AN INVESTIGATION INTO ANOMALY BASED NETWORK INTRUSION DETECTION USING SELECTED MACHINE LEARNING APPROACHES

ALAGBE, OLADAPO ABIODUN

(19PCH02040)

NOVEMBER, 2021

AN INVESTIGATION INTO ANOMALY BASED NETWORK INTRUSION DETECTION USING SELECTED MACHINE LEARNING APPROACHES

BY

ALAGBE, OLADAPO ABIODUN

(19PCH02040)

B.Tech. Computer Science, Ladoke Akintola University of Technology, Ogbomosho

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE (M.Sc.) DEGREE IN MANAGEMENT INFORMATION SYSTEMS IN THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY.

NOVEMBER, 2021

ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfilment of the requirements for the award of the degree of Master of Sciences in Management Information Systems in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria.

Mr. John A. Philip

(Secretary, School of Postgraduate Studies) Signature and Date

Prof. Akan B. Williams

(Dean, School of Postgraduate Studies) Signature and Date

DECLARATION

I, **ALAGBE, OLADAPO ABIODUN (19PCH02040)**, declare that this research was carried out by me under the supervision of Dr Isaac A. Odun-Ayo of the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria. I attest that the dissertation has not been presented either wholly or partially for the award of any degree elsewhere. All sources of data and scholarly information used in this dissertation are duly acknowledged.

ALAGBE, OLADAPO ABIODUN

Signature and Date

CERTIFICATION

We certify that this dissertation titled "**AN INVESTIGATION INTO ANOMALY BASED NETWORK INTRUSION DETECTION USING SELECTED MACHINE LEARNING APPROACHES**" is an original research work carried out by **ALAGBE, OLADAPO ABIODUN (19PCH02040)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria under the supervision of Dr Isaac A. Odun-Ayo. We have examined and found this work acceptable as part of the requirements for the award of Master of Science in Management Information Systems.

Dr. Isaac A. Odun-Ayo

(Supervisor) Signature and Date

Dr. Olufunke O. Oladipupo

(Head of Department) Signature and Date

Prof. Olusegun Folorunso

(External Examiner) Signature and Date

Prof. Akan B. Williams

(Dean, School of Postgraduate Studies) Signature and Date

DEDICATION

To God almighty, the source of all wisdom and strength, and without whose help none of these could have been accomplished.

ACKNOWLEDGEMENTS

I am deeply indebted to God almighty, the Author and Finisher of my faith, Jesus Christ, the one who gave His all for me, and the Holy Spirit, the Comforter and Spirit of Truth. Without them, I am nothing.

My deep and sincere gratitude also goes to the Chancellor, Dr David Oyedepo and the members of the Board of Reagents of Covenant University Ota for the vision and mission of the University.

My earnest gratitude goes to my supervisor, Dr. Isaac A. Odun-Ayo, for his support, fatherly advice, availability, and not giving up on me throughout this project.

I wish to express my appreciation to the Head of Department, Dr. Olufunke O. Oladipupo, for her motivation and advice towards completing this project.

Big thanks to the Post Graduate Coordinator, Dr. Aderonke A. Oni, for her support and encouragement. Special appreciation goes to the entire faculty and staff of the Computer and Information Sciences department for their encouragement and support and critiques, and contributions during the various presentations. May God bless and keep you all.

To my friends (Arch. Olufemi Adeniji, Arch. (Ms.) Oluwaseun Ogunlola, Arch. (Ms.) Oluwatoyin Olawale, Ms. Oreva, Arch. Olumide Oshikoya, Arch. (Ms.) Tunde Tumo, and Engr. Jeremiah Oyewole), thank you for being amazing friends. To my NITDA colleagues (Mr. Ayobami Shofadekan, Mr. Henry Ogbu, and Engr. Samson Oruma), thank you for being the best colleagues one could wish for. To my sweetheart (Temiloluwa Agbede), thank you for being my prayer partner and for believing in me even when I do not believe in myself.

I want to acknowledge my mother, Pastor (Mrs.) Florence B. Alagbe for her prayers, support, encouragement and belief in me, even during these past two years. To my brothers (Mr. Olayinka Alagbe and Mr. Olanrewaju Alagbe), thank you for your encouragement, support and endless prayers.

Last but certainly not least, I wish to specially acknowledge the National Information Technology Development Agency (NITDA) for their sponsorship and support during this entire program. Special thanks to Ms. Anastasia, our Scholarship Liaison Officer.

TABLE OF CONTENTS

CONTENT			Page
COVER PAGE			**i**
TITLE PAGE			**ii**
ACCEPTANCE			**iii**
DECLARATION			**iv**
CERTIFICATION			**v**
DEDICATION			**vi**
ACKNOWLEDGEMENTS			**vii**
LIST OF TABLES			**xii**
LIST OF FIGURES			**xiii**
ABBREVIATIONS			**xvii**
ABSTRACT			**xx**
CHAPTER ONE: INTRODUCTION			**1**
1.1	Background to the Study		1
1.2	Statement of the Problem		3
1.3	Aim and Objectives		4
1.4	Research Methodology		4
1.5	Significance of the Study		5
1.6	Scope of the Study		6
1.7	Organisation of the Dissertation		6
CHAPTER TWO: LITERATURE REVIEW			**7**
2.1	Introduction		7
2.2	Conceptual Review		8
	2.2.1	Anomalies	8
	2.2.2	Approaches to Anomaly Detection	10

2.2.3 Machine Learning and Deep Learning 13
2.2.4 ML/DL Based Anomaly Detection 28
2.2.5 ML/DL Anomaly Detection Results Categories 29
2.2.6 Categories of ML/DL Anomaly Detection Algorithms 30
2.2.7 Networking 32
2.2.8 Computer and Information Security 39
2.3 Methodological Review 56
2.3.1 Machine Learning and Deep Learning Process 56
2.3.2 Machine and Deep Learning Evaluation Metrics 59
2.3.3 Machine Learning and Deep Learning Frameworks 62
2.4 Application Review 63
2.4.1 Network Intrusion Detection Systems 63
2.4.2 Fraud Detection 64
2.4.3 Medical and Public Health Anomaly Detection 65
2.4.4 Equipment Failure Detection 66
2.4.5 User Entity Behaviour 66
2.4.6 Others 66
2.5 Related Works 67
2.5.1 Gaps in Literature 69
CHAPTER THREE: METHODOLOGY 73
3.1 Introduction 73
3.2 Existing Model 74
3.3 Implemented Model 74
3.4 Machine Learning/Deep Learning Methodology 75
3.4.1 Dataset Acquisition 75
3.4.2 Data Pre-processing 76

3.4.3 Feature Selection 76
3.4.4 Classifier and Model Design 77
3.4.5 Model Evaluation 78
3.5 Model Development Environment 78
3.5.1 Integrated Development Environment 78
3.5.2 Programming Languages 83
3.5.3 Packages and Libraries 84
3.6 Implementation 85
3.6.1 Exploratory Data Analysis 85
3.6.2 Data Analysis 86
3.6.3 Data Pre-processing 87
3.6.4 Feature Selection and Engineering 93
3.6.5 Model Creation and Evaluation 105
CHAPTER FOUR: RESULTS AND DISCUSSION 108
4.1 Introduction 108
4.2 Dataset Testbed Description 108
4.2.1 Dataset Description 114
4.3 Results 114
4.3.1 Models with Imbalanced Dataset 114
4.3.2 Models with Balanced Dataset 123
4.3.3 Models with Autoencoder and Balanced Dataset 131
4.4 Summary of Results from the Evaluation Metrics 143
4.4.1 Precision, Recall and F1-score 143
4.4.2 Macro-Average Summaries 143
4.5 Discussion of Findings 148
4.5.1 Model Conclusion 150

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION **151**

5.1 Summary 151

5.2 Conclusion 151

5.3 Contributions to Knowledge 152

5.4 Recommendations 152

REFERENCES **153**

APPENDIX **165**

Appendix A: Performance Charts for Attack Classes 165

Appendix B: Model Accuracy, AUC, Train and Predict Time Charts 173

LIST OF TABLES

Table	Title of Table	Page
1.1	Mapping of objectives and methodology	4
2.1	Taxonomy of Anomaly Detection Systems Approaches	12
2.2	Summary of selected cybersecurity intrusion detection datasets	24
2.3	Confusion Matrix	60
2.4	Gaps identified in literature	70
3.1	Class distribution of individual datasets before clean-up	90
3.2	Class distribution of individual datasets after clean-up	91
3.3	Distribution of the individual classes after initial pre-processing	92
3.4	Pre-data balancing attack classes frequency distribution	103
4.1	Attack distribution of the CIC-IDS2017 dataset	109
4.2	Packet Header Descriptors	110
4.3	Packet Label	110
4.4	Interarrival times	110
4.5	Network Identifiers	111
4.6	Flow timers	111
4.7	Packet Flag Features	111
4.8	Flow file descriptors	112
4.9	Sub-flow descriptors	113
4.10	Frequency distribution of attack classes post-data balancing	124
4.11	Summary of macro averages of Precision, Recall, f1-score, accuracy, AUC, training time and prediction time.	144
4.12	Precision score summary	145
4.13	Recall score summary	146
4.14	F1-score summary	147

LIST OF FIGURES

Figure	Title of Figure	Page
2.1	Conceptual overview of the content structure of chapter two	7
2.2	Relationship between AI, ML, and DL	14
2.3	Autoencoder network architecture	18
2.4	Recurrent Neural Network Structure	19
2.5	Typical DNN architecture	20
2.6	Typical Convolutional Neural Network (CNN)	20
2.7	Typical GAN architecture	21
2.8	Classical Intrusion Detection System	28
2.9	Cloud computing service model management responsibilities	35
2.10	CIA Triad and their attacks	41
2.11	The Lockheed Martin Cyber Kill Chain Model	50
2.12	Data science process workflow	56
2.13	Machine learning pipeline architecture	57
2.14	Deep learning pipeline architecture	57
3.1	Methodology workflow	73
3.2	Existing ML/DL NIDS Architecture	74
3.3	Implemented research model	75
3.4	Typical structure of an Autoencoder	76
3.5	COUSS algorithm	77
3.6	Local System information	79
3.7	Google Colab NVIDIA GPU Information	80
3.8	Google Colab Block Storage information	81
3.9	Google Colab CPU Architecture information	81
3.10	AWS EC2 CPU Architecture	82
3.11	AWS EC2 t3.2xlarge RAM information	83
3.12	Jupyter notebook IDE interface	83
3.13	Datasets’ ingestion process	86
3.14	List of features present in the dataset (column headings)	86
3.15	Code output for some features of the first five observations in the dataset	87

3.16	Code excerpt for identifying missing values	87
3.17	Distribution of data samples with missing values	88
3.18	Code excerpt for identifying and handling duplicate values	88
3.19	Distribution of data samples with duplicate values	89
3.20	Distribution of data samples with infinite values	89
3.21	Code excerpt for identifying infinite values	90
3.22	Dirty dataset vs cleaned dataset sizes	91
3.23	Code Sample for displaying class sample distribution	92
3.24	Code output displaying the datatypes of some of the features of the dataset	93
3.25	Output for `df.describe()` on the initial dataset	94
3.26	Code excerpt for displaying high collinearity among features	94
3.27	Output of `.corr()` command showing multicollinearity	95
3.28	Features with no values	95
3.29	Code excerpt for identifying highly collinear features	96
3.30	List of identified features to be dropped due to high collinearity	96
3.31	Code excerpt for variance filtering	97
3.32	Features determined to be zero variance features.	97
3.33	Features retained after running the variance filtering code	98
3.34	Bivariate plots of top five (5) features and the target label	98
3.35	Correlation matrix showing features with minimal collinearity to the target variable	99
3.36	Code excerpt for extracting features and target labels	100
3.37	Mapping of target classes to respective numerical values	100
3.38	Code excerpt for achieving dimensionality reduction using TSNE	101
3.39	TSNE visualisation of 2D representation of label clusters	102
3.40	Code excerpt to plot the sample distributions between Attack and Benign classes	102
3.41	Distribution of attack and benign categories against the total dataset	103
3.42	Pre-data balancing attack class distribution	104
3.43	Code excerpt for cleaned dataset split	104
3.44	The output of the cleaned dataset splitting process	105

4.1	Classification report, Accuracy score and AUC for the logistic regression model on imbalanced data	115
4.2	Confusion matrix of the logistic regression model on imbalanced data	115
4.3	ROC-AUC curve for the logistic regression model on imbalanced data	116
4.4	Classification report for the decision tree model on imbalanced data	116
4.5	Confusion matrix for the decision tree model on imbalanced data	117
4.6	ROC-AUC curve for decision tree on imbalanced data	117
4.7	Classification report for SVM model on imbalanced data	118
4.8	Confusion matrix for SVM model on imbalanced data	119
4.9	ROC-AUC curve for SVM model on imbalanced data	119
4.10	The architecture of the Multilayer Perceptron (MLP) model	120
4.11	Classification report for MLP model on imbalanced data	121
4.12	Confusion matrix for MLP model on imbalanced data	121
4.13	ROC-AUC curve for MLP model on imbalanced data	122
4.14	ROC-AUC summaries for all models on imbalanced data	122
4.15	Frequency of post-balancing attack classes	123
4.16	Confusion matrix for logistic regression model on balanced data	124
4.17	Classification report for logistic regression on balanced data	125
4.18	ROC-AUC curve for logistic regression on balanced data	125
4.19	Classification report for decision tree model on balanced data	126
4.20	Confusion matrix for decision tree model on balanced data	126
4.21	ROC-AUC curve for decision tree model on balanced data	127
4.22	Classification report for SVM model on balanced data	128
4.23	Confusion matrix for SVM model on balanced data	128
4.24	ROC-AUC curve for SVM model on balanced data	128
4.25	Multilayer Perceptron (MLP) model architecture for balanced data	129
4.26	Classification report for MLP model on balanced data	130
4.27	Confusion matrix for MLP model on balanced data	130
4.28	ROC-AUC curves for models on balanced data	131
4.29	Encoder architecture of the deep autoencoder (DAE) model for feature extraction	132
4.30	Decoder architecture of the deep autoencoder (DAE) architecture	133
4.31	Summary of autoencoder model	134

4.32	Train-test loss to determine the point of best fit in the autoencoder model	135
4.33	Encoder model for feature extraction	135
4.34	Display output of `X_train_encode` data	136
4.35	Display output of `X_test_encode` data	136
4.36	Classification report of LR model with balanced data and autoencoder	137
4.37	Confusion matrix of LR model with balanced data and autoencoder	137
4.38	ROC-AUC curve of LR model with autoencoder and balanced data	138
4.39	Confusion matrix of DT model with balanced data	138
4.40	Classification report of DT model with autoencoder and balanced data	139
4.41	ROC-AUC curve of DT model with autoencoder and balanced data	139
4.42	Stratified shuffle split of `X_train_encode` dataset for training SVM model	140
4.43	Classification report of SVM model using autoencoder and stratified balanced training data	141
4.44	Confusion matrix of SVM model using autoencoder and stratified balanced training data	141
4.45	ROC-AUC curve of SVM model using autoencoder and stratified balanced training data	142
4.46	Summaries of the ROC-AUC curves for all the models using autoencoder and balanced data	142

ABBREVIATIONS

ADS	Anomaly Detection System
AE	Autoencoder
ANN	Artificial Neural Networks
AUC	Area Under Curve
C2	Command and Control
CAGR	Compound Annual Growth Rate
CNN	Convolutional Neural Networks
CSP	Cloud Service Provider
DAE	Denoising Autoencoder
DBN	Deep Belief Networks
DDOS	Distributed Denial of Service
DHCP	Dynamic Host Configuration Protocol
DL	Deep Learning
DNS	Domain Name System
DOS	Denial of Service
DT	Decision Tree
EDLN	Ensemble of Deep Learning Networks
EL	Ensemble Learning
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GAN	Generative Adversarial Networks
GMM	Generative Mixture Models
HIDS	Host-based Intrusion Detection System
HTTP	Hypertext Transfer Protocol
HTTPS	Secure Hypertext Transfer Protocol
IaaS	Infrastructure-as-a-Service
ICS	Industrial Control Systems
IDS	Intrusion Detection System

IIoT	Industrial Internet of Things
IOA	Indicators of Attack
IOC	Indicators of Compromise
IoE	Internet of Everything
IoMT	Internet of Medical Things
IoT	Internet of Things
IPS	Intrusion Prevention System
IT	Information Technology
KNN	K-nearest Neighbour
M2M	Machine-to-Machine Communication
MCC	Matthew's Correlation Coefficient
MITM	Man-in-the-Middle attack
ML	Machine Learning
NADS	Network Anomaly Detection System
NB	Naïve Bayes
NFV	Network Function Virtualization
NIDS	Network Intrusion Detection System
NOC	Network Operations Centre
OT	Operational Technology
PaaS	Platform-as-a-Service
PAYG	Pay as You Go
PCA	Principal Component Analysis
PLC	Programmable Logic Controllers
R2L	Remote-to-Local
RAM	Random Access Memory
RF	Random Forest
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
SaaS	Software-as-a-Service
SAE	Sparse Autoencoder
SCADA	Supervisory Control and Data Acquisition
SDN	Software Defined Networking

SIEM	Security Information and Event Management
SLA	Service Level Agreement
SMTP	Simple Mail Transfer Protocol
SOAR	Security, Orchestration, Automation and Response
SOC	Security Operations Centre
SQL	Structured Query Language
SQLi	Structured Query Language Injection
SVM	Support Vector Machines
TCO	Total Cost of Ownership
TCP/IP	Transmission Control Protocol/Internet Protocol
TN	True Negative
TP	True Positive
U2R	User-to-Root
VAE	Variational Autoencoder
WLAN	Wireless Local Area Network
XSS	Cross Site Scripting

ABSTRACT

Early detection of attacks and indicators of compromise is critical in identifying and mitigating the actions of attackers and threat actors. Various approaches have been used to achieve prompt detection of such errant behaviours, all to varying degrees of success. Machine Learning (ML) techniques have been mainly successful in detecting activities within networks that deviate from expected patterns compared to other statistical approaches. However, these detection methods require further improvement due to their detection inconsistencies and high false alarm rates. This study presents a network anomaly detection model that utilises Deep Autoencoders (DAE) for feature extraction and machine learning techniques for classification. This model is capable of detecting various forms of network-based attacks. The CIC-IDS2017 dataset, which consists of different malware and attack categories as observed in modern networks, was used to train and evaluate the performances of various machine learning techniques, and the best performing technique was chosen. The methods evaluated include the Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) based models. These models were created using machine learning and deep learning workflows. The performances of the four (4) models were compared against each other, using iterations of the dataset that were imbalanced, balanced, and balanced with dimensionality reduction performed using the Deep Autoencoder. Based on a comparison with existing works, it was determined that the developed model performed comparatively well using metrics like the Receiver Operating Characteristics (ROC) Area Under Curve (AUC), Precision and Recall. The results obtained from the study indicates that the Decision Tree model outperforms other approaches explored.

Keywords: ***Anomaly Detection, Deep autoencoder, Logistic Regression, Support Vector Machine, Decision Tree, Multilayer Perceptron, Network Intrusion Detection***