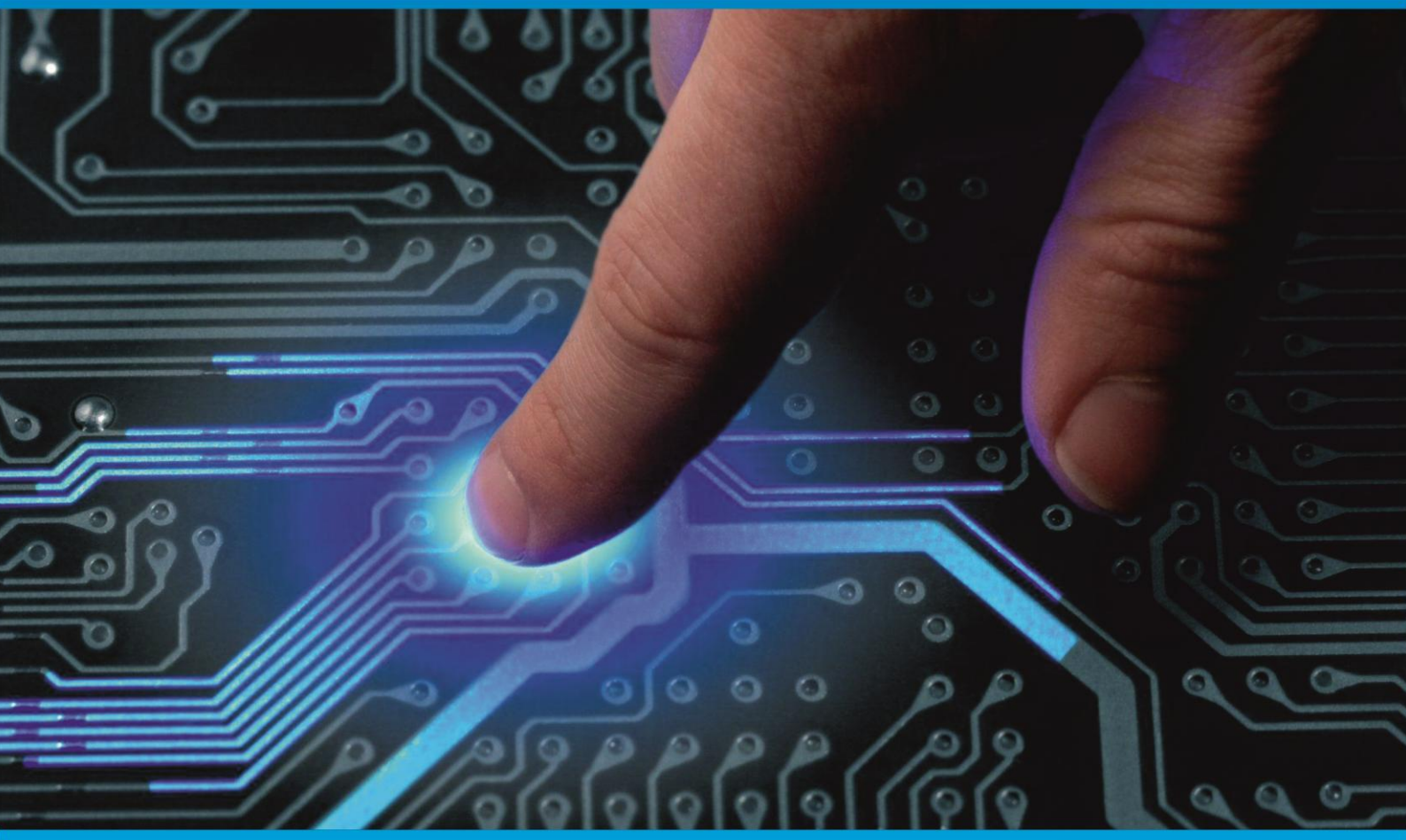
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905031 |

Malware Detection Techniques by Machine Learning

Indrajeet Kumar Yadav[1], Ashutosh Bhardwaj [2], Ms.Rajeshwari Gundla[3]

Student, School of Engineering, Ajeenkya D Y Patil University, Pune, Maharashtra, India [1]

Student, School of Engineering, Ajeenkya D Y Patil University, Pune, Maharashtra, India [2]

Assistant Professor, School of Engineering, Ajeenkya D Y Patil University, Pune, Maharashtra, India [3]

ABSTRACT: In the present computerized world the greater part of the counter malware instruments are mark based which is insufficient to distinguish progressed obscure malware viz. transformative malware. In this paper, we study the recurrence of opcode event to identify obscure malware by utilizing AI strategy. For the reason, we have utilized kaggle Microsoft malware order challenge dataset. The best 20 highlights acquired from fisher score, data acquire, acquire proportion, chi-square and symmetric vulnerability include choice strategies are thought about. We likewise considered different classifiers accessible in WEKA GUI based AI instrument and tracked down that five of them (Random Forest, LMT, NBT, J48 Graft and REPTree) distinguish the malware with practically 100% precision.

KEYWORDS: Metamorphic, Anti-malware, WEKA, Machine Learning.

I. INTRODUCTION

A program/code which is intended to enter the framework without client approval and makes a forbidden move is known as malevolent programming or malware [1]. Malware is a term utilized for Trojan Horse, spyware, adware, worm, infection, ransomware, and so forth As the distributed computing is drawing in the client step by step, the workers are putting away huge information of the clients and consequently tricking the malware designers. The dangers and assaults have additionally expanded with the expansion in information at Cloud Servers. Figure 1 shows the best 10 windows malware detailed by speedy mend [2].

Malwares are grouped into two classifications - original malware and second era malware. The classification of malware relies upon what it means for the framework, usefulness of the program and developing component. The previous arrangements with the idea that the construction of malware stays same, while the later expresses that the keeping the activity with no guarantees, the design of malware changes, after each cycle bringing about the age of new design [3]. This dynamic normal for the malware makes it harder to distinguish, and isolate. The main strategies for malware recognition are mark based, heuristic based, standardization and AI. In past years, AI has been an appreciated methodology for malware safeguards.

In this paper, we research the AI procedure for the characterization of malware. In the following segment, we examine the related work; segment 3 depicts our methodology exhaustively, segment 4 incorporates exploratory results and segment 5 contains derivation of the paper.

II. LITERATURE SURVEY

Grouping malware has been of incredible interest to scientists all through the world, inferable from the impact popular for network protection. Online protection issues have become public issues [9], and AI as well as blockchain [13], IoT [13], and cloud advances including heterogeneous customer networks [5, 14] have been utilized for battling against them. Android [16] requires insurance from Malicious PE documents can cause information spillage [19] and different risks to the security level.

Ren and Chen [19] have contrived another graphical investigation procedure for exploring malware similarity. This strategy changes over dangerous PE documents into neighborhood entropy pictures for noticing inner highlights of malware and afterward standardizes nearby entropy pictures into entropy pixel pictures for arranging malware. Zhang and Luo [14] have proposed a conduct put together examination strategy based with respect to the technique level connection relationship of use's preoccupied API calls.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905031 |

Mahmood Yousefi-Azar has shown work very like what we have done. He has advanced a strategy that he named 'Malytics' which comprises of three sections: extricating highlights, estimating likeness, and ordering everything. The three sections are introduced by a neural organization [19]15 with two secret layers and one single yield layer. The creator could accomplish an exactness of 99.45% [19]. Rushabh Vyas has chipped away at four unique kinds of PE records and has removed 28 highlights, pressing, imported DLLs and capacities from them. He could accomplish 98.7% location rates utilizing AI [15]. Erdogan Dogdu has introduced a paper wherein a shallow profound learning-based component extraction technique named as word2vec is utilized to show any given malware dependent on its opcodes. Characterization is finished utilizing slope help. They have utilized k-crease cross-approval for approving the model execution without trading off with an approval split. He has effectively accomplished an exactness of (96%) [18].

Muhammad Ijaz has utilized two procedures: static and dynamic to separate the highlights of records.[4] Under static component he could accomplish a precision of 99.36% (PE documents) and under unique instrument he could accomplish an exactness of 94.64% [16]. In static investigation, the executable record is broke down on structure bases without executing it in a controlled climate.[6] In unique examination, malware conduct is broke down in a powerful controlled climate[8]. Past this, in the most recent innovation space, information combination models have likewise been ready for malware discovery [10].

III. METHODOLOGY

To distinguish the obscure malware utilizing AI procedure, a stream graph of our methodology is appeared in fig. 1 . It incorporates preprocessing of dataset, promising element choice, preparing of classifier and discovery of cutting edge malware.

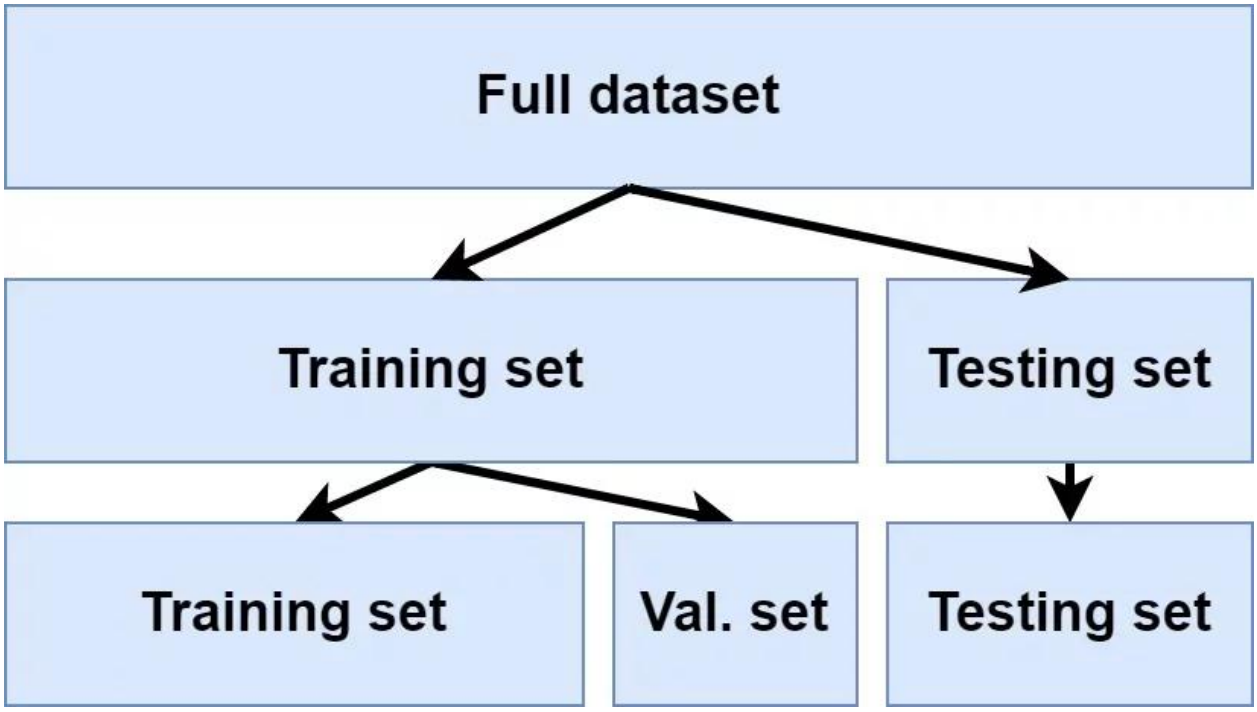


Fig. 1: Machine learning Train/Test split method [21]

3.1 Building the dataset

Microsoft delivered roughly half terabyte for kaggle Microsoft Malware Classification Challenge (2015) [14] containing malware (21653 get together codes). We downloaded malware dataset from kaggle Microsoft and gathered generous projects (7212 records) for the windows stage (checked from virustotal.com) from our school's lab. In our test, we found that as dataset develops, there is an issue of adaptability. This issue builds time intricacy, stockpiling necessity and diminishes framework execution. To beat these issues, decrease of informational collection is essential. Two methodologies can be utilized for information decrease viz. Case Selection (IS) and Feature Selection (FS). In our methodology, Instance Selection (IS) is utilized to decrease the quantity of examples (columns) in dataset by choosing

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905031 |

most proper occurrences. Then again, Feature Selection is utilized for the determination of most pertinent credits (highlights) in dataset These two methodologies are powerful in information decrease as they channel and spotless, loud information which brings about less capacity, time intricacy and improve the precision of classifiers [17] [18].

3.2 Data Preparation

From the prior investigations [12] we have discovered that opcodes contain a more significant portrayal of the code, so in proposed approach, we use opcodes as highlights. Malware dataset contains 21653 get together codes of malware portrayal, a mix of 9 distinct families, i.e., Ramnit, Lollipop, Kelihos_ver3, Vundo, Simda, Tracur, Kelihos_ver1, Obfuscator.ACY, Gatak. Gathered considerate executables dismantled utilizing objdump utility accessible in Linux framework to get the opcodes. In the malware dataset, we have tracked down that greatest size of get together code is 147.0 MB, so all the favorable gathering over the 147.0 MB are not considered for the investigation. From prior examinations, we found that there are 1808 special opcodes [12] so in our methodology, there are 1808 highlights for AI. At that point the recurrence of each opcode in each malware and the benevolent record is determined. After that in each malware and considerate record complete opcodes weight is determined. At that point it is seen that there are 91.3 % malware document and 66 % benevolent record which contains opcodes weight under 40000. So to keep up the extent of malware and amiable every one of the records under 40000 weight is chosen. After this progression, 19771 and 4762 malware and kind records are left for examination. The subsequent stage is to eliminate boisterous information from malware for that we have determined the malware and amiable records in the 500 time frames weight. Those spans in which there are no favorable documents, malware records are likewise erased around there. In this manner further stretches 100, 50, 10 and 2 of opcodes loads are made as demonstrated in Table 1 to eliminate the commotion from malware. At last, dataset contains 6010 Malware and 4573 generous documents[1,3,5,7,9].

Sr. No.	Opcode Weight inter	Number of malware files	Number of benign files
1	1-50	39	12
2	51-100	11	3
3	101-150	11	18
4	201-250	33	1
5	251-300	43	1
6	351-400	26	2
7	401-450	18	1
8	451-500	23	33
9	451-500	11	8
10	501-550	31	5
11	551-550	129	38
12	601-650	368	24
13	601-650	356	12
14	651-700	303	7
15	701-750	111	15
16	751-850	73	45
17	851-900	193	62

Table 1: Opcodes loads demonstrated

3.3 Feature Selection

Highlight determination is a significant piece of AI. In proposed approach, there are 1808 highlights among them many don't give to the exactness and even decline it. In our difficult decrease of highlights is critical to looking after precision. Consequently we initially utilized Fisher Score (FS) [19] for include choice and later four more component choice methods were likewise contemplated. The five element choice technique utilized in this methodology what capacities as indicated by the channels approach [20][19]. In this strategy, relationship of each element with the class (Malware or kindhearted) is measured, and its commitment to order is determined. This technique is free of any characterization calculation dissimilar to covering approach and permits to think about the exhibition of various classifiers[11][12]. In this methodology, Fisher Score (FS), Information Gain (IG), Gain Ratio (GR), Chi-Square (CS) and Uncertainty Symmetric(US) is utilized.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905031 |

3.4 Training of the Classifiers

After the component choice, following stage is to track down the best classifier for the identification of cutting edge malware.[13] Subsequent stage is to think about various classifiers on FS, IG, GR, CS and US utilizing top 20 highlights[14]. We examined nine classifiers viz. Choice Stump, Logistic Model Tree (LMT), Random Forest, J48, REPTREE, Naïve Bayes Tree (NBT), J48 Graft, Random Tree, Simple CART accessible in WEKA [18]. WEKA is an open source GUI based AI instrument [15]. We run every one of these classifiers on each element choice method utilizing 10–overlap cross-approval to prepare the classifiers [20]. Shows the exactness of every classifier concerning highlight determination strategy[17]. Obviously Fisher score technique is best in among all and got precision 100 % if there should be an occurrence of Random Forest, LMT, NBT and Random Tree. So in our proposed, Fisher Score performs better compared to different strategies viz. Data Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty and Chi-Square [15].

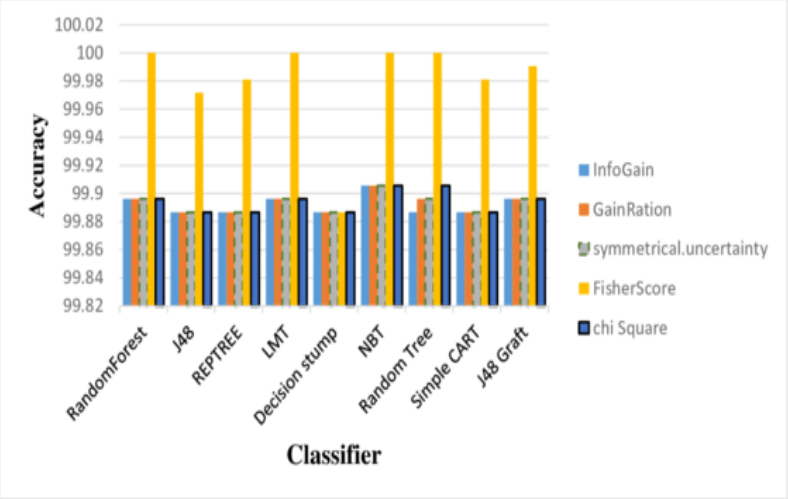


Fig 2: Fisher score technique report

3.5 Unknown Malware Detection

In a prior area, we have seen that Random Forest, LMT, NBT, J48 Graft and Random Tree accomplished greatest precision, so we chose these five classifiers for profundity examination [14]. We have arbitrarily chosen 3005 malware and 2286 generous projects which are half of the generally dataset[19].

IV. EXPERIMENTAL RESULTS

As referenced in area 3, malware is as of now in get together code just benevolent are dismantled. At that point opcodes event is determined for all malware and favorable projects. In next commotion from malware, information is eliminated by making a time frame weight for example 500, 100, 50, 10, 5 and 2 for malware and kindhearted records. Span in which there are no kind records, malware documents are erased[6]. To track down the predominant aspects or to eliminate insignificant element we utilized five element choice techniques and found that there are 20 highlights which are ruling in the characterization cycle. Shows that Fisher Score outflanks among five element determination strategies [2] [4].

Bogus Negative: the no. of malware recognized as amiable. Table 2 shows the outcome acquired by the best 5 classifiers. The investigation shows that the chose five classifiers precision is pretty much same.

V. CONCLUSION

In this paper, we have introduced a methodology dependent on opcodes event to improve malware identification precision of the obscure progressed malware [2]. Code muddling method is a test for signature based procedures utilized by cutting edge malware to avoid against malware devices [5]. Proposed approach utilizes Fisher Score strategy for the component determination and five classifiers used to uncover the obscure malware [7]. In proposed approach Random backwoods, LMT, J48 Graft, and NBT distinguish malware with 100% precision which is superior to the exactness (99.8%) detailed by Ahmadi et al [3]. (2016). In future, we will execute proposed approach on various datasets and will act in the profound examination for the grouping of cutting edge pernicious programming [8].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905031 |

REFERENCES

[1] A. Sharma and S. K. Sahay, "Evolution and Detection of Polymorphic and Metamorphic Malware: A Survey," International Journal of Computer Application, vol. 90, no. 2, pp. 7–11, 2014.
[2] E. S. Solutions and Q. Heal, "Quick Heal Quarterly Threat Report | Q1 2017," 2017 url:http://www.quickheal.co.in/resources/threat-reports . [Accessed: 13-june-2017].
[3] A. Govindaraju, "Exhaustive Statistical Analysis for Detection of Metamorphic Malware," Master's project report, Department of Computer Science, San Jose State University, 2010.
[4] M. G. Schultz, E. Eskin, and S. J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables," 2001.
[5] D. Bilar, "Opcodes As Predictor for Malware," International Journal of Electronic Security and Digital Forensics, vol. 1, no. 2, pp. 156–168, 2007.
[6] Y. Elovici, A. Shabtai, R. Moskovitch, G. Tahan, and C. Glezer, "Applying Machine Learning Techniques for Detection of Malicious Code in Network Traffic," Annual Conference on Artificial Intelligence. Springer Berlin Heidelberg, pp. 44–50, 2007.
[7] R. Moskovitch, D. Stopel, C. Feher, N. Nissim, N. Japkowicz, and Y. Elovici, "Unknown malcode detection and the imbalance problem," Journal in Computer Virology, vol. 5, no. 4, pp. 295–308, 2009.
[8] R. Moskovitch et al., "Unknown malcode detection using OPCODE representation," Intelligence and Security Informatics. Springer Berlin Heidelberg, vol. 5376 LNCS, pp. 204–215, 2008
[9] I. Santos, J. Nieves, and P. G. Bringas, "Semi-supervised learning for unknown malware detection," International Symposium on Distributed Computing and Artificial Intelligence. Springer Berlin Heidelberg, vol. 91, pp. 415–422, 2011.
[10] I. Santos, F. Brezo, X. Ugarte-Pedrero, and P. G. Bringas, "Opcode sequences as representation of executables for data-miningbased unknown malware detection," Information Sciences, vol. 231, pp. 64–82, 2013.
[11] A. Shabtai, R. Moskovitch, C. Feher, S. Dolev, and Y. Elovici, "Detecting unknown malicious code by applying classification techniques on OpCode patterns," Security Informatics, vol. 1, no. 1, p. 1, 2012.
[12] A. Sharma and S. K. Sahay, "An effective approach for classification of advanced malware with high accuracy," International Journal of Security and its Applications, vol. 10, no. 4, pp. 249–266, 2016.
[13] S. K. Sahay and A. Sharma, "Grouping the Executables to Detect Malwares with High Accuracy," Procedia Computer Science, vol. 78, no. June, pp. 667–674, 2016.
[14] Kaggle, "Microsoft Malware Classification Challenge (BIG 2015)" Microsoft, URL: https://www.kaggle.com/c/malwareclassification , [Accessed : 10/December/2016].
[15] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, and G. Giacinto, "Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification," ACM Conference Data Application Security Priv., pp. 183–194, 2016.
[15] Ahmadi, Mansour, et al. "Novel feature extraction, selection and fusion for effective malware family classification." *Proceedings of the sixth ACM conference on data and application security and privacy*. 2016.
[16] J. Drew, M. Hahsler, and T. Moore, "Polymorphic malware detection using sequence classification methods and ensembles," EURASIP J. Inf. Secur., vol. 2017, no. 1, p. 2, 2017.
[17]Drew, Jake, Michael Hahsler, and Tyler Moore. "Polymorphic malware detection using sequence classification methods and ensembles." *EURASIP Journal on Information Security* 2017.1 (2017): 1-12.
[18] J. Derrac, S. García, and F. Herrera, "A first study on the use of co evolutionary algorithms for instance and feature selection," International Conference on Hybrid Artificial Intelligence Systems. Springer Berlin Heidelberg, pp. 557–564, 2009.
[19] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning Artificial intelligence, vol. 97, no. 1–2, pp. 245–271, 1997.
[20] T. R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, no. 5439, pp. 531–537, 1999.
[21] T. G. Dietterich, "Machine learning in ecosystem informatics and sustainability," IJCAI, pp. 8-13 2009.
[22] link of website (21/4/21):-
https://www.machinecurve.com/index.php/2020/11/16/how-to-easily-create-a-train-test-split-for-your-machine-learning-model/

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details