MACHINE LEARNING APPLICATIONS IN MALWARE CLASSIFICATION: A META-ANALYSIS LITERATURE REVIEW

Tjada Nelson, Austin O'Brien and Cherie Noteboom

Beacom College of Computer & Cyber Sciences, Dakota State University, Madison, South Dakota

ABSTRACT

With a text mining and bibliometrics approach, this study reviews the literature on the evolution of malware classification using machine learning. This work takes literature from 2008 to 2022 on the subject of using machine learning for malware classification to understand the impact of this technology on malware classification. Throughout this study, we seek to answer three main research questions: RQ1: Is the application of machine learning for malware classification growing? RQ2: What is the most common machine-learning application for malware classification? RQ3: What are the outcomes of the most common machine learning applications? The analysis of 2186 articles resulting from a data collection process from peer-reviewed databases shows the trajectory of the application of this technology on malware classification as well as trends in both the machine learning and malware classification fields of study. This study performs quantitative and qualitative analysis using statistical and N-gram analysis techniques and a formal literature review to answer the proposed research questions. The research reveals methods such as support vector machines and random forests to be standard machine learning methods for malware classification in efforts to detect maliciousness or categorize malware by family. Machine learning is a highly researched technology with many applications, from malware classification and beyond.

KEYWORDS

Malware, Malware Classification, Machine Learning.

1. INTRODUCTION

Machine learning is a technology that has been at the forefront of academic research since its inception in the 1990s [1]. Applications from Machine learning is defined as the capacity for a system to learn from a problem-specific data source to identify patterns in that data build that provide insight around that data [2]. The technology behind machine learning enables a wide range of efficiency for computer systems across many disciplines. Therefore, machine learning has been the subject of many papers in academia. As of December 2022, Google Scholar, an academic article search engine and aggregator, returns over 5.4 million results for the term "machine learning." Through this research, machine learning has successfully been applied to several real-world applications ranging from healthcare to gaming.

Machine learning can be divided into three types: supervised, unsupervised, and reinforcement. Supervised machine learning takes test data representing known desired results and is used to

David C. Wyld et al. (Eds): NLDML, CCITT, EDUTEC, AUEN -2023
pp. 113-124, 2023. IJCI – 2023 DOI:10.5121/ijci.2023.120109

produce a system that predicts future results. For example, supervised machine learning is the method to perform classification or face recognition. Unsupervised machine learning takes data and tries to identify clusters to classify similar or related data points and insights. This machine learning type builds recommendation systems, anomaly detection, and tracking buy habits in customer transactions. Reinforcement machine learning allows the system to operate and then notify it when it makes mistakes, so it learns to avoid them. This machine learning type is used to create video game AI and operation simulations. Machine learning has proved to be very useful in solving real-world problems in all these application types.

A significant problem in cybersecurity is malware which plays a major role in cybercrime. AV-TEST institute reported more than 1 billion infected files in 2021, showing the sheer volume of malware cybersecurity analysts contend with year over year [3]. Malware, by definition, cause harm to systems and inflict other damages, such as financial costs, to victims. Cybercrime costed victims around $6 Trillion USD in 2021, further highlighting the need for this problem to be addressed [4]. The term malware is derived from the words malicious, and software based on its usage to cause damage on a target system. Malware is used by cybercriminals to conduct operations and run illicit businesses focused on the perpetration of cybercrime [6].

Analyzing malware allows defenders to detect infections and detect protections. There are two types of analysis commonly used to investigate malware, static and dynamic. Static analysis entails investigating a malware sample without executing it while dynamic analysis is observing the malware activity while it is running [7]. There are various techniques used for static analysis such as file fingerprinting, extraction of hard coded strings, file format, anti-virus scanning, packer detection and disassembly [8]. These static analysis techniques rely on the binary representation of the program to provide a safe and manual approach to analysis. File fingerprinting and file format techniques provide characteristics about the program. anti-virus scanning and packer detection offer identification and classification of the program and its contents. Hard coded strings and disassembly offer detailed extraction of data harboured within the program. Static analysis could provide insight on the information that is missed during dynamic analysis do to program paths not executed [9]. It allows for the classification of these malware to get a general understanding of the nature of the malware versus an in-depth analysis of the binary.

Ultimately allowing defenders to determine if a binary is malicious or not and then use that information to protect their systems. Although static analysis is generally safer than dynamic analysis the original source code of the analyzed program is usually not available. Techniques such as machine learning against this static information provides additional efficiencies in the analysis of these binaries. The malware must be detected to prevent or recover from malware infections. One method of this detection is malware classification, which identifies malicious software and its functionality from the assortment of software present on a system. Understanding if malware is on a system and the functionality of that malware helps an analyst provide the proper corrective actions to minimize the risk of further damage incurred by the malicious event. Machine learning is particularly suited for this type of classification and has been applied in mass to the problem of malware classification.

However, with the wide variety of machine learning applications, the impact of machine learning on this problem is still in question. Wagstaff proposes three focus areas to address the impact of machine learning which include benchmarking of datasets, performance measurements, and how it is applied [5]. These focus areas raise legitimate questions about how machine learning is used for malware classification and the impact of these applications. Thus the motivation for this study is to survey the current literature in the field of malware classification and machine learning to

identify the most common uses and applications. This would contribute to the current body of knowledge by building a foundation for future research around the current applications
To this end, this study will survey the volume of machine learning literature directed toward malware classification. Furthermore, this study aims to gather data to answer the following research questions: 1. Is the application of machine learning for malware classification growing? 2. What is the most common machine learning application used for malware classification? 3. What are the outcomes of the most common machine learning applications? Using the information extracted from this research to answer the above questions will also provide an up-to-date state-of-the-art on machine learning in malware classification applications.

2. METHODOLOGY

2.1. Search Strategy and Data Collection

We have defined three research questions to guide our meta-analysis of research publications on machine learning and malware classification. We then developed search queries targeting four databases, Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers Xplore (IEEE Xplore), Science Direct, and Web of Science. The query leveraged for these databases includes any references to the terms "machine learning" and "malware classification" in the metadata of a publication **("All Metadata": malware classification) AND ("All Metadata": machine learning)**. The results from this query were filtered down to publications with dates from 2017 to September 2022. The reason for this constraint allows the research to focus on the current state versus historical applications.

The authors reviewed the titles of the articles from this 2186 sized article dataset to determine the relevance of the titles toward malware classification and machine learning, removing any outliers. 372 articles were removed, leaving a total of 2186 unique articles gathered. Based on additional searches performed authors were able to determine these search times and filters provided the most suitable dataset for performing meta-analysis against terms in the article metadata. Metadata from these articles were collected, including but not limited to publication title, authors, abstract, keywords, and publication year. These fields were extracted by importing all the dataset articles into the citation tool Zotero by exporting all of the articles to a CSV format. The collected metadata enables this research to focus on the articles most relevant to answering the proposed research questions. Beyond this study, this article collection will be used in future literature reviews.

2.2. Data Analysis

From this metadata collection, we used the text mining process described in figure 1 to complete the analysis. This text mining process was informed by the work of Zheng et. al. however the implementation of this mining process was executed by python instead of Vos viewer software [10]. All stop words were removed, and the metadata was tokenized to prepare word cloud and N-gram analysis. With more insight into the trends and themes of the dataset, the N-gram analysis was leveraged to identify terms of interest, such as machine learning applications for literature review.

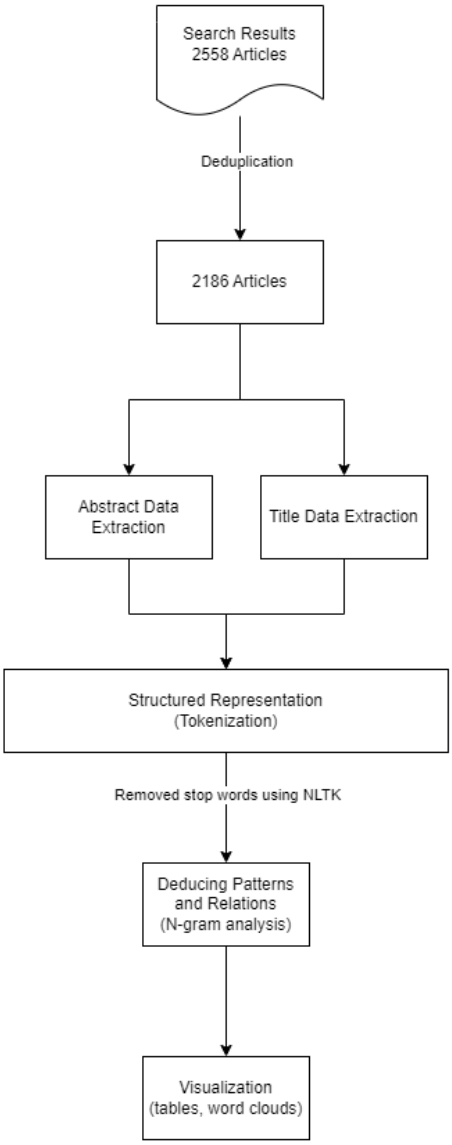

Figure 1. Text Mining Process.

3. RESULTS

3.1. Exploratory Data Analysis

Initial descriptive data analysis was conducted using our article data set to gain insights around current trends. This analysis included 1. Determining the number of articles published per year, 2. Calculating word frequency of top ten words, 3. Generating a word cloud, and 4. Performing N-gram analysis.

Figure 2 shows the number of articles published each year from 2017 to September 2022. Based on these yearly publishing numbers, there is a clear trend upward year over year, with a drop in 2022 due to the year not being completed by the time of data collection. The upward trendline answers our first research question: R1. Is the application of machine learning for malware

classification growing? Based on these results, we see a clear upward trend for the application of machine learning for malware classification.

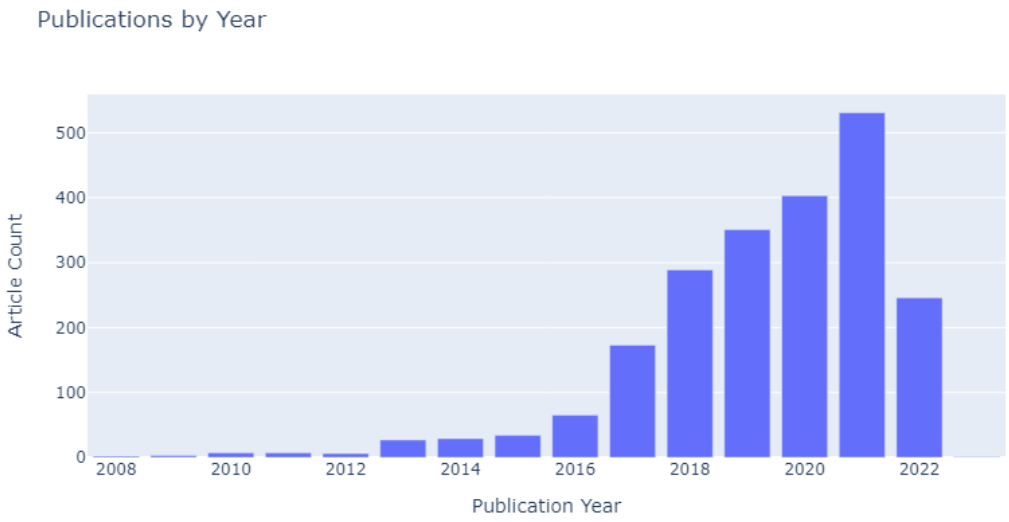

Figure 2. Number of articles published in each year.

A subsequent review of the top 10 words in the title field of metadata for the collection after removing stop words using the python libraries Word Cloud and NLTK used for stop words was performed. Table 1 shows the top 10 words from this data set. Looking at the results, ‘malware,’ ‘learning,’ and ‘classification’ were some of the most used terms. Terms such as ‘detection’ are also listed in this top 10 list, which could suggest that a detection system is the most common result of machine learning for malware classification.

Table 1. Top 10 most frequent title words.

Word	**Count**
malware	812
learning	735
detection	673
using	541
classification	503
machine	406
android	302
based	297
network	280
deep	262

Another version of word frequency analysis conducted in this study is the Top 10 most frequent words in the abstracts of the collected data. Table 2 shows the results of this analysis, were as expected, ‘malware,’ ‘machine,’ ‘learning,’ and ‘classification’ are all frequently used. In addition, as identified with the title frequency, ‘detection’ is also frequently used.

Table 2. Top 10 most frequent abstract words.

Word	Count
malware	5075
feature	3473
learning	3413
model	2911
classification	2825
detection	2593
method	2494
machine	2253
data	2213
network	1920

Word clouds provide an opportunity to visually observe the more extensive frequency data set, which can lead to additional findings that were not readily apparent in the previous tables. Word clouds represent volume by the size of the specific term instead of a word count value. This approach provides a visual cue for trends and insights. Figure 3 shows a word cloud of the most frequently used words in the abstracts of the data set. We can confirm that ‘malware,’ ‘feature,’ ‘learning,’ and ‘model’ are the most frequently used words based on the size of the associated text of the word.

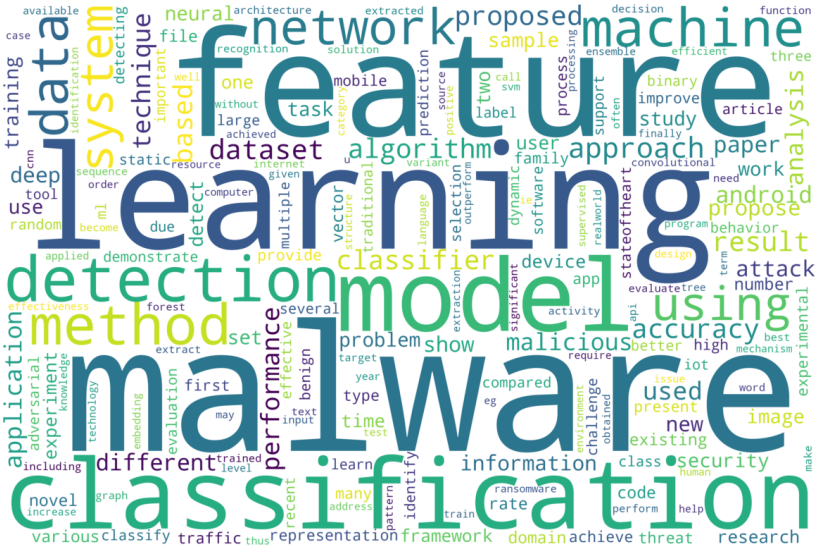

Figure 3. Word cloud of most frequent abstract words.

3.2. N-Gram Analysis of Abstract Terms

N-gram analysis takes combinations of terms within a text dataset to determine the frequency of pairs of words seen together. The results of this type of analysis allow the researcher to observe more sophisticated themes from a dataset. The aggregated abstract data from this study was placed in the NLTK n-gram library to identify these patterns in bi-gram, trigram, and 4-gram analysis. Table 3 displays the top 20 bi-gram terms within the abstract words. As expected, sets such as (“machine,” “learning,”) or (“malware,” “detection”) have the highest count. Looking down at this table, we find sets that provide insight around machine learning applications such as (“neural,” “network”), (“deep,” “learning”), (random, forest), (support, vector), (vector, machine)

and (convolutional, neural). These terms suggest the types of machine learning applications used within malware classification. This dataset also suggests that android malware is a prominent subject of this type of study based on the counts within the n-gram analysis.

Table 3. Top 20 bi-gram analysis from abstract words.

Bi-gram Words	**Count**
(machine, learning)	1695
(malware, detection)	969
(neural, network)	706
(deep, learning)	581
(android, malware)	437
(experimental, result)	388
(learning, algorithm)	383
(result, show)	368
(learning, model)	363
(malware, classification)	300
(random, forest)	296
(feature, selection)	263
(learning, technique)	260
(support, vector)	258
(proposed, method)	249
(vector, machine)	244
(malware, family)	239
(learning, method)	223
(convolutional, neural)	220
(classification, accuracy)	207

Table 4 provides a more straightforward depiction of these sets of terms by showing the top 10 tri-grams from the abstract words in the dataset. Finally, looking at the machine learning applications inferred from this table data, (“support,” “vector,” “machine”) and (“convolutional,” “neural,” “network”) are presented as prominent applications for malware classification using machine learning.

Table 4. Top 10 tri-gram analysis from abstract words.

Tri-gram Words	**Count**
(machine, learning, algorithm)	272
(support, vector, machine)	239
(convolutional, neural, network)	218
(machine, learning, technique)	195
(machine, learning, model)	168
(experimental, result, show)	163
(android, malware, detection)	155
(deep, learning, model)	129
(deep, neural, network)	127
(machine, learning, classifier)	118

The four-gram analysis of the abstract words are displayed in Table 5. This table shows Support Vector Machines and Convolutional Neural Networks at the dataset's top four-gram and most prominent machine learning application. Towards the bottom of the top 10 list are decision tree random forests and recurrent neural networks, also machine learning applications.

Table 5. Top 10 four-gram analysis from abstract words.

Four-gram Words	**Count**
(support, vector, machine, svm)	88
(convolutional, neural, network, cnn)	81
(experimental, result, show, proposed)	35
(convolutional, neural, network, cnns)	28
(using, machine, learning, technique)	24
(machine, learning, deep, learning)	22
(decision, tree, random, forest)	21
(microsoft, malware, classification, challenge)	21
(support, vector, machine, classifier)	20
(recurrent, neural, network, rnn)	19

3.3. Distribution of Machine Learning Applications

Based on the N-gram analysis conducted during this study, we can address research question 2, addressing the most common machine learning applications used for malware classification. The study observes random forest decision trees (which will be shortened to random forest to generalize this application), support vector machines, convolutional neural networks, and recurrent neural networks as the most common applications of machine learning in malware classification. Figure 4 shows the number of articles containing the designated terms, which concluding support vector machines are the most common machine learning application for malware classification.

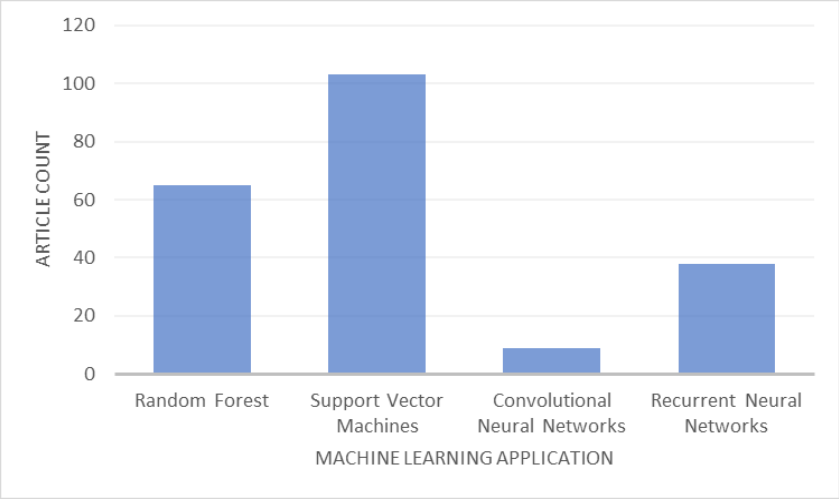

Figure 4. Number of articles with the specific machine learning application

3.3.1. Random Forest

Random forest is a classifier based, supervised machine-learning algorithm which leverages multiple decision trees to identify patterns [11]. Using this model takes a training set and a test set to build a classifier for the model. In many of the applications observed in this study, static features of the software are used to determine whether it is malware or not. In a study by Bowen Sun et al., they used static features with several machine learning algorithms, including support vector machines, decision trees, random forests and others, finding they were able to get the highest F1 score which was the highest of all tests conducted [12]. The Zhang et al. study also resulted in random forest as the highest-performing algorithm using the Ember dataset. The Zhang et al. study also noted that random forest is the most cost-effective algorithm they tested [13]. The dataset from this study found a few articles that targeted this approach against android malware versus the traditional computer system malware; a finding supported by the N-gram analysis findings. In total, 66 articles from the dataset specifically call out random forest in the abstract, making it the second most common machine learning application for malware classification.

3.3.2. Support Vector Machine

Support vector machines (SVM) is a learn-by-example machine learning algorithm, essentially using a large dataset of a particular pattern to determine how to find that pattern. It is very good at binary classification but struggles with granularity [14]. In our use case, SVM would be useful for identifying if a binary was malicious but not the category of maliciousness of the binary. Gravilut et. al, achieved a 99% accuracy and 0.75% false positive rate using a Boosted SVM model using static features of the binary as their data set. Wang and Wu were able to generate promising results by using SVM to detect packed executables, a common characteristic of malware; this study offered a use case for malware classification that differed from static feature extraction [15]. In the study “Malware Family Classification using Active Learning by Learning” by Chen et al., the learning algorithms using byte and assembler data from binaries were used to create an SVM model. The group found that using SVM with Active Learning by Learning (ALBL) could effectively perform malware classification [16]. Furthermore, 103 articles from the dataset directly mentioned SVM, the highest frequency count of the terms explored. Based on this number, SVM would be the most common machine learning application for malware classification, addressing research question 2.

3.3.3. Convolutional Neural Network

Based on early findings from the study of biological vision, convolutional neural networks (CNN) uphold that notable aspect of the neural network’s data set, making it a good solution for image and natural language processing [17]. In our dataset, only 9 articles directly mention CNN as a model in their abstract. This image-based processing application represented several of the articles in the data set. Fathurrahman et al., leveraged CNNs to perform malware classification on embedded systems using images of malware from the Malevis dataset, finding an approach that was 19.22% accurate and could be run on an embedded IoT system effectively [18]. Another notable article from the dataset containing CNN in the abstract is “Malware Classification using Deep Convolutional Neural Networks” by Kornish et. al, found using transfer learning to retain the deep convolutional neural network. However, they noted that the article’s architecture had little practical application [19]. Lin et. al, used RNN to classify IoT malware in their study; this research used opcode sequence to detect patterns using RNN.

3.3.4. Recurrent Neural Network

Recurrent neural networks (RNN) have a simple neural network structure with a built-in feedback loop allowing it to be used for forecasting. This algorithm is typically used to detect patterns in a data sequence [20]. For example, with malware classification, Shukla et. al, took the approach of visualizing the behavior of the malware in the form of a heat map and leveraging RNN to achieve a 94% accuracy score. There were 37 articles in the dataset which mentioned RNN in the abstract of the articles around malware classification, and images were used as the dataset in the machine learning process.

3.4. Research Themes

During the research of this study, Android malware emerged as a trending topic in the dataset. A significant number of articles in the data set showed up in the top 5 frequently used bi-grams from the n-gram analysis. Mobile malware, including Android-based malware, has grown in market share compared to other malware types despite the general malware decline [21]. The shift in the malware landscape and the size of the mobile market are apparent trends. These trends could explain why android malware classification solutions are turning to machine learning increasingly.

Several articles extracted static features of malware using pe and malware analysis tools to create a dataset for machine learning. Another feature set for datasets was images of the binary using processing tools. These two feature sets were the most prevalent within the dataset versus others, such as function call graphs and real-time activity. Future research in function call graphs and real-time activity as machine learning features is worthwhile exploring in the future.

4. DISCUSSION

From the analysis of 2186 articles around the application of machine learning for malware classification, it is evident that machine learning provides clear advantages in this task. Furthermore, machine learning can be performed in a variety of ways using different algorithms and datasets.

4.1. R1: Is the application of machine learning for malware classification growing?

This study's first research question sought to understand if the usage of machine learning for malware classification is increasing. The study answers this question by creating the chart using the distribution of articles by publication year within the dataset displayed in Figure 2. This chart shows an apparent increase in publications year over year until 2022, which was the current year at the time of the study. The information from this chart shows an apparent growth in articles concerning malware classification and machine learning, answering research question 1.

4.1.1. R2: What is the most common machine learning application used for malware classification?

This study used an N-gram analysis of abstract text terms against the dataset to understand the more complex topics discussed within these articles. This research analysis revealed several terms associated with machine learning applications and algorithms. Extracting the machine learning algorithms from the N-gram results, the study determined the most common by string searching those terms across the whole dataset. This effort found support vector machines (SVM) as the most common application of machine learning in malware classification. 103 articles

within the dataset mentioned SVM in the abstract, with the following more frequent application being random forest with 66 article mentions.

4.1.2. R3: What are the outcomes of the most common machine learning applications?

This study aimed to answer the third research question around identifying the outcomes of the typical machine learning applications used for malware classification. This question would provide insight into the contributions to the body of knowledge this area of study has provided. Based on the literature reviews guided by the N-gram analysis results, these typical machine learning applications' outcomes are malware detection and family categorization. The outcomes were shared across all the machine-learning applications explored in this research. This outcome was also prevalent in the title word analysis, with detection as one of the top terms present in a title.

5. CONCLUSIONS

In conclusion, this study reviewed 2186 articles about malware classification and machine learning. We found that machine learning for malware classification is steadily growing in popularity. The most common machine learning applications used for this purpose are support vector machines and other classifiers, primarily used for malware detection and malware family categorization. These findings highlight the potential of machine learning as a powerful tool in the fight against malware and highlight the importance of continued research in this area.

REFERENCES

[1] Ö. Çelik, "A Research on Machine Learning Methods and Its Applications," *J. Educ. Technol. Online Learn.*, Sep. 2018, doi: 10.31681/jetol.457046.

[2] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[3] M. Asam *et al.*, "Detection of Exceptional Malware Variants Using Deep Boosted Feature Spaces and Machine Learning," *Appl. Sci.*, vol. 11, no. 21, p. 10464, Nov. 2021, doi: 10.3390/app112110464.

[4] F. A. Aboaoja, A. Zainal, F. A. Ghaleb, B. A. S. Al-rimy, T. A. E. Eisa, and A. A. H. Elnour, "Malware Detection Issues, Challenges, and Future Directions: A Survey," *Appl. Sci.*, vol. 12, no. 17, p. 8482, Aug. 2022, doi: 10.3390/app12178482.

[5] A. Submission, "Machine Learning that Matters," p. 6.

[6] M. Lindorfer, A. Di Federico, F. Maggi, P. M. Comparetti, and S. Zanero, "Lines of malicious code: insights into the malicious software industry," in *Proceedings of the 28th Annual Computer Security Applications Conference on - ACSAC '12*, Orlando, Florida, 2012, p. 349. doi: 10.1145/2420950.2421001.

[7] A. Afianian, S. Niksefat, B. Sadeghiyan, and D. Baptiste, "Malware Dynamic Analysis Evasion Techniques: A Survey." arXiv, Nov. 03, 2018. Accessed: Jan. 11, 2023. [Online]. Available: http://arxiv.org/abs/1811.01190

[8] M. Parekh and G. Kulkarni, "A Survey on 'Malware Analysis Techniques, its Detection and Mitigation.,'" vol. 08, no. 08, 2021.

[9] W. Aman, "A Framework for Analysis and Comparison of Dynamic Malware Analysis Tools," *Int. J. Netw. Secur. Its Appl.*, vol. 6, no. 5, pp. 63–74, Sep. 2014, doi: 10.5121/ijnsa.2014.6505.

[10] D. Zeng, C. Noteboom, K. Sutrave, and R. Godasu, "A Meta-Analysis of Evolution of Deep Learning Research in Medical Image Analysis".

[11] S. Yoo, S. Kim, S. Kim, and B. B. Kang, "AI-HydRa: Advanced hybrid approach using random forest and deep learning for malware classification," *Inf. Sci.*, vol. 546, pp. 420–435, Feb. 2021, doi: 10.1016/j.ins.2020.08.082.

[12] B. Sun, Q. Li, Y. Guo, Q. Wen, X. Lin, and W. Liu, "Malware family classification method based on static feature extraction," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, Dec. 2017, pp. 507–513. doi: 10.1109/CompComm.2017.8322598.

[13] S.-H. Zhang, C.-C. Kuo, and C.-S. Yang, "Static PE Malware Type Classification Using Machine Learning Techniques," in *2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, Tainan, Taiwan, Aug. 2019, pp. 81–86. doi: 10.1109/ICEA.2019.8858297.

[14] W. S. Noble, "What is a support vector machine?," *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.

[15] T.-Y. Wang and C.-H. Wu, "Detection of packed executables using support vector machines," in *2011 International Conference on Machine Learning and Cybernetics*, Jul. 2011, vol. 2, pp. 717–722. doi: 10.1109/ICMLC.2011.6016774.

[16] C.-W. Chen, C.-H. Su, K.-W. Lee, and P.-H. Bair, "Malware Family Classification using Active Learning by Learning," in *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, Feb. 2020, pp. 590–595. doi: 10.23919/ICACT48636.2020.9061419.

[17] G. W. Lindsay, "Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future," *J. Cogn. Neurosci.*, vol. 33, no. 10, pp. 2017–2031, Sep. 2021, doi: 10.1162/jocn_a_01544.

[18] A. Fathurrahman, A. Bejo, and I. Ardiyanto, "Lightweight Convolution Neural Network for Image-Based Malware Classification on Embedded Systems," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jan. 2022, pp. 12–16. doi: 10.1109/ISMODE53584.2022.9743111.

[19] D. Kornish, J. Geary, V. Sansing, S. Ezekiel, L. Pearlstein, and L. Njilla, "Malware Classification using Deep Convolutional Neural Networks," in *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Oct. 2018, pp. 1–6. doi: 10.1109/AIPR.2018.8707429.

[20] R. M. Schmidt, "Recurrent Neural Networks (RNNs): A gentle Introduction and Overview." arXiv, Nov. 23, 2019. Accessed: Dec. 31, 2022. [Online]. Available: http://arxiv.org/abs/1912.05911

[21] European Union Agency for Cybersecurity., *ENISA threat landscape 2022: July 2021 to July 2022*. LU: Publications Office, 2022. Accessed: Jan. 11, 2023. [Online]. Available: https://data.europa.eu/doi/10.2824/764318