Improved Network Intrusion Detection System Using Hybridized Feature Selection Methods

Olakunle Titus FADEYI
LCU/PG/ 000824

Being a MSc Post-field Presentation Submitted to the Department of Computer Science, Faculty of Natural & Applied Sciences, Lead City University, Ibadan, Oyo State, Nigeria

In Partial Fulfilment of the Requirements for the Award of Master of Science Degree (MSc) in Computer Science and Information Science

2023

Certification

This is to certify that Olakunle Titus FADEYI with matriculation number LCU/PG/000824 carried out this research work titled "Development of an Improved Network Intrusion Detection System Using Hybridized Feature Selection Methods" in the Department of Computer Science, Faculty of Natural & Applied Sciences, Lead City University, Lead City University, Ibadan, Oyo State, for the award of Master of Science Degree (MSc) in Intrusion Detection and that this has not be previously submitted.

Dr. Ayoade A.M
(Supervisor)

Date

Dr. Sakpere W.
(Head of Department)

Date

Dedication

This research is dedicated to the Glory of the Lord Almighty who saw me through the realization and completion of the degree.

Acknowledgement

My greatest gratitude goes to the Almighty God who allay my fears and saw me through the journey of my pursing my Master degree in Lead City University and to accomplish it. I would like to show my appreciation to my previous supervisor in person of Dr. (Mrs.) Achimugu O., who was highly instrumental in the initial starting of this project. Her support, criticism and encouragement really pushed me in the initial commencement of this project. I would profoundly appreciate Dr Achimugu P.O., who was the *catalyst* to seeing me go beyond Chapter 1 write-up. Not leaving out Dr. Yara P.O who was the previous Co- supervisor, his concerns and his '*right push*' cannot be undermined in the continuation of this project to completion.

My sincere appreciation also goes to my current Supervisor in person of Dr. Ayoade A.M who truly I am highly indepted to as his efforts, criticism, encouragement and supervision which I can not undermined and underestimate. Sessions with him really made a difference in bringing this '*long tenure*' to a speedy end. My appreciation goes to various people such as my colleagues, friends etc who assisted me with advise, inputs and way forward while navigating the muddy waters while trying to complete the thesis when even the end of the tunnel hasn't been sighted.

I would like to appreciate my father in person of Engr. Fadeyi T.A who in the initial period of my starting this journey was very skeptical of the university and worried later of the long time it was taking to finish it. He was disappointed in my 'throwing in the towel' and his look (I would not forget that in a hurry) pushed me back to getting back on the saddle and see to its completion. I would be glad to show him I finished it after all. To my brother, Mr. Fadeyi A. who at the initial time was always on my neck to finish and move on to the next stage, thank you for not giving up on me.

Having pen down my appreciation, I hereby want to state that though the above-mentioned institutions and persons have assisted me in the process of this research work, I take full responsibility for the errors, if any, found in the work

Abstract

The usage of Machine Learning (ML) and Feature Selection have been implemented in the development of Intrusion Detection System (IDS). From the review of the literature, developing an effective IDS requires large amount of data with many features. Some of these features are not important in the operation of the IDS which slows down the detection of threats. Therefore in this thesis, an IDS which can detect threat, has reduced features and is able to obtain result was developed. Machine learning was incorporated in training of the model using three machine learning algorithm; hybrid decision trees, Naives Bayes (NB) and Random Forest (RF). This was categorized into 3; Dataset Loading and Preprocessing, improved Intrusion Detection System and testing and evaluating the developed system. These three stages saw the total number of columns to 143 in number, after some processes were carried out on it, such as the hot-encoding category features and the SelectKBest techniques which reduced the columns to 15 best columns. After the correlation matix was conducted on the final sub dataset, it shows that features with NaN values have zero correlation with other related features in each of the sub dataset. Features with near zero variance, missing values >25% and those that has high correlation between two numerical variables. With these features having minimal discriminatory power, they were therefore removed from both sub dataset. This reduced columns shows that logistic regression model built was approximately 0.8377, the accuracy score of the K-nearest model was approximately 0.7538, the accuracy score of the DecisionTreeClassifier model was approximately 0.8127, the accuracy score of the LinearSVC model was approximately 0.8101. The developed IDS using Feature Selection technique significantly improved the performance of the Network Intrusion Detection System towards learning accuracy, reduce learning time, and simplify learning results.

Keywords: Machine Learning, hybrid decision tree, hot-encoding category feature

Word Count: 295 words

Table of Content

Content		Page
Title Page		i
Certification		ii
Dedication		iii
Acknowledgement		iv
Abstract		vi
Table of Contents		vii
List of Tables		xii
List of Figures		xiii
List of Acronyms		xv
Chapter One: Introduction		
1.1	Background to the Study	1
1.2	Statement of the Problem	3
1.3	Aim and Objectives of the Study	4
1.4	Significance of the Study	4
1.5	Scope of the Study	5
1.6	Limitation of the Study	5
1.7	Operational Definition of Terms	6
1.8	Outline of the Thesis	7
	Endnotes	8
Chapter Two: Literature Review		
2.1	Intrusion Detection Systems	10
2.2	Principles of IDS	12

2.3	Basic Functions of IDS	13
2.4	Evaluation Metrics of IDS	15
2.5	Challenge of IDS	16
2.6	Network-Based IDS	18
2.6.1	Security Features of NIDS	18
2.6.2	Related Work on Network-Based IDS	20
2.6.3	Evaluation of Network-Based IDS	27
2.7	Host-Based IDS	27
2.7.1	Security Features of HIDS	28
2.7.2	Related Work on Host-Based IDS	29
2.7.3	Evaluation of Host-Based IDS	36
2.8	Network Behavioral Analysis	37
2.8.1	Security Features of Network Behavioral Analysis	38
2.8.2	Related Work on Network Behavioral Analysis	39
2.8.3	Evaluation of Network-Based Analysis	42
2.9	Intrusion Detection Methodologies	43
2.9.1	Signature-Based Model	44
2.9.1.1	Related Work on Signature-Based Model	44
2.9.1.2	Evaluation of Signature-Based Model	47
2.9.2	Anomaly-Based Model	48
2.9.2.1	Related Work on Anomaly-Based Model	49
2.9.2.2	Evaluation of Anomaly-Based Model	52
2.10	Intrusion Detection Approaches	53
2.10.1	Statistical-Based IDS	53

2.10.2	Rule-Based IDS	54
2.10.3	Heuristic-Based IDS	55
2.10.4	Pattern-Based IDS	55
2.10.5	Cloud-Based IDS	55
2.10.6	Machine Learning-Based IDS	56
2.11	ML-based NIDS Observation	60
2.11.1	ML/DL Approach used for NIDS	60
2.12	Software-defined Networking (SDN) Based NIDS	61
2.12.1	SDN Architecture and Applications	62
2.12.2	SDN-based NIDS Observation Using ML/DL	63
2.13	NIDS Using Artificial Intelligence/Machine Learning	65
2.14	General Evaluation of Intrusion Detection Systems	76
2.15	Datasets	77
2.16	KDDcup99	77
2.17	Kyoto 2006	78
2.18	NSL-KDD	78
2.19	UNSW-NB15	79
2.20	CICIDS2017	79
2.21	Summary of Gaps in Literature Reviewed	80
	Endnotes	82

Chapter Three: Methodology

3.1	Research Approach	95
3.2	System Design	96
3.3	Research Methods	99

3.3.1	Data Pre-processing	99
3.3.2	Feature Selection	100
3.4	Classifiers Used for the Model	100
3.4.1	Random Forest Classifier	100
3.4.2	Hybrid Decisions Tree	101
3.4.3	Naive Bayes Classifier	101
3.4.5	K-Nearest Neighbors	102
3.5	Cross Validation	104
3.6	Tools and Environment	105
3.6.1	Anaconda	106
3.6.2	Jupyter Notebook	106
3.6.3	Scikit-Learn	106
3.6.4	NumPy	106
3.6.5	Matplotlib	107
3.6.6	Pandas	107
3.6.7	System Configuration	107

Chapter Four: Result and Discussions of Findings

4.1	Overview of the Experiment	108
4.2	Experiment on the Dataset Loading and Preprocessing	108
4.3	Result of Feature Selection	114
4.4	Result of Training and Testing the Intrusion Detection Classifiers	117
4.4.1	Logistic Regression	118
4.4.2	K-Nearest Model	119
4.4.3	Decision Trees	119

4.4.4	Linear Support Vector Classification (Linear SVC)	120
4.4.5	Neural Network Model	121
4.4.6	Random Forest	122
4.5	The Improved Network Intrusion Detection System (ImNIDS)	123
4.6	Evaluation (Discussion of Result)	123
Chapter Five: Conclusion		
5.1	Summary of Findings	127
5.2	Conclusion	128
5.3	Contribution to Knowledge	128
5.4	Recommendation	128
5.5	Suggestion for Further Studies	129
Bibliography		130
Appendix		143
Source Code		144
Biodata		166
The University Compliance Certification		168

List of Tables

Table	Title	Page
2.1	Confusion Matrix	15
3.1	Feature Names and Data Types	97
3.2	Sample View of Train Dataset	98
3.3	Sample View of the Test Dataset	98

List of Figures

Figure	Title	Page
1.1	Internet Use in the U.S	144
2.1	Classification of Intrusion Detection System	12
3.1	Sequence of actions for the Hybrid model	96
3.4	Random Forest in Operation	101
3.5	Example of KNN	102
3.6	Cross Validation Process	104
3.7	System Flow Diagram	105
4.1	Snapshot of NSL-KDD Dataset Being Loaded Into the Jupyter Notebook	109
4.2 a	Snapshot of the Loaded Dataset from the First Column	109
4.2 b	Snapshot of the Loaded Dataset from the Last Column	110
4.3	Train Dataset Attack Type Classification	111
4.4	Test Dataset Attack Type Classification	111
4.5	Snapshot Showing the Modified Dataset from the Last Column	112
4.6	A Utility Function for Dummy Variable Creation	112
4.7	Application of Utility Function for Dummy Variable Creation	113
4.8	Result of One-Hot Encoding on the Categorical Features	113
4.9	The Final Pre-Processed Dataset	114
4.10	Correlation Matrix on Train Sub Dataset	115
4.11	Dropping the Irrelevant Features from the Training Sub Dataset	116
4.12	Selecting Best 15 features from the train sub dataset	116
4.13	Printing Out the Best 15 Features	117
4.14	Splitting the Dataset	118

4.15	Result of the Logistic Regression Model	118
4.16	Result of the K-Nearest Model	119
4.17	Result of The Decision Tree Model	120
4.18	Result of the Linear Svc Model	121
4.19	Result of the Neural Network Model	122
4.20	Result of the Random Forest Model	123
4.21	Classifiers Verses Accuracy Score	124
4.22	Detection of DOS Attack by ImNIDS	125
4.23	Detection of PROBE Attack by ImNIDS	126

List of Acronyms

Acronyms	Meaning
ABC	Artificial Bee Colony
AI	Artificial Intelligent
AIDS	Anomaly-based Intrusion Detection System
ANN	Artificial Neural Network
APT	Advanced Persistent Threats
BNID	Behaviour Based Network Intrusion Detection.
CNN	Cloud Network Node
CPDT	Correlation based Partial Decision Tree Algorithm
DDoS	Distributed Denial of Service Attack
DM	Data Mining
DNN	Deep Neural Network
DoS	Denial of Service
DTA	Decision Tree Algorithm
FN	False Negative
FP	False Positive
HIDS	Host-based Intrusion Detection System
IDPS	Intrusion Detection and Prevention System
IDS	Intrusion Detection System
IP	Internet Protocol
I-SiamIDS	Improved Siam-Intrusion Detection System
ML	Machine Learning

NBA	Network Behavioral Analysis
NBAS	Network Behavior Analysis System
NIDPS	Network-based Intrusion Detection and Prevention System
NIDS	Network-based Intrusion Detection System
NSL-KDD	Network Security Laboratory- Knowledge Discovery in Database
OS	Operation System
PCA	Principal Component Analysis
SDN	Software Defined Network
Siamese-CNN	Siamese Convolutional Neural Network
SMS	Short Message Services
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TN	True Negative
TP	True Positive
TVM	Tenant Virtual Machine
UDP	User Datagram Protocol
Vanilla-CNN	Vanilla Convolution Neural Network