SDN Deep Learning Based Anomaly Detection Projesi- SAGA

Veri Seti Analiz Raporu
Canadian Institute for Cybersecurity

AÇIK KAYNAK VERİ SETİ KULLANAN PROJELER

1. CCCS-CIC-AndMal-2020

- **Projenin Konusu:** Android kötü amaçlı yazılım tehdidi öngörüsü
- **Dataset:** CCCS-CIC-AndMal-2020 dataset
- **Makale:** Abir Rahali, Arash Habibi Lashkari, Gurdip Kaur, Laya Taheri, Francois Gagnon, and Frédéric Massicotte, “DIDroid: Android Malware Classification and Characterization Using Deep Image Learning”, 10th International Conference on Communication and Network Security, Tokyo, Japan, November 2020.
- **Makale Özet:** Makaleye erişim sağlanamamıştır.

2. CIRA-CIC-DoHBrw-2020

- **Proje Konusu:** HTTPS üzerinden DNS
- **Dataset:** CIRA-CIC-DoHBrw-2020
- **Makale:** "Mohammadreza MontazeriShatoori, Logan Davidson, Gurdip Kaur, and Arash Habibi Lashkari, "Detection of DoH Tunnels using Time-series Classification of Encrypted Traffic", The 5th IEEE Cyber Science and Technology Congress, Calgary, Canada, August 2020.
- **Makale Özet:** Makaleye erişim sağlanamamıştır.

3. CICMalDroid2020

- **Proje Konusu:** Android veri seti üzerinde kötü amaçlı yazılım sınıflandırması
- **Dataset:** CICMalDroid2020
- **Makale:** Samaneh Mahdavifar, Andi Fitriah Abdul Kadir, Rasool Fatemi, Dima Alhadidi, Ali A. Ghorbani, "Dynamic Android Malware Category Classification using Semi-Supervised Deep Learning", The 18th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC), Aug. 17-24, 2020.
- **Makale Özet:** Bu makalede, yarı denetimli derin öğrenme metotları kullanılarak android verileri için kötü amaçlı yazılım sınıflandırılması yapılmıştır. Kötü amaçlı yazılım sınıflandırması dört grup altında toplanmıştır. Bunlar; adware (reklam ile gelen kötü amaçlı yazılım), banking malware (banka hesaplarına ulaşmaya çalışan kötü amaçlı yazılım), SMS malware (SMS ile gelen kötü amaçlı yazılım) ve mobile riskware (risk taşıyan uygulamalar)dır. Bu gruplara girmeyen diğer bütün uygulamalar benign (iyi huylu) olarak tanımlanmıştır.
    - Yapılan deneylerdeki amaç ise bir uygulamayı benign (iyi huylu) veya olmadığını saptamaktır. Bu amaçla ilk olarak dört farklı makine öğrenmesi sınıflandırma yöntemi (RF, DT, SVM, KNN) ve üç farklı derin öğrenme metodu (LP, DNN, PLDNN) ile simülasyonlar yapılmıştır. En yüksek doğruluk oranını veren PLDNN (Pseudo Label Deep Neural Network) ile simülasyon için seçilmiştir. Daha sonra ise farklı gizli katman (hidden layer) ve boğum(node) sayıları ile deneyler yapılmış ve bu değişkenler için optimal sayılara ulaşılmıştır.

TABLE VI
CONFUSION MATRIX OF CATEGORY CLASSIFICATION

		Prediction Category				
		Adware	**Banking**	**SMS**	**Riskware**	**Benign**
Real Category	**Adware**	0.85	0.02	0.02	0.05	0.07
	Banking	0.0	0.96	0.03	0.01	0.0
	SMS	0.0	0.0	1.0	0.0	0.0
	Riskware	0.01	0.0	0.0	0.98	0.01
	Benign	0.01	0.0	0.0	0.01	0.98

    - Bu tablo PLDNN yöntemiyle her bir Android uygulama kategorisi için hesaplanan confusion matrisini göstermektedir. Tabloda 1000 etiketlenmiş veri seti için true positive rate (TPR) oranları gösterilmektedir. Tablodan görüldüğü üzere, SMS malware 100% bir TPR ile tahmin edilmiştir. Riskware ve Benign 98% ile, adware ise 85% oranı ile en düşük TPRye sahiptir. Bu durum, Adware uygulamaların Benign ve Riskware ile benzer yapıda olması ile açıklanabilir.
    - Sonuç olarak veri seti boyutu küçük olmasına rağmen, PLDNN yönteminin kötü amaçlı yazılım tahmini kategorileştirmesi için yüksek performans göstermiştir.

4. CIC-Darknet2020

- **Proje Konusu:** Darknet trafik tespiti ve sınıflandırması
- **Dataset:** CICDarknet2020, ISCXTor2016 and ISCXVPN2016, dataset
- **Makale:** Arash Habibi Lashkari, Gurdip Kaur and Abir Rahali, "DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning", 10th International Conference on Communication and Network Security, Tokyo, Japan, November 2020.
- **Makale Özet:** Bu araştırma çalışması, sırasıyla Tor ve VPN trafiğini kapsayan eksiksiz bir darknet veri kümesi oluşturmak için ISCXTor2016 ve ISCXVPN2016 olmak üzere iki genel veri kümesini birleştirerek VPN ve Tor uygulamalarını birlikte darknet trafiğinin gerçek temsilcisi olarak tespit etmek ve karakterize etmek için yeni bir teknik önermektedir. (Makalenin giriş kısmı hariç erişim sağlanamamıştır.)

5. Investigation of the Android Malware (CIC-InvesAndMal2019)

- 9 numaralı proje ile ilişkilidir.
- **Proje Konusu:** Android kötü amaçlı yazılım tehdidi saptaması
- **Dataset:** CICInvesAndMal2019 dataset
- **Makale:** Laya Taheri, Andi Fitriah Abdulkadir and Arash Habibi Lashkari, "Extensible Android Malware Detection and Family Classification Using Network-Flows and API-Calls", The IEEE (53rd) International Carnahan Conference on Security Technology, India, 2019.
- **Makale Özet:**

Dataset:	**Evaluation (Testing set)**		
Features:	**Network_Flows**		
Scenario:	**A (Malware Binary)**	**B (Malware Category)**	**C (Malware Family)**
Algorithm:	**RF**	**RF**	**RF**
Precision(%):	85.80	49.90	27.50
Recall(%):	88.30	48.50	25.50
	First part of the CICAndMal2017 [16] dataset		

Evaluation (Testing set)		
Permissions+Intents	**API_Calls+Network_Flows**	
A (Malware Binary)	**B (Malware Category)**	**C (Malware Family)**
RF	**RF**	**RF**
95.30	83.30	59.70
95.30	81.00	61.20
Second part of the CICAndMal2017 dataset		

  - Bu araştırmada, CICAndMal2017 veri kümesinin statik bölümdeki izinler ve amaçlar (Permissions+Intents) gibi yeni özellik kümelerini ve dinamik bölüme eklenen API çağrılarını içeren ikinci bölüm ayrı incelenmiştir. Kullanılan algoritma Random Forest (RF)tır.
  - İlk olarak 9 numaralı projede sonuçları verildiği gibi Network_Flows özellik seti ile çalışmalar yapılmıştı. Bu makalede ise amaç, önceki dinamik özellikleri (80 ağ akışı) API çağrılarının sıralı ilişkileriyle birleştirerek kötü amaçlı yazılım kategorisini ve malware family (aile sınıflandırma) performansını iyileştirilmektir.
  - Değerlendirme sonuçlarına göre, sonuçların kötü amaçlı yazılım ikili sınıflandırmasında %95,3 duyarlılık (% 7,5 iyileştirme), kötü amaçlı yazılım kategorisi sınıflandırmasında % 81,0 duyarlılık (% 32,5 iyileştirme) ve kötü amaçlı yazılım ailesi sınıflandırmasında % 61,2 duyarlılık (% 35,7 iyileştirme) elde edilmiştir.

6. DDoS Evaluation Dataset(CIC-DDoS2019)

- **Proje Konusu:** DDoS saldırıları için farklı eğitim ve test veri seti kullanılarak makine öğrenmesi yöntemleri de kullanılarak tahmin etme
- **Dataset:** CICDDoS2019 dataset
- **Makale:** Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy", IEEE 53rd International Carnahan Conference on Security Technology, Chennai, India, 2019
- **Makale Özet:** Bu projede, öncelikle mevcut veri kümeleri kapsamlı bir şekilde gözden geçirilip DDoS saldırıları için yeni bir sınıflandırma önerilmektedir. İkinci olarak, mevcut tüm eksiklikleri gideren CICDDoS2019 adında yeni bir veri seti oluşturulmaktadır. Üçüncü olarak, oluşturulan veri kümesini kullanarak, farklı ağ akışı özelliklerine dayalı yeni bir algılama ve aile sınıflandırması yaklaşımı önerilmektedir. Son olarak, farklı DDoS saldırı türlerini ağırlıklarıyla tespit etmek için en önemli özellik setleri sağlanmaktadır.
    - 12 DDoS saldırısını (NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag, WebDDoS, SYN ve TFTP) dahil eden 12 Ocak gününü eğitim veriseti için seçilen gün, 7 saldırının olduğu (PortScan, NetBIOS, LDAP, MSSQL, UDP, UDPLag ve SYN) 11 Mart gününü ise test veri seti için seçilen gün olarak atanmıştır.
    - Sonuçlar kısmında ise veri seti dört algoritma üzerinde simülasyon edilmiştir. Bunlar ID3 (eğitim veri kümesi üzerinde decision tree oluşturur), Random Forest (decision tree ve ensemble learning birlikte kullanan bir makine öğrenmesi yöntemidir), Naive Bayes (Bayes teoremini esas alan olasılıksal bir sınıflandırma aracıdır) ve Multinomial Logistic Regression (logistic regression yöntemini multiclass problemler için kullanan bir sınıflandırma metodudur).
    - Eğitim sonuçlarına bakılarak, tablodan ID3 algoritmasının %78 oranında doğruluk, %65 oranında duyarlılık ve iki metriğin harmonic bir kombinasyonu olan F1 ölçüsü için de %69 oranı ile kullanılan algoritmalar içerisinde en yüksek sonucu verdiği görülmektedir.
    - Tabloda Pr doğruluk (precision) anlamına gelmekte olup, Pr =TP/(TP+FP) şeklinde hesaplanmaktadır. Rc duyarlılık (recall or sensitivity) anlamına gelmekte olup, Rc=TP/(TP+FN) formülü ile hesaplanmaktadır. F1 ise F-Measure olarak Pr ve Rc metriklerinin harmonik kombinasyonudur ve F1 = 2/(1/Pr +1/Rc) olarak hesaplanabilir.

Table V: The Performance Examination Results

Algorithm	Pr	Rc	F1
ID3	0.78	0.65	0.69
RF	0.77	0.56	0.62
Naïve Bayes	0.41	0.11	0.05
Logistic regression	0.25	0.02	0.04

7. IPS/IDS dataset on AWS (CSE-CIC-IDS2018)

- 8 numaralı proje ile aynı makaleyi kullanmaktadır.
- **Proje Konusu:** profiles and attacks
- **Dataset:** CSE-CIC-IDS2018 dataset
- **Makale:** Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.
- **Makale Özet:** Bu projenin temel amacı, ağda görülen olayların ve davranışların soyut temsillerini içeren kullanıcı profillerinin oluşturulmasına dayalı olarak izinsiz giriş tespiti (intrusion detection) için çeşitli ve kapsamlı bir kıyaslama veri seti oluşturmak için sistematik bir yaklaşım geliştirmektir.

8. Intrusion Detection Evaluation Dataset (CIC-IDS2017)

- **Proje Konusu:** Ağ Anomallik Tespiti
- **Dataset:** CICIDS2017 dataset
- **Makale:** Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018
- **Makale Özet:** Saldırı Tespit Sistemleri (IDS'ler) ve Saldırı Önleme Sistemleri (IPS'ler), karmaşık ve sürekli büyüyen ağ saldırılarına karşı en önemli savunma araçlarıdır. Güvenilir test ve doğrulama veri kümelerinin eksikliği nedeniyle, anormallik tabanlı saldırı tespit yaklaşımları, tutarlı ve doğru performans gelişimlerinden eksiktir.
    - Bu makalede, farklı saldırı çeşitlerinin tespiti ve önlenmesi farklı algoritmalarla ile tespit edilmesi incelenmiştir. Saldırı çeşitleri brute force, heartbleed, botnet, DoS, DDoS, Web ve Infiltration olmak üzere yedi maddede gruplanmıştır.
    - Sonuçlarda ise yedi farklı algoritma kullanılmıştır. En yüksek doğruluk sonucunun RF ve ID3 algoritması ile, en yüksek duyarlılık sonucunun ID3 algoritması ile elde edildiği gözlemlenmektedir. Karmaşıklık olarak ise en düşük zaman alan algoritma Naive-Bayes'tir.

Table 4: The Performance Examination Results.

Algorithm	Pr	Rc	F1	Execution (Sec.)
KNN	0.96	0.96	0.96	1908.23
RF	0.98	0.97	0.97	74.39
ID3	0.98	0.98	0.98	235.02
Adaboost	0.77	0.84	0.77	1126.24
MLP	0.77	0.83	0.76	575.73
Naive-Bayes	0.88	0.04	0.04	14.77
QDA	0.97	0.88	0.92	18.79

9. Android Malware Dataset (CIC-AndMal2017)

- **Proje Konusu:** Android uygulamaları için kötü amaçlı yazılım tespiti
- **Dataset:** CICAndMal2017 dataset
- **Makale:** Arash Habibi Lashkari, Andi Fitriah, A. Kadir, Laya Taheri, and Ali A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification", In the proceedings of the 52nd IEEE International Carnahan Conference on Security Technology (ICCST), Montreal, Quebec, Canada, 2018.
- **Makale Özet:** Bu yaklaşımda, gelişmiş kötü amaçlı yazılım örneklerinin çalışma zamanı davranışını değiştirmekten kaçınmak için hem kötü amaçlı hem de zararlı uygulamalarımızı gerçek akıllı telefonlarda çalıştırılmaktadır. RF, KNN ve DT algoritmaları ile simule edilmiştir.
    - Senaryo A: Bu senaryo, uygulamanın kötü amaçlı olup olmadığına dair binary sonuç vermektedir. Sonuçlara göre, üç algoritma da benzer sonuçlar vermektedir. Başarı olasılığı Precision (kesinlik) ortalaması yaklaşık % 85'tir ve hem eğitim hem de değerlendirme seti için% 88 Recall (Rc) vardır.
    - Senaryo B: Senaryo B, kötü amaçlı yazılım kategorisi sınıflandırmasının sonuçlarını göstermektedir. Tüm test algoritmaları için ortalama Precision ve Recall % 50'den azdır. Değerlendirme testinde aynı prosedür kullanılmıştır.
    - Senaryo C: Bu senaryo, uygulamanın kötü amaçlı olup olmadığını alt kategoriler ile tespit etmektedir. Bu senaryo Senaryo B gibi düşük algılama sonuçları sergilemektedir.

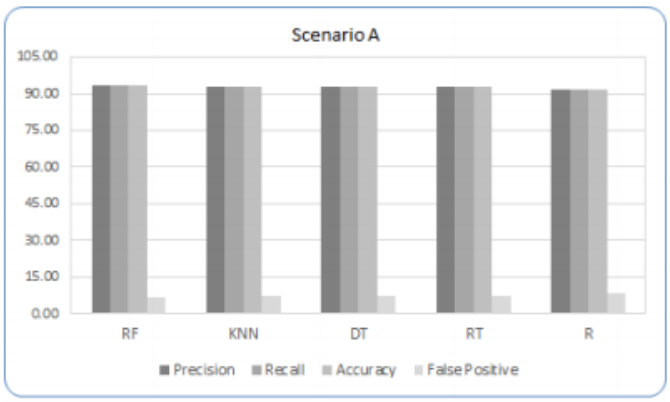
Table IV: Analysis results with 3 classifiers

Dataset:	**Training (10-fold cross validation)**									**Evaluation (Testing set)**								
Scenario:	**A (Malware Binary)**			**B (Malware Category)**			**C (Malware Families)**			**A (Malware Binary)**			**B (Malware Category)**			**C (Malware Families)**		
Algorithm:	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT	RF	KNN	DT
Precision (%):	84.00	83.60	85.10	46.50	45.70	46.50	22.00	21.50	21.00	85.80	85.40	85.10	49.90	49.50	47.80	27.50	27.24	26.66
Recall (%):	87.50	87.30	88.00	45.50	44.80	44.70	21.50	21.60	21.40	88.30	88.10	88.00	48.50	48.00	45.90	25.50	23.74	20.06

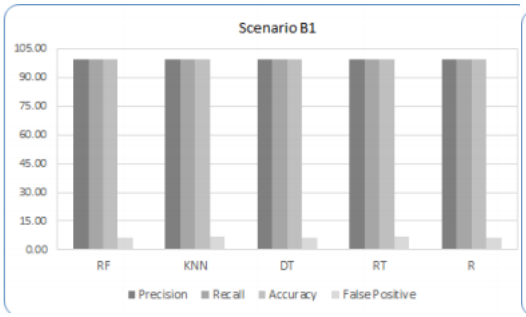
10. Android Adware and General Malware Dataset (CIC-AAGM2017)

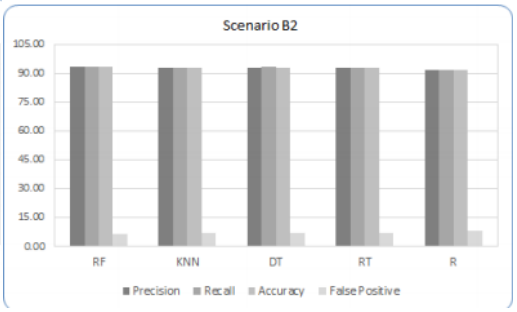
- **Proje Konusu:** Android kötü içerikli yazılım tespiti
- **Dataset:** CICAAGM dataset
- **Makale:** Arash Habibi Lashkari, Andi Fitriah A. Kadir, Hugo Gonzalez, Kenneth Fon Mbah and Ali A. Ghorbani, "Towards a Network-Based Framework for Android Malware Detection and Characterization", In the proceeding of the 15th International Conference on Privacy, Security and Trust, PST, Calgary, Canada, 2017.
- **Makale Özet:** Sofistike ve gelişmiş Android kötü amaçlı yazılımı, kötü amaçlı yazılım analisti tarafından kullanılan emülatörün varlığını belirleyebilir ve buna yanıt olarak, algılanmaktan kaçınmak için davranışını değiştirebilir. Bu sorunun üstesinden gelmek için Android uygulamalarını gerçek cihaza yüklenmiştir ve ağ trafiği incelenmiştir.

    - **Senaryo A:** Bu senaryoda, tüm genel kötü amaçlı yazılımlar ve reklam yazılımları kötü amaçlı yazılım olarak etiketlenir ve veri kümesinin iyi huylu ve kötü amaçlı yazılım olmak üzere iki sınıfı vardır. Senaryo A, bu senaryo için beş ortak sınıflandırıcı yürütmenin sonuçlarını gösterir. Sonuçlar, sadece seçilen dokuz özelliği % 93 doğrulukla kullanan RF algoritmasının en yüksek algılama oranına sahip olduğunu göstermektedir. Beş algoritmanın tamamının ortalaması yaklaşık % 93 algılama doğruluğunu gösterir ve hepsi için başarı olasılığı (Precision) yaklaşık % 92'dir.

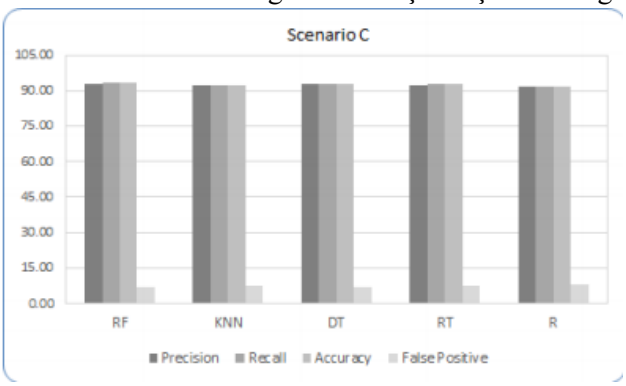


    - **Senaryo B:** İlk kötü amaçlı yazılım karakterizasyonunu içermektedir, genel kötü amaçlı yazılım ve zararsız uygulamalar arasındadır. Senaryo B1, seçilen beş makine öğrenimi algoritması için senaryo B'deki ilk karakterizasyonun sonucunu sunar. Tüm test algoritmaları için ortalama doğruluk oranı % 99'dan fazladır. % 0,0065 FP oranı da yüksek doğru algılama ve çok düşük yanlış etiketlemeyi gösterir. B2 senaryosu ise, B senaryosundaki ikinci karakterizasyonun sonucunu sunar bu da iyi huylu (B2) dan reklam yazılımlarını tespit etmektir. Sonuç, beş makine öğrenimi algoritmasının tümünün çok yüksek doğrulukta doğru etiketlemeyi ortalama% 93,3 ve çok düşük FP ile ortalama olarak yaklaşık% 0,06 elde ettiğini göstermektedir.





- **Senaryo C:** Bu senaryo tüm kategorilerin tespitini ve karakterizasyonunu içermektedir. Bunlar, reklam yazılımı, genel kötü amaçlı yazılım ve zararsızdır. Beş makine öğrenimi algoritmasında maksimum% 0,08 FP oranıyla ortalama% 92 doğruluk sunar. Ortalama olarak tüm algoritmalar için başarı olasılığı % 92'dir.

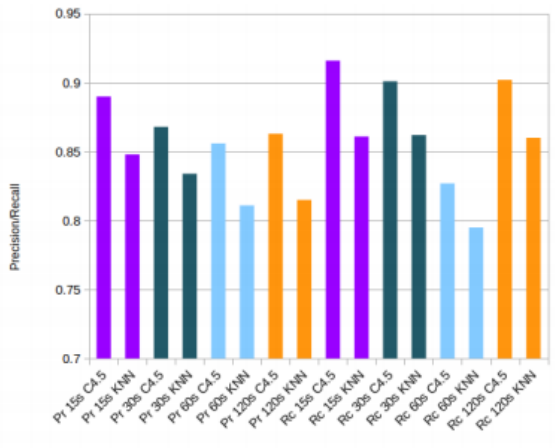


11. DoS dataset (application-layer) 2017

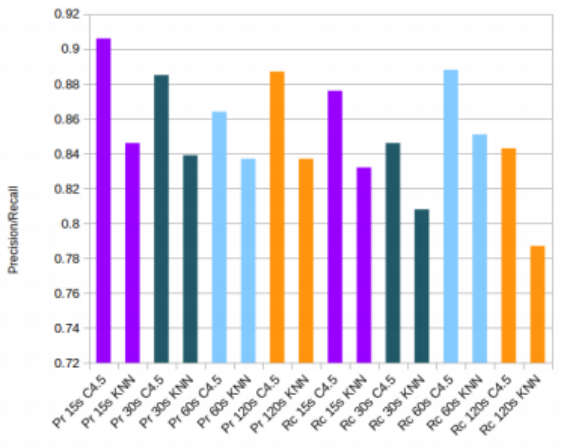
- **Proje Konusu:** DoS atakları
- **Dataset:** DoS veriseti, araştırma niteliğinde bir makaledir.
- **Makale:** Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. "Detecting HTTP-based Application Layer DoS attacks on Web Servers in the presence of sampling." Computer Networks, 2017.
- **Makale Özet:** Bu çalışmada çeşitli uygulama katmanı DoS saldırıları ve parametrik olmayan CUSUM algoritmasına dayalı genel bir algılama yaklaşımı önerir.

12. VPN-nonVPN dataset (ISCXVPN2016)

- **Proje Konusu:** VPN ve nonVPN trafik kategorizasyonu
- **Dataset:** ISCXVPN2016
- **Makale:** Gerard Drapper Gil, Arash Habibi Lashkari, Mohammad Mamun, Ali A. Ghorbani, "Characterization of Encrypted and VPN Traffic Using Time-Related Features", In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP 2016) , pages 407-414, Rome, Italy.
- **Makale Özet:** ISCX projesinde gerçek dünya trafiğinin temsili bir veri kümesi oluşturmak için, veri kümesinin çeşitlilik ve miktar açısından yeterince zengin olmasını sağlayan özellikler(features) tanımlanmıştır. İki çeşit oturum ile; düzenli ve VPN üzerinden veri seti yapıldığı için toplamda 14 trafik kategorisi verisetine dahil edilmiştir; VOIP, VPN-VOIP, P2P, VPN-P2P, vb.
    - **Senaryo A:** Bu senaryonun amacı, şifrelenmiş trafiği VPN kimliği ile karakterize etmektir, ör. VPN (VPN-VOIP) aracılığıyla VOIP ve sesli aramalar arasında ayrım yapabilir. Sonuç olarak, 14 farklı trafik türüne, 7 normal şifreli trafiğe ve 7 VPN türüne sahip olunmuştur. Bu senaryoda karakterizasyon iki adımda gerçekleşmektedir. İlk olarak, VPN ve VPN olmayan trafiği birbirinden ayrılması ve ardından her bir trafik türünü ayrı ayrı karakterize edilmesi (VPN ve VPN Olmayan). Bunu yapmak için de veri seti iki farklı veri setine bölünmüştür: biri düzenli şifrelenmiş trafik akışlarıyla, diğeri VPN trafik akışlarıyla olmak üzere. Şekilde Precision (doğruluk) ve Recall (hassasiyet) değerleri görülmektedir.

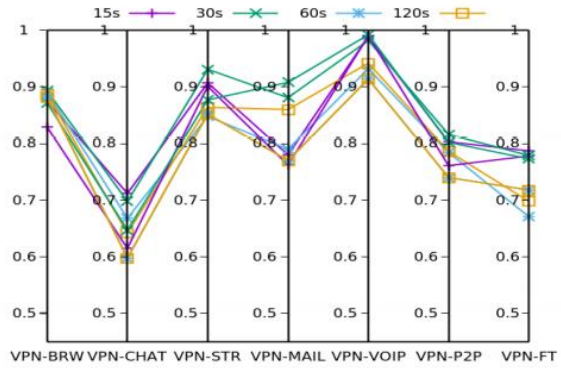


(a) Scenario A VPN Precission and Recall

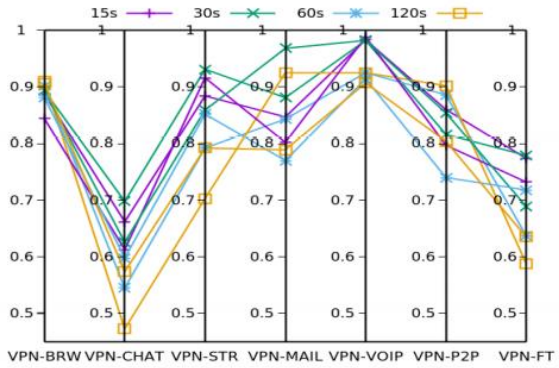


(b) Scenario A NON-VPN Precission and Recall

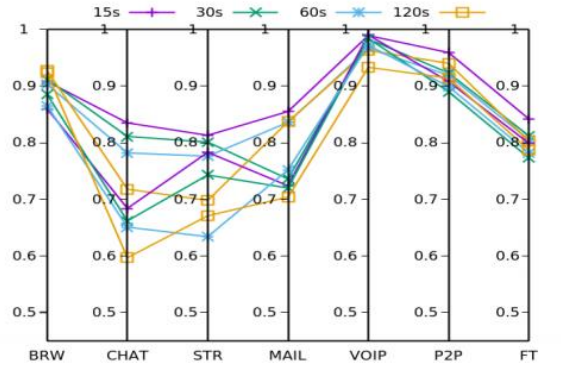
- **Senaryo B:** Bu senaryoda, karakterizasyonu tek adımda yapmak için karma bir veri kümesi kullanılmaktadır. Sınıflandırıcının girdisi, düzenli olarak şifrelenmiş trafik ve VPN trafiğidir. Çıktı olarak da Senaryo A'daki 14 farklı (7 şifreli ve 7 VPN) kategori kullanmaktadır.



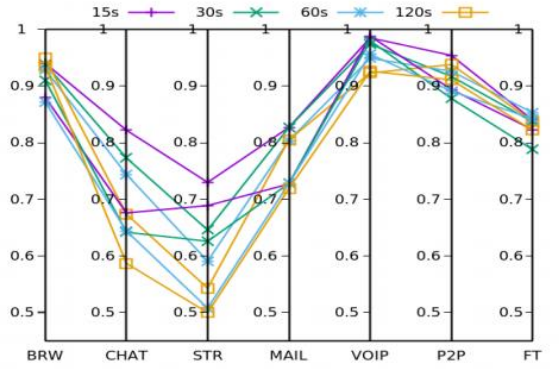
(a) ScenarioA VPN Precision



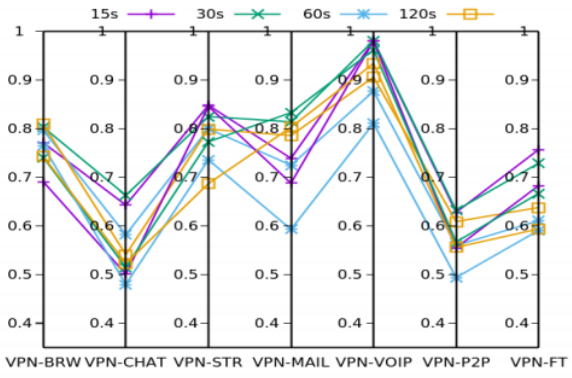
(b) ScenarioA VPN Recall



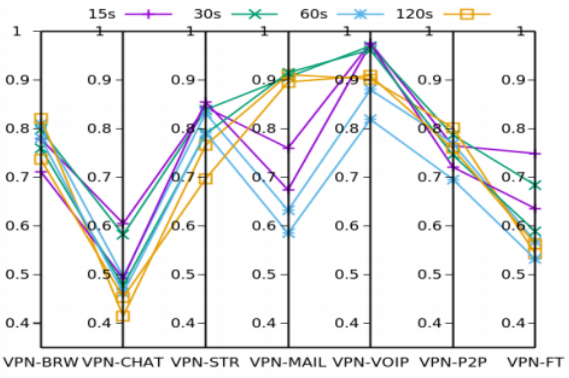
(c) ScenarioA Non-VPN Precision



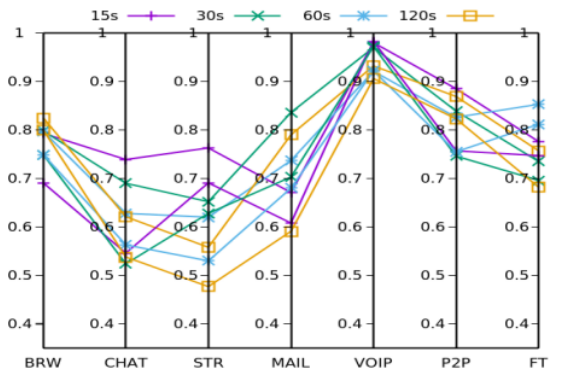
(d) ScenarioA Non-VPN Recall



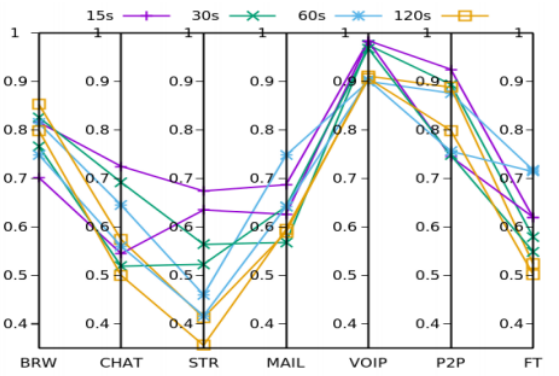
(e) ScenarioB VPN Precision



(f) ScenarioB VPN Recall



(g) ScenarioB Non-VPN Precision



(h) ScenarioB Non-VPN Recall

- **Senaryo A Sonuçları Analizi:** Şekil 2'de, trafiği VPN ve VPN Olmayan olarak sınıflandırdığımız, senaryo A'nın ilk bölümünün doğruluk (Pr) ve duyarlılık (Rc) sonuçlarına sahibiz. Akış zaman aşımı (ftm) değerleri ile sınıflandırıcıların performansı arasında doğrudan bir ilişki olduğunu görebiliriz. Özellikle, C4.5 VPN trafik sınıflandırıcısının Kesinliği (Pr), 120 saniye kullanarak 15 saniye kullanarak 0.890'dan 0.86'ya ve VPN olmayan trafik için Pr, 0.906'dan 0.887'ye düşer. VPN trafiğinin 0,848'den 0,815'e ve VPN olmayan trafik durumunda 0,846'dan 0,837'ye düştüğü KNN algoritması durumunda da benzer bir davranış görebiliriz. En iyi sonuçlar C4.5 algoritması ve 15s ftm: VPN için 0.89 ve NonVPN için 0.906 kullanılarak elde edilir. Bu, zamanla ilgili özellikleri kullanarak VPN'i VPN olmayanlardan 15 saniyelik bir gecikmeyle (bir akış oluşturmak için gereken süre) ayırt edebileceğimiz anlamına gelir. Bu sonuçlar, VPN ve VPN olmayan trafik sınıflandırması için zamanla ilgili özellikler kullanıldığında, daha kısa zaman aşımı değerleri kullanmanın doğruluk oranını artırdığını göstermektedir. Senaryo A'nın ikinci bölümü, VPN ve VPN olmayan trafiğin ayrı ayrı karakterizasyonuna odaklanır. En iyi sonuçlar (ortalama Pr) C4.5 ve 15s ftm ile elde edilir; sırasıyla VPN ve VPN olmayan sınıflandırıcılar için 0.84 ve 0.89dur.
- **Senaryo B Sonuçları Analizi:** Sonuçlar şu şekilde yorumlanır; C4.5 algoritması kullanıldığında, VPN-Browsing, VPN-Mail ve Mail'in 15 saniyelik Pr değeri sırasıyla 0.771, 0.739, 0.671'dir ve 120 saniye ile elde edilen 0.809, 0.786, 0.79'dan daha düşük değerlerdir. KNN sonuçları da benzerdir, VPN-Browsing, VPN-Chat ve VPN-Mail trafiği kategorilerinin Pr'si 15 saniye için (0.691, 0.501, 0.688) şeklindedir. 120 saniye (0.743, 0.501, 0.688) ile elde edilen doğruluktan daha küçüktür. Ayrıca, farklı ftm değerlerinden en yüksek ortalama doğruluk, C4.5 için 0.783 ve KNN algoritmaları için 0.711, Senaryo A'daki en iyi değerlerden yaklaşık 0.5 puan daha düşüktür.

13. Tor-nonTor dataset (ISCXTor2016)

- **Proje Konusu:** Zaman özellikleri kullanılarak Tor trafik kategorizasyonu
- **Dataset:** ISCXTor2016
- **Makale:** Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun and Ali A. Ghorbani, "Characterization of Tor Traffic Using Time Based Features", In the proceeding of the 3rd International Conference on Information System Security and Privacy, SCITEPRESS, Porto, Portugal, 2017.
- **Makale Özet:** Gerçek dünya trafiğini temsil eden bir veri kümesi oluşturmak için özellik seti (feature set) tanımlanmıştır. Tarayıcı trafiği veri seti için üç kullanıcı ve sohbet, posta, FTP, p2p, vb. gibi iletişim bölümleri için iki kullanıcı oluşturulmuştur. Tor dışı trafik için (non Tor) 12. Projedeki VPN projesinden iyi huylu(benign) trafiği alınmıştır.

  - **Senaryo A:** Bu senaryoyu oluşturmak için 2 farklı veri kümesini, bu belgede sunulan Tor veri kümesini ve Draper-Gil ve diğerleri tarafından oluşturulan şifrelenmiş trafiğin veri kümesi birleştirilmiştir. Akışlar oluşturulmuştur ve her veri kümesinden önerilen zamana dayalı özellikler çıkarılmıştır. Daha sonra, Tor veri kümesindeki tüm akışlar Tor olarak ve Draper-Gil ve diğerlerinden gelen tüm akışları etiketlenmiştir.
  - **Senaryo B:** Bu senaryoda, yalnızca bu yazıda sunulan Tor veri kümesini kullanılmıştır. Ağ geçidinde yakalanan .pcap dosyalarından akışları oluşturulmuştur ve bunları iş istasyonunda (workstation) yürütülen uygulamaya göre (Gözatma, Ses, CHAT, Posta, P2P, FT, VOIP ve Video) etiketlenmiştir.
  - Tablo 3'te elde edilen sonuçlar, her durumda C4.5 ve KNN'nin alt sınır olan Zero R'den daha iyi olduğunu göstermektedir. Sonuçlardan, daha uzun zaman aşımı değerlerinin (120 sn veri kümesi) daha kısa olanlardan (örneğin, 15 sn veri kümesi) daha iyi sonuçlar sağladığı görülmektedir, ancak bu eğilim aynı zamanda daha uzun zaman aşımı değerlerinin sonuçlarımızı alt sınıra (Zero R) yaklaştırdığını gösteriyor.

Table 3: Training results for Scenarios A and B.

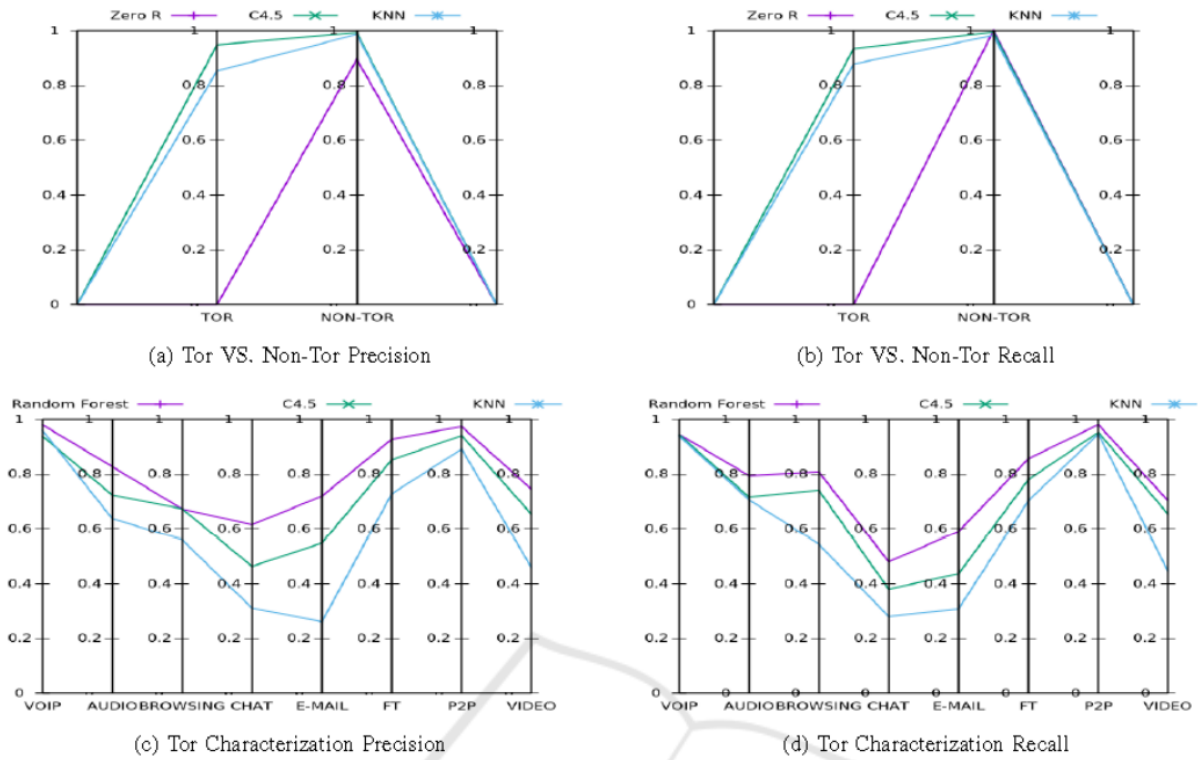
Scenario A

	Zero R				C4.5				KNN			
	SE+BF		IG+RK		SE+BF		IG+RK		SE+BF		IG+RK	
	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC
10s.	0.777	0.881	0.777	0.881	0.950	0.950	0.973	0.973	0.940	0.940	0.953	0.953
15s.	0.801	0.895	0.801	0.895	0.976	0.976	0.987	0.987	0.967	0.967	0.971	0.970
30s.	0.871	0.933	0.871	0.933	0.979	0.979	0.987	0.987	0.975	0.975	0.976	0.976
60s.	0.922	0.960	0.922	0.960	0.985	0.986	0.990	0.990	0.981	0.981	0.983	0.983
120s.	0.951	0.975	0.951	0.975	0.988	0.988	0.990	0.991	0.985	0.985	0.988	0.988

Scenario B

	Random Forest				C4.5				KNN			
	SE+BF		IG+RK		SE+BF		IG+RK		SE+BF		IG+RK	
	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC	PR	RC
10s.	0.760	0.762	0.842	0.840	0.728	0.732	0.790	0.790	0.675	0.676	0.702	0.704
15s.	0.833	0.831	0.841	0.836	0.797	0.798	0.796	0.796	0.688	0.691	0.704	0.707
30s.	0.799	0.799	0.808	0.808	0.760	0.760	0.754	0.756	0.656	0.660	0.664	0.666
60s.	0.744	0.748	0.750	0.754	0.696	0.698	0.690	0.695	0.612	0.611	0.615	0.618
120s.	0.725	0.728	0.741	0.743	0.665	0.664	0.674	0.675	0.595	0.600	0.607	0.609

SE+BF is CfsSubsetEval+BestFirst
IG+RK is Infogain+Ranker
PR is Precision
RC is Recall



(a) Tor VS. Non-Tor Precision

(b) Tor VS. Non-Tor Recall

(c) Tor Characterization Precision

(d) Tor Characterization Recall

Figure 3: Precision and Recall of Validation experiments.

- **Senaryo A Sonuçları Analizi:** Test sürecindeki en iyi kombinasyonların kesinliğini ve geri çağrılmasını hesaplamak için yapılan analizlerin sonuçları şekilde görülmektedir. Her bir sınıf (Tor, Tor olmayan) için kesinlik ve geri çağırma değerleri gösterilmektedir. Zero R sınıflandırıcısı herhangi bir Tor örneğini algılamayacağı anlamı çıkarılırken, Makalede önerilen C4.5 sınıflandırıcısı tüm Tor örneklerinin % 93,4'ünü algılayabilir (Recall) ve bir örneği Tor olarak her etiketlediğinde, % 94,8 başarı olasılığına (Precision) sahiptir. Tor olmayan örneklerle ilgili olarak, Zero R sınıflandırıcı Tor olmayan örneklerin % 100'ünü algılayacaktır (her şeyi Tor olmayan olarak etiketler, geri çağırma = 1) ve etiketleri % 89,5 doğru olacaktır. C4.5 sınıflandırıcısının da Tor olmayan örneklerin % 99,4'ünü duyarlı tespit eder, % 99,2 doğru olacaktır.
- **Senaryo B Sonuçları Analizi:** Sonuçlar, 8 sınıfın her biri için kesinlik ve geri çağırma değerlerini gösteren Şekil 3'de sunulmuştur. Şekilde açıkça görüldüğü gibi, en iyi sonuçlar Random Forest ile elde edilir ve ağırlıklı ortalama hesaplandığında, kesinlik için 0.843, 0.788, 0.705 ve geri çağırma için 0.838, 0.790, 0.705 (Random Forest, C4.5 ve KNN sırasıyla) sonuçları bulunur.

14. URL dataset (ISCX-URL2016)

- **Proje Konusu:** Kötücül URL tespiti
- **Dataset:** URL dataset
- **Makale:** Mohammad Saiful Islam Mamun, Mohammad Ahmad Rathore, Arash Habibi Lashkari, Natalia Stakhanova and Ali A. Ghorbani, "Detecting Malicious URLs Using Lexical Analysis", Network and System Security, Springer International Publishing, P467--482, 2016.
- **Makale Özeti:** Kötü amaçlı URL'lerin saldırı türlerine göre tespit edilmesi ve sınıflandırılması için sözcüksel analizin (lexical analysis) bu URL'lerin proaktif tespiti için verimli olduğu gösterilmektedir. Ayrıca, kötü amaçlı URL türlerini hedefleyen gizleme tekniğinin türünü anlamak için gizleme tekniklerinin kötü amaçlı URL'ler üzerindeki etkisi de incelenmektedir.
    - Kötü amaçlı URLler dört grupta incelenmiştir; benign (iyicil), spam, phishing (kimlik hırsızlığı, e-dolandırıcılık) ve defacement (tahrifat, içerik bozma) URLdir.
    - Verilerin sınıflandırılması iki grup algoritma aracılığıyla yapılır. K-Nearest Neighbours algoritması (KNN, komşu gruplara katkılarına göre ağırlık verilerek sınıflandırma için kullanılan bir örüntü tanıma yöntemidir) ve ağaç tabanlı sınıflandırıcılar (C4.5 ve RandomForest, sonuçları ağaç olarak sunmak için kullanılır) kullanılmıştır.
    - Tabloda Pr doğruluk (precision) anlamına gelmekte olup, Pr =TP/(TP+FP) şeklinde hesaplanmaktadır. Rc duyarlılık (recall or sensitivity) anlamına gelmekte olup, Rc=TP/(TP+FN) formülü ile hesaplanmaktadır.
    - Sonuçlara bakılarak, ortalama %97 doğruluk ve duyarlılık ile tespit yapılmıştır.

Table 3. Classification results based on lexical features (multi-class)

Dataset	Labels	C4.5		KNN		RF	
		Pr	Rc	Pr	Rc	Pr	Rc
Multi class	Spam	0.96	0.971	0.96	0.97	0.962	0.986
	Phishing	0.92	0.856	0.92	0.85	0.926	0.928
	Malware	0.96	0.97	0.96	0.97	0.979	0.983
	Defacement	0.93	0.97	0.93	0.97	0.969	0.973
	Average	0.94	0.94	0.94	0.94	**0.97**	**0.97**

15. Android Botnet dataset

- **Proje Konusu:** Android Botnets
- **Dataset:**
- **Makale:** Andi Fitriah A. Kadir, Natalia Stakhanova, Ali A. Ghorbani, "Android Botnet: What URLs are telling us", 9th International Conference on Network and System Security (NSS), November 3-5, 2015,New York City, USA.
- **Makale Özet:** Bu yazıda, Android platformunun ilk görünümünden bu yana tespit edilen Android botnetlerinin Komuta ve Kontrol (C&C) ve yerleşik URL'lerinin derin bir analizi sunulmaktadır. Hem statik hem de dinamik analizleri görselleştirmeyle birleştirerek, analiz edilen botnet ailelerinin çoğunluğu arasındaki ilişkileri ortaya çıkarır ve her kötü amaçlı altyapıya ilişkin bir içgörü sunulmaktadır.

16. Botnet dataset

- **Proje Konusu:** Botnets
- **Dataset:** ISOT dataset , ISCX 2012 IDS dataset
- **Makale:** Beigi, Elaheh Biglar, et al. "Towards effective feature selection in machine learning-based botnet detection approaches." Communications and Network Security (CNS), 2014 IEEE Conference on. IEEE, 2014.
- **Makale Özet:** Bu yazıda, mevcut botnet tespit çalışmalarında kullanılan akış temelli özellikler gözden geçirilerek ve göreceli etkinlikler değerlendirilmektedir. Uygun bir değerlendirme sağlamak için, çeşitli botnet izleri ve arka plan trafiği içeren bir veri kümesi oluşturulmaktadır.

17. Android validation dataset

- **Proje Konusu**: Android uygulamalarının benzerliğini saptama
- **Dataset:** Veri seti, farklı kaynaklardan alınan 72 uygulamayı ve bu uygulamaların her birinin 10 farklı transformasyonunu kullanarak toplamda 792 uygulamadan oluşmaktadır.
- **Makale:** Gonzalez, Hugo, Natalia Stakhanova, and A. Ghorbani. "Droidkin: Lightweight detection of android apps similarity", Proceedings of the 10th Securecomm, 2014.
- **Makale Özet:** Bu yazıda, Android uygulamaları benzerliğinin tespiti için sağlam bir yaklaşım olan DroidKin sunulmaktadır. DroidKin, ona eşlik eden ikili ve meta verilerden türetilen bir dizi özelliğe dayanarak, çeşitli gizleme düzeyleri altındaki uygulamalar arasındaki benzerliği tespit etmek için kullanılır. DroidKin, uygulamalar arasındaki benzerlikleri belirleyen ve bunların ilişkilerini belirleyen analizler gerçekleştirir.

18. Intrusion Detection Evaluation Dataset (ISCXIDS2012)

- **Proje Konusu:** Aykırılık tespiti
- **Dataset:** Bu makale deneysiz bir makaledir. Veri seti üretim yöntemlerini ele almaktadır.
- **Makale:** Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, Ali A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection", Computers & Security, Volume 31, Issue 3, May 2012, Pages 357-374.
- **Makale Özet:** Bu yazıda, gerekli veri setlerini üretmek için sistematik bir yaklaşım ele alınmaktadır. Temel kavram, izinsiz girişlerin ayrıntılı açıklamalarını ve uygulamalar, protokoller veya alt düzey ağ varlıkları için soyut dağıtım modellerini içeren profiller kavramına dayanmaktadır. Gerçek izler, HTTP, SMTP, SSH, IMAP, POP3 ve FTP için gerçek trafik oluşturan aracı profilleri oluşturmak üzere analiz edilir.