Klassifikasiya metodlarının zərərli proqramların aşkarlanmasına tətbiqi

Elşən Bağırov
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
elsenbagirov1995@gmail.com

***Xülasə*— Son zamanlarda zərərli proqramların çeşidinin və mürəkkəbliyinin artımı, fərdi məlumatların təhlükə ilə üz-üzə qalması onların aşkarlanması və analiz edilməsi istiqamətində tədqiqatların aparılmasını labüd edir. Bu tədqiqat işində zərərli proqramların analizi və aşkarlanması metodları tədqiq olunmuş, 138047 sayda zərərli və zərərsiz fayllardan əldə edilən əlamətlərdən düzəlmiş baza üzərində python mühitindən istifadə edərək klassifikasiya metodları tətbiq edilmiş, qiymətləndirmə metrikaları vasitəsilə metodların səmərəliliyi analiz edilmişdir.**

Açar sözlər— zərərli proqram; virus; avtivirus; siqnatura; statik analiz; dinamik analiz; klassifikasiya; maşın təlimi;

I. GİRİŞ

İnformasiya texnologiyalarının, internetin əhatə dairəsinin durmadan genişlənməsi, eləcə də zərərli proqram istehsalçılarının artımı kompüterdə arzuolunmaz fəaliyyətlərə səbəb ola bilən zərərli proqramların çoxalmasına və belə proqramların informasiya təhlükəsizliyinə olan təhdidlərdən birinə çevrilməyə öz təsirini göstərmişdir [1]. Nəticədə zərərli proqramlar müasir informasiya cəmiyyətinin əsas problemlərindən birinə çevrilmişdir [2].

Əksər zərərli proqram təminatları rəsmi olmayan veb saytlardan axtarılan faylların, xüsusilə icra oluna bilən formatlı faylların internet üzərindən endirilməsi zamanı sistemə daxil olur. Zərərli fayl sistemdə öz yolunu tapan kimi xarakterik davranışına görə sistemin iş qabiliyyətinin aşağı salınması, fərdi məlumatların oğurlanması, kritik verilənlərin ələ keçirilməsi və ya dəyişdirilməsi, iri və xırda məbləğdə pul mənimsəmə, xidmətdən imtina kimi arzuolunmaz fəaliyyətləri həyata keçirməyə müvəffəq ola bilir [3, 4].

Zərərli proqramlardan qorunmaq üçün əsas müdafiə xətti antiviruslar hesab edilməkdədir [5]. Bir neçə asanlıqla aşkar oluna bilən məlum zərərli proqramlar vardır ki, antivirus tərəfindən rahat bir şəkildə aşkarlanıb təmizlənə bilir. Çünki antivirusların tərkibində zərərli proqramları unikal identifikasiya edən siqnaturalardan ibarət olan, mütəmadi olaraq yenilənməsinə ehtiyac yaranan baza mövcuddur [6]. Lakin yeni meydana çıxan, o cümlədən, metamorfizm, paketləmə və ya polimorfizm kimi antiviruslardan yayınma üsulları tətbiq edilmiş zərərli proqramların aşkarlanmasında antiviruslardan istifadə səmərəsizdir. Belə növ təkmilləşdirilmiş zərərli proqramların aşkarlanması nisbətən mürəkkəb və çətindir [7, 4].

Zərərli proqramların aşkarlanması və analizi üçün müxtəlif tədqiqat işləri aparılmışdır. Zərərli proqramların kəmiyyət və keyfiyyət baxımından artımı nəticəsində analiz və aşkarlama üçün həmin proqramlardan statik və ya dinamik üsulla əldə edilmiş əlamətlərdən ibarət verilənlər bazası üzərində maşın təlimi metodlarının tətbiqi məsələlərinə son illər daha çox diqqət ayrılmışdır.

II. ƏLAQƏDAR TƏDQİQATLAR

[2]-də 3086 sayda icra oluna bilən formatda olan 2136 zərərli və 980 zərərsiz fayllardan PE (*ing.* portable executable) başlığı, DLL (*ing.* Dynamic Link Library) adları və DLL funksiya adları kimi əlamətlər çıxarılaraq baza yaradılmış və Naive Bayes, Qərar ağacı, Təsadüfi Meşə, SVM və digər klassifikasiya metodları tətbiq edilmişdir.

[8]-də zərərli proqram təminatlarının aşkarlanması üsullarından olan hibrid analiz yanaşması istifadə edilmişdir. Statik analiz üçün icra oluna bilən fayllardan klassifikasiya məqsədilə əlamətlər çıxarılmışdır. Dinamik analiz yanaşmasında isə NtTrace adlanan idarəedici virtual mühitdə icra oluna bilən faylın davranışı müşahidə edilmişdir.

[9]-da zərərli proqramlar tərəfindən yerinə yetirilən zərərli fəaliyyətləri təsvir edən davranış əsaslı əlamətlər modeli təklif edilmişdir. İlkin olaraq idarə olunan virtual mühitdə zərərli proqramların davranışı analiz edilmiş və API (*ing.* Application Programming Interface) müraciətlərinin izləri tutularaq qeydiyyata alınmışdır. Bu izlər əsasında proqram davranışının zərərli və ya zərərsiz olduğunu müəyyən etmək üçün Qərar Ağacı, Təsadüfi Meşə və SVM kimi klassifikasiya metodları tətbiq edilmişdir.

[10]-da Chatchai L. və b. zərərli proqram fayllarının məzmunundan çıxarılmış n-gram əsaslı ardıcıl əlamətlərdən istifadə edərək yaradılmış baza əsasında üç müxtəlif klassifikasiya metodunu (qərar ağacları, çoxlaylı persepteron və SVM) tətbiq etmişlər. Eksperimentlərin nəticəsində ən yüksək dəqiqlik SVM metodunda (96.64 % dəqiqliklə və n=4 olduqda) alınmışdır.

[11]-də Şerif B. və b. təkmilləşmiş zərərli proqramların beş seçilmiş əlaməti vasitəsilə zərərli proqram nümunələrindən ibarət bazadan “stuxnet” virusunun proqnoz edilməsi üçün reqressiya modelləri işlənən çoxölçülü yanaşma təklif etmişlər. Analiz nəticəsində məlum olmuşdur ki, təkmiləşdirilmiş zərərli

DOI: 10.25045/NCInfoSec.2019.49

proqram növlərinin bəzi əlamətləri arasında yüksək korrelyasiya mövcuddur.

[5]-də Hemant R. və b. zərərli proqramların klassifikasiyası üçün supervizorlu və supervizorsuz öyrənmə metodlarını tətbiq etmişlər, əlamətlər vektoru olaraq “opkod tezliyi” götürülmüşdür. Bu zaman təsadüfi meşə alqoritminin digər alqoritmlərdən üstün dəqiqlik göstərdiyi müşahidə edilmişdir.

III. ZƏRƏRLİ PROQRAMLAR VƏ ONLARIN NÖVLƏRİ

Zərərli proqram termini ingilis dilindən “malicious” (zərərli) və “software” (proqram təminatı) terminlərinin deduksiya olunması nəticəsində yaranmışdır [1, 6, 12]. “KasperskyLab” şirkəti tərəfindən verilmiş tərifdə zərərli proqramların qanuni istifadəçinin kompüterini yoluxdurmaq və ona müxtəlif yollarla zərər vurmaq üçün hazırlanmış proqramlar olduğu deyilir. Son zamanlarda ən ciddi təhdidlərdən biri məhz zərərli proqramlar hesab edilir.

Çox vaxt istənilən növ zərərli proqramı “virus” kimi qəbul edirlər [6]. Halbuki viruslar zərərli proqramların ən sadə növü hesab edilir və davranışına görə zərərli proqramlar öz müxtəlifliyi ilə zəngindir. Çox rast gəlinən zərərli proqramlara viruslar, soxulcanlar, troyanlar, casus proqramlar, reklam proqramları, girov proqramları, rutkitlər, bekdorlar, keyloqqerlər misal göstərilə bilər [4, 12-14]. Antivirus sistemlərinin test edilməsi üzrə ixtisaslaşan “AV-Test” İnstitutunun dediyinə görə gün ərzində 350 minə yaxın yeni növ zərərli proqramlar qeydə alınır və xarakterinə görə təsnifatlaşdırılır [15].

A. Viruslar. Həcmi kilobaytlarla ölçüləcək səviyyədə olan viruslar ən sadə zərərli proqram növü hesab edilir. Viruslar istifadəçinin razılığı olmadan yüklənir və işə düşür, bioloji virus kimi digər faylları yoluxduraraq özünün nüsxəsini yaradır və çoxalır [4, 10, 12]. Virusların kompüterdəki fəallığını əks etdirən əsas təzahürlər aşağıdakılardır:

- parolların və digər kritik informasiyanın oğurlanması;
- yaddaş həcmini israf edərək kompüterin iş qabiliyyətini aşağı salmaq;
- verilənlərin silinməsi və ya modifikasiyası;
- spamların yayılması.

B. Soxulcanlar. Viruslarla çox oxşardır. Fərq odur ki, soxulcanlar şəbəkə üzərindən yayıla bilir və öz nüsxəsini digər kompüterlərə sərbəst yaya bilirlər [1, 4, 13]. Soxulcanların axın kimi yayılması nəticəsində yaddaş, rabitə kanalları yüklənir və sistem bloka alınır. Viruslardan fərqli olaraq şəbəkə mühitində yayılma imkanına malikdir. İlk kompüter soxulcanı 1988-ci ildə Kornel Universitetinin aspirantı Robert Tappan Morris tərəfindən yaradılmışdı. Soxulcan böyük sürətlə şəbəkədə yayılaraq internet qovşaqlarının 10%-ni sıradan çıxara bilmişdi [1]. Morris soxulcanının mənbə kodu Kompüter Tarixi Muzeyində çevik disketdə qorunub saxlanılır. Digər məşhur soxulcanlara misal olaraq Sasser, My Doom, Blaster, Melissanı göstərmək olar [4].

C. Troyanlar. Funskional cəhətdən qanuni proqram təminatı kimi görünən, lakin işə düşdükdə faydalı funksiyalarla bərabər, arxa planda öz zərərli funksiyalarını da yerinə yetirirlər [1, 6]. Troyanlar pirat proqramlar, yoluxmuş fayllar, sosial mühəndislik və digər üsullarla yayılırlar. Kaspersky Lab tərəfindən troyanların kompüterdə yerinə yetirdiyi fəaliyyətlərə və təbiətinə görə bəzi növləri aşağıda verilmişdir.

- Troyan-casus: casus proqramları toplusu olub, əsasən kritik verilənləri və parolları oğurlamağı hədəf alır.
- Troyan-yükləyici: avtomatik olaraq uzaqdakı serverə qoşularaq zərərli proqramları kompüterə yükləyir.
- Troyan-dropper: zərərli proqramları endirir, quraşdırır və işə salır.
- Troyan-proksi: işə salınması ilə bədniyyətlinin yoluxmuş sistemi proksi serverə çevrilir [6].

D. Casus proqramları. Səciyyəvi fəaliyyətlərinə istifadəçinin internet axtarış tarixçəsinin izlənməsi və istifadəçi profilinə uyğun olan reklamlarla yenidən qayıtması aiddir [13]. Bir şəxsin və ya təşkilatın xəbəri olmadan tez-tez istifadə olunan veb-saytlar, kredit kart məlumatları, klaviaturadan daxil edilmiş klavişlərin loqları, elektron poçt ünvanları və digər məlumatları istifadəçidən xəbərsiz hücumçuya göndərə bilir. Eyni zamanda kompüter üzərində idarəetməni ələ ala bilir. Casus proqramları hər hansı proqram məhsulunun sınaq versiyası və ya pulsuz versiyasının endirilməsi və quraşdırılması zamanı aktivləşə bilər [4, 6]. Reklam proqramları və keyloqqerlər casus proqramlarına daxil edilirlər.

E. Girov proqramları. Troyanlardan istifadə etməklə kompüter sistemini və ya verilənləri şifrləyir və istifadəçinin öz verilənlərinə daxil olmasını məhdudlaşdırır, müəyyən müddət bitənədək sistemi “dondurur” [13]. Qurbanın şifrəni əldə etməsi üçün interfeys yaradılır və ona rəqəmsal kriptovalyuta ilə (əsasən bitkoin) ödəniş üsulu təklif edilir. Girov proqramına misal olaraq 2017-ci ildə böyük sürətlə yayılan “WannaCry” kiberhücum hadisəsini göstərmək olar.

IV. ZƏRƏRLİ PROQRAMLARIN ANALİZİ VƏ AŞKARLANMASI ÜSULLARI

Zərərli proqramların aşkarlanması prosesi kompüter təhlükəsizliyi kontekstində həmişə maraq doğurmuşdur. Son zamanlarda zərərli proqramların geniş surətdə artımı və təkmilləşməsi aşkarlama metodlarının müxtəlifliyinə səbəb olmuşdur [15].

Zərərli proqramların analiz metodları əsasən statik və dinamik olmaqla iki kateqoriyaya bölünür [13, 16]. Bundan başqa hər iki metodun birlikdə istifadə edilməsi nəticəsində yaranan hibrid metod da tətbiq edilir.

A. Statik analiz metodu. Statik analizdə icra oluna bilən (*ing*. executable) fayl işə salınmadan struktur səviyyədə, yəni kod səviyyəsində analiz edilir [3, 16]. Bunun üçün müəyyən vasitələrdən istifadə etməklə (Məsələn, python mühitində yazılmış “PEFİLE” alətindən) icra edilə bilən fayllardan statik parametrlərin çıxarılaraq əlamətlərdən ibarət bazanın

yaradılması və üzərində analizlərin aparılması prosesi yerinə yetirilir. Statik analizin əsas üstünlüyü zərərli proqramın bütün mümkün davranış ssenarilərinin üzə çıxarıla bilməsidir.

Statik şəkildə zərərli proqram analizi üçün mühüm bir metod olaraq antivirusların istifadə etdiyi siqnatura əsasında aşkarlama metodunu göstərmək olar. Antiviruslar yeni daxil olmuş icra oluna bilən faylın zərərli olub olmadığını yoxlamaq üçün həmin faylın statik üsulla əldə edilmiş siqnaturası ilə məlum zərərli faylların siqnaturaları bazasından uyğunluq tapmağa cəhd edir. Əgər uyğunluq yaranarsa antivirus onu zərərli proqram kimi qeydə alır [13].

$$A(P) = \begin{cases} zərərli, s \in P \\ zərərsiz, s \in !P \end{cases}$$

Burada, P – zərərli və ya zərərsiz olduğu müəyyənləşdirilən proqram, A(P) – P proqramının aşkarlayıcı funksiyası, s – P proqramının siqnaturasıdır.

Statik analiz metodunun əsas üstünlüyü siqnaturaların unikallığı baxımından dəqiqliyin maksimum təmin edilməsidir. Üsulun çatışmazlığı isə əvvəllər məlum olmayan zərərli proqramları, eyni zamanda mövcud zərərli proqramlar üzərində “paketləmə”, “polimorfizm”, “metamorfizm” kimi antiviruslardan yayınma üsulları tətbiq edilmiş zərərli proqramları aşkarlamaqda çətinlik çəkməsidir [13]. Əksər kommersiya antivirus vendorları məhz siqnaturaya əsaslanan aşkarlama metodunu tətbiq edir. Mütəmadi olaraq yeni növ zərərli proqramlar aşkar edildikdən sonra siqnatura bazası yenilənməyə məcburdur.

B. Dinamik analiz metodu. Statik analizdən fərqli olaraq dinamik analiz metodunda izolə edilmiş mühitdən (simulyator, emulyator, sandbox, virtual maşın) istifadə edərək şübhəli faylın davranışı monitorinq edilir [3, 13, 17]. Nəticədə faylın davranış atributları (Məsələn, icra olunan zaman zərərli proqramın API müraciətləri, dəyişdirilmiş, silinmiş və ya oxunmuş reqistrlər, İP ünvan və DNS sorğuları, URL-lərə giriş, şəbəkədə aktivliyi və s.) əldə edilir [3].

Zərərli proqram təminatlarının dinamik olaraq analizi üçün izolə edilmiş mühitlərdən ən əlverişlisi “sandbox”lardır [3]. Sandboxlara misal olaraq “Cuckoo”, “SNDBOX”, “Falcon”, “ViCheck”-i göstərmək olar. Sandboxlar zərərli proqramların davranışını loqlaşdıra bilən agentlərlə təchiz olunmuşdur. Zərərli proqramın sandobxda icra edilməsi zamanı agent loqlaşdırmanı başladır, icra müddəti bitdikdən sonra mühiti əvvəlki vəziyyətə qaytara bilir. Bu zaman əsas sistemdən təcrid olunduğu üçün ona heç bir zərər gəlmir.

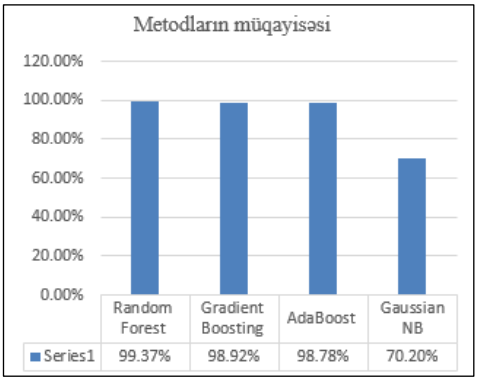
Üsulun əsas üstünlüyü əvvəllər məlum olmayan zərərli proqramları aşkarlaya bilməsidir. Həmçinin, statik analizlə müqayisədə sürətlidir. Əsas çatışmazlığı isə onun dəqiqliyindədir [13]. Belə ki, burada birinci və ikinci növ səhvlərin olması ehtimalı mövcuddur.

Statik və dinamik analizin birlikdə istifadə olunduğu hibrid metodlardan istifadə genişlənməkdədir. Belə ki, statik olaraq aşkarlanmadan yayınmaq üçün zərərli proqramlara yayınma üsullarının tətbiqi və dinamik olaraq aşkarlanmanın dəqiqliyi kimi problemlərin mövcud olduğunu nəzərə alaraq hibrid modeldə zərərli proqram dinamik mühitə salınaraq statik əlamətlər əldə edilir və analiz olunur [17].

C. Maşın təlimi metodları. Zərərli proqramlar genişlənməkdə davam edir və müxtəlif növ yeni əlamətlər meydana çıxır. Təkcə statik və ya dinamik analizin tətbiq edilməsi ilə zərərli proqramlarla mübarizə aparmaq səmərəsizdir. Buna görə də tədqiqatçılar zərərli proqramları aşkarlamaq üçün maşın təlimi metodlarının tətbiqinə diqqət ayırmaqdadırlar.

V. EKSPERİMENTLƏR

Bu işdə 96724 zərərli və 41323 zərərsiz fayldan statik üsulla əldə edilmiş 138047 sətrdən ibarət, “CSV” (comma seperated values) faylında saxlanılan verilənlər bazası üzərində “Təsadüfi meşə”, “Gradient Boosting”, “Adaboost”, “GaussianNB” metodları tətbiq edilmiş və qiymətləndirmə nəticəsinə şəkil 1-də əks olunmuş nəticələr əldə edilmişdir. Tədqiqatlar python mühitində (python3 versiyası quraşdırılmış Jupyter Notebook veb tətbiqindən istifadə etməklə) aparılmışdır.



Şəkil 1. Metodların müqayisəsi.

Tədqiq edilmiş baza öyrənmə və testetmə üçün 3/7 nisbətində iki hissəyə bölünmüşdür. Metodların qiymətləndirilməsi nəticəsində ən yüksək dəqiqlik verən metodun Təsadüfi meşə alqoritmi olduğu məlum olmuşdur.

NƏTİCƏ

Zərərli proqramların analiz edilməsi üçün iki ən mühüm yanaşma – statik və dinamik analiz metodları, onların fərqli, üstün və zəif cəhətləri təhlil edilmişdir. Zərərli proqramların təbiətinə görə müxtəlifliyini nəzərə alaraq demək olar ki, dinamik analiz mühiti onun davranışını tam əks etdirə bilmə bacarığına sahib olmaya bilər. Birincisi, zərərli proqramlar istifadə olunan mühitin müxtəlifliyinə uyğun olaraq fərqli davranış göstərə bilər. Bundan başqa son zamanlar zərərli proqramlar üzərində izolə edilmiş mühitdə öz davranışını dəyişdirmək xüsusiyyətinin tətbiq edilməsi halları da müşahidə edilməkdədir. Zərərli proqramı sandbox mühitində dinamik olaraq icra edən zaman ilk növbədə agenti axtarış etməsi ilə bu qənaətə gəlmək olar. Zərərli proqram təcrid olunmuş mühitin əlamətlərinə rast gəldikdə özünü zərərsiz fayl kimi apara bilir.

Baxılan işdə maşın təlimi metodları tətbiq edilmişdir. Qiymətləndirmə nəticəsində tədqiq edilmiş baza üçün Təsadüfi meşə alqoritmi yüksək dəqiqlik göstərmişdir.

Bir çox tədqiqatçılar tərəfindən zərərli proqramların aşkarlanması və analizi üçün müxtəlif metodların işlənməyinə baxmayaraq statistikaya əsasən zərərli proqramların sayında hələ də ümumi halda azalma müşahidə edilməməkdədir. Zərərli proqram istehsalçıları aşkarlama metodlarının təkmilləşdirilməsinə cavab olaraq yeni əlamətlər yaradırlar. Buna görə də aşkarlama dəqiqliyinin artırılması üçün istifadə olunan əlamətlərin yenidən baxılmasına ehtiyac yaranır.

İSTİNADLAR

[1] R. Əliquliyev, Y. İmamverdiyev, “İnformasiya təhlükəsizliyi insidentləri”, İnformasiya Texnologiyaları nəşriyyatı, 2012, 219 s.

[2] Y. İmamverdiyev, E. Kərimov, “Zərərli proqramların aşkarlanması üçün maşın təlimi metodlarının tətbiqi”, Proqram mühəndisliyinin aktual elmi-praktiki problemləri I respublika konfransı, 17 may 2017-ci il, s. 166-169.

[3] M. Ijaz, M. Durad, M. Ismail, “Static and dynamic analysis using machine learning”, 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 8-12 January 2019, pp. 687-691.

[4] P. Vinod, R. Jaipur, M. Gaur, “Survey on malware detection methods”, Proceedings of 3rd Hackers’ Workshop on Computer and Internet Security, 17-19 March 2009, pp. 74-79.

[5] H. Rathore, S. Agarwal, S. Sahay et al., “Malware detection using machine learning and deep learning”, International Conference on Big Data Analytics, 18-1 December 2018, pp. 402-411.

[6] H. Tayyab, H.Aqib, “A survey of analysis of malware infection”, 2015, 40 p.

[7] Y. Ye, D. Wang, T. Li et al., “An intelligent PE-malware detection system based on association mining”, Journal in Computer Virology, vol. 4, 2008, pp. 323–334.

[8] S. Kumar, N. Aggarwal, C. Krishna et al., “Malicious data classification using structural information and behavioral specifictions in executables”, Proceedings of Recent Advances in Engineering and Computational Sciences (RAECS), 6-8 March 2014, pp. 1-6.

[9] H. Galal, Y. Mahdy, M. Atiea, “Behavior-based features model for malware detection”, Journal of Computer Virology and Hacking Techniques, vol. 12, 2016, pp. 59-67.

[10] O. Sornil, L. Chatchai, “Malware Classification Using N-grams Sequential Pattern Features”, International Journal of Information Processing and Management(IJIPM), vol. 4, 2013, pp. 59-67.

[11] Ş. Bahtiyar, M. Yaman, C. Altınığne, “A multi-dimensional machine learning approach to predict advanced malware”, Computer Networks, vol. 160, 2019, pp. 118-129.

[12] S. Bragen, “Malware detection through opcode sequence analysis using machine learning”, MS Thesis, Gjovik University College, 2015, 67 p.

[13] K. Chumachenko, “Machine learning methods for malware detection and classification”, Bachelor’s Thesis, 2017, 93 p.

[14] M. Siddiqui, “Data mining methods for malware detection”, PhD thesis, University of Central Florida, 2008, 111 p.

[15] “Av-Test malware statistics”, https://www.av-test.org/en/statistics/malware/

[16] C. Andrade, C. Mello, “Malware automatic analysis”, BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, September 2013, pp. 681-686.

[17] J. Yan, Y. Qi, Q. Kao, “Detecting Malware with an Ensemble Method Based on Deep Neural Network”, Security and Communications Networks, vol. 1, 2018, pp. 1-16.

APPLICATION OF CLASSIFICATION METHODS FOR MALWARE DETECTION

Elshan Baghirov

Institute of Information Technology of ANAS, Baku, Azerbaijan

elsenbagirov1995@gmail.com

Abstract— In recent years, the increasing variety and complexity of malware, facing personal data with threat requires research towards detection and analyzing. This study investigated the methods of analyzing and detecting malware, applied classification methods on dataset based features extracted from 138047 malware and benign files using the python framework, and the effectiveness of the methods have been analyzed using the evaluation metrics.

Keywords— *malware; virus; antivirus; signature; static analysis; dynamic analysis; classification; machine learing*