TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

Malware Predictor using Machine Learning Techniques

[1]Jaevier A. Villanueva, [2]Dr. Roben Juanatas, [3]Dr. Luisito L. Lacatan
[1][2][3] AMA University
[1]jaevier.villanueva@gmail.com ,[2] rjuanatas@gmail.com , [3]lllacatan@amaes.edu.ph

Article Info
Volume 82
Page Number: 5665 - 5674
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 27 January 2020*

Abstract:
Malware has always been a threat to the computer world, but with fast growth in the use of the internet, malware severely affects the computer world. Malware predictors and detectors are critical tools in defense against malware. The existing malware detectors and predictors have been created, the effectiveness of these detectors and predictors depend upon the techniques being used. This study is specifically, addressed the following objectives: (1) propose a model to predict malware behavior using machine activity data; (2) apply the random forest algorithm in predicting malicious behavior. In this study, applied research is being used; this is the stage in the life cycle of the study in which we recognize how well we have used our knowledge to solve a pressing problem and to produce predictable results. In the proposed work of this research, a useful machine learning model using a random forest is developed and implemented with a malware data base. The proposed multi-layer machine learning model is used for training and predictive malware analysis on multiple parameters, including error factor, accuracy rate, and overall performance. The result of the model from the evaluation measures provides a high accuracy rate and a lesser mean absolute error value. There are very few parameters in the random forest as well, and these can be optimized using generalization theory without having to separate validation set during training.

Keywords: *Malware Detection, Prediction, Random Forest, Machine learning, Algorithm*

I. INTRODUCTION

Digital Technology has a part of today's life in which the worlds of business and education depend on technology and its applications. However, these advances have also opened up opportunities for the attacking community, and within a few years, malware has become the primary security threat, affecting computers and the network widely.[1].

Malicious software is known as malware. Computer malware is a program that, when executed, reproduces and infects a computer, poses a threat to the integrity of the system. The scope of the malware harm could be anywhere from removing files, destroying software to reformatting the hard disk. The malware will spread the systems in a variety of ways. One way is to download from the internet, and once the malware finds its way to the systems, the action will begin. Some of the time, the malware won't harm the system; otherwise, it could affect the performance and cause an overload method. On the other hand, some malware is hidden in the process, which is difficult to detect by the current malware detection. Based on the above challenges, it is important to carry out a more in-depth analysis to understand the malware for better detection and predictability.

[2] On the feasibility of online malware detection with performance counters, this was discussed by machine learning to detect discrepancies between

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

normal and malicious performance counter measurements during execution. Detection techniques proposed in[3],[4] Aware processor: a platform for efficient online malware detection, referred to as conventional machine-based detection, aim at finding a classifier that distinguishes malicious and benign information. In an attempt to differentiate between the actions of benign and malicious data, a trained classifier is a single model. The problem, however, is that most malware is introduced into programs that are otherwise harmless. For a specific program, its execution could be either benign or malicious depending on whether the malware is installed

activated, making it difficult to classify the software during practice, the classifier is also really the classifier is also searching for the option of benevolent and malicious examples in the training set and may lead to undesirable false positives and false negatives.

We propose to determine whether a file maliciously uses a behavior-based model to create a method that could be used in an end-user solution. Nonetheless, benign and malicious files contain a wide range of code and potential actions, our assumption is that malicious action starts rapidly once the malicious file starts to run because this decreases the overall runtime of the program. Eventually, use the Random Forest algorithm to predict malicious activity and check the ability of our system to detect various malware and variants that it has not seen before.

II. LITERATURE REVIEW

Related work is divided into two sections, with the first description of malware family research and malware detection. The second part is the study of machine learning methods, including structural support.

a. Malware Families

Malware identification is a complex process. Code that enables the device to be operated illegally is malicious. Malware comes in a variety of types and classes. These are usually classified on the basis of their method of spread and actions performed on the infected machine using the specified malicious program.

The virus is a malicious program that spreads from one program to another or from one computer to another by injecting its code into another program.[5]

Worm. It is a self-replicating program that spreads from one machine to another by spreading a copy of itself through a network without user permission[5].

Trojan horse hides them by pretending to be something real. Trojans usually crash data or attempt to collect private information, including financial data. The code, too.[5],[6]

Spyware is any program mounted on the device without the knowledge of the user. It is a hybrid term for technology that tracks and collects personal information back to the invader so that the invader can use the information taken in a suspicious manner. This generally gets into the process when it takes open source or experimental code and runs on the system without the user's knowledge, changes the configuration of your browser, or adds hostile browser toolbars..[7][8]

Scareware is a malware that is protected as a free or preliminary enemy of virus programming or some other free online trick. This appears to be implemented by the user when making false security programming, opening connections, or visiting a pernicious page. After it has been developed, it assembles all the data stored on your computer that could be sold to other cyber criminals.[9]

Adware encourages verified software that automatically runs, shows, or duplicates ads to a device after malicious programming is implemented or the application is used. This bit of code is usually set to a free downloaded application. The most

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

common source of adware programs is free games, including peer-to-peer clients. [9]

The botnet is remotely controlled by stand-alone technology. It's normally a zombie system under general command for any network infrastructure.[9]

Ransomware is a particularly malicious form of malware that limits a person's access to their computer and demands payments to restore functionality. Since then, the attack has been streamlined and professionalized. It is known to be highly lucrative, with past claims valued at hundreds of millions of dollars a year.[10]

b. Machine Learning Techniques

Machine learning is stated to a set of methods that automate the analysis of extensive – scale data. In this study, we the researcher considers classification functions where machine learning models are intended to learn mappings between information area and a predefined set of yields called classes. For instance, the data collection domain and the classes malicious or benign when the errand of attention is malware detection in documents. Techniques like support vector machines [11], and more recently, deep learning [12], and revisiting neural networks architectures [13] are common choices to learn supervised model data.

[14] An online cloud anomaly detection program using a single-class support vector machine was pointed out. One SVM formulation category is used at the level of the hypervisor. Security and stability are important functions in cloud infrastructure; hence, the cloud should be able to react to unknown threats. The virtual machine used in the experiment uses per VM methods to perform the analysis, which helps in the detection of malware.

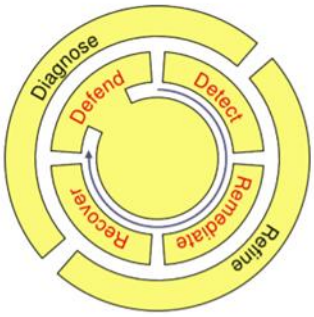

Fig 1. The High level description of the D2 R2 + DR network resilience architecture 2[14].

The above framework consists of two modes of operation. The outer loop which detects the vulnerability of the current configuration and then refines the overall system and resilience strategy. Often, an internal control loop that involves defending the system, detecting faults and irregularities, remedying them, and eventually recovering from detected faults..[14]

[15] presented a scalable, clustering approach that identifies and clusters malware. An extended ANUBIS system is used with taint tracking for analyzing the behavior. The limiting factor of this approach is trace dependence. Another issue is dynamic data tainting; so malicious binary could be injected. Some malware triggers only during some actions or events such as setting up a time to explore, and so on these forms significant limitations for this framework.

[16] pointed out the analysis of static and dynamic malware approaches. The static method is achieved without running system whereas dynamic while running. The study highlights the significant pros and cons of each technique with different usage of it.

The static method uses manual inspection, which fails due to evasion, obfuscation, and polymorphism technique. On the other hand, the advanced static method provides a better result with more specific details about the malicious programs. The integration of these two approaches results in a better solution to the detection of malware activities.

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

[17] proposed a behavioral-based framework for malware analysis. It improves behavior-based analysis for suspicious programs allowing end-users to get a secure environment which makes them feel executing directly on the former environment. The system calls are used both security lab and potential victim programs.

[15] proposed the system to detect malware in an automated way by reducing the rate of false identification of malware. The usage of dynamic analysis is to detect the presence of malicious behavior. The architectural view of the proposed system is depicted in the figures below, which mainly consist of the analysis module and the classification module. The analysis module mainly does preprocess tasks such as generating data suitable for the classifier tool. Whereas, classification modules deal with the preprocessed data to perform the proper classification in order to differentiate each sample into its corresponding neighbors.

Machine Learning is one of the advanced technologies that have worked in this direction.. Many research focused on building research frameworks[18][19][20] to acquire static features[21],[22] and to identify malware families[23],[24],[25]. The architecture depends on the machine learning to detect unknown malicious information, without the need to eliminate the obfuscation, has been presented in[19], using text classification methods it has been shown to improve the accuracy of the obfuscated samples.

One of the best known machine learning algorithms is the Random Forest. This requires almost no data planning and simulation but usually results in an incorrect result — random collection of decision trees, making accurate predictions.

The main idea is to evolve multiple decision trees based on the data set's independent subsets. The node, the n variable out of the set of features, is selected randomly, and the best split on these variables is found. According to[26], the algorithm can be described as follows: (1) Multiple treasury is based on approximately two thirds of the training data (62.3 per cent). Data shall be selected randomly. (2) Some parameter predictors are selected randomly from all variables. At that point, the best partition of the selected variables is used to break the node. Second, the number of variables chosen is the square root of the absolute number of all the predictors for classification and steady for all three (3) The frequency of misclassification is determined when the rest of the data is used. The total error rate is determined as the average out - of-bag error rate. (4) Every qualified tree shall, on the basis of its own vote, give its own classification result. As a result, the class earned the highest number of votes. Transformation highlights the fundamental change in our world due to the pervasive existence and dissemination of digital technologies. We have entered the fourth industrial revolution, building on the past three, but using new "full-powered" digital technologies, allowing technology development and diffusion even faster than before [27]. The process of digital transformation has been extensively studied in a wide range of educational and industrial fields. The digital world provides a different background to the old battle between content-creating companies and distribution-providing companies. In 2009, Apple managed to achieve the highest competitive advantage of digital convergence. The price of depreciation was rapidly experienced by its competitors [28]. A new global economy, characterized by dynamism, customization, and intense competition, is developing, and the cornerstones for success are the integration of knowledge, technology, and innovation into products and services.

For several activities and organizational results in the 1990s, "Transformation" became a buzzword. The concept has been divided into four clusters. (1) Re-Engineering: improving overall organizational efficiency while only partially addressing better workforce engagement ; (2) Re-engineering: improving efficiency without necessarily improving organizational capacity to achieve its long-term goals and opportunities ;(3)

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

Renewing: improving efficiency, efficiency and innovation through empowerment of employees without a clear focus on the desired outcome and (4) Regenerating: improving existing processes and fundamentally re-examining the direction and portfolio of available opportunities. Digital innovation is perceived as new and relies on digital technology as the process and outcome of digital innovation. It is a question of combining digital technology in new ways or with physical components that enable socio-technical changes and create new value for consumers [25], [28]. Digitalization, digital innovation, and digital transformation are closely related and connected in different ways. Digital innovation is perceived as new and relies on digital technology as the process and outcome of digital innovation. It is a question of combining digital technology in new ways or with physical components that enable socio-technical changes and create new value for consumers. Based on the features of digital transformation and connections from digitalization and digital innovation, defining digital transformation when digitalization or digital innovation is implemented over time to allow significant changes in how business is performed, is leading to a substantial transformation of an organization or sector.

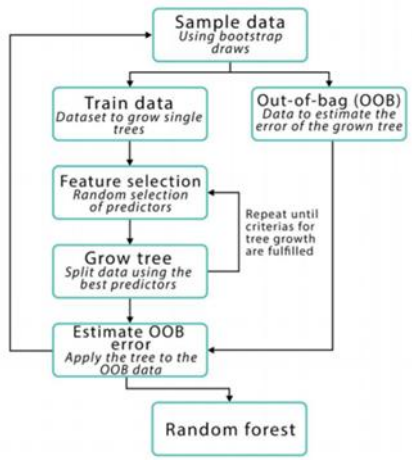

Figure 2. Random Forest Scheme

Since RFT are efficient, multi-class, and it is ready to deal with adequate attribute space, they have been generally utilized in a few domains, for example, a real-time face recognition [26], bioinformatics [27], and there is also some research in the medical domain, for example [18][19],[20].

The author suggests the use of random forests to forecast malicious activity as they can process and collect change data over time as well as raw input values. San, B, man. Et., al tried to identify random forest malware families, to classify malware into families based on static features derived using features from three dimensions, including byte code features, assembly code features, and PE features. The choice and merger method received a score of 93.56% by random forest classifier.

III. METHODOLOGY

The researchers have used applied research; this is the stage in the life-cycle process in which we consider how well we have used our expertise to find a solution to an unyielding problem and to produce predictable outcomes. According to (Thoms W. Edgar & David O. Manz), this type of research is an important aspect of cyber security. Safety is mostly the application of knowledge in order to achieve the desired results. The findings of applied research, therefore, help to obtain the knowledge needed to advise and implement the process more effectively.

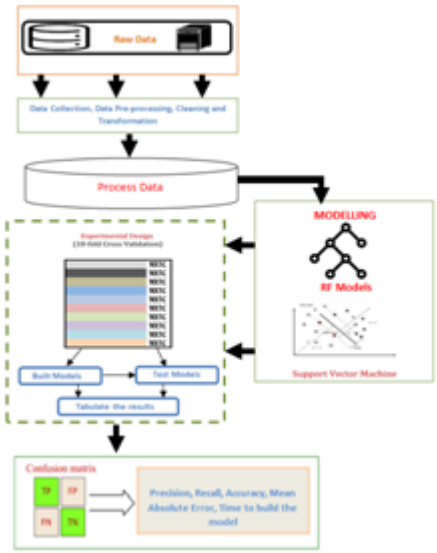

Figure 3. The overview of the methodology used

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

This section describes the methods used to establish the proposed malware prediction as outlined in fig. 3. Two common methods random forest, vector support machines under the function are used to construct prediction models and compared to each other using 10-fold cross-validation hold-out samples As a result, two distinct types of classification models were developed in this analysis. The performance measures of classification are determined using a 10-fold cross-validation methodology. In this experimental methodology, the data set is partitioned into 10 equal-sized different subsets. For each experiment, nine subsets were used for training and one part is used for testing. For each of the 2 model types, this process is repeated for 10 times. Instead, test results are aggregates to represent the model's "unbiased" performance estimate.

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \qquad (1)$$

Equation 1. Cross-validation [21]

As shown in the Eq section. (1) The performance measurement (PM) is averaged over k-folds (we set the value of k to 10 in this study). At the Eq. (1), CV is a cross-validation measure k is the amount of folds used and PM is the quality metric for each fold. Weka 3.6.8, a popular data mining toolkit, was used to describe and validate the proposed methodology.

Data preparation

Preparation of data is a collection of techniques that integrate, clean standardize, minimize, and transform data such as a selection of features. The final dataset is supposed to be accurate and usable for training and testing activities after data processing tasks are performed. Quality ranges are generally similar through the frequency distribution (see fig. 4).

Figure 4. Frequency distribution of benign and malicious samples.

Classification: Machine learning Approach

Random Forests: is an algorithm that learns from a poor template (like DT) to create a more robust one with a minimum cost to avoid over-adjustment. The forest is built using well-known techniques of bootstrap. Bootstrapping's main idea is to merge learning models by increasing the overall outcome of the classification. In order to achieve this, the following steps are carried out after a dataset X_N of size N is split with the bootstrap technique [22] :

1) Draw a random sample of the size n bootstrap (randomly select n samples from the replacement array of training)

2) Build a bootstrap test decision tree. At every node:

• Pick d features at random without removing them

• Separate the node by using the correct split function

3) Repeat steps from 1 to k times.

4) Add the prediction by each tree to assign a majority vote to the class label:

• Choose D(XN)= Xb with non-replacement N samples.

• Create a bootstrap dataset B in 1, N with Eq 2 being estimated with the previous assumptions.

$$\text{P rob(K)} = \text{P rob(K)} = \frac{N!}{\left(K!\,(N-K)\right)\left(\frac{1}{N}\right)K\,\frac{N-1}{N}} \;; 0 \le k \le N \qquad (2)$$

Equation 2. Probability Estimation [21]

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

Support Vector Machine (SVM). It is an algorithm projecting each observation as a point into an n-dimensional space (when it comes to the number of features of the dataset) where each feature is a value of a given coordinate. The main advantages of this algorithm are that the generalization error is minimized, and the overfitting consequences are effectively avoided by measuring the correct decision boundaries with large margins. Parallel to the decision boundaries are balanced to optimize the difference, positive and negative hyperplanes. See below formula 3-4 explaining the context of the SVM:

$$w_0 + w^T x_{pos} = 1 \quad (3)$$

$$w_0 + w^T x_{neg} = -1 \quad (4)$$

Where w0 is a vector of input and where w^T is a vector of weight. The problem of generalization is solved when subtraction is applied to equations 3–4, as shown in equation 5.

$$w^T (x_{pos} - x_{neg}) = 2 \quad (5)$$

Equation 5 is then standardized using W as shown in Equation 6

$$||w|| = \sqrt{\sum_{j=1}^{m} w_j^2}. \quad (6)$$

Equation 7 provides the solution from Equation 6

$$\frac{w^T (x_{pos} - x_{neg})}{||w||} = \frac{2}{||w||} \quad (7)$$

On the left side of Equation 6, the distance between the positive and negative hyperplanes, which is the margin meant to be maximized, is denoted. Eventually, as shown in Equation 8, SVM's objective function becomes the most excellent maximization of its margin

$$w_0 + w^T x^i \geq 1 \; if \; y^{(i)} = 1 \quad (8)$$

Through maximizing w under the right classification criteria Equations, it is possible to rewrite 4-5 as shown in Equations 3-4 as shown in equation 9-10.

$$\frac{2}{||w||} \quad (9)$$

It is inferred from Equations 9–10 that malware tests in the positive hyperplane and the good ones in the negative hyperplane will be put in our case study.

$$w_0 + w^T x^i < 1 \; if \; y^{(i)} = -1 \quad (10)$$

Evaluation measure

Through evaluating the difference between the class it generates for a given input and the class it should generate, the execution of the classifier can be evaluated. They used common metrics. (1) We use a confusion matrix to quantify the classification results, which describes all possible predictive outcomes and has the following form:

		Actual Class	
		1	−1
Predicted Class	1	TN	FN
	−1	FP	TP

Figure 4. Cofusion Matrix used

A "positive" performance in our experiments is one in which an anomaly is observed by the detector, i.e. a category of-1 is produced. We infer from this that if the classifier produces a-1 during malware execution, a true positive (TP) is possible; otherwise it is regarded as a false positive (FP). Negative results often arise when routine activity is detected by the detector. When malware does not run, an output of 1 is a true negative (TN); otherwise, it is regarded as a false negative (FN). A number of output matrixes can be derived from the uncertainty matrix shown in equation 11.

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

$$Accucary = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F\ measure = 2\ x\ \left(\frac{Precision\ x\ Recall}{Precision + Recall}\right)$$

Accuracy is the degree to which all newly checked data specimens are correctly identified by the detector, while accuracy is a proportion of the exact number of the positive classification. The recall metric is a test of the ability of the detector to identify an anomaly correctly; The final metrics are the harmonic mean (F measure) which provides a more rounded calculation of the detector output by counting to some degree for all effects.

IV. RESULTS AND DISCUSSION

A. *Experimental result*

The results of the models on all presentation investigations were reported in Table No.1 with a mean and standard deviation of the particular execution measured in each cell. On average, Random Forest received a higher accuracy rate (89,0323%) and a lower mean absolute error value (0,1097) compared to SVM with an accuracy rate of (85,4839%) and a mean absolute error value of (0,1532) respectively. Random Forest (RF) obtained a higher accuracy rate against the SVM as RF uses statistical learning theory to look for a regularized hypothesis that matches well with the available data in the absence of overfitting. RF also has very few parameters, which can be optimized using the principle of generalization without the need to separate validation set during learning. The rated significance of the predictor factors was also investigated in this paper to assess the relative contribution of each to the predictor models. To understand the relative importance of the features used in the investigation, Figure 3 shows the results of the selection of attributes performed using the Correlation Ranking Filter as assignment Evaluator and Attribute Ranking as search tool on prepared expectation models, in which the comparative significance. When the significant variables are determined for each of the 2 models, for positioning purposes they are collected and combined. of the input features in predicting the output was calculated.

Figure 5. rank of (25) selected attributes.

Similarly, the rank of each predictor factor is shown in Figure 5. Each variable is an essential predictor in the success rate of malicious and benign data.

B. *Discussion and Conclusion*

The effects of this investigation that, if proper preprocessing strategies are applied to sufficiently data sets with the rich arrangement of variables, investigation techniques are equipped to anticipate the success of high accuracy (nearly 90 percent) of malicious or benign data. From the usability point of view, both RF and SVM had improved expectations, and one might decide to use other classification techniques on the grounds that contrasted with RF and SVM, some of them may depict an increasingly straightforward template structure such as (DT) Decision trees. A unique quality of this paper is that it gives the meaning of the attributes used in prediction modeling to the rank ordered. In particular, the impact / review investigation is

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

applied to predictive models in order to identify their comparative value (i.e., additive contribution) in predicting the output factor. To create the final list of variable-value sets, the affectability values of all variables over each of the 2 template kinds are collected. These awareness helps improve models and helps decision-makers understand that factors are most critical in improving malware detection. Achieving an exam project is strongly dependent on the wealth (amount and dimensionality) of the data which contributes to the wonder that is being considered. Despite the fact that this investigation used an enormous example of information (covering some open-source information) with a somewhat rich collection of highlights, more data and more factors that help to improve the examination/forecast results. Potential future headings of this report include (i) expanding the prescient techniques for integrating troops (joining / intertwining models); (ii) updating data sources by including overview-based information from institutional reviews. An analytics project's success depends heavily on the resources (quantity and dimensionality) of the information describing the considered phenomenon. Although this study used a large sample of data (covering many open-source data) with a rather rich set of features, more data and more variables could potentially help enhance the results of the analysis / prediction. Potential future directions of this research include (i) expanding predictive modeling methods to include ensembles (combining / fusing techniques model) and (ii) improving sources of information by including data from academic survey-based studies.

Decision trees specifically illustrate the reasoning process for various forecast outcomes, supporting a common hypothesis, while RF and SVM are statistical models that do not provide such an unmistakable viewpoint on how they do what they do. (Which are carefully designed and painstakingly guided for the purposes of expectations), and perhaps in particular (iii) the sending of the data frame as a guide of selection for business and instruction, with the goal of assessing the advantages and disadvantages of the frameworks for creating and better adapting to the needs of individuals.

Table 1. Ten Folds Cross-Validation Classification Performance measure of the Models

Model	Accuracy	MAE	TP Rate	FP Rate	Precision	Recall	F-measure	Time taken to build the model(seconds)
RANDOM FOREST	89.0323%	0.1097	0.890	0.890	0.890	0.890	0.942	0.06
SMO/SVM	85.4339%	0.1532	0.855	0.611	0.853	0.855	0.854	0.8

ACKNOWLEDGMENT

We might want to thank the unknown commentators and the members of the School of the Graduate Studies at AMA University, the Philippines, for their feedback on this work.

REFERENCES

[1] L. I. U. Wu, L. I. U. Ke, R. E. N. Ping, and D. Hai-Xin, "Study and Detection of malware based on behavior," no. 60203044, pp. 39–42, 2011

[2] J. Demme et al., "On quality counter on online malware detection feasibility," Proc. Int. Int. Symp. Symp. Software. Computer. Architecture, pp. 559–570, 2013.

[3] D. R. Ellis, K. S. Attwood, J. G. Aiken, and S. D. Tenaglia, "A worm identification strategy," WORM' 04-Proc. 2004 Activities of the ACM. Rapid Malcode, pp. 43–53, 2004.

[4] M. I. Gorelik, N. Abu-Ghazaleh, and D. Ozsoy, C. Donovick. Ponomarev, "Malware-aware processors: an active online malware detection system," IEEE 21st Int 2015. Symp. Symp. Strong Perform. Software. Computer. Archit. 2015 HPCA, pp. 651–661, 2015.

[5] R. Sharp, "Malware Introduction," Netw. Safe.

January - February 2020
ISSN: 0193 - 4120 Page No. 5665 - 5674

Secure. Laboratory check, pp. 331–363, 2015.

[6] J. Blount, Tauritz, and Tauritz, and S. A. Mulder, "Adaptive rule-based malware detection using training classification systems: concept proof," Proc. Int. Int. Software. Computer. Softw. Appl. For instance, pp. 110–115, 2011.

[7] R. Tian, "An Integrated Identification and Detection of Malware," 2011.

[8] S. S. Plus and P. P. Gaikwad, "Efficient Malware Detection Trust-based Voting Method," Procedia Comput. Sci, the flight. 79, pp. 657–667, 2016.

[9] K. Mathur, S. Hiranwal, "Computer Science and Software Engineering.pdf, 2013 International Journal of Applied Research," Int. J. Adv. Res. Comput. Sci, that's it. Softw. Eng., vol. 3, no. 4, pp. 422–428, 2013.

[10] C. Simoiu, J. Bonneau, C. G. Symantec, and S. Goel,"' I have been told to buy a program or risk my computer. I have ignored it': a ransomware report," USENIX Symp. Personal open. Safe. Secure. 2019. 11–13 August 2019, St. Clara, California, United States, 2019.

[11] C. "Support-Vector Networks," vol. Cortes and V. Vapnik. 297, pp. 273-297.

Q. Le, O. Boydell, the initials of B. Mac, and M. Scanlon, "Machine learning at the shallow end: non-domain expert malware detection," Digit. Investigation. Vol. 26, pp. In 2018, S118–S126.

[13] Y. Gal, "Pillar I of Bayesian Deep Learning: Deep Learning."

[14] M. R. Watson, A. K. Marnerides, N. Shirazi, A. Mauthe, and D. Hutchison, "Cloud Computing Network Malware Detection," vol. 5971, c, pp. 1-14, 2015.

[15] A. Sujyothi as well as S. Acharya, "Digital Environmental Complex Malware Analysis and Detection," Int. J. Mod, you remember. Educ. Educ. Software. Computer. Sci, the plane. 9 Number 3, pp. 48-55, 2017.

[16] S. Yusirwan, and I, and Y. Prayudi. Riadi, "Use Static and Dynamic Analysis System to Apply Malware Security," Int. J. Comput. Appl., plane. 117, number 6, pp. 11-15, 2015.

[17] L. R. Paleari, Martignoni, and D. Bruschi, "A cloud-based behavior-based malware analysis framework How static analysis is analyzed and detected today is either too burdensome or impossible," no. December 2009, pp. 14–18

[18] N. Nissim, R. Moskovich, L. Rokach, Y. Elovici, "New active learning methods for improved detection of PC malware in windows operating system," Expert Syst. Appl., plane. 41, no. 13, pp. 5843–5857, 2014.

[19] J. Z. Kolter as well as M. A. Maloof, "Learning to detect wild executables that are malicious," J. He. Mach. Know how to do it. Res., the plane. 7, paragraph 11, pp. 2721–2744, 2006.

[20] I. S. Shiaeles, Baptista, and N. Kolokotronis, "A Machine Learning and Binary Visualization-based Novel Malware Detection Program," 2019 IEEE Int. Conf. Conf. Simple. Common. It's working. (The ICC Job, pp. 1-6, 2019.

[21] R. Islam, L. Batten, R. Tian, and S. Versteeg, "String-based malware identification and function selection functionality," Proc. Second Faith in Cybercrime. Software. Computer. It's working. 2010 CTC, pp. 9–17, 2010.

[22] D. Uppal, R. Sinha, V. Mehra, and V. Jain, "API sequence extraction-based malware detection and classification," Proc. Int. 2014. Conf. Conf. Adv. Adv. Software. Computer. Commun. Computer science, ICACCI, 2014, pp. 2337–2342, 2014.

[23] A. Lakhotia, C. Miles, A. Walenstein, A. Singh, "VILO: A fast learning classifier for malware triage nearest-neighbor," J. Software. Computer. Virol, the plane. 9, section 3, pp. 109–123, 2013.

[24] R. S. Pirscoveanu, S. S. Hansen, M. Stevanovic, T. M. T. Larsen, and J. M. Pedersen, "Malware Behavior Analysis: Machine Learning Type Classification."

[25] K. Rieck, Trinius, Willems, Willems, T. Holz, "2011 Automatic machine learning evaluation of malware behavior," pp. 1–30, 2011.pdf.

[26] G. Biau, E. Scornet, "a guided tour of the random forest," study, vol. 25, No. 2, pp. 7-227, 2016.

[27] D. Kirat and G, of course. Vigna, "MalGene: Automatic Evasion Signature Extraction of Malware Research," Ccs, pp. 769–780, 2015.

.

[28] S. Iii, "Cross-Validation K-Fold," 2009. J. "Boot (news)," Obes, Hebebrand. Facts, the flight. 3, number 6, pp. 343–4, 2010.