

Comparative Performance Evaluation of Random Forest on Web-based Attacks

Oluwaseye Abayomi ADEYEMI

LCU/PG/001144

**Being a MSc Thesis Submitted to the Department of Computer Science, Faculty of
Natural and Applied Science, Lead City University, Ibadan, Oyo State, Nigeria**

**In Partial Fulfillment of the Requirements for the Award of Master of Science Degree
(MSc) in Computer Science**

2023

Certification

This is to certify that Oluwaseye Abayomi ADEYEMI with Matriculation Number LCU/PG/001144 carried out this research work title ‘Comparative Performance Evaluation of Random Forest on Web-based Attacks’ in the Department of Computer Science, Faculty of Natural and Applied Science, Lead City University, Ibadan, Oyo State, for the award of Master Degree (MSc) in Computer Science and that this has not been previously submitted.

.....

Dr. Ajani Azeed WAHEED

Supervisor

.....
Date

.....

Dr. Wilson SAKPERE

Head of Department

.....
Date

Do Not Copy, Lead City University, Nigeria

Dedication

This project work is dedicated to Almighty God, the author and finisher of my faith who in his infinity mercy has kept me thus far.

Do Not Copy, Lead City University, Nigeria

Acknowledgement

First and foremost, I want to express my gratitude to Almighty God for the commencement and completion of this Degree and for making this research work a success; all glory belongs to Him. Special thanks goes to my humble and hardworking supervisor, Dr. Ajani Azeez WAHEED, who is also the Postgraduate Coordinator of Computer Science Programme, for his tireless effort, effective guidance, constant encouragement, constructive criticism, and willingness to listen and provide suggestions whenever needed, all of which contributed to the success of this work. I also have to appreciate Head of Department Computer Science, Dr. Wilson SAKPERE, for his guidance, encouragement, and support, toward the success of this work. I am very grateful. My humble appreciation goes to all the academic and non-academic staff of the department for teaching, nurturing, and influencing knowledge that will be of extreme value to me in my future endeavours.

Unforgettably, I would love to thank my entire family, most especially my mother- Mrs. Ogunkemi Oluwafunmilayo, my wife- Mrs. Adeyemi Blessing, my son's- Adeyemi Eriifeoluwa, Adeyemi Ireayomide, for their prayers, love and unflinching support in all aspects of my life, you are wonderful. My appreciation will be incomplete without mentioning my brothers and sisters, colleagues, friends, and classmates (Computer Science M.Sc. Class of 2021/2022), who have contributed in one way or the other toward the success of this work. Thank you, and God bless you all.

“ Though the above-mentioned institutions and persons have assisted in the process of this research work, I alone stand responsible for the errors, if any, found in the work”.

Abstract

As human resources try to break into networks, control systems, and steal information with the help of expanding data communication paths and protocols, cyber intrusions are currently on the rise. The majority of typical online attack methods are thoroughly researched and documented. Countries, corporations, people, and vital infrastructures that depend on information technology for daily operations have suffered financial losses, the loss of personal information, and economic harm as a result of web-based intrusion. However, foreseeing an attack before it happens can aid in its prevention. This research proposes a predictive model for web-based attacks and a performance comparison of random forest with and without feature selection to secure the availability, integrity, and secrecy of networks, computer systems, and their data. The CIC-Bell-IDS2017 dataset, which includes typical and contemporary intrusion attacks, served as the raw data source for the proposed model. A python-based programming environment and interface for Anaconda Navigator, Jupyter Notebook, was used to create the predictive models. Performance evaluation and comparative analysis were conducted, and the results demonstrate that, once big data analytics (feature scaling and feature selection) were applied to the dataset, the models' prediction accuracies improved, creating a potential intrusion detection system. The outcome yielded excellent accuracy and model development times in both cases, with 97% and 98% precision for both sets and model development times of 35 seconds for the raw set and 15 seconds for the reduced set, which is an important factor when deploying machine learning models in a real-time setting. Random Forest is more computationally expensive than Correlation feature Selection-based classifiers, but having higher predictive accuracy, according to a comparison. Both of these methods work well and each has advantages and disadvantages. The use of big data analytics (PySpark) was found to help machine learning models perform better, resulting in better intrusion detection system.

Keywords: Web Based Attacks, Random Forest, Correlation Feature Selection,

Word Count: 300

Tables of Contents

Contents	Page
Certification	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Tables	ix
List of Figures	x
List of Acronyms	xi
Appendix	xii
Chapter One: Introduction	
1.1 Background to the Study	1
1.2 Statement of the Problem	5
1.3 Aim and Objectives of the Study	6
1.4 Scope of the Study	6
1.5 Significance of the Study	6
1.6 Limitation of the Study	7
1.7 Operational Definition of Terms	8
Endnotes	10
Chapter Two: Literature Review	
2.1 Conceptual Review	12
2.1.1 Predictive Model	12

2.1.2	Cyber Security	13
2.1.3	Web-Based Attacks	14
2.1.4	Random Forest Algorithm	15
2.2	Theoretical Framework	17
2.3	Review of Related Works	21
2.4	Summary of Literature Reviewed	54
	Endnotes	55

Chapter Three: Methodology

3.1	Overview of Research Approach	62
3.2	System Design	64
3.2.1	Data Collection and Description of Dataset	64
3.2.2	Pre-processing / Data Cleaning	65
3.2.2.1	Dropping Unwanted Column	66
3.2.2.2	Removing infinite value and replacing nan value or whitespace	66
3.2.2.3	Label Encoding	67
3.2.3	Data Splitting for Raw Dataset	67
3.2.4	Development of the Predictive Model Raw Dataset	67
3.2.5	Evaluation of the Raw Model	70
3.2.6	Feature Selection	71
3.2.7	Development of Predictive Model for Reduced Dataset	74
3.2.8	Evaluation of the Reduced Dataset	74
3.3	Requirement Specifications	74
3.3.1	Hardware Implementation Tools	74

3.3.2	Software Implementation Tools	74
3.4	Research Methods	75
3.4.1	Data Collection Methods	75
	Endnotes	76
Chapter Four: Results and Discussion of Findings		
4.1	Implementation and Evaluation	77
4.2	Predictive Model Development	77
4.3	Experimental Results	78
4.3.1	Performance Evaluation without Feature Selection	78
4.3.2	Performance Evaluation with Feature Selection	79
4.4	Discussion of Findings	80
Chapter Five: Conclusion		
5.1	Conclusion	82
5.2	Recommendation and Future Works	82
5.3	Contribution to Knowledge	83
Bibliography		85
Appendix		94
Bio-data		135
The University Compliance Certification		139

List of Tables

Table	Title	Page
3:1	Distribution of Attacks	59
3:2	Confusion Matrix for Binary Classification	64
4:1	Random Forest without Feature Selection for the Raw Dataset	73
4:2	Random Forest with Feature Selection for the Reduced Dataset	73
4:3	Comparative Analysis of the Performance Evaluation	74

Do Not Copy, Lead City University, Nigeria

List of Figures

Figures	Title	Page
2:1	Machine Learning Algorithm Types	18
2:2	SQL Injection Attack Web-Based Application Security	24
2:3	Percentage of Websites Vulnerable to different cyberattacks	25
3:1	System Architecture	57

Do Not Copy, Lead City University, Nigeria

List of Acronyms

Abbreviation	Meaning
A I	Artificial Intelligence
M L	Machine Learning
ISTR	Internet Security Treat Report
SNSes	Social Networking Services
RF	Random Forest
DT	Decision Tree
BOF	Buffer Over Flow
XSS	Cross Site Scripting
RSMT	Robust Software Modeling Tool
CFS	Correlation Feature Selection

Do Not Copy, Lead City University, Nigeria

List of Appendix

Appendix	Title	Page
Appendix A:	Initial Data Input and Modeling	88
Appendix B:	Exploratory Data Analysis, Data Processing and Model Training	97
Appendix C:	Correlation Analysis	101
Appendix D:	Export Model Data	121
Appendix E:	Model Fitting and Testing Predictions	124

Do Not Copy, Lead City University, Nigeria

Do Not Copy, Lead City University, Nigeria