

A NEW INTELLIGENT CLASSIFICATION MODEL TO DETECT PHISHING EMAILS

BY:

MAJED AYIASH ALENIZI

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF:

MASTER OF SCIENCE IN INFORMATION SECURITY

TO:

INFORMATION SECURITY DEPARTMENT

THE COLLEGE OF COMPUTER AND INFORMATION SECURITY  
NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES

MARCH 2018

SUPERVISED BY:

DR. HUSSEIN Y. ABU MANSOUR

NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES

# Table of Contents

Abstract .....	ii
المستخلص .....	iii
Dedication .....	iv
Acknowledgment .....	v
Table of Contents .....	vi
List of Figures .....	ix
List of Tables .....	x
List of Acronyms .....	xi
Chapter 1: Introduction .....	1
1.1 Motivation .....	1
1.2 Problem statement .....	5
1.3 Research objectives and scope .....	5
1.4 Thesis Contributions .....	6
1.4.1 Proposing a new hybrid feature selection method .....	6
1.4.2 Presenting new emergent classification models .....	7
1.4.3 Adapting the emergent classification models to Phishing email problem .....	7
1.4.4 Empirical study on phishing emails detection: .....	7
1.4.5 Intensive literature review on phishing emails classification .....	7
1.5 Research methodology .....	8
1.6 General structural design .....	9
1.7 Thesis outline .....	11
Chapter 2: Background and literature review .....	12
2.1 Background .....	12
2.1.1 Electronic mails artifacts .....	12
2.1.2 Social engineering: .....	15
2.2 Timeline of phishing .....	18
2.3 Common Phishing emails attacks .....	20
2.3.1 Deceptive phishing .....	20
2.3.2 Spear Phishing .....	20
2.3.3 Clone Phishing Emails .....	20
2.4 Phishing Email Cost reports .....	22
2.5 Phishing attack, a real case .....	23
2.6 Common Phishing emails detection approaches .....	23
2.6.1 Data Mining .....	25
2.7 Knowledge Discovery and Data mining (KDD) .....	28
2.8 Literature Review .....	31

2.8.1	Related work .....	31
2.8.2	Summary of related works approaches .....	40
2.8.3	Summary .....	42
Chapter 3:	The Proposed Feature Selection Method .....	43
3.1	Introduction.....	43
3.2	Current Feature selection methods .....	44
3.2.1	Wrapper Methods.....	45
3.2.2	Filter Methods .....	45
3.2.3	Embedded Methods .....	46
3.3	Impact of Feature selection on classification accuracy .....	46
3.4	The Proposed feature selection method .....	48
3.4.1	Information Gain Feature Selection Algorithms (IG):.....	49
3.4.2	Genetic Algorithm Feature selection approach (GA): .....	50
3.5	Evaluation and experimental results .....	52
3.5.1	Data collection .....	52
3.5.2	Tools .....	53
3.5.3	Classification Model .....	53
3.6	Evaluation and Results Discussion: .....	55
3.6.1	Results Discussion .....	55
3.7	Summary .....	58
Chapter 4:	The Adaptation of the Emergent Classification Model to Phishing Email Detection: A Case Study .....	59
4.1	Introduction.....	59
4.2	Phishing emails problems .....	59
4.3	Phishing email detection main phases .....	60
4.3.1	Pre-process phase: .....	60
4.3.2	Learning and forming the classifier system .....	61
4.3.3	Evaluation Phase .....	63
4.4	Empirical Study and Experiments .....	64
4.4.1	Data set collection.....	65
4.4.2	Feature Extraction .....	65
4.4.3	Information Gain.....	66
4.4.4	Genetic Algorithm .....	67
4.4.5	The emerged classification systems with the proposed feature selection method .....	70
4.5	Evaluation and experimentation results .....	70
4.6	Summary .....	73
Chapter 5:	Conclusions and future work .....	74
5.1	Conclusions.....	74
5.1.1	Proposing a new hybrid feature selection method .....	75

5.1.2 Presenting new emergent classification models.....	75
5.1.3 Adapting the emergent classification models to Phishing email problem.....	75
5.1.4 Empirical study on phishing emails detection .....	75
5.1.5 Intensive literature review on phishing emails classification .....	75
5.2 Future Work.....	76
Bibliography .....	77
Appendix.....	83

## List of Figures

Figure 1. 1 Unique phishing emails reports on half of 2017 .....	2
Figure 1. 2 Phishing life cycle .....	3
Figure 1. 3 General Architecture of Automatic email classification .....	4
Figure 1. 4 General Structural Design .....	9
Figure 1. 5 Phishing emails empirical study.....	10
Figure 2. 1 The process of sending emails.....	13
Figure 2. 2 Types of phishing .....	17
Figure 2. 3 Phishing emails life cycle (Suganya 2016) .....	18
Figure 2. 4 Targeted industry sectors in second quarter of 2017 according to APWG. ....	22
Figure 2. 5 Sequential structure of KDD model process. ....	28
Figure 2. 6 KDD processing model steps .....	29
Figure 3. 1 A General Framework of Feature Selection for Classification .....	47
Figure 3. 3 The Pseudocode of the Information Gain method is depicted .....	50
Figure 3. 4 The Pseudocode of the Genetic Algorithm method .....	51
Figure 3. 5 Pseudocode of KNN.....	54
Figure 3. 6 Process of KNN.....	54
Figure 3. 7 Relative accuracy.....	57
Figure 4. 1 JAVA statements .....	66
Figure 4. 2 Genetic algorithm work.....	69
Figure 4. 3 Ratio of False alarms triggered when IG selection method .....	72

## List of Tables

Table 2. 1 Worldwide Daily Email Traffic, 2017-2021.....	14
Table 2. 2 Worldwide Email User, 2017-2021 .....	14
Table 2. 3 Annual Phishing rates (APWG).....	19
Table 2. 4 Total number of unique phishing reports (campaigns) received, according to APWG... .....	19
Table 2. 5 Summary of related works approaches .....	40
Table 3. 1 Dataset collection details .....	52
Table 3. 2 Average accuracy of ten different data set.....	56
Table 3. 3 Number of attributes with IG and Proposed method .....	57
Table 4. 1 Feature description.....	67
Table 4. 2 Performance of the tested classifiers in measurements (precision, recall and accuracy rates).....	71
Table 4. 3 Empirical evaluation results of different machine learnings .....	72
Table 4. 4 Features of data Set in ARFF file .....	83

## List of Acronyms

Acronyms	Explanations
ANN	Artificial Neural Networks.
URL	Uniform Resource Locator.
KDD	Knowledge Discovery and Data Mining
DM	Data Mining
APWG	Anti-Phishing Work Group
FT	False True
FP	False Positive
TP	True Positive
TN	True Negative
IG	Information Gain
GA	Genetic Algorithm
KNN	k-Nearest Neighbors algorithm
TV	Television
E-mails	Electronic Mails
CC	Carbon Copy
MIME	Multipurpose Internet Mail Extension
PDF	Portable Document Format
USB	Universal Serial Bus
CD-ROMS	Compact Disc Read-Only Memory
PC	Personal Computer
URL	Universal Resource Locator
US	Untied States
HTML	Hypertext Markup Language
SA	Saudi Arabia
IP	Internet Protocol
SVM	Support Vector Machine
CWL	Confidence Weighted Liner
NN	Neural Network
WEKA	Waikato Environment for Knowledge Analysis
CSV	Comma Separated Values
SMO	Sequential Minimal Optimization
LR	Logistic Regression
CART	Classification And Regression Tree
BART	Bayesian Additive Regression Trees
RF	Random Forest

<b>Acronyms</b>	<b>Explanations</b>
MP	Multilayer Perceptron
MNN	Multilayer Neural network
SMS	Short Message Service
J48	Decision Trees algorithm
Rnd	Random Forest
PDENF	Phishing Dynamic Evolving Neural Fuzzy Framework
GP	Genetic Programming
PNN	Probabilistic Neural Net
GMDH	Group Method of Data Handling
DT	Decision Trees
MP	Multilayer Perceptron
MLP	Multilayer Perceptron
NB	Naive Bayes
ARFF	Attribute-Relation File Format