Connection Attempt Failures

The connection attempt failure is stated when the connection is made to the unused IP addresses or to the closed target ports [68]. Within a short period time, if the host receives large number of connection failure packets, it is said that particular host is infected by Internet worm.

Various approaches are proposed by different authors [68][43] to detect the existence of Internet worms. They are divided into signature based and anomaly based detection schemes. Among the two methods, Anomaly based schemes provide better detection for the newly appearing worms in the Internet and its literatures are discussed in chapter 2. Under Anomaly based detection schemes, various methods are discussed and out of them Machine Learning Methods provide faster detection accuracy for rapidly changing Internet worms. Once the malicious activities of Internet worms are detected, it is to be stopped or blocked to recover the systems from further infection and is discussed in the containment section below.

1.7.2. Containment Schemes

After the detection scheme, the Internet worms existing in the network are found [68]. The detected worms need to be removed, to prevent the network. This process of preventing the systems from further infection through eliminating the detected worms is referred as containment. The containment of Internet Worms is performed through two different ways namely, slow down and blocking.

Slow Down

The initial and important step in containment is to slow down the spread of worms [68]. When the rate of spreading is very large, then the Internet worms infect a large number of systems in a short period of time. Hence, before the process of blocking, the important process to be performed is to slow down the spread of worms.

Blocking

The blocking is the step performed with slow-down process [43]. The system would have a threshold, when it is met first time, then the slow-down process would be performed. If the threshold attainment to the maximum again, which would infect the system, then the blocking concept is applied to stop the worm infecting the system.

The blocking is being generally categorized into two types namely, content blocking and address blocking.

Content Blocking

If the anomalous contents match the Malcode signatures or the irregular patterns, then that particular packets with malicious contents are blocked [43].

Address Blocking

The traffic created by the infected/victim host is identified and the traffic from that particular computer address is dropped [43]. Based on the character of Internet worms, the defense scheme is framed.

Based on the challenges created by Internet worms, it is necessary to create a defense mechanism to prevent the network from attack. With the discussed damages and losses in the *Research Motivation* section, the problem statement is stated.

1.8. Problem Statement

Given the propagation of Internet worms, device a defense mechanism to increase the detection and classification accuracy based on the worms characteristics and containment to prevent from further infection.

1.9. Objectives of the Thesis

With the above discussed security threats and damages in *Research Motivation* section, the objectives of this research work is formulated. The objectives are formed after the study of *Literature*, to overcome the existing methods limitations. The primary objective of the research work is to device a defense mechanism achieving better detection and containment of Internet worms. The secondary objectives of the thesis are:

- Improve the detection and classification accuracy
- Reduce the memory utilization
- Minimize time consumption
- Increase precision value
- Maximize recall value
- Enhance containment rate

A Three-Step Methodology is proposed with four contributions to meet the above objectives and they are discussed in following chapters. The significant contributions of the research are discussed below.

1.10. Significant Contributions of the Thesis

The contributions involved for detection and containment of Internet worms are based on

Contribution 1: Detection of malcode from the incoming executable files - ***PMR Method***
Contribution 2: Detection of malicious contents of packet payload - ***DFF Method***
Contribution 3: Detection of illegal traffic to unused addresses IP addresses - ***ECB Method***
Contribution 4: Detection of failures during connection attempts – ***kEA Method***

Contribution 1: Detection of malcode from the incoming executable files - PMR Method

In contribution one, ***PMR Method*** is proposed. PMR method is a combination of **P**rincipal Component Analysis, **M**ulticlass Support Vector Machine and **R**abin Footprint Algorithm. Here, the malcodes existing in the downloaded programs are detected and classified under vulnerable classes using **P**rincipal Component Analysis with **M**ulticlass Support Vector Machine (***PMSVM***) method. The classified Malcode packets are blocked using content blocking method named **R**abin **F**ootprint **A**lgorithm (***RFA***). The proposed ***PMR (*****PM**SVM with **R**FA***)*** method, blocks all the detected Malcode packets. The experiments are performed with the dataset collected from Internet. The results obtained during detection are compared with the existing Support Vector Machine (SVM) and further blocked. Though unknown Malcodes are blocked in contribution one, self-carried worms propagate and infect the network through packet payload information exploiting the vulnerable applications and are explained in contribution two.

Contribution 2: Detection of malicious contents of packet payload - DFF Method

In contribution two, ***DFF Method*** is proposed. DFF method is a combination of **D**eterministic Finite Automata, **F**uzzy Logic Classifier and **F**ilter-Ary Sketch. Using this method, repeated contents occurrence named *payloads* in the packets is detected and classified using **D**eterministic Finite Automata with **D**elayed Dictionary Compression and **F**uzzy Logic Classifier (***DDF***) method. The classified payload packets are blocked using

Filter-**A**ry **S**ketch (***FAS***). The proposed ***DFF (*****D**DF with **F**AS***)*** method, blocks all the detected payload packets. The experimentation is done using the collected dataset and the results obtained by proposed DDF is compared with existing *G*eneral *F*requent-common *G*ram *S*earching (*GFGS*) method. Payload detection has its limitation when the packet is encrypted. During those circumstances, Blind scan worms generate an arbitrary number of scans to unused addresses creating Botnet propagation and its detection has been explained in contribution three.

Contribution 3: Detection of illegal traffic to unused IP addresses - ECB Method

In contribution three, ***ECB Method*** is proposed. ECB method is a combination of **E**nhanced **C** 4.5 Algorithm and **B**lacklist method. This method detects the illegal traffic created by unused address using **C** 4.5 with **P**earson correlation **C**o-efficient (***CPC***) method. The classified malicious IP addresses are blocked using **B**lacklist method. The proposed **ECB (E**nhanced **C** 4.5 with **B**lacklist**)** method**,** blocks all malicious IP addresses. The experimentation is done using the collected dataset from web link. The results obtained for detection is compared with the existing Reduced Error Pruning Tree (REP Tree) method. There are also some few worms perform their transmission through the non-existing IP address and closed ports. During those transmissions, the connection attempt fails and are explained in contribution four.

Contribution 4: Detection of failures during connection attempts – kEA Method

In contribution four, ***kEA Method*** is proposed. kEA method is a combination of **k**ernelized **E**xtreme Learning Machine with **A**utomated Worm Containment Algorithm (**kEA**). The proposed method detects and classifies anomalous traffic based on connection attempts failures using **k**ernelized **E**xtreme **L**earning **M**achine (***kELM***) method. The classified malicious IP addresses are blocked using **A**utomated **W**orm **C**ontainment (***AWC***) steps. The proposed **kEA (k**ernelized **E**LM with **A**WC**)** method**,** blocks all malicious IP addresses. The experimentation is done using the collected dataset from web sources. The results obtained for detection is compared with the existing C 4.5 method.

The above four contributions perform the detection of Internet Worms based on Anomaly Schemes and containment based on Blocking Schemes. Moreover, the detection is done based on the Internet worm characteristics.

1.11. Organization of the Thesis

This thesis is mainly divided into eight chapters and is framed around the research objectives. The organization of the thesis is as follows.

Chapter 1 presented the basis for the research work.

Chapter 2 presents the related works on the Internet worm detection based on Anomaly schemes. Various containment methods used to block the Internet worms are discussed.

Chapter 3 describes the research design based on the ***Three-Step Methodology***. The chapter discusses the four different contributions proposed using three-step methodology.

Chapter 4 presents the detection of Malcode in the incoming programs based on Unknown Signatures and containment using ***PMR Method***. Experiments conducted and results obtained are presented.

Chapter 5 discusses the detection and containment of Internet Worms based on packet payloads using ***DFF method***. The results obtained through experiments are presented.

Chapter 6 presents the detection of unused addresses created by Internet worms based on illegal traffic and containment using ***ECB method***. Experimental results obtained are presented.

Chapter 7 describes the detection of malicious IP addresses based on Connection attempt failures and containment using ***kEA method***. The experiments performed with the different traffic traces and the results achieved are presented.

Chapter 8 provides the achievements of four contributions, its limitations and further research directions.

1.12. Chapter Summary

Internet worm attacks are causing serious and challenging threats in the network and communication security. Internet worms are malicious software that replicates and injects the systems within short period of time. This active propagation of Internet worms

leads to construction of bot or zombies in the network. This chapter discussed the various types of Internet worms. The characteristics of Internet worms are categorized based on target finding and transfer phases in worms' life cycle, since the worms will be active in Internet during these two phases based on their characteristics effective detection, the defense mechanism framed for the research worm is based on the characteristics on Internet worms. The objectives of the research are formulated and the contributions of the thesis are discussed.

The Review of Literature for the research work is discussed in next chapter.

CHAPTER 2

REVIEW OF LITERATURE

Internet worms are causing high security threats and heavy financial damages in the world for the past 20 years. To overcome these damages and to defend against these attacks, effective defense mechanism is necessary. In the defense mechanism detection and containment of Internet worms provides safe network.

For detection, there are two types of approaches existing namely, signature based and anomaly based. Among both detection schemes, anomaly based detection scheme provide better detection on newly appearing known and unknown worms. So, the research work is based on anomaly detection and this chapter discusses the study on existing anomaly-based approaches. Among various anomaly based approaches, Machine Learning Approaches provide the detection faster and hence some of the significant contributions in the literature on Machine Learning approaches for worm detection are discussed in this chapter. As the proposed methodology is framed with four contributions based on Unknown signature for Malcode detection, payload based, illegal traffic and connection attempt failures, the study is made in depth on these four factors only.

Not only the detection provides the secure network, the detected attacks have to be stopped or blocked from further infections. Basically, Containment approaches stop and block the anomalous infection in the network. So to provide effective defense for the network, containment is necessary and the study on containment approaches is also discussed in this chapter.

2.1. Anomaly-based Methods other than Machine Learning Methods

To defend the network against Internet worm attack, worm detection and containment scheme is applied. Detection approach is divided into signature and anomaly schemes based on their parameters. Among the two detection schemes, anomaly based provides better detection on newly appearing known and unknown worms. Some of the different anomaly-based detection methods used for detection of Internet worms is shown in figure.2.1 below.

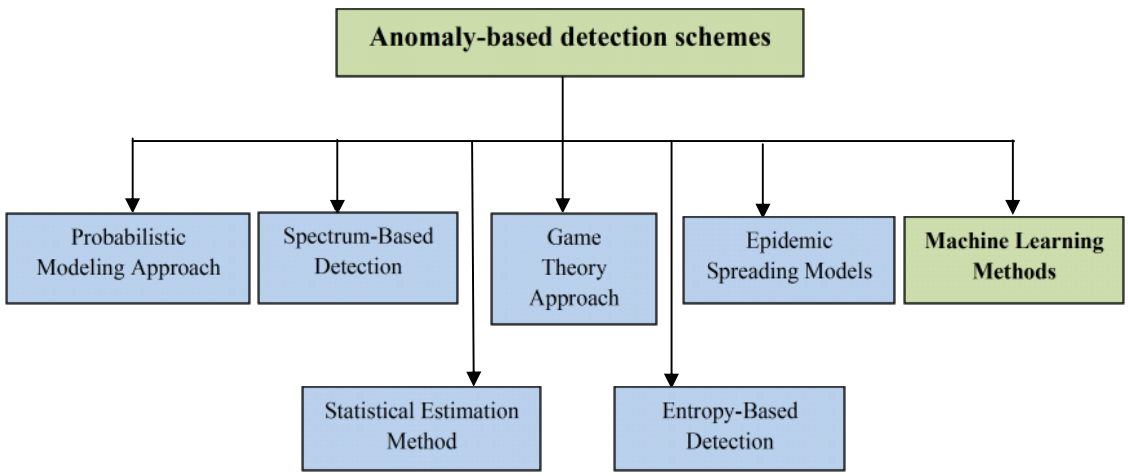

Figure.2.1. Anomaly-based Schemes for Internet Worm Detection

The various existing anomaly based detection approaches are discussed in detail below:

i) Probabilistic Modeling Approach

Quian Wang et al. [45] applied the probabilistic modeling method and a sequential growth model to classify the bot worms under infection family tree and detect the number of infected hosts. The mathematical analysis shows that the geometric distribution with parameter 0.5 is the line curve of the number of children in the worm family tree. The various scanning worms such as localized and permutation scanning worms are detected. The infected hosts are detected using estimated Poisson distribution parameter. The authors use this statistical estimation to detect and capture the infected hosts at the earlier propagation stage itself.

ii) Spectrum-Based Detection

Wei Yu et al. [61] proposed a spectrum-based detection scheme for the detection of C-worms. C-worm is one of the classes of Active worms. Using two domains, the traffic created by C-worms are detected; they are the time domain and frequency domain. In frequency domain, Power Spectral Density (PSD) and Spectral Flatness Measure (SFM) are applied for distinguishing normal traffic and abnormal worm traffic. The authors used

the closed-loop control with the spectrum-based detection to create fake traffic analysis, to free the worm attacker to propagate and then they are detected.

iii) Statistical Estimation Method

Qian Wang et al. [44] proposed a method based on moments, maximum likelihood and linear regression estimators in the statistical estimation method to detect the infection time of host and worm infection sequence. Mean Square Error (MSE) is used to find the host infection time and it is shown automatically. The proposed estimator is applied to various scanning worms like random scanning worms and localized scanning worms. The estimators detect infected zero hosts and recover the network.

iv) Game Theory

Wei Yu et al. [62] proposed a model that formulates the relationship between the propagator and the defender of worms named Game Theoretic formulation. A new type of worm named "Self-disciplinary worms" is discussed by the authors and the authors classified the self-disciplinary worms into static and dynamic. Threshold based scheme and trace-back schemes provide effective result for static self-disciplinary worms. The worm propagator is forced not to spread the worms if trace back interval $t_B > t_E$. To provide effective defense against dynamic self-disciplinary, integration of trace-back, threshold-based and spectrum-based schemes are framed. To stop the propagation, worm propagator stops its spread when $f_m^o(t_E - t_B) \leq m$ and the defender chooses $T_R = \infty$ and $T_m = T_m^0$.

v) Entropy-Based Detection

Wei Yu et al. [60] proposed an attack-target Distribution Entrophy-based Dynamic detection scheme, to defend against varying scan rate worms. To divide and categorize the worm traffic from non-worm traffic, attack-target distribution is applied with dynamic decision adoption. The authors build the varying scan rate worm model and developed the detection scheme using various statistical features like entropy of detection, data identity and anomalous behavior capture. Decision adaption is applied dynamically to achieve better detection. Table 2.1 below lists some of the existing anomaly based approaches for Internet worm detection.

Table.2.1. Existing Anomaly based Internet worm Detection Approaches

Year	Author	Countermeasure Schemes	Observations
2012	Qian Wang et al.	***Probabilistic Modeling*** Assessment Strategies for Forensic Analysis	Detected 9.10% of assessed bot ratio using random assessment method and achieved 22.36% of better bot detection.
2011	Wei Yu et al.	***Spectrum-based Scheme*** Detection Scheme	Achieved mean scan rate of 70/min with time and mean values 1239 and 1161 min. Maximal infection ratio of 0.03 is detected.
2011	Qian Wang	***Statistical Estimation Method*** Destination Detection and Defenses	MME achieved improved accuracy of 6.9%.
2011	Wei Yu et al.	***Entropy-based*** Detection	Achieved the detection time of 240 time units and the Maximum Infection Ratio of 0.004.
2010	Wei Yu et al.	***Game Theory*** Detection and Forensic analysis	Achieved increase in False Positive rate from 1% to 8% and decrease in Maximum infection rate from 36% to 12%.

vi) Epidemic Spreading Models

Bimal Kumar Mishra et al. [8] proposed a mathematical model and formulated Susceptible – Exposed – Infectious – Susceptible with Vaccination (SEIS-V) epidemic transmission model of worms in a computer network with natural death rate. Runge–Kutta Fehlberg fourth-fifth order methods are employed to solve and simulate the system. The modified reproductive number R_v is applied to measure the stability of the result. This proposed model proved that if reproduction number $R_v \leq 1$, the worm free equilibrium result is globally stable in the feasible region. Otherwise it is unstable. To analyze the antivirus software this model is highly useful.

Bimal Kumar Mishra and Samir Kumar Pandey [9] formulated the *e-epidemic SEIRS model,* for detecting the transmission of worms through vertical transmission in computer network. The authors used Runge–Kutta Fehlberg fourth–fifth order method to

analyze the system and notice the behavior of the susceptible, and infectious nodes with respect to time. If $R_0 \leq 1$, the worm free equilibrium is globally stable in the feasible region and the worms fade out from the network. Otherwise it is stable and the worms persist at a constant endemic level.

Fangwei Wang. [21] proposed a *SEIQV model*, an improved version of the SIR model for worm mitigation. The authors introduced five states and eight state transitions in the SEIQV model. The author considered the network-delay factor as important consideration which tends to cause the infection rate of active worms initially. The worm-free equilibrium point is examined. The basic reproduction number is derived from the set of differential equation which measures the guideline for effective worm defense. The quarantine rate λ depends on accuracy and detection speed of detection algorithm. Network delay and the vaccinated rate are observed as 0.011s and 179, 203, 247 s respectively.

Ossama et al. [39] proposed a *VEISV (vulnerable – exposed – infectious – secured – vulnerable) network worm attack model*. This model is used for determining the effects of security countermeasures on worm propagation. The model is applied to find the accurate positions for these functional hosts and their replacements in state transition. Using the reproduction rate, global stability of a worm-free state and local stability of a unique worm-epidemic state are derived. This model worked out the local stability for worm as R_0 = 2.62.

Xiaoming Wang [64] proposed a *EiSIRS model,* an expanded model of iSIRS. This model is studied for the dynamics of worm propagation. It maintains the sleep and work interrupting schedule policy for sensor nodes. The EiSIRS model is based on the epidemic theory and the differential equation. This epidemic theory describes the process of worm propagation in a WSN. The result is shown that the number of nodes in infectious working state changes with time when the communication radius of nodes is different and spreading threshold is set to λ= 714.

Yu Yao et al. [72] proposed a model called *SIDQV model(susceptible, infected, delayed, quarantined, vaccinated*). Worm detection considers the time delay, which is influenced by the size of the time window. A worm propagation model with time delay under quarantine defense is constructed. This time delay leads to Hopf bifurcation and

make the worm propagation system unstable. Thus, the analysis of Hopf bifurcation helps to ensure that the worm propagation system is stable and can help in the elimination of Internet worms. When the time delay τ remains less than τ0 by decreasing the window size of the IDS, the proposed model detects the worm effectively.

Yu Yao [73] proposed a method called the *pulse quarantine strategy* for worm control. In this strategy, a constant quarantine measure is used on the infectious anomaly detection method and then they are quarantined. The pulse quarantine strategy improves the efficiency of quarantine measure but also reduces the condition that worm propagation system must stabilize at the infection-free periodic equilibrium. The author proved that the basic reproduction number R_1 of the pulse quarantine model is less than one, which satisfies its stability condition.

Table.2.2 below lists the various existing epidemic models for Internet worm detection proposed by different authors.

Table.2.2. Existing Epidemic Spreading Models

Year	Author	Methods	Parameters used	Observations
2010	Fangwei Wang et al	SEIQV model	The quarantine rate, vaccinated rate	The quarantine rate λ depends on accuracy and detection speed of detection algorithm. Network delay and the vaccinated rate are 0.011s and 179, 203, 247 s.
2010	Xiaoming Wang	Epidemic theory	Infectious detection, spreading threshold	The result proved that the number of nodes in infectious working state changes with time when the communication radius of nodes is different and spreading threshold λ= 714.
2011	Bimal Kumar Mishra and Samir Kumar Pandey	Runge–KuttaFehlberg fourth–fifth order method	Unique Endemic Equilibrium	A unique endemic equilibrium is globally stable in the interior of the feasible region and the worms persist at a constant endemic level.

Year	Author	Methods	Parameters used	Observations
2012	Yu Yao	Constant Quarantine Measure	Infection-Free Periodic Equilibrium	The basic reproduction number R1 of the pulse quarantine model is less than one, which satisfies its stability condition
2012	Ossama et al	VEISV (vulnerable – exposed – infectious – secured – vulnerable)	Reproduction Rate	This model confirmed that local stability for worm is R_0 = 2.62 .
2013	Yu Yao et al	SIDQV model(susceptible, infected, delayed, quarantined, vaccinated)	Time delay	Time delay τ should remain less than τ_0 by decreasing the window size of the IDS, the proposed model detected the worm effectively.
2014	Bimal Kumar Mishra , Samir Kumar Pandey	Runge–KuttaFehlberg fourth-fifth order methods	Worm free Equilibrium	Proposed model proved that if reproduction number $R_v \leq 1$, the worm free equilibrium is globally stable in the feasible region and the worms fade out from the network, whereas if $R_v > 1$, the worm free equilibrium is unstable

Various anomaly based detection approaches discussed in the literature for Internet worm detection are discussed above. Among various approaches, the Machine Learning methods detects the rapid changing new worms quickly. The different Machine Learning methods proposed for Internet worm detection are discussed below.

2.2. Anomaly-based Method using Machine Learning Methods

Machine learning approaches detect the Internet worms quickly. Some of the existing approaches are discussed below:

Asaf Shabtai et al. [5] applied the machine learning classifiers on static features for the detection of malicious code. The classifiers are applied to learn patterns in the binary code files in order to classify new (unknown) files. A classifier is a rule-set that learns from

a given training set, which includes examples of both malicious and benign files. Support Vector Machine, Logistic Regression, Random Forest, Artificial Neural Networks, Decision Trees, Naive Bayes, BDT and BNB are the classification algorithms used. The highest performance is achieved with less than 33.3% in the training set, and accuracy above 95% is achieved.

Dima Stopel et al. [15] proposed an Artificial Neural Network (ANN) to detect the presence of computer worms based on measurements of computer behavior. The author defined the different features that can be measured in the computer during its operation. Trained supervised ANN is proposed with the known worms. After training, the binary patterns of hidden neuron outputs are extracted. With the extracted output, behavior of new pattern is designed to detect new worms. The average accuracy of detecting the unknown worm is 90%.

Igor Santos et al. [26] proposed an Anomaly based method to detect unknown malware using data mining algorithms. This method is based on the frequency of the appearance of opcode sequences. This method is used to extract the feature from the opcode sequences and several data-mining based classification algorithms are applied to detect unknown malware. These techniques are used for mining the relevant opcode and assess the frequency of each opcode sequence. The accuracy result is 95% and the true positive ratio is also 95%.

Nir Nissim et al. [37] proposed a framework based on new Active Learning (AL) methods (Exploitation and Combination) designed for acquiring unknown malware. This framework adopts for efficiently updating PC antivirus tools on a daily basis. The two AL methods are used to acquire a larger number of malware daily. Exploitation is based on SVM classifier principles, oriented towards selecting the examples that are probably the most malicious. The combination method lies between SVM-Margin and Exploitation. The detection rate of Exploitation is achieved with 2.6 times and 7.8 times more than the existing AL method. This work showed efficiency in daily improvement

Pedro Casas et al. [41] proposed UNIDS, an Unsupervised Network Intrusion Detection System capable of detecting unknown network attacks without using any kind of signatures, labeled traffic, or training model. UNIDS uses novel unsupervised outliers

detection (EA4RO) approach based on Subspace Clustering (DBSCAN) and Multiple Evidence Accumulation techniques to pin-point different kinds of network intrusion attacks and rank the degree of abnormality of traffic flows. Detection accuracy is 96% for Dos attack and 35% for NIDS system. Similarly probe detection accuracy is 100% and 87% in detecting R2L, U2R. More than 90% of the attacks can be correctly detected.

Robert Moskovitch et al. [48] proposed a worm detection scheme based on machine learning techniques. The author focuses on the feasibility of accurately detecting unknown worm activity in individual computers while minimizing the required set of features collected from the monitored computer. The mean detection accuracy exceeded 90%, and for specific unknown worms the accuracy reached above 99%, while maintaining a low level of false positive rate at 0.005.

Some of the Machine Learning Methods applied for Internet worm detection are Decision trees, Naïve Bayes, Artificial Neural Networks and Support Vector Machine classifiers.

Machine Learning Methods are also applied for Internet worm detection based on Unknown Signature for Malcode detection, Packet Payload, Illegal Traffic and Connection Attempts Failure. They are discussed below.

2.2.1. Malcode detection

Internet worms spread in the network in the form of malcodes inside the programs. To detect that character of Internet worms, some of the approaches are existing in the literature and are discussed below.

Asaf Shabtai et al. [6] presented a new pattern called OpCode n-gram patterns for the inspected files. Patterns extracted from the files after disassembly are used as features for the classification process. The classification process is used to detect the unknown malware within a set of suspected files and later included in antivirus software as signatures. Support Vector Machine, Logistic Regression, Random Forest, Artificial Neural Networks, Decision Trees, Naïve Bayes, BDT and BNB are used as classification algorithms. The method achieves an accuracy level higher than 96% with TPR above 0.95 and FPR approximately 0.1.

Nir Nissim et al. [36] proposed a conceptual method for detecting the unknown computer worm activity. Four feature-ranking measures were used to reduce the number of features required for classification. The support vector machine is applied to the resulting feature subsets. In addition, an active learning is used as a selective sampling method to increase the performance of the classifier and improves its robustness in the presence of misleading instances in the data. The method achieved mean detection accuracy greater than 90 %, accuracy above 94 % and low false-positive rate.

Robert Moskovitch et al. [49] presented a methodology for the detection of unknown malicious code. It examines the concept from text categorization, based on *n*-grams extraction from the binary code and feature selection. Artificial Neural Networks, Decision Trees, Naïve Bayes and their boosted versions, BDT and BNB, as well as SVM with three kernel functions are used as classification algorithms for classifying the malicious codes in this methodology. Proposed approach achieved greater than 95% accuracy.

Various approaches based on Unknown signatures for Malcode detection have been discussed and the next section discusses the payload based detection.

2.2.2. Payload based detection

Internet worms spreading in the form of payloads are detected by various authors and those existing methods under Machine Learning are discussed below.

Aruna Jamdagni et al. [3] proposed a 3-Tier Iterative Feature Selection Engine (IFSEng) for feature subspace selection. Principal Component Analysis (PCA) technique is used for the preprocessing of data. Mahalanobis Distance Map (MDM) is used to discover hidden correlations between the features and between the packets. The author also proposed a real-time Payload-based Intrusion Detection System (RePIDS) that integrates a 3-Tier IFSEng and the MDM approach. Mahalanobis Distance (MD) dissimilarity criterion is used to classify each packet as either a normal or an attack packet. The method has achieved better F-values at 0.9958 for DARPA 99 dataset and 0.976 for Georgia Institute of Technology dataset respectively with 0.85% false alarm rate and 1.3 times higher throughput.

A.N.M. Ehtesham Rafiq et al. [7] proposed a string search algorithm suitable for the hardware of the deep packet classification. The proposed algorithm is based on modified Boyer–Moore algorithm, and it requires a reduced number of operations. This algorithm finds all the malicious occurrences of a pattern in the text. The time complexity of this proposed work is O(n).

Irfan Ahmed and Kyung-Suk Lhee [24] proposed a content-classification scheme that analyses the executable contents in the incoming packets. It first analyzes the packet payload to see if it contains multimedia-type data (such as avi, wmv, jpg). If not, then it classifies the payload either as text-type (such as text, jsp, asp) or executable. The classification of contents is done using Distance-based algorithm which contain learning and identification algorithms. The proposed scheme has achieved low rate of false negatives with 4.69% and 2.53% of false positives.

Ning Weng et al. [35] developed a pattern matching engine for deep packet pre-filtering and finite state encoding. The author applied the Memory Efficient Pattern Match, State Collapsing, DFA splitting and Character-aware state coding. It achieves better performance in terms of speed (1000 times), by utilizing deep packet pre-filtering and novel finite state encoding.

Paul C. Van Oorschot [42] proposed a monitoring system which detects repeated packets in network traffic. It uses Bloom Filters With Counters (BFWC). It filters out a small number of common pre identified non-worm repeated packets in the dataset of network traffic, such that they are not processed by the BFWC. The system analyzes traffic in routers on a network. The false alarm rate is less than 15 packets per second because the Bloom-table is reset for every second.

Roberto Perdisci et al. [47] proposed a McPAD (Multiple classifier Payload-based Anomaly Detector), a new accurate payload-based anomaly detection system that consists of an ensemble of one-class classifiers. It combined multiple one class SVM classifiers that are based on descriptions of the patterns in different feature space using majority voting rule. The combination of classifiers is trained on different feature spaces to effectively exploit the different pattern representations. The method has achieved 100% accurate

detection of network attacks and bears shell-code in the malicious payload by increasing the Bayesian detection rate.

Wen-Chen Sun et al. [59] proposed a novel Rough Set Worm Detection (RSWD) scheme which extends well developed Rough Set Theory (RST) to detect zero-day polymorphic worms. It provide a minimum set of filtering rules to network barrier equipments, such as a firewall, to block the worm spreading. In this scheme, all attack packets are generated from some specific worm program and attack the same vulnerability of the victim hosts. This scheme contains clustering adjacency vector algorithm for each arriving packet to save the RST computation effort and has matching value to classify possible worm traffic before performing rule induction. It has suspicious cluster step to avoid false alarm yield by flash crowds. The RSWD module could detect the worm propagation within 17 seconds and produce a precise blocking rule by exhibiting 100% true positive rate and 99.82% accuracy rate.

The approaches proposed for Internet worm detection based on payload has been discussed above and the next section discusses the illegal traffic based detection.

2.2.3. Illegal traffic detection

The Internet worms propagating in the network and creating illegal traffic from unused addresses are detected by existing approaches and some of them detected using Machine learning are discussed below.

Eduardo Feitosa et al. [16] developed an Orchestration-oriented Anomaly Detection System (Heuristic orchestration algorithm) to detect attacks (coordinated or not), intrusions and anomalies at an earlier stage. It is based on controlled collaboration among different techniques to provide an effective anomaly detection system. OADS approach makes decision by achieving more than 95% of alerts.

Kim et al. [23] developed a ZASMIN (Zeroday-Attack Signature Management Infrastructure) system for network attack detection. This system provides early warning at the moment the attacks start to spread on the network. To block the spread of the cyber attacks, the system is automatically generating a signature that could be used by the network security appliance such as IPS. This system is suitable for suspicious traffic

monitoring, attack validation, polymorphic worm recognition, signature generation for unknown network attack detection. ZASMIN system has attained a low false rate.

Noriaki Kamiyama et al. [38] proposed an optimum design method of cache capacity and location for minimizing the amount of P2P traffic within the networks of transit ISPs when they allocate caches over their networks. By applying the proposed method to 31 backbone networks of actual commercial transit ISPs, the authors determine cache efficiency and clarify the network topological structure for which caches are effective in reducing P2P traffic. Transit ISPs reduces P2P traffic within its network by about 50% to 85%.

The next section discusses the detection of Internet worm traffic based on the failures during the connection attempt

2.2.4. Connection attempt failures

Internet worm detection based on connection failures are detected by various proposed approaches and some of them detected through Machine Learning are discussed below.

Asaf Shabtai et al. [4] proposed a Knowledge-based Temporal Abstraction for detecting previously unencountered instances of known malicious classes based on their temporal behavior. Time-stamped security data are continuously monitored within the target computer system or network and then processed. Automatically-generated temporal abstractions are monitored to detect suspicious temporal patterns. These patterns are compatible with a set of predefined classes of malware as defined by a security expert employing a set of time and value constraints. The scan rate was set at 5 probes or 10 probes, and the vulnerable population size at 15% or 37%. The scanning rate for detecting worm instances is 420 at 5 probes or 1,200 at 10 probes.

Chun-Ying Huang et al. [13] provided an effective solution to detect bot hosts within a monitored local network using c4.5 algorithm. A bot has a differentiable failure pattern because of the botnet-distributed design and implementation. Hence, by monitoring failures generated by a single host for a short period, it is possible to determine whether the host is a bot or not by using a well-trained model. The approach detects bot hosts with more than 99% accuracy and false positive rate at 0.5%.

Syed Ali Khayam et al. [54] proposed a two Joint Network-Host based anomaly detection techniques that detect self-propagating malware in real-time by observing deviations from a behavioral model derived from a benign data profile. This malware detection technique is employed perturbations in the distribution of keystrokes that are used to initiate network sessions. Higher accuracy with almost 100% detection rates and very low false alarm rates are achieved.

Yang XinYu et al. [67] proposed a detection and location method, DLAL (Detection and Location Algorithm against the Local-worm) to detect the local-worm. This method can respectively locate the high-speed worm and the low-speed worm, according to their different scanning rates. Network traffic is generated by scanning at the rate of 40 packetes/sec and 20 packets/sec. Ordinate represents the traffic, which is expressed by the number of packets per 0.1 sec. The average rate after scanning is higher than the normal traffic. Table.2.3 below lists the different detection approaches by various authors for the detection of Internet worms based on Unknown Signatures for Malcode Detection, Payload, illegal traffic and connection attempt failures.

Table.2.3. Exisitng Machine Learning Approaches for Internet Worm Detection Based on Anomaly

Year	Authors	Methods used	Parameters used	Observations
		Malcode-classification		
2012	Asaf Shabtai et al.	Support Vector Machine, Logistic Regression	Accuracy, TPR and FPR	Achieved accuracy level higher than 96% with TPR above 0.95 and FPR approximately 0.1.
2012	Nir Nissim et al.	Random Forest, Artificial Neural Networks, Decision Trees, Naïve Bayes, BDT and BNB	Detection Accuracy	Mean detection accuracy in excess of 90 %, an accuracy above 94 % low false positive rate is achieved

Year	Authors	Methods used	Parameters used	Observations
2012	Robert Moskovitch	Support Vector Machine, Artificial Neural Networks, Decision Trees, Naïve Bayes and their boosted versions, BDT and BNB, as well as SVM with three kernel functions	False-Positive Rate, Accuracy	Greater than 95% accuracy through use of training set that has malicious file content of less than 33.3% is achieved.
		Packet payload		
2004	EhteshamRafiq et al	String Search Algorithm, Boyer Moore Algorithm	Packet classification	Time complexity of this proposed work is O(n).
2006	Paul C. van Oorschot	Bloom Filters With Counters(BFWC)	False alarm rate	False alarm rate is less than 15 packets per second because its Bloom-table is reset every second.
2009	Wen-Chen Sun et al	Rough set worm detection (RSWD) scheme	True positive rate	Detects the worm propagation within 17 seconds and produce a precise blocking rule exhibiting 100% true positive rate and 99.82% accuracy rate.
2009	Roberto Perdisci et al	SVM classifiers	Accuracy, Bayesian detection rate	Achieves100% accurate detection of network attacks that bear shell-code in the malicious payload by increase the Bayesian detection rate
2011	Irfan Ahmed and Kyung-sukLhee	Distance-based algorithm	False Negative Rate	The proposed scheme has low rate of false

Year	Authors	Methods used	Parameters used	Observations
				negatives and positives (4.69% and 2.53%, respectively).
2012	NingWeng et al	DFA splitting, Memory Efficient Pattern Match		Achieves better performance in terms of speed (1000 times)by utilizing deep packet pre-filtering and novel finite state encoding.
2013	ArunaJamdagni et al	Mahalanobis Distance Map (MDM), Payload-based IntrusionDetection System	False Alarm Rate	Achieves better F-values, 0.9958 for DARPA 99 dataset and 0.976 for Georgia Institute of Technology dataset respectively, with only 0.85% false alarm rate and 1.3 times higher throughput.
		Illegal Traffic		
2008	Kim et al	IPS	Low False Rate	System has achieved low false rate.
2012	Eduardo Feitosa et al	Heuristic Orchestration Algorithm	False alarm rate	Achieved more than 95% of alerts.
	Noriaki Kamiyama et al	Networks of Transit ISPs	P2P traffic rate	Transit ISPs reduces P2P traffic within its network by about 50% to 85%.

Year	Authors	Methods used	Parameters used	Observations
		Connection failure		
2008	YANG XinYu et al	Location algorithm against local-worm	Scanning rate, Average rate	Network traffic is generated by scanning at the rate of 40 packetes/sec and 20 packets/sec. Ordinate represents the traffic which is expressed by the number of packets per 0.1 sec. The average rate after scanning is higher than the normal traffic.
2010	Asaf Shabtai et al	Knowledge-based Temporal Abstraction	Scanning Rate	The scan rate was set at 5 probes or 10 probes, and the vulnerable population size at 15% or 37%. The scanning rate for detecting worm instances is 420 at 5 probes or 1,200 at 10 probes.
2011	Syed Ali Khayam et al	Joint Network-Host Based Anomaly Detection Techniques	Detection Rate, False Alarm Rate.	Higher accuracy with almost 100% detection rates and very low false alarm rates.
2014	Chun-Ying Huang et al	C 4.5 Algorithm	Detect bot hosts, Accuracy	Detect bot hosts with more than 99% accuracy and false positive rate is lower than 0.5%

The above section discussed various existing anomaly based approaches proposed by different authors. Moreover, the various Machine Learning approaches proposed by authors for Internet worm detection are also discussed in detail. The detection approaches with containment techniques provide effective defense mechanism for secure network. So, the study is made for the containment approaches also. The following section discusses the existing worm containment methods.

2.3. Existing Containment Methods

The containment approaches are used to block the detected Internet worms and to secure the network from further infection of detected attacks. Some of the proposed methods for containment of Internet worms are discussed below.

Ram Dantu et al. [46] proposed novel security architecture called Automated Defense System (ADS) based on the feedback-control theory to automate the defense against worms. ADS is based on a multi loop, feedback-control system. A state-space feedback control model controls the spread of worms by measuring the velocity of the number of new connections that an infected host makes. The objective of ADS is to slow a worm by controlling the total number of connections made by an infected host. The proposed method contained the spread of Internet worms within few minutes and to a small percentage of hosts (5 percent). The infection rate is minimized up to 0.9% of hosts per second.

Sarah H. Sellke et al. [51] presented a branching process model to characterize the worm propagation and the model is to detect similar type worms and then extended to Local Preference Scanning Worms (LPS). The model leads to develop an automatic containment strategy which can prevent the spreading of worm in the earlier stage. In detecting similar type of worms, it finds an exact condition that determines whether the spreading of worm could be stopped and obtains the hosts infected by worms.

Xuxian Jiang et al. [66] presented a worm break-in provenance information to propagate it along information flows in the Operating System (OS) level for tracing worm. The authors have not been fully utilized the break-in provenance information for worm investigation. They have presented process coloring, a provenance-preserving approach to worm alerts, as well as worm break-in and contamination tracing. In this approach, "*color*"

- a unique system wide identifier, is associated with every potential worm break point. The color will be either directly succeeded by any spawned child process or diffused indirectly through the processes' actions along the information flows between processes or between processes and objects. As a result, any process or object affected by a colored process will be infected with the same color. To preserve the provenance of such influence, the corresponding log entry will also record the color. Process colors, as recorded in the log entries, reveal valuable information about possible worm break-ins and contamination actions. This coloring is suitable for networked server hosts running multiple service processes. Process coloring is a generic, extensible mechanism that may be applied to other types of malware. An internal cache (16 KBytes) is maintained to amortize the overall disk write operations and it has incurred very small additional system overhead.

Yuanyuan Zeng et al. [71] proposed a per-process containment structure on each host that scrutinizes the runtime behavior on each process and accordingly allocate the process a suspicion level generated by an SVM. Subsequently, each suspicion level is fixed with the limiting threshold. The average overhead for the CPU is 10.5%, memory of 14.5% and disk at 4.7%. The suspicion levels of 10 processes are generated within half a second. The false positive of per-process scheme that provides 4.15%, is the best one. The static scheme produces 20.14% false positives.

Guangsen Zhang et al. [22] proposed a Cooperative Internet worm containment and gossip based aggregation using decentralized information sharing. This consists of epidemic algorithms to distribute the information about infection and to obtain the quasi-global knowledge about attack behaviors. The model characterizes the relationships between the level of knowledge in the distributed system and about the accuracy of attack detection. The containment of worm propagation is acceptable, even for a gossip interval of 10 minutes.

Fabio Soldo et al. [18] proposed an Optimal Source-Based Filtering to filter out the worm. This reduces the collateral damage significantly, i.e., by 50%., while the communication overhead increases only linearly with the overall number of filters available.

Liming Zheng et al. [28] proposed a Malicious packets blocking algorithm to block the malicious packets. An anomaly detection system is proposed based on the Filter-ary-Sketch, in which traffic is recoded online and the anomalies are detected, based on one class support vector machine. When an anomaly is detected, malicious buckets identified according to the KL distance between the observed sketch and the forecasting sketch. Finally, malicious packets are blocked using a packet blocking algorithm. The computational complexity of blocking is *O(M+1)*. It can scale to high-speed networks, but the error comes from some normal packets as it uses the feature value identified as malicious.

Xufei Zheng et al. [63] proposed a Cloud based benign Re-WAW model. It has two revised Worm-Anti-Worm (WAW) models that are used for cloud-based benign worms. These Re-WAW models are based on the law of worm propagation and the two-factor model. The cloud based benign Re-WAW model is to achieve effective worm containment. Another is the two-stage Re-WAW propagation model, which follows active and inactive switching defending techniques based on the ratio of benign worms to malicious worms. This model is intended to avoid the network congestion and other potential risks caused by the active scan of benign worms. Table.2.4 lists the existing containment approaches discussed above.

Table.2.4. Existing Containment Approaches

Year	Authors	Methods	Parameters used	Observations
2008	Sellke et al	LPS Worm Containment System	Time deployment	Worm dies out completely in 750 minutes.LPS worm containment system is very effective when there is a 100 percent deployment. When there is only a partial deployment, it protects the local networks and provides global benefit.
2010	Guangsen Zhang et al	Cooperative Internet worm containment and gossip based aggregation	Time	Containment of worm propagation is acceptable even for a gossip interval of 10 minutes.

Year	Authors	Methods	Parameters used	Observations
2012	XufeiZheng et al	Cloud based benign Re-WAWmodel	Time	Slows down containment trend after the switching time of 31.8 seconds but does little effect on the overall containment.
2012	Liming Zheng et al	Malicious packets blocking algorithm	Time	The computational complexity of blocking is *O(M+1)*.Scale to high-speed networks but error comes from that since some normal packets also uses the feature value identified as malicious.
2012	Fabio Soldo	Optimal Source-BasedFiltering	Containment Rate , communication overhead	This reduces the collateral damage significantly, i.e., by 50%., while the communication overhead increases only linearly with the overall number of filters available.
2013	Rrushi et al	Botnet Containment	Time	It detects a botnet outbreak at its very early stage, thereby it can enable a timely botnet containment

The above section discussed various anomaly based detection approaches proposed by different authors. The existing containment methods are also discussed. There are few limitations existing in the discussed methods and they are discussed below.

2.4. Observations due to literature

From the above study, it observed that there are few limitations found in the existing Machine Learning detection approaches and containment techniques. Some important observations due to the review of literature are

- The worms containing new Malcodes in the program have limitations in detecting the relevant and irrelevant documents.
- When new worms exhibit unknown patterns, detecting those attacks lack in speed and it consumes more time for detection.
- If large number of changes are found in the network traffic, then the detection accuracy of illegal traffic traces becomes hard

- When there are more failures found in the network from Internet, the detection accuracy in detecting large number of malicious IP addresses consume more time.
- Only few containment approaches are proposed and there are only very few approaches proposed for both detection and containment of particular Internet worms.
- There are no existing approaches found completely for four Anomaly based Detection schemes such as Unknown Signature, Payload based, illegal traffic and Connection failures and Blocking Containment schemes based on content and address blocking.

To overcome the above observed limitations from the study, the research work has proposed a three-step methodology. Thereby the research work provides

- Better detection results based on the parameters such as precision, recall, accuracy, memory utilization and time consumption.
- Improved Containment results by blocking based on Detection rate and Containment rate.

2.5. Chapter Summary

This chapter discussed briefly the various anomaly detection methods used to detect the known and unknown worms. Existing anomaly based detection methods are not effective in detection of quickly changing patterns of Internet worms. Among the anomaly based detection methods, Machine Learning Approaches provide better accuracy and performs faster. Hence, the Machine Learning approaches used for Internet worm detection are discussed. Only detection will not secure the network in future, the detected Internet worms should be blocked from further spreading. So, the study is made on the containment approaches also. Significant detection and containment approaches of worms existing are discussed in detail. Moreover, the observations due to the study are also discussed.

To overcome the limitations stated in the observations due to literature, a three-step methodology has been proposed in this research work and is discussed in the subsequent chapter.