

New Datasets for Dynamic Malware Classification

1st Buket Gençaydın
Computer Engineering
Gebze Technical University
Kocaeli, Turkey
b.gençaydin2019@gtu.edu.tr

2nd Ceyda Nur Kahya
Management Information Systems
Kadir Has University
İstanbul, Turkey
20161708007@stu.khas.edu.tr

3rd Ferhat Demirkıran
Department of Cyber Security
Kadir Has University
İstanbul, Turkey
ferhat.demirkıran@khas.edu.tr

4th Berkant Düzgün *
Management Information Systems
Kadir Has University
İstanbul, Turkey
20151709009@stu.khas.edu.tr

5th Aykut Çayır
Management Information Systems
Kadir Has University
İstanbul, Turkey
aykut.cayir@khas.edu.tr

6th Hasan Dağ
Management Information Systems
Kadir Has University
İstanbul, Turkey
hasan.dag@khas.edu.tr

Abstract—Nowadays, malware and malware incidents are increasing daily, even with various anti-viruses systems and malware detection or classification methodologies. Many static, dynamic, and hybrid techniques have been presented to detect malware and classify them into malware families. Dynamic and hybrid malware classification methods have advantages over static malware classification methods by being highly efficient. Since it is difficult to mask malware behavior while executing than its underlying code in static malware classification, machine learning techniques have been the main focus of the security experts to detect malware and determine their families dynamically. The rapid increase of malware also brings the necessity of recent and updated datasets of malicious software.

We introduce two new, updated datasets in this work: One with 9,795 samples obtained and compiled from VirusSamples and the one with 14,616 samples from VirusShare. This paper also analyzes multi-class malware classification performance of the balanced and imbalanced version of these two datasets by using Histogram-based gradient boosting, Random Forest, Support Vector Machine, and XGBoost models with API call-based dynamic malware classification. Results show that Support Vector Machine, achieves the highest score of 94% in the imbalanced VirusSample dataset, whereas the same model has 91% accuracy in the balanced VirusSample dataset. While XGBoost, one of the most common gradient boosting-based models, achieves the highest score of 90% and 80% in both versions of the VirusShare dataset. This paper also presents the baseline results of VirusShare and VirusSample datasets by using the four most widely known machine learning techniques in dynamic malware classification literature. We believe that these two datasets and baseline results enable researchers in this field to test and validate their methods and approaches.

Index Terms—Malware, dynamic analysis, machine learning, balanced, imbalanced dataset.

I. INTRODUCTION

Malware is the short term used for malicious software, also known as a harmful code intentionally developed to harm a computing system. The main aim of malware is to gain access to the system, disrupt system services, deny service, steal confidential information, and destroy resources, to name a few. They rapidly increase in numbers, which causes an

imbalance between goodware and malware and between malware families. Because of the changing nature of the malware, updated datasets are required to practice the necessary study on malware classification. The rapid increase of malware attracts the attention of many researchers; currently, there are several standard datasets such as the Maling [1], and BIG2015 [2] to use, but they do not apply for dynamic malware analysis. To our best knowledge, there are only two datasets based on API calls, [3], [4]. This paper presents two trustworthy, recent, and updated datasets for dynamic malware analysis benchmarks.

One of the datasets used in the article is created from VirusSamples. It also contains malware samples with current samples from 2018, 2019, and 2020. In addition, the diversity has increased by taking the malware from the old years together with these current malware samples. With this, two newly created and updated datasets are built based on API calls.

There are three general approaches for classifying malware: static, dynamic, and hybrid malware classification. Static technique analysis the structure of the malware, whereas dynamic technique analysis the behavior of the malware during their execution, generally using a sandbox approach. On the other hand, the hybrid techniques combine static and dynamic analysis aspects to increase performance.

This study used the dynamic malware detection technique with API calls to evaluate the machine learning techniques because both API calls and the behavioral pattern of each file are the most-frequently-used attributes with machine learning techniques in malware detection and classification.

To extract API calls from the PE (Portable Executable) file, we use a Python framework called PEfile, a multi-platform module for reading and working with PE files and extracting different information. The class of the malware is obtained via VirusTotal.

Major contributions of this paper are:

- Present two alternative data sets that researchers can use to benchmark for studies on dynamic malware family

*Corresponding author: 20151709009@stu.khas.edu.tr

classification using API calls with increased current and diversified malware samples.

- To our best knowledge, the dataset obtained from Virus-Samples contains the most up-to-date malware samples based on API calls.
- Proposing a way to use cloud platforms to overcome the daily API key limit of VirusTotal without overloading or abusing the system to build a new dataset more efficiently.
- Presentation of baseline classification results with machine learning techniques, such as Histogram-Based Gradient Boosting, Random Forest, Support Vector Machine for the new malware datasets.,
- Comparative analysis of the Machine Learning (ML) techniques to identify the best performing technique for effective malware classification and use it as a candidate technique for developing dynamic malware classification systems.

The rest of the paper is as follows: The related work and background of other research are discussed in Section II. Section III presents features of the new datasets and the main infrastructure of virus analysis. Google Cloud Platform for Multi-User Malware Data Collection is explained in Section IV. In Section V, machine learning techniques: Histogram-Based Gradient Boosting, Random Forest, Support Vector Machine, and XGBoost are described, and their results with the used dataset are given. Finally, Section VI concludes this paper.

II. RELATED WORK

In the malware detection operation, understanding malware behavior is one of the vital parts. The usage of API calls obtained with dynamic analysis is one of the most used approaches in malware analysis since they represent malware behavior [5].

API Call sequences obtained by tracing the sequences of calls are used for generating specific behavioral patterns for each file. The sequences of API Calls collected through the dynamic approaches are used with data mining, machine learning, or deep learning techniques to detect or classify malware [6].

The techniques mentioned above need to be applied on datasets created by API Call sequences.

In [3], sequences of Windows Operating System API calls are obtained within the cuckoo sandbox isolated environment for each malware file. Malware family labels are determined using unique hash codes of each malware on the VirusTotal website. In total, there are 7,107 samples, which contain hash codes of malware, Windows operating system API Call sequences, and their malware family classes.

In [4], 42,797 malware API call sequences and 1,079 goodware API call sequences are obtained via cuckoo sandbox for dynamic malware analysis. Instead of using whole API call sequences, the first 100 non-consecutive API call sequences are extracted from the parent processes to reduce complexity and detect the malicious pattern as quickly as possible. The generated dataset containing hashcodes, label (malware or

goodware), and 100 API Calls for each sample was used for binary malware classification.

III. DATASETS

The presented datasets¹ include raw malware regarding VirusTotal and VirusShare/VirusSample based classification by using MD5 hashes. VirusShare dataset is a malware repository of live malicious codes; all the samples are in zip format, with password protection. VirusSamples generated malware samples robustly with various collection methods by processing more than 150,000 malware daily. For both sites to access the malware samples, a request must be sent from within sites; after the approval, the malware hashes are provided.

VirusTotal is a service that analyzes suspicious files and facilitates real-time detection of malware content. VirusTotal inspects items with anti-virus scanners and URL/domain blacklisting services, in addition to a myriad of tools to extract signals from the studied content [9]. VirusTotal's API allows users to upload and scan files, submit and scan URLs, access finished scan reports, and make automatic comments on URLs and samples. To present a free method, the public API key is a free service, available for any website or application but limited to 500 requests per day with a rate of 4 requests per minute is used for dataset construction.

From both datasets, the malware families with less the 80 samples are excluded. To obtain a balanced version of these datasets, malware families are capped with 300 samples. Two versions (balanced and imbalanced) of two datasets (VirusShare and VirusSample) are created with those operations. VirusShare databases include the malware families: Adware, Agent, Backdoor, Downloader, Ransomware, Riskware, Trojan, Virus, Worms. VirusSample databases include the malware families: Adware, Agent, Backdoor, Trojan, Virus, Worms.

A. Dataset Construction

The dataset is provided as comma-separated values (CSV) file. Each row in the file contains name of the malware, API call and class i.e. malware type. For example; 7ff49f2f0912352416b05c010f35f402cc79feed,"IntersectRect, GetCurrentProcess, GetVersion",Virus.

- 1) In each ZIP file obtained from VirusSamples and VirusShare, malware samples are represented with their MD5 hashcodes. MD5 hashcodes allow searching and analyzing these files without uploading. For every sample in the ZIP files, the hashcodes are written into a text file with groups of 500. Moreover, every group is numbered from 0 to n.
- 2) Hashcodes in text files are sent to VirusTotal system for determining their malware family. Then the analysis results are written to a file. If any of the hashcode files does not exist in the VirusTotal system, the response

¹https://github.com/khas-ccip/api_sequences_malware_datasets

can be empty.

- 3) VirusTotal returns approximately 70 malware detection application results. These 70 application results are checked; the most frequently repeating name gives the malware family. It can be spyware, trojan, virus, ransomware, etc. The hashcodes are grouped by their malware family.
- 4) After grouping the malware hashes, the next step is to extract the malwares from the ZIP file. According to their malware family output from the VirusTotal query, it generates malware family folders, and each of them contains live malware with password protection “infected”. To extract the API calls from the grouped malwares, the Python module PEfile is used.
- 5) Datasets with three sections; the hashcodes of malware files, API calls from PEFile library in Python, and the malware type from the VirusTotal API, are built-in CSV format.

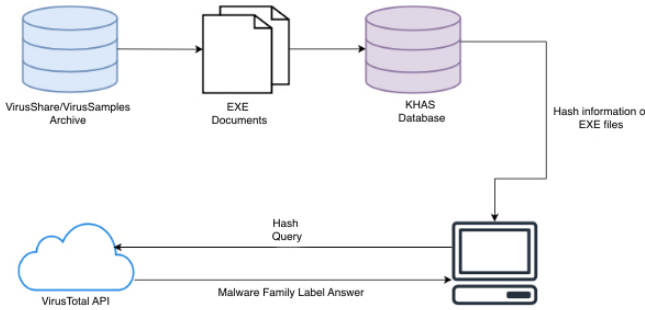


Fig. 1. Main infrastructure.

IV. GOOGLE CLOUD PLATFORM FOR MULTI-USER MALWARE DATA COLLECTION

The system has a daily 3 API key limit which allows 1,500 hash analysis [11], therefore we developed a new approach. To increase the number of analyses, the systems that ordinarily run on virtual machines moved into the Google Cloud Platform. The three main benefits of using the Google Cloud Platform are, the system on the cloud is an exact copy of the machine running on the virtual with better computing power and capacity, there is no IP conflict in the platform, which provides an increasing number of analyses by opening multiple virtual environments and, the cloud allows downloading and analyzing large files that belong to different years [12]. Typical cloud systems do not allow downloading large files unless a firewall rule is created; therefore, the firewall rule must be established before opening the virtual machine on the cloud using Debian/Linux as the operating system. Later on, the zipped system can be uploaded into the cloud platform to run the virtual machine. The image of the machine has to be

taken after the installation part to reduce the repeating steps by using the multiple machine feature of the cloud. Thus, the analysis continues without any interruption while opening 5 or 6 machines into the cloud system by using 3 API keys for each one.

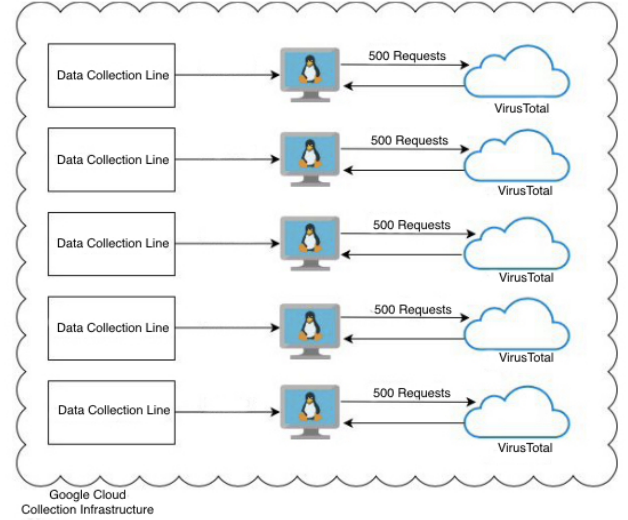


Fig. 2. Google Cloud Data Collection Infrastructure.

Figure 2 shows the created approach called “Google Cloud Data Collection Infrastructure” in this study. With this method, an IP address is assigned to each machine, thus preventing conflict on IPs. With this approach, if a single API key is used on each machine, 500 hashes can be sent to VirusTotal to have the hashcodes analyzed for each of them.

V. MODELS FOR VIRUSSHARE AND VIRUSSAMPLE DATASETS

A. Histogram-based Gradient Boosting (HGB)

Gradient boosting classification trees are becoming popular machine learning algorithms due to their ability in a wide range of implementation domains and results in easier management of model complexity by using tree depth and the number of trees [13].

Unlike the Gradient Boosting Classifiers, Histogram-based Gradient Boosting Classifier [14] estimators are considerably faster when applied for full-grown and big datasets. One of the key features defined in this estimator is that it provides native support for missing values found in the dataset. During training, the tree grower learns at each split point whether samples with missing values should go to the left or right child, based on the potential gain. When predicting, samples with missing values are assigned to the left or right child consequently. If no missing values are detected in the dataset, samples with missing values are matched to most child samples.

Tables I and IV show the classification study results of Histogram-based Gradient Boosting, Random Forest, Support Vector Machine, and XGBoost models on the balanced VirusShare and VirusSample datasets. Tables II and V presents

TABLE I
CLASSIFICATION REPORT OF HISTOGRAM-BASED BOOSTING, RANDOM
FOREST, SUPPORT VECTOR MACHINE, XGBOOST FOR
VIRUSSHARE-BALANCED DATASET

	HGB	RF	SVM	XGB
Malware Types	F1-score	F1-Score	F1-Score	F1-Score
Adware	0.79	0.74	0.80	0.78
Agent	0.52	0.50	0.50	0.52
Backdoor	0.94	0.90	0.92	0.92
Downloader	0.73	0.74	0.76	0.76
Ransomware	0.70	0.59	0.70	0.74
Riskware	0.47	0.55	0.44	0.47
Trojan	0.72	0.72	0.79	0.79
Virus	0.92	0.75	0.92	0.90
Worms	0.83	0.68	0.83	0.86

TABLE II
CLASSIFICATION REPORT OF HISTOGRAM-BASED BOOSTING, RANDOM
FOREST, SUPPORT VECTOR MACHINE, XGBOOST FOR
VIRUSSHARE-IMBALANCED DATASET

	HGB	RF	SVM	XGB
Malware Types	F1-score	F1-Score	F1-Score	F1-Score
Adware	0.55	0.45	0.57	0.59
Agent	0.29	0.30	0.30	0.30
Backdoor	0.84	0.84	0.86	0.84
Downloader	0.43	0.43	0.68	0.71
Ransomware	0.86	0.80	0.86	0.86
Riskware	0.40	0.40	0.55	0.67
Trojan	0.92	0.84	0.93	0.93
Virus	0.95	0.72	0.96	0.95
Worms	0.62	0.40	0.63	0.67

TABLE III
ACCURACY OF HISTOGRAM-BASED BOOSTING, RANDOM FOREST,
SUPPORT VECTOR MACHINE, XGBOOST FOR VIRUSSHARE DATASET

Balanced Data	HGB	RF	SVM	XGB
Accuracy	0.78	0.73	0.79	0.80
Weighted Avg.	0.78	0.72	0.79	0.79

Imbalanced Data	HGB	RF	SVM	XGB
Accuracy	0.88	0.77	0.89	0.90
Weighted Avg.	0.87	0.76	0.88	0.89

the classification study results for the same models on the imbalanced version of the VirusShare and VirusSample datasets. Table III shows the accuracy and weighted average results of Histogram-based Gradient Boosting, Random Forest, Support Vector Machine, and XGBoost models of VirusShare dataset; Table VI presents the accuracy and weighted average results of the same models of VirusSample dataset.

B. Random Forest (RF)

Random Forest is a popular machine learning algorithm that evolved into a common non-parametric method, which can be applied to classification and regression problems [15], [16]. RF algorithm creates multiple decision trees and then combines them to get more accurate predictions. RF adds additional randomness to the model when growing decision trees and finds the best feature among a random subset of features. As a result, there is a broader diversity which leads to better results in the model [17]. Hence, the RF algorithm is used

for the balanced and imbalanced versions of VirusShare and VirusSample datasets while presenting the results.

C. Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm that can be mostly used in classification problems on complex datasets. This article uses SVM as a non-parametric classifier that is not constrained by a set number of parameters and excludes any prior knowledge or assumptions about the data distribution. It maps data from the inputs and outputs in the training set. Its working logic uses a line to separate the points on the plane and ensures that the points of the two classes are at the maximum distance. [18]. For this separation to occur, it creates decision boundaries, or in other words, hyper-planes. The advantages of using this model are that it works more effectively in high-dimensional spaces, uses a subset of the training points, and results in the more productive use of memory in the decision function, and different kernel functions can be specified in cases where cases decision functions are versatile. In addition, the use of SVM allows us to have high accuracy rates and to work with a large number of independent variables [19]. Thus, the SVM algorithm is used for the balanced and imbalanced versions of VirusShare and VirusSample datasets while presenting the results.

D. Extreme Gradient Boosting (XGBoost)

XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. XGBoost is an ensemble learning method. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. XGBoost minimizes a regularized (L1 and L2) objective function that combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity (in other words, the regression tree functions). The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees combined with previous trees to make the final prediction [20], [21]. Another gradient boosting method XGBoost algorithm, is used for the balanced and imbalanced versions of VirusShare and VirusSample datasets while presenting the results.

VI. CONCLUSION

To sum up, this paper presents new datasets called VirusShare and VirusSample for dynamic malware classification using API calls. Furthermore, we have also analyzed multi-class malware classification performance of the balanced and imbalanced version of these two datasets using Histogram-based gradient boosting, Random Forest, Support Vector Machine, and XGBoost models, the most common classifiers in the literature of dynamic malware classification tasks. Our results show that Support Vector Machine and XGBoost are more successful than Random Forest and Histogram-based boosting models. For example, XGBoost classifier achieves 80% accuracy for the balanced VirusShare and 90% accuracy

TABLE IV
CLASSIFICATION REPORT OF HISTOGRAM-BASED BOOSTING, RANDOM FOREST, SUPPORT VECTOR, XGBOOST MACHINE FOR VIRUSAMPLE-BALANCED DATASET

	HGB	RF	SVM	XGB
Malware Types	F1-score	F1-Score	F1-Score	F1-Score
Adware	0.62	0.67	0.70	0.70
Agent	0.18	0.18	0.18	0.18
Backdoor	0.89	0.89	0.90	0.89
Trojan	0.95	0.88	0.96	0.96
Virus	0.99	0.80	0.98	0.99
Worms	0.77	0.43	0.77	0.76

TABLE V
CLASSIFICATION REPORT OF HISTOGRAM-BASED BOOSTING, RANDOM FOREST, SUPPORT VECTOR, XGBOOST MACHINE FOR VIRUSAMPLE-IMBALANCED DATASET

	HGB	RF	SVM	XGB
Malware Types	F1-score	F1-Score	F1-Score	F1-Score
Adware	0.89	0.92	0.96	0.91
Agent	0.91	0.91	0.91	0.91
Backdoor	0.95	0.95	0.94	0.95
Trojan	0.75	0.72	0.76	0.76
Virus	0.95	0.79	0.97	0.98
Worms	0.92	0.68	0.92	0.90

TABLE VI
ACCURACY OF HISTOGRAM-BASED BOOSTING, RANDOM FOREST, SUPPORT VECTOR MACHINE, XGBOOST FOR VIRUSAMPLE DATASET

Balanced Data	HGB	RF	SVM	XGB
Accuracy	0.90	0.82	0.91	0.90
Weighted Avg.	0.89	0.81	0.91	0.90

Imbalanced Data	HGB	RF	SVM	XGB
Accuracy	0.93	0.84	0.94	0.94
Weighted Avg.	0.93	0.83	0.94	0.94

for the imbalanced version of the same dataset. On the other hand, the Support Vector Machine classifier achieves 91% and 94% accuracy for the imbalanced and balanced version of the VirusSample dataset, which contains the latest malware samples. This paper presents these results as baseline scores for two new datasets in dynamic malware classification.

We also describe how to collect updated malware samples using cloud infrastructure efficiently. We have extracted hash values of malware files, API calls using the PEFile library in Python, and the malware type from the VirusTotal API to create a dataset. Our cloud infrastructure allowed us to collect 14,616 for VirusShare and 9,795 samples for VirusSample datasets. We sincerely hope that these two datasets will enable researchers to test their methods and compare the results available in this paper and the literature.

ACKNOWLEDGMENT

This work is supported by The Scientific and Technological Research Council of Turkey under the grant number 118E400.

REFERENCES

- [1] Nataraj, L., Karthikeyan, S., Jacob, G., & Manjunath, B. S. (2011, July). Malware images: visualization and automatic classification. In Proceedings of the 8th international symposium on visualization for cyber security (pp. 1-7).
- [2] Ronen, R., Radu, M., Feuerstein, C., Yom-Tov, E., & Ahmadi, M. (2018). Microsoft malware classification challenge. arXiv preprint arXiv:1802.10135.
- [3] F. O. Catak and A. F. Yazı, "A Benchmark API Call Dataset for Windows PE Malware Classification," May 2019, [Online]. Available: <http://arxiv.org/abs/1905.01999>
- [4] A. Oliveira and R. J. Sassi, "Behavioral Malware Detection Using Deep Graph Convolutional Neural Networks A Preprint," TechRxiv, Nov. 2019. doi: 10.36227/TECHRXIV.10043099.V1.
- [5] A. Fujino, J. Murakami, and T. Mori, "Discovering Similar Malware Samples Using API Call Topics," 2015.
- [6] Y. Ki, E. Kim, and H. K. Kim, "A novel approach to detect malware based on API call sequence analysis," International Journal of Distributed Sensor Networks, vol. 2015, Jun. 2015, doi: 10.1155/2015/659101.
- [7] VirusSamples. Available online: <https://www.virusamples.com/>.
- [8] VirusShare. Available online: <https://virusshare.com/>.
- [9] VirusTotal. Available online: <https://support.virustotal.com/hc/en-us/articles/115002126889-How-it-works>.
- [10] Carrera, E. (2016, 03 26). pefile. Retrieved 06 12, 2021, from Pypi: <https://pypi.org/project/pefile/>
- [11] VirusTotal. 2021. Public vs Premium API. [online] Available at: <https://developers.virustotal.com/reference>.
- [12] A. K. Marnerides, M. R. Watson, N. Shirazi, A. Mauthe and D. Hutchison, "Malware analysis in cloud computing: Network and system characteristics," 2013 IEEE Globecom Workshops (GC Wkshps), 2013, pp. 482-487, doi: 10.1109/GLOCOMW.2013.6825034.
- [13] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- [14] Guryanov, A. (2019, July). Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In International Conference on Analysis of Images, Social Networks and Texts (pp. 39-50). Springer, Cham.
- [15] Steinberg, D. (2009). CART: classification and regression trees. In The top ten algorithms in data mining (pp. 193-216). Chapman and Hall/CRC.
- [16] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- [17] Probst, P., Wright, M.N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9.
- [18] Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. In Data mining techniques for the life sciences (pp. 223-239). Humana Press. ISO 690
- [19] Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567. ISO 690
- [20] Çayır, U., Yenidoğan, I., & Dağ, H. (2019). Use Case Study: Data Science Application for Microsoft Malware Prediction Competition on Kaggle. Proceedings Book, 98.
- [21] Zhang, Y., Huang, Q., Ma, X., Yang, Z., & Jiang, J. (2016, August). Using multi-features and ensemble learning method for imbalanced malware classification. In 2016 IEEE Trustcom/BigDataSE/ISPA (pp. 965-973). IEEE.