

**Characteristics based Detection of Internet worms using Combined
Machine Learning Methods and Worm Containment**

By

**S. Divya
(12PHCSF001)**

Supervisor

Dr. G.Padmavathi

**A Thesis Submitted to
Avinashilingam Institute for Home Science and Higher Education
for Women, Coimbatore - 641 043**

**In partial fulfilment of the requirements for the degree of
Doctor of philosophy in Computer Science**

June 2015

CERTIFICATE

This is to certify that the thesis entitled “**Characteristics based Detection of Internet worms using Combined Machine Learning Methods and Worm Containment**” submitted to the Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for the award of the degree of **Doctor of Philosophy in Computer Science**, is a record of original research work done by **S. Divya**, during the period of her study in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, under my supervision and guidance and the thesis has not formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or similar title to any candidate of any University.

**Signature of the
Head of the Department**

Signature of the Supervisor

Signature of the Dean

DECLARATION

I hereby declare that the thesis titled “**Characteristics based Detection of Internet worms using Combined Machine Learning Methods and Worm Containment**” is the result of investigations carried out by me in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, under the supervision and guidance of **Dr. G. Padmavathi**, Professor and Head, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, University, Coimbatore, and that it has not been submitted for the award of any Degree / Diploma / Associateship / Fellowship or similar title of any University or Institute.

Signature of the Supervisor

Signature of the Candidate

ACKNOWLEDGEMENT

First and foremost, I am extremely thankful to the **GOD ALMIGHTY** for his graces and blessings bestowed on me.

I would like to place on record my reverential gratitude to **Late Ayya Dr. T. S. Avinashilingam Avl.**, Founder, President and First Chancellor of Avinashilingam University for Women, Coimbatore for providing the temple of learning and I owe my sincere and humble gratitude to **Late Amma Dr. Rajammal P. Devadas Avl., M.A., M.Sc., Ph.D.** Former Chancellor, Avinashilingam University for Women, Coimbatore for their heavenly blessings.

I record my sincere thanks to **Dr. T. S. K. Meenakshi Sundaram**, Chancellor, Avinashilingam University for Women, Coimbatore for providing the infrastructure for the conduct of the research.

I express my sincere and heartfelt thanks to **Dr. Sheela Ramachandaran, M.Sc., PG. Dip., Ph.D.**, Vice Chancellor, Avinashilingam University for Women, Coimbatore for providing necessary facilities and resources for the successful completion of this research work.

I am gratefully indebted to **Dr. (Mrs.) A. Venmathi**, M.Sc., Dip. Ed., M.Phil., Ph.D., Registrar (i/c), Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for continuous help to carry out this research programme successfully.

I also record my thanks to **Dr. Mrs. Parvathi**, Dean, Faculty of Science, University for Women, Coimbatore for motivation, constant encouragement, care and advice throughout the research.

I feel extremely privileged and fortunate to have worked under the able guidance and professional supervision of my guide, **Dr. G.Padmavathi, M.Sc., M.Phil., Ph.D.**, Professor and Head, Department of Computer Science, Avinashilingam University for Women, Coimbatore. Her constant encouragements, motivation, valuable and constructive suggestions and untiring support have guided me to gain and explore deep knowledge in the research field.

She helped me to define my research goals and showed the way to achieve them. Her able guidance, systematic approach guided me to put my best possible efforts in completing my work and documentation. Her sympathetic friendly nature, timely counseling, willingness to help at anytime, anywhere and at any situation during the entire period of the study have molded this research into a reality.

My special thanks to **Dr. E. George Dharma Prakash Raj**, Department of Computer Science and Engineering, Bharathidasan University, Trichy, for providing encouraging and constructive feedback for my research work.

I record my gratitude to all the **Faculty members** of Department of Computer Science and **Staff Members** of Computer Center, Avinashilingam University for Women, Coimbatore for their encouragement and support.

I also thank **my entire family** for their patience, wishes, prayers, encouragement and constant support extended to me at all the times throughout my career. I also thank all my **Friends** for their support and encouragement.

S. Divya

ABSTRACT

With the rapid growth of Internet today, many Internet based applications are evolving. Internet is an open network accessed by all. The major challenge in Internet is security. Many attacks and vulnerabilities affect the network. Among the various attacks, Internet worms are vulnerable because they infect a large number of hosts within a short period of time and from that infected hosts they further initiate attacks like distributed denial-of-service, phishing and spyware through their propagation. Internet worms like “Code-Red” in 2001, “Slammer” in 2003, “Witty”/ “Sasser” in 2004, Storm in 2007, Conficker in 2008 and StuxNet in 2010-2012 have created prominent damages to the hosts. Within the period of five years, 4,00,000 computers got infected due to Blaster worm in 2003. Conficker worm damaged approximately 13 million IP addresses. This number may increase year by year. To overcome these damages and to defend against these attacks, effective defense mechanism is necessary. Therefore, detection and containment of Internet worms are the need of the day. For Internet Worm detection, there are two main approaches existing namely, signature based and anomaly based. Out of the two, anomaly based detection schemes provide better detection on newly appearing worms. There are various anomaly based detection approaches exist in the literature such as Probabilistic modeling, Spectrum based, Statistical estimation, Game theory, Epidemic spreading and Machine learning methods. Among these approaches, Machine Learning methods provide faster detection accuracy for rapidly changing Internet worms. Containment methods are applied to prevent the network from further infection after detection.

Based on the challenges created by Internet worms, the objectives of this research work are formulated after studying significant literatures. Though Internet worms are detected using different Machine Learning Approaches, the detection based on the characteristics of worms provides for better detection of new unknown worms. The characteristics refer to the nature of worms and it makes the detection effective at the initial stages of the propagation itself.

A Three-Step Methodology is proposed with four contributions to meet the objectives of the thesis. The Internet worm detection is done using the combined Machine Learning Methods based on anomaly detection schemes and Containment based on

blocking schemes. The proposed **Principal Component Analysis with Multiclass Support Vector Machine and Rabin Footprint Algorithm (PMR)** detects the Malcode existence in the downloaded programs based on unknown signatures. The detected and classified Malcode programs are contained to prevent from further infection. The proposed **Deterministic Finite Automata with Fuzzy Logic Classifier and Filter-Ary Sketch (DFD)** performs detection and containment of malicious contents in packets based on payload. The proposed **Enhanced C 4.5 Algorithm and Blacklist (ECB)** detects and contains the unused addresses based on illegal traffic. The proposed **kernelized Extreme Learning Machine with Automated Worm Containment Algorithm (kEA)** is used for detection and containment of malicious traffic from non-existing IP addresses based on connection attempt failures.

The proposed methods are implemented using Java NetBeans IDE 7.4 and Microsoft SQL Server. The parameters used for evaluation are Memory Utilization, Time Consumption, Precision Value, Recall Value, Accuracy, Detection Rate and Containment Rate. In contribution one, the detection accuracy achieved by the proposed **PMSVM** is improved by 13.57% and all the detected Malcode programs are blocked using **PMR** with 100% containment rate. The time taken to contain the detected programs is 200ms. In contribution two, the proposed **DDF** method achieved better detection accuracy with improved 0.23% and proposed **DDF** method achieved containment rate with 100% and the time consumed to block is 1300ms. In contribution three, the proposed **CPC** method for detection of illegal traffic achieved detection accuracy improved by 14.47%, and proposed **ECB** method provides containment with 100% of blocking all detected malicious traffic within 20 ms. In contribution four, the proposed **kELM** method achieved detection accuracy improved by 23.67%. Finally, the proposed **kEA** method blocks all the detected malicious IP addresses with 100% containment at the time span of 33ms. The four contributions based on combined Machine Learning Methods provide better detection and containment of newly appearing Internet worms entering the networks.

The proposed research methodology can be applied for other characteristics of Internet worms like hitlist, topological and web search target finding worms with polymorphic and metamorphic payload schemes. The proposed methodology can be integrated with the hardware devices at the Network Intrusion Detection System to handle real attacks affecting the network.

TABLE OF CONTENTS

Chapter No.	Title	Page No.
1	INTRODUCTION	1
	1.1. Worm Attack	2
	1.2. Categories of Worm	3
	1.3. Research Motivation	4
	1.4. Internet Worm Attack	5
	1.5. Types of Internet Worms	5
	1.6. Characteristics of Internet Worms	7
	1.7. Internet Worm Defense Mechanisms	9
	1.7.1. Detection Schemes	10
	1.7.2. Containment Schemes	11
	1.8. Problem Statement	12
	1.9. Objectives of the Thesis	12
	1.10. Significant Contributions of the Thesis	13
	1.11. Organization of the Thesis	15
	1.12. Chapter Summary	15
2	REVIEW OF LITERATURE	17
	2.1. Anomaly-based Methods other than Machine Learning Methods	18
	2.2. Anomaly-based Method using Machine Learning Methods	24
	2.2.1. Malcode detection	26
	2.2.2. Payload-based detection	27
	2.2.3. Illegal traffic detection	29
	2.2.4. Connection attempt failures detection	30
	2.3. Existing Containment Methods	35
	2.4. Observations due to literature	38
	2.5. Chapter Summary	39
3	PROPOSED METHODOLOGY	40
	3.1. Steps involved in the Proposed Methodology	41
	3.2. Specific Contributions of the Thesis	42
	3.3. Chapter Summary	46

Chapter No.	Title	Page No.
4	DETECTION AND CONTAINMENT OF SELF PROPAGATING MONOMORPHIC CHARACTERISTIC WORMS BASED ON UNKNOWN SIGNATURES USING THE PROPOSED PMR METHOD	47
	4.1. Introduction	48
	4.2. Steps of the Proposed Contribution One	48
	4.2.1. Analysis and Preprocessing	49
	4.2.1.1. Principal Component Analysis	50
	4.2.2. Detection and Classification of Internet Worms	51
	4.2.2.1. Multiclass Support Vector Machine	51
	4.2.2.2. Kernel Function	53
	4.2.2.3. Radical Basis Function Kernel	53
	4.2.2.4. Selective Sampling	54
	4.2.3. Containment of Internet Worms	55
	4.2.3.1. Rabin Footprint Algorithm for Malcode Containment	55
	4.3. Flow diagram of the Proposed Contribution One – PMR Method	56
	4.4. Steps involved in the Proposed PMR Method	58
	4.5. Pseudo code of PMR Method	59
	4.6. Experimental Setup and Results	60
	4.7. Chapter Summary	66
5	DETECTION AND CONTAINMENT OF SELF PROPAGATING MONOMORPHIC CHARACTERISTIC WORM BASED ON PAYLOAD INFORMATION USING THE PROPOSED DFF METHOD	67
	5.1. Introduction	68
	5.2. Steps of the Proposed Contribution Two	68
	5.2.1. Analysis and Preprocessing	70
	5.2.1.1. Regular Expression	70
	5.2.1.2. Patterns to split Regular Expressions	71
	5.2.2. Detection and Classification of Internet Worms	72
	5.2.2.1. Matching Regular Expression with DFA	72
	5.2.2.2. Deterministic Finite Automata (DFA)	73

Chapter No.	Title	Page No.
	5.2.2.3. Delayed-Dictionary Compression (DDC)	74
	5.2.2.3.1. Encoder	74
	5.2.2.3.2. Stateless Compression Algorithm	75
	5.2.2.3.3. Decoder	76
	5.2.2.4. DFA Matching to detect Payload	76
	5.2.2.5. Fuzzy Logic Classifier	78
	5.2.3. Containment of Internet Worms	79
	5.2.3.1. Filter-Ary Sketch Method	79
	5.3. Flow diagram of the Proposed Contribution Two – DFF Method	80
	5.4. Steps involved in the Proposed DFF Method	82
	5.5. Pseudo code of DFF Method	82
	5.6. Experimental Setup and Results	84
	5.7. Chapter Summary	88
6	DETECTION AND CONTAINMENT OF SECOND-CHANNEL PROPAGATING MONOMORPHIC CHARACTERISTIC WORM BASED ON ILLEGAL TRAFFIC USING THE PROPOSED ECB METHOD	89
	6.1. Introduction	90
	6.2. Steps of the Proposed Contribution Three	90
	6.2.1. Analysis and Preprocessing	92
	6.2.1.1. Attribute Vector Selection	92
	6.2.2. Detection and Classification of Internet Worms	93
	6.2.2.1. C4.5 Algorithm	93
	6.2.2.2. Entropy and Gain Ratio	94
	6.2.2.3. Pearson’s Correlation Coefficient	95
	6.2.3. Containment of Internet Worms	97
	6.3. Flow diagram of the Proposed Contribution Three – ECB Method	97
	6.4. Steps involved in the Proposed ECB Method	99
	6.5. Pseudo code of ECB Method	100
	6.6. Experimental Setup and Results	102
	6.7. Chapter Summary	106

Chapter No.	Title	Page No.
8	CONCLUSION AND FUTURE DIRECTIONS	124
	8.1. Summary and Conclusions	125
	8.2. Future Research Directions	126
	REFERENCES	128
	ANNEXURES	137
	Annexure I	137
	Annexure II	140
	Annexure III	143
	Annexure IV	146
	PUBLICATIONS	157

LIST OF FIGURES

Figure No.	Title	Page No.
1.1	Worm Propagation in the Network	2
1.2	Classification of Internet worms characteristics	7
1.3	Internet Worm Defense Scheme Classification	9
2.1	Anomaly-based Schemes for Internet Worm Detection	19
3.1	The Proposed Three-Step Methodology	41
3.2	Contributions of the Thesis	43
3.3	Consolidated view of the Proposed Methodology	44
3.4	Technical details of four research contributions based on the Three-Step Methodology	45
4.1	Block Diagram of the Proposed Contribution One	49
4.2	Flow diagram of the Proposed Contribution One – PMR Method	57
4.3	Experimentation Methodology for Contribution One	61
4.4	Comparison of Memory Utilization for Contribution One	64
4.5	Comparison of Time Consumption for Contribution One	64
4.6	Comparison of results for Precision Value for Contribution One	65
4.7	Comparison of results for Recall Value for Contribution One	65
4.8	Comparison of results for Accuracy for Contribution One	65
4.9	Results for Containment due to Contribution One	66
5.1	Block diagram of the Proposed Contribution Two	69
5.2	DDC- Encoder	75
5.3	Stateless Compression Algorithm for packet payload	75
5.4	DDC- Decoder	76
5.5	DFA illustrating Regular Expression	78
5.6	Graphical representation of the concept “range of truths”	78
5.7	Flow diagram of the Proposed Contribution Two – DFF Method	81
5.8	Experimental Methodology for Contribution Two	84

Figure No.	Title	Page No.
5.9	Comparison of Memory Utilization for Contribution Two	86
5.10	Comparison of Time Consumption for Contribution Two	86
5.11	Comparison of results for Precision Value for Contribution Two	86
5.12	Comparison of results for Recall Value for Contribution Two	86
5.13	Comparison of results for Accuracy for Contribution Two	87
5.14	Results for Containment due to Contribution Two	87
6.1	Block diagram of the Proposed Contribution	91
6.2.	Flow diagram of the Proposed Contribution Three – ECB Method	99
6.3.	Experimentation Methodology for Contribution Three	102
6.4	Comparison of Memory Utilization for Contribution Three	104
6.5.	Comparison of Time Consumption for Contribution Three	104
6.6.	Comparison of results for Precision Value for Contribution Three	104
6.7	Comparison of results for Recall Value for Contribution Three	104
6.8.	Comparison of results for Accuracy for Contribution Three	105
6.9	Results for Containment due to Contribution Three	105
7.1	Block Diagram of the Proposed Contribution Four	109
7.2	Flow diagram of the Proposed Contribution Four – kEA Method	116
7.3	Experimentation Methodology for Contribution Four	119
7.4.	Comparison of Memory utilization for Contribution Four	120
7.5	Comparison of Time Consumption for Contribution Four	120
7.6	Comparison of results for Precision Value for Contribution Four	121
7.7	Comparison of results for Recall Value for Contribution Four	121
7.8	Comparison of results for Accuracy for Contribution Four	121
7.9	Results for Containment due to Contribution Four	122

LIST OF TABLES

Table No.	Title	Page No.
1.1	Potential Damages caused by Internet Worms	4
1.2	Propagation Factors of Internet Worms	7
2.1	Existing Anomaly based Internet worm Detection Approaches	21
2.2	Existing Epidemic Spreading Models	23
2.3	Existing Machine Learning Approaches for Internet Worm Detection Based on Anomaly	31
2.4	Existing Containment Approaches	37
4.1	Algorithm of Multiclass SVM	53
4.2	Rabin Footprint Algorithm for Malcode Containment	56
4.3	Pseudocode of PMR Method	59
4.4.	Performance Comparison of Detection Results for Existing and Proposed PMSVM Method	63
5.1.	Patterns for Regular Expression	71
5.2.	Pseudocode of DFF Method	83
5.3	Performance Comparison of Detection Results for Existing and Proposed DDF Method	85
6.1	Attributes from Network Flow	92
6.2	Algorithm of C 4.5	94
6.3	Malicious IP Address Containment Algorithm	97
6.4	Pseudocode of ECB Method	101
6.5.	Performance Comparison of Detection Results for Existing and Proposed CPC Method	103
7.1	Failure types	111
7.2	Extracted Features from a snapshot of failures	112
7.3	Containment Algorithm	114
7.4	Pseudo code of kEA Method	118
7.5.	Performance Comparison of Detection Results for Existing and Proposed kELM Method	120

ACRONYMS

AL	Active Learning
ANN	Artificial Neural Network
ADS	Automated Defense System
AWC	Automated Worm Containment
BDDC	Basic Delayed Dictionary Compression
BFWC	Bloom Filters With Counters
CL	Connected List
CPC	C 4.5 with Pearson's correlation Coefficient
DDF	Delayed Dictionary Compression and Fuzzy Logic Classifier
DFE	Deterministic Finite Automata with Fuzzy Logic Classifier and Filter-Ary Sketch
DFA	Deterministic Finite Automaton
DDC	Delayed Dictionary Compression
DDF	Delayed Dictionary compression and Fuzzy Logic Classifier
DLAL	Detection and Location Algorithm against the Local-worm
DDOS	Distributed Denial of Service
ECB	Enhanced C 4.5 Algorithm and Blacklist
FAS	Filter-Ary Sketch
GFGS	General Frequent-common Gram Searching
IOT	Internet Of Things
IM	Instant Messaging
IFSEng	Iterative Feature Selection Engine
JSD	Jensen Shannon Divergence
KEA	kernelized Extreme Learning Machine with Automated Worm Containment Algorithm
kELM	kernelized Extreme Learning Machine
LPS	Local Preference Scanning
MDM	Mahalanobis Distance Map
MD	Mahalanobis Distance

McPAD	Multiple classifier Payload-based Anomaly Detector
MSE	Mean Square Error
MSVM	Multiclass Support Vector Machine
NFA	Non-deterministic Finite Automaton
OS	Operating System
OADS	Orchestration-Oriented Anomaly Detection System
PCC	Pearson Correlation Co-efficient
PL	Pending List
PMSVM	Principal Component Analysis with Multiclass Support Vector Machine
PCA	Principal Component Analysis
PMR	Principal Component Analysis with Multiclass Support Vector Machine and Rabin Footprint Algorithm
PMSVM	Principal component analysis with Multiclass Support Vector Machine
PSD	Power Spectral Density
RFA	Rabin Footprint Algorithm
RBF	Radical Basis Function
REP Tree	Reduced Error Pruning Tree
RePIDS	Real-time Payload-based Intrusion Detection System
RST	Rough Set Theory
RSWD	Rough Set Worm Detection
SLFN	Single Hidden Layer Feed Forward Network
SVM	Support Vector Machine
SFM	Spectral Flatness Measure
SEIS-V	Susceptible – Exposed – Infectious – Susceptible with Vaccination
SIDQV	Susceptible, Infected, Delayed, Quarantined, Vaccinated
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UNIDS	Unsupervised Network Intrusion Detection System
VEISV	Vulnerable – Exposed – Infectious – Secured – Vulnerable
WAW	Worm-Anti-Worm
ZASMIN	Zeroday-Attack Signature Management Infrastructure