables (number of features selected and the size of training data) and one dependent variable (accuracy). We initially selected 16 (sixteen) studies conducted within the last 5 (five) years for our study. We extracted the number of features used, the size of the training data and the accuracies obtained from these studies. After identifying and discarding outliers from the extracted values, we were left with 12 (twelve) studies. We conducted our analysis on these 12 studies.* ***Results***: *The result of our analysis shows that there exist a weak positive and negative relationships among the dependent and independent variables. Although, our analysis shows weak positive and negative relationships among the variables, our model is useful in predicting the accuracies of ML models given the number of features selected and the size of the training data.*

Keywords: Empirical Analysis, Accuracy Prediction, Features Selected, Training Data, ML models, Cloud Security

1. Introduction

A cloud computing paradigm allows users to connect to a shared pool of configurable resources on demand such as applications, servers, networks, etc., that may be swiftly supplied and released with minimal administration effort or service provider contact (Satyanarayana, 2012). It covers a wide range of computing areas and has grown in popularity in recent years. Furthermore, cloud computing reduces capital costs, decouples services from the underlying technology, and gives resource provisioning flexibility (Khan et al. 2013).

Despite the numerous advantages offered by cloud computing and despite it being perceived as a major business avenue in the coming years, it is still bedeviled by numerous security challenges (Singh and Chatterjee 2017; Tamilselvan 2019).

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

In recent years, machine learning (ML), a subset of artificial intelligence, has grown in popularity. It is used to teach an artificial intelligence system how to make decisions and has been applied to solve problems in many domains. The most visible benefit of ML models is that they can quickly detect patterns and trends in existing data with little or no human intervention (Tamilselvan 2019). For optimal performance, machine learning models require initial training with existing relevant data in order to enable them make smart decisions subsequently. Another important factor in a machine learning process is feature or attribute selection. Not all features of a dataset may be required by a machine learning model in making prediction. It is important for models to select only the required features for making prediction.

Machine learning (ML) techniques have proven to be very effective in providing security considering the continuous evolving nature of threats. The type of machine learning algorithm used, as well as sample size, are two important factors that have a non-trivial impact on prediction accuracy (Cui and Gong 2018). Other factors that could influence the predicted accuracy are the number of features selected and portion of dataset used for training. Many studies have conducted empirical analyses of the effects of one or more combination of these factors on predicted accuracies of ML models. However, the effect of the portion of the entire dataset that is used for training the ML model as well as the number of features extracted from the dataset in predicting the accuracy of an ML model is yet to be investigated.

Consequently, the objective of our study is to investigate if there is a significant relationship between the size of the training dataset and model accuracy, as well as the number of features used by a machine model and the accuracy obtained in ML based approaches to cloud security solutions. As a result, the following research questions were developed.

RQ1: Does using a large training dataset guarantee improved accuracy?

RQ2: Does selecting more features guarantee improved accuracy?

We aim to provide answers to these questions by proposing a model with accuracy as our dependent variable and number of features selected and the size of training dataset as our independent variables. We analyze our proposed model to determine its usefulness and see if me accept or reject our hypothesis.

2. Related Works

In this section, we explain empirical analysis and take a look of some empirical analysis studies that have been conducted to determine the relationship among dependent and independent variables. Empirical research is defined as investigation into phenomena that have been observed and measured. It can generate numerical data between two or more variables and report research based on actual observations or experiments using quantitative research methods. Empirical research is being is being applied in many fields of research today, such as Agriculture, Medicine, Science and Engineering, etc.

Many studies have recently been conducted to determine the relationship between one or more independent variables and a dependent variable. An investigation on the influence of gender in the requirement elicitation process was carried out by (Díaz et al. 2021). The goal of the study was to find out if requirements elicitation sessions were influenced by participants' gender. Similarly, (Chen and Qian 2020) looked into the impact of marine environmental regulations on the manufacturing industry's structural adjustment in China.

(Zakaria, Jun, and Ahmed 2019), (Ofoegbu, Akwu, and O 2016) and (Lapatinas 2019), investigated the effects of different factors on economic growth and development in different countries. While (Zakaria et al. 2019) investigated the effect of terrorism on the economic growth and development in Pakistan, (Ofoegbu et al. 2016) conducted an empirical study on the impact of tax revenue on Nigeria's economic development. (Lapatinas 2019),

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

on the hand, conducted an empirical investigation into the impact of the internet on economic sophistication in developing countries.

On image dataset, (Barbedo 2018), conducted an empirical analysis to determine the effect of dataset size and variety on deep learning and transfer learning effectiveness for plant disease classification. Likewise, (Dutta and Gros 2018), conducted an empirical study to assess the effect of deep learning architectural component choice and dataset size on a medical imaging task. There is a perceived relationship between online customer ratings for hotels and customer sentiments. This relationship was investigated by (Geetha, Singha, and Sinha 2017). Furthermore, the effects of the ML regression algorithm and sample size on individualized behavioral/cognitive prediction performance was investigated by (Cui and Gong 2018). Although many of these studies suggest to have investigated the effects on one or two independent variables on one dependent variable, other variables and controls were introduced during the course of the actual modeling and analysis.

3. Methodology

We propose the use of Linear Regression and Correlation in our experimental design. The justification for this is that it is useful for investigating response or dependent variable which is affected by one or more independent variables. We have two independent variables namely size of the training dataset and the number of features selected. Our goal is to investigate if it is possible to predict our dependent variable (in this case the accuracy) given these two independent variables.

3.1. Hypothesis Development

This study investigates how accuracies of ML models are affected by number of features selected and the size of training data. Accordingly, we propose the following null and alternate hypothesis:
H_0: The size of the dataset used for training have a positive effect on the accuracies obtained.
H_1: The number of features selected has a positive effect on the accuracies obtained.
The research model is explained in details in section 3.4.

3.2. Variable Design

For our experiment, we have two independent variables (size of the training dataset and the number of features selected) and one dependent variable (the accuracy obtained). We now provide a brief explanation of these variables.

1. **Accuracy:** Accuracy is defined as the ratio of correctly predicted observations to total observations. It is usually computed using the false positives and false negatives values observed from the dataset. These values obtained depends on many factors such as the ML algorithm used, and so on. For this reason, we choose accuracy as our dependent variables.

2. **Number of feature:** The datasets used in ML consists of features, which the ML models learn from in order to make predictions. Selecting features affects the performance of an ML algorithm depending on if the most relevant ones were selected or otherwise. We have chosen the number of features as one of the independent variables that could influence accuracy.

3. **Size of Training Data:** An ML model requires data to train it before it can be tested. There is a recommended percentage of the dataset that is required to be used for both training and testing. Some researches however do no adhere strictly to this recommendation. For this reason, we have chosen the size of the training data as our second independent variable that could influence the accuracy of an ML model.

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

3.3. Data Collection

We extracted data from sixteen (16) cloud security machine learning studies conducted in the last five years via simple random sampling initially. The selected studies specified the machine learning models built, the dataset used (for both training and testing), the number of features selected from the dataset as well as the accuracy obtained from the models. The studies selected and the data extracted are shown in Table 1. The dataset used by most of the studies are subsets of existing open source datasets, some of which contain hundreds of thousands of records to millions. We extract the exact number of dataset used in each study as reported. Also, the different datasets have different number of features associated with them. Whereas some studies maintain those exact features, others reduce that number with the aim of eliminating redundant features. We have extracted the number of features as reported in the studies. Table 1 shows the references of the studies selected and data extract from them.

Table 1. Data Extracted from Studies

Study Ref	Authors	Title	Publisher	Publication Type	Year	Number of Features	Training Dataset	Training Dataset ($\times 10^3$)	Accuracy
(Chiba et al. 2019)	Chiba et al.	Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms	Elsevier	Journal	2019	12	20,000	20	99.86
(Rabbani et al. 2020)	Rabbani et al.	A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing	Elsevier	Journal	2019	47	4,000	4	97
(Hizal et al. 2021)	Hizal et al.	A new Deep Learning Based Intrusion Detection System for Cloud Security	IEEE	Conference	2021	41	125973	125.97	99.88
(Alrawashdeh and Purdy 2017)	Alrawashdeh	Toward an Online Anomaly Intrusion Detection System Based on Deep Learning	IEEE	Conference	2016	122	145584	145.58	97.9
(Niyaz et al. 2015)	Niyaz et al.	A Deep Learning Approach for Network Intrusion Detection System	ICST	Conference	2016	121	125973	125.97	88.39
(Kumar et al. 2020)	Kumar et al.	Machine Learning based Malware Detection in Cloud Environment using Clustering Approach	IEEE	Conference	2020	139	1946	1.95	95.1
(Salman et al. 2017)	Salman et al.	Machine Learning for Anomaly Detection and Categorization in Multi-cloud Environments	IEEE	Conference	2017	18	82333	82.33	93.35
(Moraboena et al. 2020)	Shone et al.	A Deep Learning Approach to Network Intrusion Detection	IEEE	Journal	2018	41	494021	494.02	97.90
(Panker and Nissim 2021)	Panker et al.	Leveraging malicious behavior traces from volatile memory using machine learning methods for trusted unknown malware detection in Linux cloud environments	Elsevier	Journal	2021	26	21800	21.8	98.7
(Karatas et al. 2020)	Karatas et al.	Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset	IEEE	Journal	2020	41	286191	286.91	96.23
(Rashid et al. 2020)	Rashid et al.	Cyberattacks Detection in IoT-Based Smart City Applications Using Machine Learning Techniques	MPDI	Journal	2020	25	140272	140.27	94.2
(Abusitta et al. 2019)	Abusitta et al.	A deep learning approach for proactive multi-cloud cooperative intrusion detection system	Elsevier	Journal	2019	41	494021	494.02	92.5
(Nguyen et al. 2019)	Nguyen et al.	SeArch: A Collaborative and Intelligent NIDS Architecture for SDN-based Cloud IoT Networks	IEEE	Journal	2019	15	30000	30	96
(Singh et al. 2016)	Singh et al.	Collaborative IDS Framework for Cloud	Others	Journal	2016	41	148517	148.52	98.35
(Raju et al. 2021)	Raju et al.	Performance Enhancement of Intrusion Detection System Using Machine Learning Algorithms with Feature Selection	IEEE	Conference	2021	14	125973	125.97	76
(Verma and Ranga 2020)	Verma et al.	Machine Learning Based Intrusion Detection Systems for IoT Applications	Springer	Journal	2020	49	175341	175.34	96.73

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

In order to ensure that the data we collected do not contain outliers that may negatively affect our analysis, we adopted the box model in determining the outliers for our two variables, number of features and no of training data, as well as the dependent variable accuracy. For any outlier discovered, we discarded the entire record. The outliers discovered are shown in red color in Table 1.

Box Model for Number of Features

12 14 15 18 25 26 41 41 41 41 41 47 49 121 122 139

Median (m) = 41

Position of Q1 = .25(16 + 1) = 4.25

Position of Q3 = .75(16 + 1) = 12.75

Q1 = 18 + 0.25(25 – 18) = 19.

Q3 = 47 + 0.75(49 – 47) = 48.5

IQR = 48.5 – 19.75 = 28.75

Lower fence = 19.75 – 1.5(28.75) = – 23.375

Upper fence = 48.5 + 1.5(28.75) = 119.75

Figure 1 shows the Box Model for Number of Features

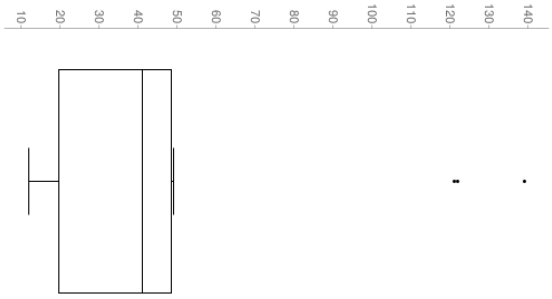

Figure 1: Box Model for Number of Features

Outliers = 121, 122, 139

Box Model for Size of training data

4 20 21.8 30 82.33 125.97 125.97 140.27 148.57 175.34 286.91 494.02 494.02

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Median (m) = 125.97

Position of Q1 = .25(13 + 1) = 3.5

Position of Q3 = .75(13 + 1) = 10.5

Q1 = 21.8 + 0.5(30 – 21.8) = 25.9

Q3 = 175.34 + 0.5(286.91 – 175.34) = 231.125

IQR = 231.125 – 25.9 = 205.225

Lower fence = 25.9 – 1.5(205.225) = – 282.8375

Upper fence = 231.125 + 1.5(205.225) = 538.9625

Fig. 2 shows the Box Model for Size of training data

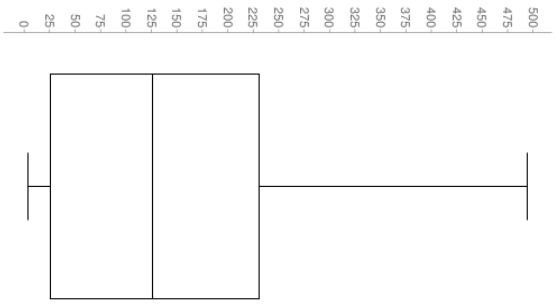

Figure 2. Box Model for Size of training data

Outliers = none

Box Model for Accuracy

76 92.5 93.35 94.2 96 96.23 96.73 97 97.9 98.35 98.7 99.86 99.88

Median(m) = 96.73

Position of Q1 = .25(13 + 1) = 3.5

Position of Q3 = .75(13 + 1) = 10.5

Q1 = 93.35 + 0.5(94.2 – 93.35) = 93.775

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Q3 = 98.35 + 0.5(98.7 – 98.35) = 98.525

IQR = 98.525 – 93.775 = 4.75

Lower fence = 25.9 – 1.5(205.225) = – 282.8375

Upper fence = 231.125 + 1.5(205.225) = 538.9625

Fig. 3 shows the Box Model for Accuracy

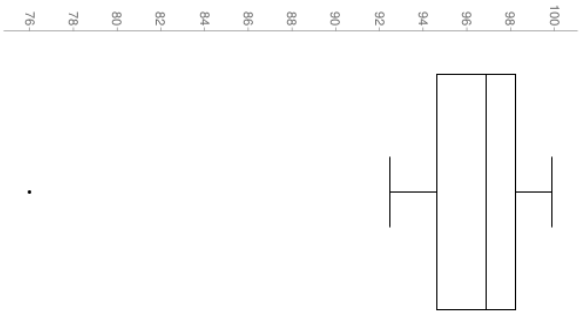

Figure 3. Box Model for Accuracy

Outliers = 76

3.4. Experimental Procedure

For our experiment, we have one dependent variable (Accuracy) and two independent variables (number of feature and number of training data). We want to see if the number of features selected and the number of training affect the value of the accuracy. As such, we propose our model as shown in equation 1.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e \ldots\ldots\ldots\ldots (1)$$

We calculate our bestfitting line and intercept using equations 2 and 3 respectively

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots\ldots (2)$$

$$\alpha = \bar{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} \ldots\ldots\ldots\ldots (3)$$

The slopes of our regression line for our two independent variables is computed with equations 4 and 5

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

$$\beta_1 = \frac{(\sum x_2^2)\ (\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \ \ldots\ldots\ldots (4)$$

$$\beta_2 = \frac{(\sum x_1^2)\ (\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \ \ldots\ldots . (5)$$

Equations 6 to 8 are used for computing the cross products of our independent variables with respect to the dependent variable.

$$\sum x_1 y = \sum x_1 y - \frac{(\sum x_1)(\sum y)}{N} \ \ldots\ldots\ldots\ldots\ldots\ldots . (6)$$

$$\sum x_2 y = \sum x_2 y - \frac{(\sum x_2)(\sum y)}{N} \ \ldots\ldots\ldots\ldots\ldots\ldots . (7)$$

$$\sum x_1 x_2 = \sum x_1 x_2 - \frac{(\sum x_1)(\sum x_2)}{N} \ \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Equations 9 to 11 are used to compute the correlation among the three variables.

$$r_{yx_1} = \frac{\sum x_1 y}{\sqrt{(\sum y^2)(\sum {x_1}^2)}} \ldots\ldots (9)$$

$$r_{yx_2} = \frac{\sum x_2 y}{\sqrt{(\sum y^2)(\sum {x_2}^2)}} \ldots\ldots (10)$$

$$r_{x_1 x_2} = \frac{\sum x_1 x_2}{\sqrt{(\sum {x_1}^2)\ (\sum {x_2}^2)}} \ldots (11)$$

In order to conduct our analysis, we construct the following table 2 showing the sum, mean, standard deviation and sum of squares of the dependent variables and independent variables.

Table 2. Sum, mean, standard deviation, cross product and sum of squares of all variables

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

	Accuracy	Number of Features	Number of Training Data			
	y	x_1	x_2	$x_1 \times y$	$x_2 \times y$	$x_1 \times x_2$
	99.86	12	20	1198.32	1996	240
	97	47	4	4559	388	188
	99.88	41	125.97	4095.08	12581.19	5164.77
	93.35	18	82.33	1680.3	7685.51	1481.94
	97.90	41	494.02	4013.9	48364.56	20254.82
	98.7	26	21.8	2566.2	2151.66	566.8
	96.23	41	286.91	3945.43	27609.35	11763.31
	94.2	25	140.27	2355	13213.43	3506.75
	92.5	41	494.02	3792.5	45696.85	20254.82
	96	15	30	1440	2880	450
	98.35	41	148.52	4032.35	14606.94	6089.32
	96.73	49	175.34	4739.77	16960.64	8591.66
SUM	1160.7	397	2023.18	38417.85	194134.13	78552.19
N	12	12	12	12	12	12
M	96.725	33.08	168.60	3201.49	16177.84	6546.02
SD	2.42	13.06	172.22	1272.45	16401.14	7369.28
SS	64.16	1874.92	326239.98			

4. Empirical Analysis and Result

In order to compute our regression coefficients and line, we compute the sums of squares, cross products and correlations among the variables using equations 4 to 11.

Using equations 6 to 8,

$$\sum x_1 y = 38417.85 - \frac{(397)(1160.7)}{12} = 18.025$$

$$\sum x_2 y = 194134.13 - \frac{(2023.18)(1160.7)}{12} = -1557.96$$

$$\sum x_1 x_2 = 78552.19 - \frac{(397)(2023.18)}{12} = 11618.65$$

Using equations 4 and 5 to compute the values of our regression coefficients,

$$\beta_1 = \frac{(326239.98)(18.025) - (11618.65)(-1557.96)}{(1874.92)(326239.98) - (11618.65)^2} = 0.05$$

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

$$\beta_2 = \frac{(1874.92)(-1557.96) - (11618.65)(18.025)}{(1874.92)(326239.98) - (11618.65)^2} = -0.007$$

We now compute the intercept using equation 3 as follows:

$$\alpha = 96.725 - (0.05)(33.08) - (-0.007)(168.60) = 96.25$$

Our bestfitting line therefore is

$$\hat{y} = 96.25 + 0.05x_1 - 0.007x_2$$

This translates to:

$$Accuracy = 96.25 + 0.05number\ of\ features - 0.007Size\ of\ Training\ Data \ldots (12)$$

Using equations 9 through 11, we compute the correlation among the variables as follows:

$$r_{yx_1} = \frac{18.025}{\sqrt{(64.16)(1874.92)}} = 0.05$$

$$r_{yx_2} = \frac{-1557.96}{\sqrt{(64.16)(326239.98)}} = -0.34$$

$$r_{x_{1x_2}} = \frac{11618.65}{\sqrt{(1874.92)(326239.98)}} = 0.47$$

5. Threats to Validity

The main limitation of this study include the number of studies selected and the period the study covers. Selecting more studies over a longer period may give a different result and we are exploring this in our future research. Other threats that may affect our findings and how we minimized them are highlighted as follows.

Conclusion Validity: This threat refers to a factor that can cause you to draw the wrong conclusion about a relationship based on a set of data. This study may suffer from insufficient data. To minimize this threat, we tried to select as many recent studies as we could find in order to generate sufficient data. We also ensured that the data are normally distribute in other not to affect the outcome of our experiment, hence leading to a wrong conclusion.

Internal Validity: This type of threat is concerned with factors that may have an impact on the causality factor. The data in this study was extracted from existing studies conducted in the last five years via simple random sampling. This way of selection may result in some bias. We addressed this threat by reducing the selection randomness as much as possible by selecting as many studies as we found that supplied the required information for our analysis.

External Validity: The extent to which the findings of a study can be applied to other situations, people, settings, and measures is known as external validity. The context of this study is ML models as applied to cloud security.

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Our findings may be generalized to ML models applied to other arears other than cloud security if they offer the same characteristics as the studies selected in this research.

6. Discussion

In this section, we provide a discussion of our results with respects to our two RQs.

6.1 RQ1: Does using a large training dataset and selecting more features guarantee improved accuracy?

From our result and analysis in section 4, we can see that there is a very weak positive relationship between accuracy and the number of features selected. Despite these weak relationships, we can see that based on our analysis, the more relevant features selected then the better the accuracy obtained. Figures 4 shows the scatter map for the correlation between accuracy and number of features selected.

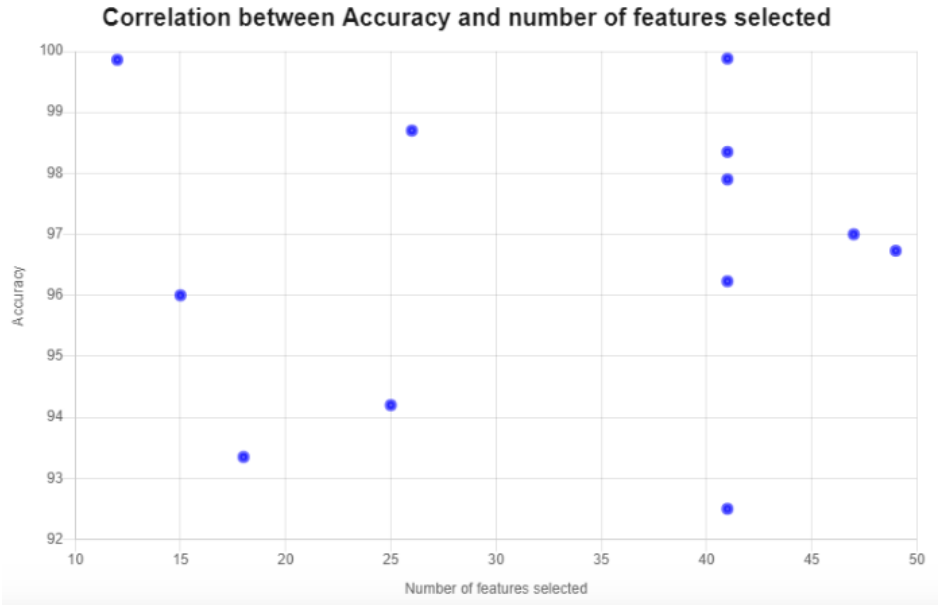

Figure 4. Correlation between Accuracy and Number of features

6.2 RQ2: Does selecting more features guarantee improved accuracy?

Also from our result and analysis in section 4, we can see that there is a weak negative relationship between the accuracy and the size of training data. Despite this weak relationship, we can see that training an ML model with a larger dataset results in improved accuracy. Figure 5 shows the scatter map for the correlation between accuracy and the size of the training data.

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

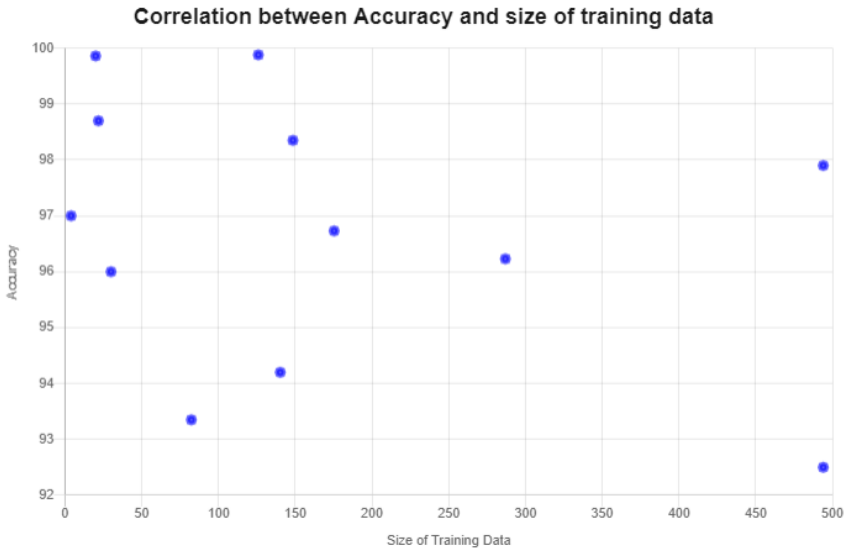

Figure 5 Correlation between Accuracy and Size of Training Data

6.3 Usefulness of the Model

Considering that the relationships among the dependent variables and the independent variables are weak, how useful then is the model we obtained in equation 12? One way to determine if our model (equation 12) is useful is to conduct an Analysis of Variance (ANOVA), to determine if we should reject our null hypothesis or not. After this, we then conduct an F-test to determine if our F-value falls within or outside this critical region at different significant levels.

For our Analysis of Variance, we need to compute the SSR as follows:

$$SSR = {\beta_1}^2 \sum {x_1}^2 + {\beta_2}^2 \sum {x_2}^2 + 2\beta_1\beta_2 \sum x_1x_2 \ldots\ldots\ldots eq.13$$

$$SSR = (0.05^2)(1874.92) + (-0.007^2)(326239.98) + (2 \times 0.05 \times -0.007)(11618.65)$$

$$SSR = 12.54$$

Table 3 shows our ANOVA.

Table 3. Anova Table

Source	Df	SS	MS	F
Regression	2	12.54	6.27	1.092
Error	9	51.62	5.74	
Total	11	64.16		

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 161-176
https://doi.org/10.56471/slujst.v4i.279
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

To see if our model is useful in predicting the accuracy of an ML model given the number of features selected and the size of the training data, we conduct an F-test. The result of the F-test at different significant levels is shown in table 4.

Table 4. F-test result

Significance Level	F-test critical value	Remark
0.01	8.02	Significant
0.05	4.26	Significant
0.1	3.01	Significant

We can see that despite the weak relationships among dependent and independent variables, our model is still quite useful since the F-value we obtained from the Analysis of Variance is not outside the critical region at different significant levels.

In conclusion, we fail to reject our H_0 and H_1. The number of features selected and the size of the training data can influence the accuracy of ML models.

7. Conclusion and Future work

This study investigated the influence of the number of features selected and the size of the training data on the accuracy of ML based approaches to the provision of security in the cloud using Ordinary Least Squares (OLS) regression. We initially selected 16 (sixteen) studies conducted in the last five years for our experiment. However, after identifying and discarding the outliers, we were left with 12 (twelve) studies. We extracted the necessary data and conducted our analysis on these 12 studies. Our analysis shows weak positive and negative relationships among the variables, our model is useful in predicting the accuracies of ML models given the number of features selected and the size of the training data. In the future, we proposed to select more studies that span over a longer period of time, and also to identify and add more independent variables in order to improve the usefulness of our model.

References

Abusitta, A., Bellaiche, M., Dagenais, M., & Halabi, T. (2019). A deep learning approach for proactive multi-cloud cooperative intrusion detection system. *Future Generation Computer Systems*, *98*, 308–318. https://doi.org/10.1016/j.future.2019.03.043

Alrawashdeh, K., & Purdy, C. (2017). Toward an online anomaly intrusion detection system based on deep learning. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 195–200. https://doi.org/10.1109/ICMLA.2016.167

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture*, *153*(July), 46–53. https://doi.org/10.1016/j.compag.2018.08.013

Chen, X., & Qian, W. (2020). Effect of marine environmental regulation on the industrial structure adjustment of manufacturing industry: An empirical analysis of China's eleven coastal provinces. *Marine Policy*, *113*(December 2019), 103797. https://doi.org/10.1016/j.marpol.2019.103797

Chiba, Z., Abghour, N., Moussaid, K., El omri, A., & Rida, M. (2019). Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. *Computers and Security*, *86*, 291–317. https://doi.org/10.1016/j.cose.2019.06.013

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage*, *178*(February), 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001

Díaz, E., Panach, J. I., Rueda, S., Ruiz, M., & Pator, O. (2021). Are requirements elicitation sessions influenced by participants' gender? An empirical experiment. *Science of Computer Programming*, *204*, 102595. https://doi.org/10.1016/j.scico.2020.102595

Dutta, S., & Gros, E. (2018). *Evaluation of the impact of deep learning architectural components selection and dataset size on a medical imaging task*. *1057911*(March 2018), 36. https://doi.org/10.1117/12.2293395

Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis. *Tourism Management*, *61*, 43–54. https://doi.org/10.1016/j.tourman.2016.12.022

Hizal, S., Cavusoglu, U., & Akgun, D. (2021). A new Deep Learning Based Intrusion Detection System for Cloud Security. *HORA 2021 - 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, 43–46. https://doi.org/10.1109/HORA52670.2021.9461285

S. Satyanarayana, "Cloud Computing : Saas," *GESJ Comput. Sci. Telecommun.*, vol. 4, no. 36, pp. 76–79, 2012.

Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset. *IEEE Access*, *8*, 32150–32162. https://doi.org/10.1109/ACCESS.2020.2973219

Khan, R., Othman, M., Madani, S. A., & Member, I. (2013). *A Survey of Mobile Cloud Computing Application Models*. 1–21.

Kumar, R., Sethi, K., Prajapati, N., Rout, R. R., & Bera, P. (2020). Machine Learning based Malware Detection in Cloud Environment using Clustering Approach. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*. https://doi.org/10.1109/ICCCNT49239.2020.9225627

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Lapatinas, A. (2019). The effect of the Internet on economic sophistication: An empirical analysis. *Economics Letters*, *174*, 35–38. https://doi.org/10.1016/j.econlet.2018.10.013

Moraboena, S., Ketepalli, G., & Ragam, P. (2020). A deep learning approach to network intrusion detection using deep autoencoder. *Revue d'Intelligence Artificielle*, *34*(4), 457–463. https://doi.org/10.18280/ria.340410

Nguyen, T. R. I. G. I. A., Phan, T. V, & Nguyen, B. T. (2019). *SeArch : A Collaborative and Intelligent NIDS Architecture for SDN-based Cloud IoT Networks*. *X*, 1–18. https://doi.org/10.1109/ACCESS.2019.2932438

Niyaz, Q., Sun, W., Javaid, A. Y., & Alam, M. (2015). A deep learning approach for network intrusion detection system. *EAI International Conference on Bio-Inspired Information and Communications Technologies (BICT)*. https://doi.org/10.4108/eai.3-12-2015.2262516

Ofoegbu, G. N., Akwu, D. O., & O, O. (2016). Empirical Analysis of Effect of Tax Revenue on Economic Development of Nigeria. *International Journal of Asian Social Science*, *6*(10), 604–613. https://doi.org/10.18488/journal.1/2016.6.10/1.10.604.613

Panker, T., & Nissim, N. (2021). Leveraging malicious behavior traces from volatile memory using machine learning methods for trusted unknown malware detection in Linux cloud environments. *Knowledge-Based Systems*, *226*, 107095. https://doi.org/10.1016/j.knosys.2021.107095

Rabbani, M., Wang, Y. L., Khoshkangini, R., Jelodar, H., Zhao, R., & Hu, P. (2020). A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing. *Journal of Network and Computer Applications*, *151*(October 2019). https://doi.org/10.1016/j.jnca.2019.102507

Raju, A. S., Rashid, M. M., & Sabrina, F. (2021). *Performance Enhancement of Intrusion Detection System Using Machine Learning Algorithms with Feature Selection*. 34–39. https://doi.org/10.1109/itnac53136.2021.9652151

Rashid, M. M., Kamruzzaman, J., Hassan, M. M., Imam, T., & Gordon, S. (2020). Cyberattacks detection in iot-based smart city applications using machine learning techniques. *International Journal of Environmental Research and Public Health*, *17*(24), 1–21. https://doi.org/10.3390/ijerph17249347

Salman, T., Bhamare, D., Erbad, A., Jain, R., & Samaka, M. (2017). Machine Learning for Anomaly Detection and Categorization in Multi-Cloud Environments. *Proceedings - 4th IEEE International Conference on Cyber Security and Cloud Computing, CSCloud 2017 and 3rd IEEE International Conference of Scalable and Smart Cloud, SSC 2017*, 97–103. https://doi.org/10.1109/CSCloud.2017.15

Singh, A., & Chatterjee, K. (2017). *Cloud security issues and challenges : A survey*. *79*(August 2016), 88–115. https://doi.org/10.1016/j.jnca.2016.11.027

Singh, D., Patel, D., Borisaniya, B., & Modi, C. (2016). Collaborative IDS framework for cloud. *International Journal of Network Security*, *18*(4), 699–709.

Tamilselvan, E. K. S. L. (2019). *ORIGINAL RESEARCH PAPER A focus on future cloud : machine learning-based cloud security*. 237–249.

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Verma, A., & Ranga, V. (2020). Machine Learning Based Intrusion Detection Systems for IoT Applications. *Wireless Personal Communications*, *111*(4), 2287–2310. https://doi.org/10.1007/s11277-019-06986-8

Zakaria, M., Jun, W., & Ahmed, H. (2019). Effect of terrorism on economic growth in Pakistan: an empirical analysis. *Economic Research-Ekonomska Istrazivanja* , *32*(1), 1794–1812. https://doi.org/10.1080/1331677X.2019.1638290