Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MALWARE PREDICTION ALGORITHM: SYSTEMATIC REVIEW

[1]MOHD NAZ'RI MAHRIN, [1]SURIAYATI CHUPRAT, [1]ANUSUYAH SUBBARAO, [2]ASWAMI FADILLAH MOHD ARIFFIN, [2]MOHD ZABRI ADIL TALIB, [2]MOHAMMAD ZAHARUDIN AHMAD DARUS, [2]FAKHRUL AFIQ ABD AZIZ

[1]Universiti Teknologi Malaysia, [2]Cybersecurity Malaysia

E-mail: [1]mdnazrim@utm.my, [1]suriayati.kl@utm.my, [1]anusya_r@yahoo.com, [2]aswami@cybersecurity.my, [2]zabri@cybersecurity.my, [2]zaharudin@cybersecurity.my, [2]fakhrul.afiq@cybersecurity.my

ABSTRACT

Malware is a threat to information security and poses a security threat to harm networks or computers. Not only the effects of malware can generate damage to systems, they can also destroy a country when for example, its defense system is affected by malware. Even though many tools and methods exist, breaches and compromises are in the news almost daily, showing that the current state-of-the-art can be improved. Hundreds of unique malware samples are collected on a daily basis. Currently, the available information on malware detection is ubiquitous. Much of this information describes the tools and techniques applied in the analysis and reporting the results of malware detection but not much in the prediction on the malware development activities. However, in combating malware, the prediction on malware behavior or development is as crucial as the removing of malware itself. This is because the prediction on malware provides information about the rate of development of malicious programs in which it will give the system administrators prior knowledge on the vulnerabilities of their system or network and help them to determine the types of malicious programs that are most likely to taint their system or network. Thus, based on these, it is imperative that the techniques on the prediction of malware activities be studied and the strengths and limitations are understood. For that reason, a systematic review (SR) was employed by a search in 5 databases and 89 articles on malware prediction were finally included. These 89 articles on malware prediction has been reviewed, and then classified by techniques proposed in detection of new malware, the identified potential threats, tools used for malware prediction, and malware datasets used. Consequently, the findings from the systematic review can serve as the basis for a malware prediction algorithm in future as malware predication became a critical topic in computer security.

Keywords: *Malware Prediction Techniques, Computer Security, Potential Threats, Malware, Malware Datasets*

1. INTRODUCTION

The threat (and the effects thereof) of malware will expand considerably in the coming years, mainly due to the improvements in techniques, goals and also the Internet's advancement. The struggle against malware spins off from different areas. It is ranging from the awareness among users to adopt security measures to the development of antimalware software by specialized companies [1][2]. This struggle also develops through the setting up of adequate security policies in different agencies and companies.

Over the past decade, there has been an increase in the number of types of malware created and this eventually leads to the existence of their effects. According to a study reported by Panda Labs, the mean number of computers infected by malware is currently 31.88%, the countries with the highest infection rates are China (52.26%), Turkey (43.59%), Peru (42.14%), and Bolivia (41.67%). On the other hand, the countries least affected are Sweden (21.03%), Norway (21.14%), and Germany (24.18%) [3]. The economic losses caused by malware in its different scenarios (government agencies, companies and individuals) are huge and have been estimated at thousands of millions of dollars per year.

The 2016 McAfee Labs Report mentioned that malware is still at large with significant new changes to the kinds of threats such as fileless attacks, exploitation of remote shell and remote control protocols, encrypted infiltrations, and credential theft which are harder to detect. In addition, this report claimed that Stuxnet and supporting Duqu, Flame, and Gauss malware have been developed to secretly target specific devices and make minor configuration changes that would result in a major impact, for example to a nuclear program. The intent was not to destroy a computer or harvest massive amounts of data, instead, it was to achieve the attackers' goals by carefully selecting the modified working systems [4].

In December 2016, Kaspersky Lab detected over 1,966,324 registered notifications on attempted malware infections that aimed to steal money via online access to bank accounts. Ransomware programs were detected on 753,684 computers of unique users; where by 179,209 computers were targeted by encryption ransomware. In addition to that, Kaspersky antivirus solution also detected 121,262,075 unique malicious objects: scripts, exploits, executable less, etc. and this could be one of the reasons why 34.2% of computer users were subjected to at least one web attack over the year [5].

Currently, the available information on malware detection is ubiquitous. Much of this information describes the tools and techniques applied in the analysis and reporting the results of malware detection but not much in the prediction on the malware development activities. However, in combating malware, the prediction on malware behavior or development is as crucial as the removing of malware itself. This is because the prediction on malware provides information about the rate of development of malicious programs in which it will give the system administrators prior knowledge on the vulnerabilities of their system or network and help them to determine the types of malicious programs that are most likely to taint their system or network. Thus, based on these, it is imperative that the techniques on the prediction of malware activities be studied plus the strengths and limitations are understood. Consequently, the purpose of this study is to address the details of malwares as mentioned above.

The goal of this paper is to report the findings of a Systematic Review (SR) which discovers about malware. In order to investigate further about malware, this paper will be structured into several sections. Section 2 provides the definitions and main concepts that are used in this report. Section 3 describes the objective of this systematic review, the research questions, the search strategy, and the selection process. The evaluation criteria and data extraction strategy are presented in Section 4. Section 5 describes the main results of the review conducted. Section 6 discuses the threats to validity. Section 7 concludes the report by summarizing the results, and highlighting some ideas on future work.

2. BACKGROUND CONCEPTS AND DEFINITIONS

2.1 Systematic Review and Snowballing

Systematic Review or SR is a method for examining a particular research topic area, or answering a particular research question. It is done by systematically identifying and evaluating all available relevant research works. All individual studies that are identified as relevant research contributing to a SR are called primary studies. In order to do SR in software engineering, a well-known guideline by Kitchenham and Brereton is followed [6].

It is crucial to correctly and clearly identify as many relevant research papers as possible when conducting a SR. The strategy is to identify the primary studies and ultimately produce the actual outcome of the review. The guideline by Kitchenham and Brereton for SR in software engineering suggests that to conduct a SR, it is advisable to begin with a database search based on a search string to be called the automatic search [6]. This guideline also recommends complementary searches, for example, doing a *manual search* on conferences proceedings, journals and references lists, and publications lists of researchers in the field.

Both automatic and manual search have limitations. Automatic search depends on the selection of databases, on database interfaces and their limitations, on the construction of search strings, and on the identification of synonyms. The manual search depends on the selection of research outlets, e.g. conferences or journals and the sources cannot be exhaustive.

Therefore, to overcome these limitations, Kitchenham and Brereton have proposed the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

snowballing search strategy as the first step to conduct the systematic review. The key actions of the snowballing search strategy are:

i. Ascertain a set of primary papers
ii. Identify further primary papers by using the reference list of each primary paper (This is called the back- ward snowballing)
iii. Distinguish further primary papers that cite the primary papers (This is called the forward snowballing)
iv. Repeat Steps 2 and 3 until no new primary papers are found

We are convinced that the snowballing search strategy complements the automatic and manual search strategies. In our SR, we define and perform a snowballing search strategy that has been developed based on a set of primary papers found from the automatic and manual search.

2.2 Sections and Subsections

Malware is the generic term used to delegate any informatics program created deliberately to carry out an illegal activity that, in many cases, is harmful to the system in which it has been lodged [7]. Malware such as Trojan, virus, worm, or spyware not only designed to infect a system but they are harmful to computer users, networks or computers in multiple ways for example high usage of CPU/memory, stealing confidential information, consume bandwidths and effect on web browsers. On the other hand, a malware prediction refers to an intelligent guess made to predict the future based on the current trend or situation [8].

3. SYSTEMATIC REVIEW METHODOLOGY

Figure 1 depicts the methodology used in conducting this systematic review. The details will be presented in the following sub sections.

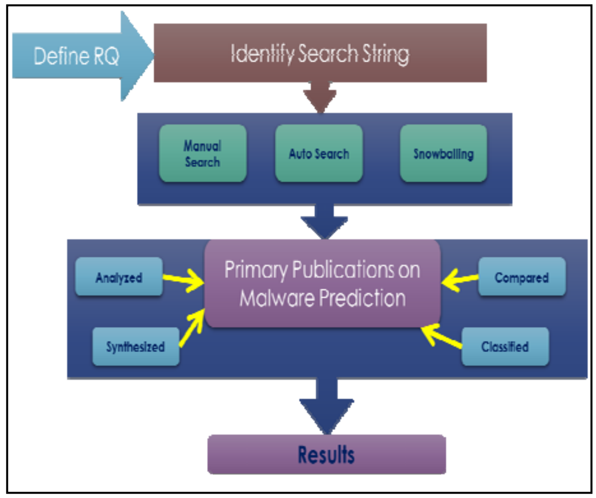

Figure 1: SR Methodology

3.1 Formulating the Research Questions

The first step conducted in this SR methodology is to derive the research question. Based on our early investigation on the problem background, we derived the following research questions:

RQ1: *What are the existing prediction techniques for malware threats/attacks?*

RQ2: *What are the potential threats that the techniques try to predict?*

RQ3: *What are the most current and established tools used for malware prediction?*

RQ4: *What are the datasets used for evaluating the proposed prediction techniques?*

3.2 Identify the Search String

By considering the identified research questions, we outlined the research keywords which include:

i. Malware OR Malicious OR Attacks OR Threat
ii. Prediction
iii. Techniques

Using the outlined research keywords, we identify the search string and used it in searching the related literature. The identified search string is:

<< ((Malware OR Malicious OR Attacks OR Threat) AND Prediction AND Techniques)>>

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3 Search String

The third step in our SR is to execute the literature search using the identified search string. We execute the search using the following search strategy:

i. Automatic search in established databases for literatures includes: IEEE Explore Digital Library, Science Direct, ACM Digital Library, Springer Link and Wiley Online Library. Google Scholars is not included in this study because it does not provide necessary elements for systematic scientific literature retrieval such as tools for incremental query optimization, export of a large number of references, a visual search builder or a history function. Besides, Google Scholar is not ready as a professional searching tool for tasks where structured retrieval methodology is necessary [9].

ii. Manual search in conferences proceedings and journals

iii. Snowballing for a complete set of primary Malware papers

iv. Search period: Jan 2010 – October 2017

We conducted the literature search using the search string identified in section 3.2 and the search result summarized in Table 1.

Table 1: Search Result

Num	Databases	URL Address	Number of Papers
1.	IEEE Explore Digital Library	http://www.ieee.org	32
2.	Science Direct	http://www.sciencedirect.com	308
3.	ACM Digital Library	http://portal.acm.org	27
4.	Springer Link	http://www.springerlink.com	254
5.	Wiley Online Library	http://onlinelibrary.wiley.com	100
	TOTAL		**721**

3.4 Applying the Inclusion and Exclusion Criteria

During the initial selection, we applied a set of inclusion and exclusion criteria based on guideline proposed by Kitchenham and Brereton as well as Khanian and Mahrin to ensure only relevant works on malware prediction were accepted into the SR [6][10]. The inclusion and exclusion criteria were applied in 6 phases as presented in Table 2.

3.5 Primary Publication Selection and Its Results

With the application of inclusion and exclusion criteria, the results of executing the search string in databases are shown at the Figure 2 in the appendix. As a result, 89 out of the 670 recovered papers were used for data extraction based on the research questions. The final selection from each database is shown in Table 3.

Table 2: Inclusion/Exclusion Criteria

Phase (P)	Inclusion/exclusion criteria
P1	Searching literature via the search string on electronic databases to cover journal articles, workshops and conference papers
P2	Excluding numbers of literature that is a short paper, a poster presentation, prefaces, editorials, slides presentation, non-English papers
P3	Removing duplicate literatures that emerge in different databases
P4	Read the full paper (the introduction, method section and conclusion)
P5	Excluding literatures that were not related to malware prediction
P6	Excluding literatures that cannot answer to two or more research questions from four research question

Table 3: List of Databases and Selected Papers

Num	Databases	Number of Papers
1.	IEEE Explore Digital Library	15
2.	Science Direct	35
3.	ACM Digital Library	12
4.	Springer Link	16
5.	Wiley Online Library	11
	TOTAL	**89**

4. EVALUATION CRITERIA AND DATA EXTRACTION STRATEGY

We evaluated each paper based on the relevance to the search keywords, which include malware, prediction and techniques. The criterion for each selected research work is that it must address the way the malware prediction technique was conducted. Then, we extracted the data from the papers to answer all the four research questions. The results of each research questions will provide guidelines for future research on malware prediction. The results of the extracted data were recorded which includes the following records of each paper:

i. Year
ii. Author
iii. Paper Title
iv. Publisher

We selected 89 articles on malware prediction from 40 journals, conferences and workshop papers. Each paper was reviewed and analyzed based on four research questions (RQs) as mentioned in Section 3.1. Distribution of articles by journals as illustrated in Table 4 shows that Journal of Computers & Security published more than 10% (9 out of 89 research papers) of the total number of papers. Journal of Security and Communication Networks published more than 8% (8 out of 89 research papers), along with, Journal of Computer Virology and Hacking Techniques (6 out of 89 papers, or 6.74%) published the second and third largest percentage of malware research papers among the journals. The most research papers were published in Computers & Security and Security and Communication Networks, because these journals focus on knowledge of the application of malware prediction systems by industry, governments and universities worldwide. Besides this, it publishes original research papers on all security areas including network security, cryptography, cyber security, etc. The emphasis is on security protocols, threats, malwares algorithms, security approaches and techniques applied to all types of information and communication networks, including wired, wireless and optical transmission platforms.

Table 4: Distribution of Articles by Journals and Conferences in which articles were published

Journal title	Number
ACM Transactions on Privacy and Security	1
ACM Transactions on Management Information Systems	1
Applied Soft Computing	1
Security and Communication Networks	8
Concurrency and Computation: Practice and Experience	2
Wireless Communications and Mobile Computing	1
International Journal of Network Management	1
Multimedia Tools and Applications	1
Journal of Communications and Networks	1
Power of Fuzzy Markup Language	1
Information Systems and e-Business Management	1
IEEE Transactions on Systems, Man, and Cybernetics	1
Journal of Network and Computer Applications	3
Computer Networks	2
Computer Communications	1
Journal of Computer and System Sciences	1
Information Science	1
Computers & Security	9
Information and Software Technology	1
In Control of Cyber-Physical Systems	1
Expert Systems With Applications	3
Neurocomputing	1
Empirical Software Engineering	1
Transactions on Emerging Telecommunications Technologies	1
Digital Investigation	1
Journal of Parallel and Distributed Computing	1
Knowledge-Based Systems	1
Future Generation Computer Systems	1
Journal of Systems and Software	2
Journal title	**Number**
Journal of Visual Languages and Computing	1
Pattern Recognition Letters	1
Computational Statistics and Data Analysis	1
Computer Fraud & Security	1
Soft Computing	3
Neural Computing and Applications	1
International Journal of Information Security	1
Journal in computer virology	5
Journal of Intelligent Information Systems	1
Journal of Computer Virology and Hacking Techniques	6
Arabian Journal for Science and Engineering	1
International Workshops	5
International Conferences	11

5. RESULT AND DISCUSSION

In this section, we describe the results of our Systematic Review study. The goal of this research is to identify the available prediction techniques for malware threats or attacks.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RQ1: What are the existing prediction techniques for malware threats/attacks?

To answer RQ1, we reviewed and classified the articles according to proposed techniques for malware prediction. The identified proposed techniques for malware prediction is presented in Table 5. The effectiveness of these techniques based on features extracted using dynamic or static analysis has been presented in the domain of malware detection and the field of malicious document detection.

Table 5: Proposed Techniques for Malware Prediction

Techniques for Malware Prediction	References
Bipartite graph	P1
API call graph	P41
Graph structure + Clustering process	P44
Control flow graph (CFG)	P43, P57
Fuzzy	P2, P15, P22, P63
Fuzzy + Association rules	P27, P37
Fuzzy+ Clustering method	P66
Network intrusion activity on computer network	P3
Markov Model	P4, P10
Markov Model + Entropy-based detection	P47
Stochastic Model	P5
Ensemble learning algorithms	P6, P25, P61, P83
Ensemble Methods + Harmony search	P59
Clustering algorithms	P7, P17, P35, P86
Clustering + Genetic algorithm	P65
Propagation model	P8
Propagation model + File relation graph + Active learning method	P52
Techniques for Malware Prediction	**References**
Honeypot technique + Association rule mining	P9
Honeypot technique	P28
Decision tree classifiers (J48, Random Forest (RF))	P11, P39, P87, P89
Decision tree + Feature selection algorithm	P78
Decision trees + Adaboost	P19
Decision trees + Support Vector Machines (SVMs)	P58
Support Vector Machine (SVM)	P29, P53, P67, P79
SVM + Interpretable string analysis	P71
SVM + graph kernels	P20, P72
Speculative execution	P12
Forecasting modeling	P14, P23
Multi Agent Systems	P18
Neural Network	P24
Application's network traffic patterns	P31
Logistic Regression	P34
Static analysis techniques + Classification algorithm	P32
Static analysis techniques	P42, P48
Static analysis + Dynamic analysis	P49
Partial matching classification algorithm	P36
AccessMiner (system-centric approach)	P40
Collaborative decision fusion	P46
Motivation Theory	P50
Text mining + Information retrieval	P51
Sequential association rule	P13
Association algorithm + connectivity metric	P16
Associative classification (Classification + Association rule)	P26
Association rule + Learning-based method	P33
Object oriented association mining + called API's	P56, P70
Sequential pattern mining + Nearest Neighbor classifier	P45
Pattern mining + Hooking	P62
Frequent pattern mining	P84
Nearest-Neighbor algorithm (KNN)	P54, P75
Naive Bayes classifier	p76
Naive Bayes classifier + Logistic regression + Threshold matching + Rank based	P80
Naive Bayes + Dimensionality reduction with Markov Blanket	P81
Classification algorithms (Decision trees, KNN, SVM, Artificial neural network, Logistic Regression, Hierarchical Clustering)	P21
Classification algorithms (Decision trees, KNN, SVM, Naive Bayes)	P30
Classification algorithms (Decision trees, SVM, AdaBoost, logistic regression)	P38
Classification algorithms (AdaBoost, Decision trees, Bayesian Network, Naive Bayes, Sequential Minimal Optimization, Logistic Regression, Bagging)	P55
Classification algorithms (Decision trees, Bayes network, KNN, multi-layer perceptron) + Anomaly-based	P64
Classification algorithms (SVM, rule learning, Decision tree classifiers (J48, Random Forest))	P68
Classification algorithms (Decision trees, SVM, KNN, logistic, Naive Bayes, Adaptive regularization of weights)	P85
Lazy associative classification algorithm + Execution-based dynamic analysis	P82
Techniques for Malware Prediction	**References**
Positive selection classification algorithm	P73
Behavior-based detection technique	P74
N Gram-based attribution method	P77
Header information technology	P88
Swarm-based approach + Stigmergic communication	P60
Hierarchical associative classification	P69

Table 5 shows the detection of malwares using different techniques in way to predict new malwares. These techniques provide the relevance of the features for discriminating between the group of searched malwares and the rest, and on the quality of training data for being unbiased and representative of malwares. Some articles [11] [12] [13] have proposed the structural feature extraction methodology for the detection of unknown malwares using machine learning algorithms. The same result was also proposed by [14] who apply classification algorithms to classify unknown malicious in documents based on structural features.

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The articles were classified by the most used techniques in malware detection as showed in Figure 3. Techniques that have been employed less than three times have been classified in "Others". It is apparent that malware prediction researches increased the employing classification algorithms such as Decision trees (14 out of 89 papers) and Support Vector Machine (SVM) (12 out of 89 papers). Among data mining techniques, also Fuzzy, KNN, Clustering and association rule mining have been used the most often in malware prediction researches (7 out of 89 papers). These techniques are able to predict the unknown, new malwares accurately, by feature selection process and feature extraction process.

Researchers select these techniques to categorize the features of malware into static features which are pertaining to installation files, dynamic features which are pertaining to the behavior of the application after installation or hybrid features which are combination of both dynamic and static features and also features extracted from executable files include printable strings, byte code n-gram, system calls, instruction sequence and opcode n-gram. On the other hand, these classification techniques extracted the features (i.e., byte sequences, printable strings, and system resource information) from malware samples via dynamic analysis or static analysis and based on the extracted features which identifies the malware automatically.

The result from reviewing articles showed that the Fuzzy, Naive Bayes, Decision trees and SVMs classifier are commonly used techniques in malware prediction research that significantly outperformed all other classifier algorithms, and is likely to perform the best. The reason is these algorithms use the information retrieved from benign software and malwares in order to obtain a benign behavior profile for the defense against unknown malware attacks. Then, every significant deviation from this profile is qualified as suspicious [15].

Figure 3: Distribution of research papers by used techniques

As showed in Figure 3, classification algorithms including Decision trees and SVM are the most popular classification techniques in malware prediction. Decision trees and SVM detect the malwares and classify them based on the identifying features and behavior of each malware. The evaluation results show the highest efficiency using Decision tree algorithms with an average overall accuracy of up to 90% [16] [17] [18]. In parallel, Decision trees and SVM have their own capabilities in term of speed, accuracy and strength to predict the malware as illustrated in the Table 6.

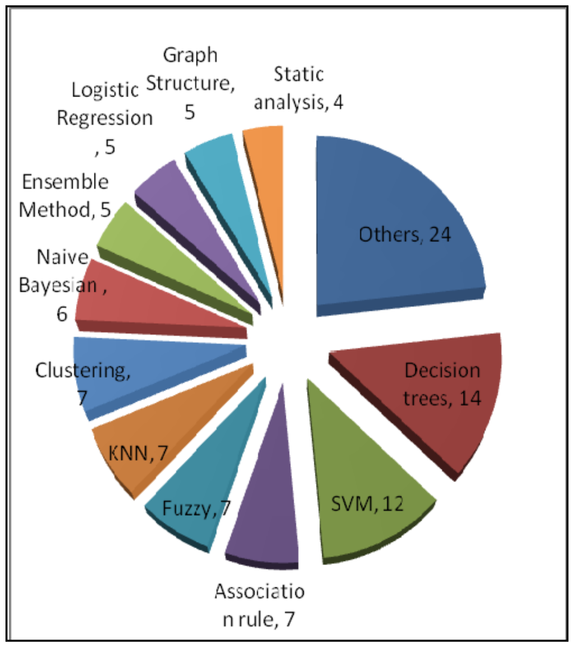

Table 6: Most popular Classification Techniques in Malware Prediction

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Consequently, Decision trees could predict the malware very fast compared to other prediction techniques. In comparison with the other machine learning methods mentioned in this study, Decision trees algorithm has the advantage that it is not a black-box model, but can easily be expressed as a set of rules. Subsequently, this technique also provides high accuracy in malware prediction. It is easy to understand as well as reduce complexity. The nearest malware technique to Decision tree is Support Vector Machines (SVM) [19].

RQ2: What are the potential threats that the techniques try to predict?

Malware causes a lot of harm to users, such as stealing personal information and using too much battery or CPU. The majority of adversaries can be involved in targeted attacks: corporations, cyber-criminals, hacktivists, online social hackers, nation states, cyber terrorists, cyber fighters, employees and script kiddies [20] [21].

The development of malware can be done through different vectors such as the sending of infected files or links to malicious web sites by e-mail messages, the use of removable devices (external hard drives, USB memory sticks, CD-ROMS, etc), malware on graphics processing units (GPUs), the downloading of infected files from malicious web sites, cyber-threats and the sending of infected files through Bluetooth, SMS and MMS. Advanced Persistent Threats (APTs) are detected by the state and enterprises to leak personal information. A brute attack is another threat in which attacker obtains information such as personal identification number (PIN) or a user password [22]. Leakage of personal data from mobile phones is a data breach. Stealing and exploiting sensitive data seem to be the outstanding characteristics of Android malware [23]. The identified potential threats in malware prediction from papers is presented in Table 7 as shown in the Appendix.

Advanced Persistent Threat (APT), one of the novel attacking models by emails on the Internet, is a very serious security problem for the computer system [24]. APT is a new generation of attack to be characterized as tailored to one specific entity and 3 out of 89 research papers have identified them (See Figure 4). Botnets are a disastrous threat because they execute malicious activities such as distributed denial-of-service, spam email, malware downloads (such as egg downloads), and spying by exploiting zombie PCs under their control [25]. As shown in Figure 4, 14 out of 89 research papers focus on detections of Botnets. Botnets infect PCs on a huge scale by initially scanning the service ports of vulnerable applications for the purpose of propagation, which is leveraged as the size of the botnet increases [25].

Malware Technique	Speed	Accuracy	Strength	Weakness
Decision Tree	Very Fast	High	Easy to understand, easy to generate rules and reduce problem complexity.	Mistake on higher level will cause all wrong result in sub tree.
Support Vector Machines (SVM)	Fast	High	Regression and density estimation results. Better performance in text classification, pattern segment and spam classification	Expensive and problem lies on the prohibitive training time

The majority of papers have identified Badware threats such as Worms (34 out of 89 papers), Trojan (28 out of 89 papers) and Backdoors (18 out of 89 papers) respectively as shown in Figure 4. A worm is a standalone malware to spread itself using a computer network and harm to the networks. Worms rely on security failures on a targeted computer to access it. Trojans are the malicious programs that stealing data, taking control of computer, and inserting malwares on to a victim's computer. Backdoors grow when networking systems and multiuser are used by many organizations. In a data access such as login for system, a backdoor involved can be in the form of a hard-coded username and password. Hackers employ backdoors to build malwares with modify code and data access. It is noted that Figure 4 shows list of most malware threats identified in the research papers.

There are many types of malware that are currently available on the Internet but worm, Trojan, backdoor, virus and botnet are the most common types of malware to be considered as the many dangerous threats for Internet users. Therefore, the most malware studies were aimed to predict these types of malwares due to causing more harms and defects to the network and operating systems in comparison to other Malwares. Security researchers combat vulnerabilities in operating systems and computer applications by designing antivirus applications and anti-malware which are used to detect malware.

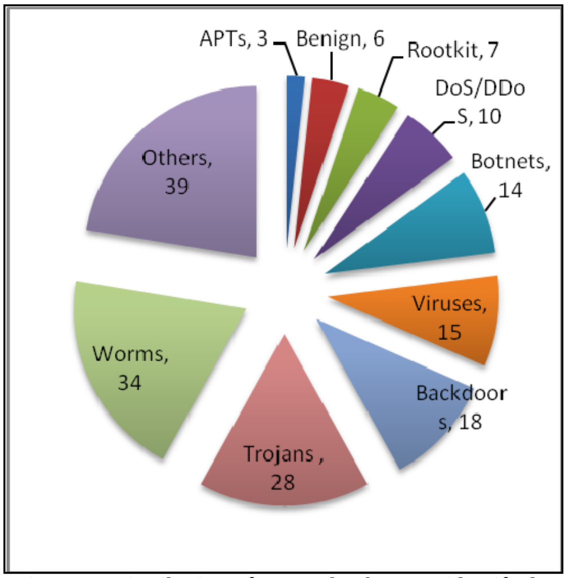

Figure 4: Distribution of papers by the most identified potential threats

RQ3: What are the most current and established tools used for malware prediction?

Security researchers employ an effective and automatic analyzing tool for identifying unknown malware attacks. To answer RQ3, we analyzed research papers to extract the established tools used for malware prediction. Established tools used for malware prediction from literatures is presented in Table 8 as shown in the Appendix.

Over the last decade, the most articles have applied machine learning classifiers for malware prediction using the Weka (22 out of 89 articles or 24.71%). Weka is an open-source data mining and machine learning toolkit to include data mining algorithms and written in JAVA programming language [12] [26]. Among the 89 articles, 15 articles or 16.85% used python programming language for performance evaluation of malware prediction techniques as shown in Figure 5. Java language has been used in 11 out of 89 papers. Java and python are the high-level programming languages that run on different platforms, such as Mac OS, Windows, and UNIX, hence more research papers consider on these languages for developing malware prediction techniques.

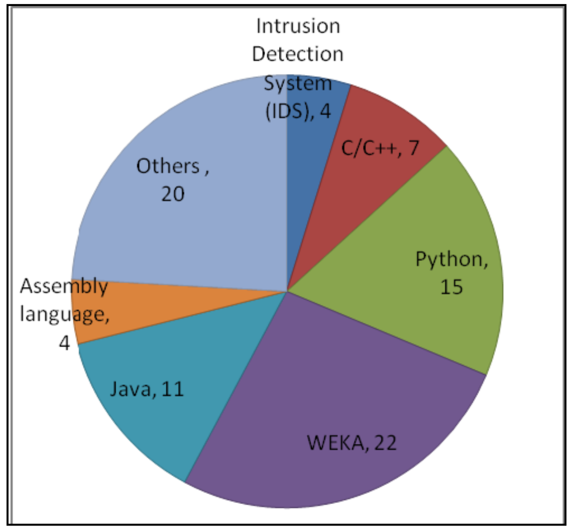

Figure 5: Distribution of Papers by the most Tools used for Malware Prediction

Java is object-oriented, class-based, concurrent, and have as few implementation dependencies as possible. An Intrusion Detection System (IDS) is a software application or tools that monitor the systems or networks in identifying unknown malicious instances and has been used 4 times (4.49%) as shown in the Figure 5.

RQ4: What are the datasets used for evaluation of prediction techniques proposed?

In order to validate the proposed methods or techniques in detecting malwares, researchers use various datasets related to malware dataset or benign software dataset to test their techniques. Consequently, it is meaningful to review the articles according to datasets used for evaluation of techniques in malware detection (RQ4). The common malware datasets used for experiments from malware studies is mentioned in Table 9 as shown in the Appendix.

We found that the most common malware datasets were malwares collected from VX-

Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Heavens website which is around 12.35% (11 out of 89 papers) and Genome dataset more than 11% (10 out of 89 papers) as shown in the Figure 6. VX-Heaven is a dataset of malware samples downloaded that the most research papers use it to validate their proposed technique in malware prediction. This dataset provides the information about malwares and computer viruses. Genome dataset is a benchmark dataset belonging to android Malware Genome project to explain examples of Android malware including Trojans. VX-Heavens and Genome datasets are free-accessed databases to be used for research purposes on malware prediction.

Researchers have also conducted experiments by collecting malwares from an installed Microsoft Windows such as Windows libraries, Windows XP, Windows 32/64-bit, DOS to validate their proposed techniques (8 out of 89 papers or 8.98%) as stated in the Figure 6. The other popular used datasets on malware detection research fields were dataset provided from anti-virus companies such as Kaspersky, Laboratory of Kingsoft, McAfee, AV vendor (8 out of 89 papers or 8.98%).

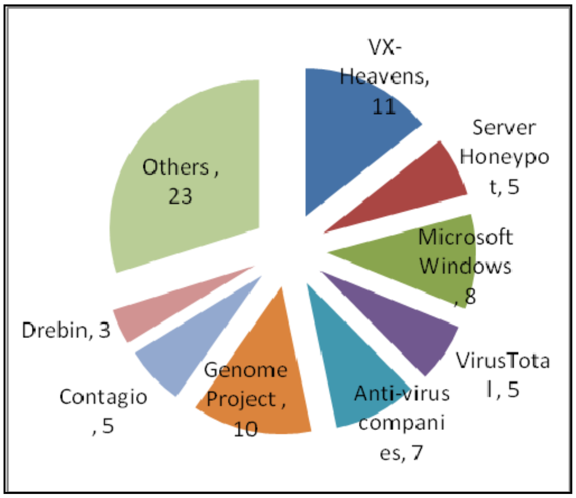

Figure 6: Distribution of Papers by the most Popular Datasets

This overall finding of this study is very significant to the malware world. The complete findings on malware prediction techniques, threats, tools and data sets may nourish information and knowledge to the researchers and technical practitioners in the industries. Besides this, the technical practitioners may apply these identified malware techniques from this study in order to identify technical problems that related to malware in their organizations. Consequently, these findings may be useful to software developers in order to analysis and develop new predictions techniques for malware. Subsequently, security organizations may use these findings for their research and development (R&D) activities as well. These findings definitely will be embarking high impacts on existing studies on malware.

6. THREATS TO VALIDITY

In this section, we discuss the threats to validity of this study according to the lessons learned on validity in SR [6] and our own experience.

6.1 The search process

To make the best use of the relevant articles returned by the search engines, we have kept the search string not too specific but still reflecting what we have wanted to search for. Moreover, the search string has been used for searching not only for the titles, abstracts but also for the full text. To minimize the possibility of missing relevant papers, we have kept our search string generic so that we cover as many relevant papers as possible (more than a thousand relevant papers found). To balance with the automatic search, we have also conducted the manual search on relevant journals and proceedings of relevant conferences. Then, to alleviate the limitations of automatic and manual search, we have adopted the snowballing strategy. Another possible threat is that we did not conduct extensive search for books related to malware prediction. However, we did include the option to search for book chapters while performing the automatic search.

6.2 Selection of primary studies

During the search and selection process was conducted, some publications might have been missed. To minimize this risk, every doubtful or "borderline" publication was being cross-checked and discussed by all the reviewers. Additionally, our clearly predefined review protocol with the inclusion and exclusion criteria has helped to reduce biasness in selecting the primary studies. The results of this SR papers are based on the data extracted and synthesized from the selected malware prediction studies.

7. LIMITATION AND ASSUMPTIONS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This study has limitation due to time constraints as described below: -

i. This study just included five (5) databases only. In order to obtain highly reliable result, it is suggested to include more databases in the future study.
ii. This study also did not investigate the applied research methods for each identified paper in this SR study in the future. The finding on research methods may provide information on how the data are gathered in each selected paper which was included in this SR study.

This study assumes that the prediction of malware could be done by investigating the techniques, threats, tools and datasets which are related to malwares. Consequently, this study was carried out based on the above assumption.

8. DISCUSSION

This study is having been carried in a way to provide a comprehensive and complete information on malwares that includes the techniques of malware prediction, threats, tools and datasets. The current available studies less discussed about the malwares completely and in depth. Besides this, most studies discussed the malware in a single perspective only such as malwares and datasets, malware algorithms, malware models and etc.

This study overcomes the gap in the current literatures by providing a comprehensive work on malware. This study covers all the available malware prediction techniques that is available in the literature by describing the technique name, details, and the sources as well. The same goes to other malware potential research such as malware threats, tools and datasets.

This study provides a wide-ranging and inclusive malware details for the practitioners and scholars to be carried in the future. Furthermore, this study aggregate data from the many databases and discuss the malware details. This study acts as a foundation in order to accomplish more researches and findings on malware prediction.

The statistical information provided in this study may guide the scholars and practitioners to have in depth investigation on malware prediction.

This study opens issues for further investigation on malwares. Malware research on mobile computing, cloud computing, securities, networks, standards and policies are needed to be carried out in future.

9. CONCLUSIONS AND FUTURE WORK

This paper reports our research effort aimed at systematically reviewing and analyzing malware prediction techniques, threats, tools and datasets. Malware is the primary choice of weapon to carry out malicious intents in the cyberspace, either by exploitation into existing vulnerabilities or utilization of unique characteristics of emerging technologies. Based on a rigorous analysis and systematic synthesis, we have presented an extensive systematic review on the malware prediction technique. The SR is based on a meticulous three-pronged search process, which combined automatic search and manual search with snowballing strategy. Using clearly predefined selection criteria, 89 malware prediction articles have been strictly selected, and then reviewed. From these primary malware prediction articles, we have extracted and synthesized the data to answer the four research questions. These 89 articles on malware prediction has been reviewed, and then classified by techniques proposed in detection of new malware, the identified potential threats, tools used for malware prediction, and malware datasets used. Among machine learning and data mining algorithms, the most employed algorithms for malware prediction are Decision trees, SVM classifier, Rule mining and Fuzzy algorithms. Researchers have also conducted several studies on data mining classifier algorithms such as K-Nearest Neighbor, Naive Bayes to identify malwares. The majority of papers have identified worms, Trojan and backdoors as serious security problem for the computer system. We also found that the most common malware datasets were malwares collected from VX-Heavens website and Genome dataset.

We further our research in Decision tree algorithm based on the finding from this Systematic Review. The efficiency of Decision tree algorithm is analyzed further by more number of industry datasets in order to predict the malware occurrences. Consequently, we are analyzing large number of datasets on how to execute the Decision trees algorithm to predict the most potential threats such as worm, Trojan, backdoor, virus and botnet as study shows that Decision trees algorithm is easily predict potential threats. Besides this, further research could be conducted on predicting the malware occurrences using Decision trees algorithm by applying statistical testing such as reliability testing and regression testing with large number of datasets. Parallel to this, we tend to analyze on how to associate set of rules in Decision tree algorithm to increase the accuracy of predicting the occurrences of malware. To conclude, we hope that this paper will supply researchers and practitioners with guidelines for future direction and insights on malware detection as a critical topic in computer security.

REFERENCES:

[1] Shabtai, A., Tenenboim-Chekina, L., Mimran, D., Rokach, L., Shapira, B., & Elovici, Y. (2014). Mobile malware detection through analysis of deviations in application network behavior. Computers & Security, 43, 1-18.

[2] Huda, S., Abawajy, J., Alazab, M., Abdollalihian, M., Islam, R., & Yearwood, J. (2016). Hybrids of support vector machine wrapper and filter based framework for malware detection. Future Generation Computer Systems, 55, 376-390.

[3] Panda Security Report, 2017. [Online][Accessed 02 December 2017]. Available at: https://www.pandasecurity.com/mediacenter/src/uploads/2017/11/PandaLabs 2017_Annual_Report.pdf

[4] McAfee Labs Report 2016. [Online][Accessed 02 February 2017]. Available at: https://www.mcafee.com/us/resources/reports/rp-quarterly-threats-sep-2016.pdf

[5] Kaspersky Lab's Cyber Security report, 2016. [Online][Accessed 02 February 2017]. Available at: https://media.kaspersky.com/documents/business/brfwn/en/The-Kaspersky-Lab-Global-IT-Risk-Report_Kaspersky-Endpoint-Security-report.pdf

[6] Kitchenham, B., and Brereton, P. (2013). A systematic review of systematic review process research in software engineering. Information and software technology, 55(12), 2049-2075.

[7] Tipton HF, Krause M. Information Security Management Handbook, 6th edn. Auerbach Publications: Boca Raton, Florida, USA, 2010.

[8] Venkatesh Jaganathan, Priyesh Cherurveettil, and Premapriya Muthu Sivashanmugam (2015). Using a Prediction Model to Manage Cyber Security Threats. The Scientific World Journal ,Volume 2015.

[9] Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. BMC Medical Research Methodology, 13, 131. http://doi.org/10.1186/1471-2288-13-131

[10] Najafabadi, M. K., and Mahrin, M. N. R. (2016). A systematic literature review on the state of research and practice of collaborative filtering technique and implicit feedback. The Artificial Intelligence Review, 45(2), 167.

[11] Vatamanu, C., Gavriluţ, D., & Benchea, R. M. (2013). Building a practical and reliable classifier for malware detection. Journal of Computer Virology and Hacking Techniques, 9(4), 205-214.

[12] Bai, J., & Wang, J. (2016). Improving malware detection using multiview ensemble learning. Security and Communication Networks, 9(17), 4227-4241.

[13] Altaher, A. An improved Android malware detection scheme based on an evolving hybrid neuro-fuzzy classifier (EHNFC) and permission-based features. Neural Computing and Applications, 1-11.

[14] Cohen, A., Nissim, N., Rokach, L., & Elovici, Y. (2016). SFEM: Structural feature extraction methodology for the detection of malicious office documents using machine learning methods. Expert Systems with Applications, 63, 324-343.

[15] Narudin, F. A., Feizollah, A., Anuar, N. B., & Gani, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. Soft Computing, 20(1), 343-357.

[16] Truong, D. T., & Cheng, G. (2016). Detecting domain-flux botnet based on DNS traffic features in managed network. Security and Communication Networks, 9(14), 2338-2347.

[17] Bocchi, E., Grimaudo, L., Mellia, M., Baralis, E., Saha, S., Miskovic, S. & Lee, S. J. (2016). MAGMA network behavior classifier for malware traffic. Computer Networks, 109, 142-156.

[18] Ye, Y., Chen, L., Wang, D., Li, T., Jiang, Q., & Zhao, M. (2009). SBMDS: an interpretable string based malware detection system using SVM ensemble with bagging. Journal in computer virology, 5(4), 283-293.

[19] Mohd Najwadi Yusoff and Aman Jantan (2011). Optimizing Decision Tree in Malware Classification System by using Genetic Algorithm. International Journal on New

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Computer Architectures and Their Applications (IJNCAA) 1(3): 694-713

[20] Canfora, G., Mercaldo, F., & Visaggio, C. A. (2016). An HMM and structural entropy based detector for Android malware: An empirical study. Computers & Security, 61, 1-18.

[21] Talha, K. A., Alper, D. I., & Aydin, C. (2015). APK Auditor: Permission-based Android malware detection system. Digital Investigation, 13, 1-14.

[22] Jiang, C. B., Liu, I., Chung, Y. N., & Li, J. S. (2016). Novel intrusion prediction mechanism based on honeypot log similarity. International Journal of Network Management.

[23] Wu, S., Wang, P., Li, X., & Zhang, Y. (2016). Effective detection of android malware based on the usage of data flow APIs and machine learning. Information and Software Technology, 75, 17-25.

[24] Sexton, J., Storlie, C., & Anderson, B. (2016). Subroutine based detection of APT malware. Journal of Computer Virology and Hacking Techniques, 12(4), 225-233.

[25] Kim, D. H., Lee, T., Kang, J., Jeong, H., & In, H. P. (2012). Adaptive pattern mining model for early detection of botnet-propagation scale. Security and Communication Networks, 5(8), 917-927.

[26] Al-Bataineh, A., & White, G. (2012, October). Analysis and detection of malicious data exfiltration in web traffic. In Malicious and Unwanted Software (MALWARE), 2012 7th International Conference on (pp. 26-31). IEEE.

APPENDIX:

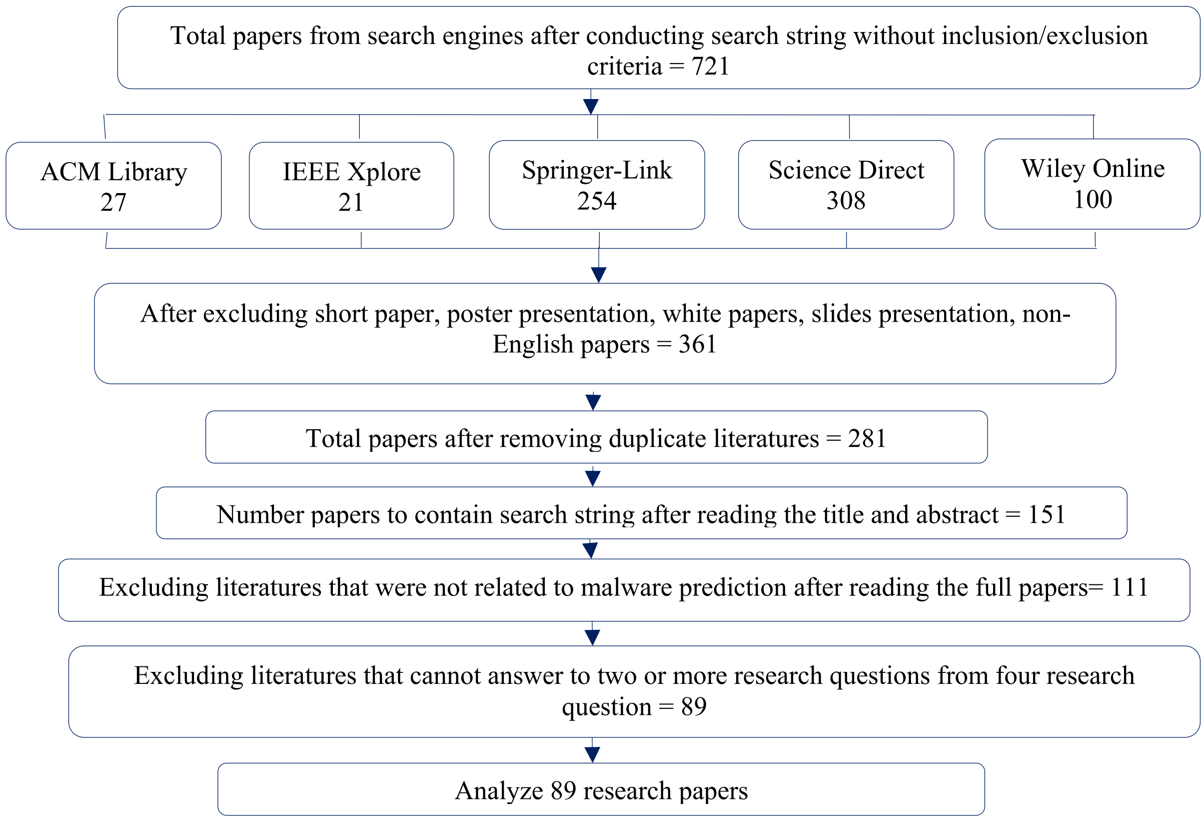

Figure 2: Search Result

Table 7: The Identified Potential Threats

Potential Threats	References
Denial of Service (DoS) attacks + Web application attacks	P2
Distributed Denial of Service (DDoS) + Worms + Spamming	P3
DoS attacks	P48
DDoS attacks	P7, P23
DDoS + brute force attacks	P9
DDoS attacks + Insecure interfaces (APIs)	P46
DoS + Mass mailing worm + P2P-Worm + Trojan + Rootkit + Backdoor + Flooder + Exploit + Constructor	P84
XML Denial-of-Service (DoS) attacks	P27, P37
Advanced Persistent Threats (APTs)	P13, P14, P80
Badware threats (Botnets)	P4, P11, P16, P44, P60, P86, P89
Badware threats (Worms)	P5, P8, P10, P28, P45, P57
Trojans	P62
Botnets + Trojans + Viruses + Backdoors + Worms	P6
Botnets + Trojans + Viruses + Backdoors + Rootkits	P22
Botnets + Trojan horses + Worms + Dropper	P40
Botnets + Trojan	P47
Botnets + Trojans + Worms + Viruses + Backdoors + Spyware	P52
Botnets + Trojans + Viruses + Backdoors + Rootkits	P63
Botnets + Worms + Bot programs	P74
Trojans + Worms + Spyware/Adware + Downloader	P12
Trojan + Worms	P18
Trojans + Worms + Virus + Backdoor + Floodor + Exploit + Rootkit	P19
Trojans + Worms + Virus + Backdoor + Adware + VirTool + Rogue + Software Bundler	P34
Trojans + Worms + Virus + Backdoor	P36
Trojans + Worms + Virus + Backdoor + Floodor + Benign	P61
Trojans + Worms + Virus + Spam + Rootkit	P21
Trojans + Worms + Virus	P24, P30, P55, P56
Trojans + Spyware	P29
Trojans + Worms + Backdoors + Spyware + Benign	P26
Trojans + Worms + Backdoors + Spyware	P71, P73
Trojans + Backdoors + Smart HDD + Winwebsec	P79
Trojans + Worms + Virus + Benign	P41
Trojans + Worms + Virus + Benign + Rootkit + Backdoor + Flooder + Exploit + Constructor	P53

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Potential Threats	References
Trojans + Worms + Virus + Spam + Rootkit	P54
Trojans + Worms + Backdoor + Infector	P58
Trojans + Worms + Backdoors + Benign	P69
Trojan horses + Worms + Backdoors	P70
Worms + Benign executable files	P76
Worms + Banker + Agent + BackDoor + Parite + Storm + SDBot	P88
Backdoor + Viruses	P15
Backdoor + Viruses + Rootkit	P75
Scareware (malware software that disrupt system and trick user into buying using credit card)	P25
Spying on users or stealing user data	P31
Information leakage in which personal data from mobile phones are leaked to attackers	P32
DNS bots on host + Spyware + data exfiltration	P33
Malicious code of malicious software	P35, P51, P78, P83
Android/Mobile malwares	P42, P64, P65, P66, P68, P82, P85, P87
Metamorphic malwares	P43, P81
Web Spam	P38
BaseBridge + FakeInstaller + DroidKungFu + Lotoor + FakeBattScar + GoldDream	P49
Cyber-threats + Malicious URLs	P50
Netbull Virus	P72
Morphing malware (such as W32.Agent, W32.Hupigon and W32.Pcclient)	P77

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 8: Distribution of Articles by Tools Used for Malware Prediction

Tools for Malware Prediction	References
Intrusion Detection System (IDS) deployed over the network	P2, P3, P13, P14
Uses standard GNU C/C++ libraries	P5, P20, P40, P44, P70, P75, P78
Automatic script tools based on a malicious log dataset within a botnet detection system	P4
Cloud computing framework based on Hadoop MapReduce	P6
HTTP server and Domain name system (DNS) service with MATLAB code	P7
HTTP server	P35, P39, P60
MATLAB code	P8, P10
Python programming language	P9, P17, P18, P22, P24, P47, P49, P52, P53, P54, P63, P87
Python language + Weka machine learning tool	P11
Python language + Java Language	P21
GOLDENEYE (a new dynamic analysis tool) consists of Python code and C library	P12
WEKA tool + HTTP Server	P38
Java Language	P32, P45, P51 ,P86
Monkey tools[1] develop by WEKA tool	P85
WEKA[2] machine learning tool	P19, P25, P27, P30, P31, P33, P37, P55, P59, P61, P64, P68, P71, P73, P76, P81, P82, P83, P88, P89
Java + HTML libraries by calling an Apache web server through AJAX interfaces	P23
VisualFML Tool[3] which completely programmed in Java	P15
Android Emulator (Java Agent Development Framework (JADE) implemented in Java)	P29
Dalvik software to the Java bytecodes (JAR file) using dex2jar Dalvik is a discontinued process virtual machine (VM) in Google's Android operating system	P46
Debugging tools WinAPIOverride32[4] and JavaScript language	P62
Jadx tools[5] for generate Java source code from Android	P65
FastFluxMonitor (FFM) system[6]	P16

[1] http://developer.android.com/tools/help/monkeyrunnerconcepts.html
[2] Data mining software in Java. http://www.cs.waikato.ac.nz/ml/weka/
[3] VisualFML Tool is a development environment for fuzzy-inference-based systems.
[4] (http://jacquelin.potier.free.fr/winapioverride32/)
[5] (https://github.com/skylot/jadx)
[6] A new dynamic analysis tool to use fast flux networks as contextual features to illuminate the evolution and dynamic relationships among IPs, domains, nameservers, and ASes

Windows Application Programming Interfaces (APIs)	P26, P67
Delphi programming language	P41
DroidAnalyzer tool on Linux OS	P42
Android application (APK) Auditor system[7]	P48
Web 2.0 Tools on Mobile devices	P50
Malware analysis tools (GFI Sandbox and Norman Sandbox)	P58
Assembly language programs	P43, P77, P80, P84
PE-Explorer tool (Assembly codes)[8]	P57
LibLinear package[9]	P69

[7] a learning-based lightweight system to be used by Android devices and generates a new approach for malware detection
[8] (http://www.pe-explorer.com/peexplorer-tour-di sassembler.htm)
[9] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 9: Malware Datasets used for Experimental

Malware Datasets	References
ISP Networks	P1, P44
DARPA dataset	P2, P33
Distributed Intrusion Detection System (DShield)[10] logs	P3
Network trace data collected in a dormitory at the Korean University	P4
C/C++ based malware codes	P5, P74
VX-Heavens[11] malware repository	P6, P19, P30, P36, P41, P43, P45, P73, P83
Malware dataset gathered from VX-Heaven + onlinedown.net + download.com	P67
VX-Heavens + Server Honeypot[12] (a network of real computers for attackers which logs collected from a Honeynet project)	P53
Server Honeypot from Honeynet project	P9, P17, P29, P54
Data collected from Mobile devices	P31, P51
Real local area network (Pseudo-random Domain Names (PDN) dataset and Legitimate Domain Names (LDNs) dataset)	P11
Malwares extracted from Anubis[13] services	P40
Malware dataset provided from Anubis services + Offensive Computing[14]	P12
Malwares from the websites Offensive computing + “vx.netlux.org”	P56
Simulation of a security game (iCTF)	P14
Malwares collected from Malicia Project[15]	P79
PE-files and Malwares collected from an installed Microsoft Windows (7 studies)	P18, P20, P57, P58, P59, P61, P62, P72
Scareware malware database of Lavasoft[16]	P25
Malwares gathered from VirusTotal[17] server	P34, P35
Malwares gathered from VirusTotal + Contagio[18]	P55
Web spam downloaded from Webb Spam Corpus[19]	P38
Dataset created from an anti-virus company (such as Kaspersky, Laboratory of Kingsoft, AV vendor, McAfee Center)	P24, P26, P52, P69, P70, P71, P75, P88
Malwares collected from VirusShare[20]	P32

[10] Available at: http://www.dshield.org
[11] http://vxheavens.com/ OR http://vx.netlux.org/
[12] http://amunhoney.sourceforge.net/
[13] http://anubis.iseclab.org
[14] http://www.offensivecomputing.net/
[15] http://malicia-project.com/ Accessed 21 Sept 2015
[16] http://lavasoft.com
[17] https://www.virustotal.com
[18] http://contagiodump.blogspot.com
[19] http://www.webbspamcorpus.org/
[20] https://www.virusshare.com/

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Benign and malicious applications extracted from VirusShare + Aptoide[21]	P65
Malwares collected from Android Malware Genome project[22]	P42, P46, P49, P64, P66, P68
Malwares collected from Genome Project + Contagio	P82, P87
Malwares gathered from Genome project + Contagio + Drebin[23] + VirusTotal	P48, P86
Malwares gathered from Drebin dataset	P47
Malwares from malwaretips[24] + Acer eDC company in Taiwan	P63
Malwares collected from VirusSign[25]	P76, P84

[21] http://www.aptoide.com/
[22] http://www.malgenomeproject.org
[23] https://www.sec.cs.tu-bs.de/~danarp/drebin/index.html
[24] http://malwaretips.com/
[25] http://www.virussign.com