www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 9, No: 4, 2022 (1572-1582)

El-Cezerî Journal of Science and Engineering
Vol: 9, No: 4, 2022 (1572-1582)
DOI : 10.31202/ecjse.1209586

ECJSE

Makale / Research Paper

Siber Güvenlik Amaçlı Büyük Veri Görselleştirme: BETH Veri Seti

Aytaç DOĞANAY[a], Abdullah ORMAN[b*], Murat DENER[c]

[ac] Bilgi Güvenliği Mühendisliği, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, ANKARA
[b] Bilgisayar Teknolojileri Bölümü, Teknik Bilimler MYO, Ankara Yıldırım Beyazıt Üniversitesi, ANKARA

Received/Geliş: 24.11.2022 **Accepted/Kabul:** 04.12.2022

Öz: Bu çalışmada siber güvenlik amaçlı büyük veri görselleştirmesi ile ilgili gerçekleştirilen literatür taranarak örnek bir veri seti üzerinde veri görselleştirme uygulaması yapılmıştır. Gerçekleştirilen görselleştirme çalışması literatürdeki benzeri olan bir çalışma ile uygulamalı olarak karşılaştırılmıştır. Karşılaştırma neticesinde, bu çalışmada sunulan kriterler kullanılarak görselleştirmenin uygulanması halinde oluşacak görsellerin kullanıcı (insan) tarafından çok daha rahat şekilde okunabileceği ve bu özelliği ile siber saldırı tespitlerini kolaylaştıracak görselleştirme yapılabileceği ortaya konmuştur. Çalışmada güncel bir siber saldırı veri seti olan BETH veri seti üzerinde Principle Component Analysis (PCA) yöntemi uygulanmıştır.

Anahtar Kelimeler: BETH, büyük veri görselleştirme, KDD99, MDS, PCA, siber güvenlik.

Big Data Visualization for Cyber Security: BETH Dataset

Abstract: In this study, a data visualization application was made on a sample data set by scanning the literature on big data visualization for cyber security purposes. The visualization study carried out was practically compared with a similar study in the literature. As a result of the comparison, it has been revealed that if the visualization is applied using the criteria presented in this study, the visuals that will be created can be read more easily by the user (human) and that visualization can be made that will facilitate the detection of cyber attacks with this feature. In the study, Principle Component Analysis (PCA) method was applied on the BETH data set, which is a up-to-date cyber attack data set.

Keywords: BETH, big data visualization, KDD99, MDS, PCA, cyber security.

1. Giriş

Gelişen teknoloji daha çok verinin dijital ortama taşınması sonucunu doğurmuştur. Gün geçtikçe artan sayıda cihazın dijital ortama taşınması neticesinde birbirine Internet veya intranet ağları ile bağlı platformlarda daha çok veri barındırılmaya ve işlenmeye başlanmıştır. Birbiriyle haberleşmeye başlayan ve ağ ortamında daha çok veri üretmeye başlayan Internet of Things (IoT) cihazlarının da katılımıyla bahse konu platformlarda tutulan ve işlenen verinin boyutları devasa hale gelmiştir. 2018 yılında Wang ve arkadaşları tarafından yapılan araştırmada 2011 yılında bile sadece sağlık büyük verisi miktarının 150 Exabyte'ı geçtiği vurgulanmıştır [1]. Yapılı, yarı yapılı ya da yapısız olarak tutulan bu devasa boyutlardaki verilere Büyük Veri adı verilmektedir. Büyük veriler genellikle akıllı telefonlar, kişisel bilgisayarlar, trafik kameraları ve sensörler gibi farklı cihazlardan toplanır [2]. Bu teknolojik gelişmelerin mükemmel gelişmeler olmaları yanında, büyük verilerin depolanması ve işlenmesi, kullanıcı gizliliğinin korunması ve hassas bilgilerin güvenliğinin sağlanması sorunları ortaya çıkarmaktadır [3].

Bu makaleye atıf yapmak için
Doğanay, H.A., Orman, A., Dener, M., *"Siber Güvenlik Amaçlı Büyük Veri Görselleştirme: BETH Veri Seti"* El-Cezerî Fen ve Mühendislik Dergisi 2022, 9(4); 1572-1582.

How to cite this article
Doğanay, H.A., Orman, A., Dener, M., "Big Data Visualization for Cyber Security: BRTH Dataset" El-Cezerî Fen ve Mühendislik Dergisi 2022, 9(4); 1572-1582.
ORCID ID: [a]0000-0003-4816-4373; [b]0000-0002-3495-1897; [c]0000-0001-5746-6141

Bu çalışma ICAIAME 2022'de sunulmuştur.

Büyük veri analitiği büyük verinin karakteristik özellikleri dikkate alınarak uygulanmaktadır. Büyük verinin geleneksel karakteristikleri 3V olarak ortaya konmuştur [4] ve bunlar Volume, Variety ve Velocity yani sırasıyla Hacim, Değişkenlik ve Hızdır. Büyük verinin gelişim döngüsü içerisinde bu geleneksel karakteristiklere iki kavram daha eklendiği ve karakteristiklerin 5V olarak kabul edildiği görülmektedir [5]. Eklenen iki V Veracity ve Value, yani sırasıyla Doğruluk ve Değerdir.

Büyük veri endüstri, ticaret, sosyal medya, eğitim, sağlık, tarım gibi birçok alanda doğru işlendiği takdirde karın artırılması, servis kalitesinin artırılması, harcamaların düşürülmesi, bilgiye erişimin hızlandırılması gibi sayısız faydalar sunabilmektedir. Siber güvenliğin sağlanması için büyük veriden faydalanan çalışmalarda da etkili ve başarılı sonuçlar alındığı görülmektedir [6]. Özellikle makine öğrenmesi, derin öğrenme ve yapay zekâ teknolojilerinin gelişiminin büyük veri analitiğine sağladığı katkılar neticesinde siber saldırı ve sızma girişimlerinin tahmin doğrulukları ve saldırıları önlemek için alınan önlemlerin etkili olma yüzdeleri artmıştır.

Büyük verinin özellikle hacim ve hız karakteristikleri dikkate alındığında güvenlik duvarı logları, işletim sistemi logları, sızma tespit ve önleme sistemleri logları, bal küpü logları ve diğer SIEM (Security Information and Event Management) sistemlerine ait loglar çok hızlı akan bir trafik ortamında devasa boyutlarda veri yığını oluşturabilmektedir. Statista'nın verilerine göre engellenen günlük siber saldırı sayısının 2016 yılında 229.000, 2017 yılında 611.000 ve 2018 yılında 953.000 olduğu görülmektedir [7]. Atağın engellenmesine kadar geçen sürede saldırıya ait milyonlarca kayıt oluşmakta ve bu kayıtlar büyük veri özelliği taşımaktadır. Bu büyük veri yığını siber saldırıların karakteristiklerinin, kaynaklarının, çeşitlerinin, intervallerinin, frekanslarının ve yıkıcılık özelliklerinin büyüklüğünün ne derecede olduğuna dair çok kıymetli veriler içermektedir. Siber saldırılar ve saldırı girişimleri esnasında oluşan logların analiz edilerek saldırı karakteristiklerinin tespit edilmesi neticesinde mevcut siber saldırı tespit ve önleme sistemlerinin yetenekleri artırabilmektedir. Bu durum, kurum ve kuruluşların siber ortamda güvenlik kalkanlarını sağlamlaştırmaktadır.

Bowie Üniversitesinde gerçekleştirilen bir çalışmanın sonucuna göre büyük veriyi başarılı şekilde kullanan işletmelerin %84'ü siber saldırıları durdurmayı başarmıştır [8]. Tek bir kurumda bile siber güvenliğe ilişkin verilerin sayısı milyarları bulabilmektedir. Siber saldırıların ve sızma girişimlerinin tespit edilmesi için siber güvenliğe ilişkin verilerin analizinde çeşitli yöntemler ve araçlar kullanılmaktadır. Yapılan analizin ve çıkarılan sonuçların isabetli olması için doğru yöntem ve araçların kullanılması da önemlidir. Hacim ve çeşitlilik karakteristiği açısından da büyük veri kavramını bütünüyle karşılayan siber güvenlik loglarının isabetli şekilde analiz edilmesi yöntemlerinden birisi de büyük verinin görselleştirilmesi yöntemidir. Literatürde, büyük veri niteliğindeki siber güvenlik verilerinin görselleştirilmesini konu alan birçok çalışma mevcuttur. Artık verilerin ve tekrarlanan verilerin çokluğu büyük veri görselleştirmesini olumsuz etkilerken makine öğrenmesi ve yapay zekâ kullanılması ise siber güvenlik önlemlerine destek olacak şekilde olumlu ölçüde katkı sağlamaktadır. Bununla birlikte literatürdeki büyük veri çalışmalarında ortaya konulan modellerin test edilmesi için veri setlerinin kullanılması da siber güvenlik için büyük veri işlenmesinde hangi yöntemin daha verimli olduğunun anlaşılmasını sağlamaktadır. Ancak gerçekleştirilen literatür taramasından da anlaşılacağı üzere incelenen çalışmalarda kullanılan veri setlerinin oldukça eski olduğu görülmektedir. Teknoloji ilerledikçe siber saldırganlar güvenlik çerçevelerine meydan okumak için özel ve hassas verilerin güvenliğini tehlikeye atacak yeni teknikler geliştirmektedirler [9]. Yeni saldırı ve sızma girişimi teknikleri loglarda da yeni kayıtların oluşması anlamına gelmektedir. Bu açıdan bakıldığında literatürde gerçekleştirilen çalışmalarda uygulanan model testlerinin yeni ve güncel veri setlerinin üzerinde yapılmasının siber güvenlik için alınacak önlemleri de güncelleyeceği değerlendirilmektedir. Çalışmamızın ana fikrini oluşturan bu değerlendirme doğrultusunda ikinci bölümde büyük veri görselleştirme kullanan literatür taraması

gerçekleştirilmiş ve çalışmanın literatürden farkı ortaya konmuş, büyük veri görselleştirme araç ve yöntemleri ile bu çalışmalarda kullanılan veri setleri incelenmiş ve sonuç olarak 2021 yılında oluşturulmuş olması itibari ile güncel bir siber güvenlik büyük veri seti olan BETH Veri Seti [10] üzerinde görselleştirme uygulaması gerçekleştirilmiştir. Üçüncü bölümde çalışmada kullanılan veri seti (BETH) tanıtılmış, dördüncü bölümde bu çalışmada kullanılan BETH veri seti ile güncel literatürde kullanılan oldukça eski veri setleri karşılaştırılmış, beşinci bölümde BETH veri seti Principle Component Analysis (PCA) kullanılarak görselleştirilmiş ve eski tarihli bir veri seti olan KDD veri setinin görselleştirilmesiyle karşılaştırılarak altıncı bölümde sonuçlar sunulmuştur.

2. İlgili Çalışmalar

Siber saldırı ve sızma girişim tespitlerinin yapılabilmesi için siber güvenlik ile ilişkili logların bu saldırılara ilişkin anomalileri içermesi beklenir. Anomali tabanlı değerlendirme yapabilmek için anomali içeren veri setleri üzerinde çalışılmalıdır. Bu veri setleri akan canlı veri seti olabileceği gibi kayıt altına alınmış veri setleri de olabilmektedir. Bazı veri setleri veri güvenliği açısından halka açık şekilde paylaşılmamaktadır. Birçok veri seti ise özellikle akademik çalışmaların geliştirilmesi için açık erişime sunulmaktadır.

Blue Gene/L, Thunderbird, Redstorm, Liberty ve Spirit isimli loglar dünya sıralamasında ilk 500 içerisinde olan gerçek süper bilgisayar sistemlerinden elde edilmiş logları içeren veri setleridir [11]. Araştırmacılar tarafından 2021 yılında gerçekleştirilen geniş ölçekli anomali tespiti için mekânsal havuzlama yöntemi kullanan bir çalışmada sonuçların test edilmesi için Blue Gene/L veri setinden elde edilen ve 4,747,963 log kaydı içeren BGL veri seti kullanılmıştır [12]. 2020 yılında gerçekleştirilen bir çalışmada araştırmacılar tarafından loglar üzerinde denetimsiz anomali tespiti modeli Thunderbird veri seti üzerinde gerçekleştirilmiştir [13]. 2017 yılında gerçekleştirilmiş bir araştırmada RedStorm veri setinin makine öğrenimi kullanılarak HPC (High Performance Computer) sistemleri uygulamalarındaki performans değişikliklerinin tanılanması konusunda kullanıldığı görülmektedir [14]. Liberty veri setinin K-en yakın komşular makine öğrenimi algoritması ile log tabanlı anomali tespiti için 2020 yılına ait bir çalışma kullanıldığı görülmektedir [15]. 2021 yılında gerçekleştirilen başka bir araştırmada log verisindeki anomalilerin taksonomisi konusu araştırılmış ve Thunderbird, Spirit ve BGL veri setlerinin kullanıldığı görülmüştür [16]. Blue Gene/L, Thunderbird, Redstorm, Liberty ve Spirit veri setlerinin oluşturulmaya başlanma tarihleri oldukça eskidir [17]. Sayılan veri setlerinin oluşturulma tarihleri ve yapısal bazı bilgileri Tablo 1 olarak sunulmuştur. Literatür taraması neticesinde son 5 yıl içerisindeki güncel çalışmalara bakıldığında halen çok eski veri setlerinin kullanıldığı görülmektedir. Aynı veri setlerinin kullanıldığı bazı güncel akademik çalışmalar Tablo 2 olarak sunulmuştur.

Tablo 1. Blue Gene/L, Thunderbird, Redstorm, Liberty ve Spirit veri setleri.

Veri Seti	Üretim Yılı	Üretici	Boyut (GB)	Mesaj Sayısı	Alarm Sayısı
Blue Gene/L	2005	IBM	1207	4747963	348460
Thunderbird	2005	Dell	24367	211212192	3248239
RedStorm	2006	Cray	29990	219096168	1665744
Spirit	2005	HP	30289	272298969	172816564
Liberty	2004	HP	22820	265569231	2452

Tablo 2. Blue Gene/L, Thunderbird, Redstorm, Liberty ve Spirit veri setleri kullanan akademik çalışmalar

Veri Seti	Araştırmacı	Yılı	Karşılaştırma
Blue Gene/L	Hirakawa ve ark.	2021	2005 tarihli eski veri seti kullanılmıştır.
Thunderbird	Farzad ve ark.	2020	2005 tarihli eski veri seti kullanılmıştır.
RedStorm	Tuncer ve ark.	2017	2006 tarihli eski veri seti kullanılmıştır.
Spirit	Wang ve ark.	2020	2005 tarihli eski veri seti kullanılmıştır.
Liberty	Wittkopp ve ark.	2021	2004 tarihli eski veri seti kullanılmıştır.

Siber güvenlik için büyük verinin görselleştirilmesinin önemini vurgulayan ve bu konunun çalışıldığı birçok araştırma mevcuttur. Bu çalışmaların birçoğunda modeller yine veri setleri üzerinde test edilmektedir. Üzerinde çalışılan KDD99, ISCX ve NSL-KDD gibi birçok eski tarihli veri setinin güncel çalışmalarda kullanıldığı görülmektedir. Makine öğrenimi tekniği kullanarak sızma tespiti için MapReduce tabanlı akıllı model kullanılan ve 2021 yılında gerçekleştirilen bir çalışmada araştırmacıların çalışmasında görselleştirme modülü bulunduğu ve KDD99 veri seti kullanıldığı görülmektedir [18]. Yine 2021 yılında gerçekleştirilen Kurumsal Sistem Yönetiminde güvenlik olayları tespiti için büyük veri ve derin öğrenmenin kullanımını konu alan bir çalışmada ortaya konan sızma tespiti gerçekleştiren modelin görselleştirme modülüne sahip olduğu ve KDD99 veri setinin kullanıldığı görülmektedir. [19]. 2020 yılında gerçekleştirilen saldırılar ve anomali sınıfları aracılığıyla denetimsiz algoritmaların eğitimli seçimi hakkında gerçekleştirilen çalışmada ISCX veri setinden faydalanmışlardır [20]. Ağ trafiğinin yeniden inşasına dayalı şifreli trafik sınıflandırması konulu bir çalışmada yine ISCX veri seti kullanılmıştır [21]. Büyük veri çerçevesini kullanarak sızma girişim tespiti için topluluk tabanlı ölçeklendirilebilir bir yaklaşım modeli konulu 2021 tarihli çalışmada araştırmacılar NSL-KDD veri setini kullanmışlardır [22]. Araştırmacılar, veri analizi üzerinden DDOS saldırısı tabanlı sızma girişimi tespit sistemi önerileri konulu 2021 tarihli olup 2022 yılında yayınlanacak olan çalışmalarında NSL-KDD veri setini kullanmışlardır [23]. KDD99 veri seti 1999 yılında [24], ISCX veri seti 2012 yılında [25] ve NSL-KDD veri seti ise KDD99 veri setindeki bazı hataların düzeltilmesi amacıyla 2009 yılında üretilmiştir [24]. KDD99, NSL-KDD ve ISCX veri setlerine ilişkin bilgiler Tablo 3 ve bu veri setlerinin kullanıldıkları çalışmalar Tablo 4 olarak sunulmuştur.

Tablo 3. KDD99, NSL-KDD ve ISCX veri setleri bilgileri

Veri Seti	Üretim Yılı	Üreten
KDD99	1999	UCI
NSL-KDD	2009	Tavallaee ve ark.
ISCX	2012	UNB

Tablo 4. KDD99, NSL-KDD ve ISCX veri setleri kullanılan çalışma tarihleri

Veri Seti	Araştırmacı	Yılı	Karşılaştırma
KDD99	Asif ve ark.	2021	1999 tarihli eski veri seti kullanılmıştır.
KDD99	Lee ve Lee.	2021	1999 tarihli eski veri seti kullanılmıştır.
ISCX	Zoppi ve ark.	2020	2012 tarihli eski veri seti kullanılmıştır.
ISCX	Ma ve ark.	2021	2012 tarihli eski veri seti kullanılmıştır.
NSL-KDD	Sahu ve ark.	2021	2009 tarihli eski veri seti kullanılmıştır.
NSL-KDD	Pande ve ark.	2021 (2022)	2009 tarihli eski veri seti kullanılmıştır.

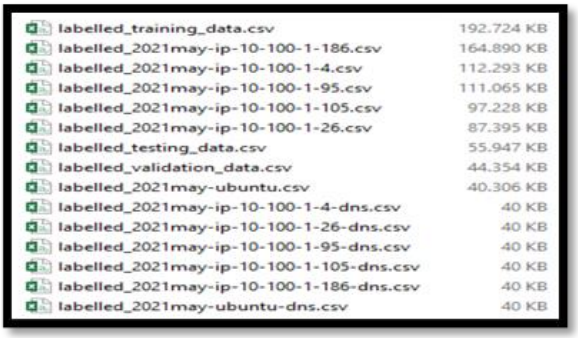
Literatür taraması göstermektedir ki güncel büyük veri ve makine öğrenmesi çalışmalarında halen eski veri setleri sıklıkla kullanılmaktadır. Yeni tür siber saldırı ve sızma girişimlerinin tespiti için

daha güncel veri setlerinin kullanılması çalışmaların etkinliğini artıracaktır. Bu çalışmada literatürden farklı olarak 2021 yılında oluşturulmuş güncel ve sızma girişim kayıtları içeren BETH veri setinin siber güvenlik amaçlı olarak büyük veri görselleştirilmesinde kullanılabilirliği irdelenmiştir.

3. BETH Veri Seti

BETH veri seti [10] anomali tespiti araştırması için araştırmacılar tarafından gerçek siber güvenlik verisi içeren bir veri seti olarak 5 Mayıs 2021 tarihinde toplanarak ICML 2021 Uncertainty and Robusness in Deep Learning çalıştayında Kate Highnam, Kai Arulkumaran, Zachary Hanif ve Nicholas R. Jennings isimli araştırmacılar tarafından sunulmuştur. BETH veri seti Internet üzerinde CSV formatında indirilebilir haldedir [27].

İndirilen dosyanın sıkıştırılmış hali 39,7 MB ve açılmış hali 928 MB'tır. Sıkıştırılmış dosya açıldığında içerisinden CSV uzantılı 15 dosya çıkmaktadır. Dosya isim ve boyut bilgileri Şekil 1'deki gibidir.



Şekil 1. BETH Veri Seti dosyaları

BETH veri seti büyük bir bulut sağlayıcısı üzerinde art arda olmayan 5 saat boyunca 23 Bal küpünden temin edilmiştir. Araştırmacılar oluşturdukları veri setini UMAP (tek tip manifold tahminleme ve gösterimi) yöntemi ile görselleştirmişler ve eğitim ve test setlerinin örtüştüğü ve veri seti içerisindeki kötücül ve kötücül olmayan veri kayıtlarının da uyumluluk gösterdiğini vurgulamışlardır. Araştırmacılar tarafından çalışmalarında [10] BETH veri seti ile ilgili toplam veri sayıları belirtilmiştir. BETH veri setinin Şekil 1'deki dosyaları yukarıdan aşağı (1, 2, 3, …) olarak numaralandırılmıştır. Bu numara sırası ile csv uzantılı dosyalar içerisindeki metinler sütunlara dönüştürülerek sahip oldukları özellik isimleri, ilerleyen zamanlardaki siber güvenlik çalışmalarında kullanabilecek olan araştırmacılar için referans olması amacıyla Tablo 5 olarak sunulmuştur. Tablo 5 değerlendirildiğinde veri seti toplamda 28 tekil özellik içermektedir.

Tablo 5. BETH Veri Setinin özellik çıkarım tablosu, Şekil 1 sıralı

1, 7, 8	**2, 3, 4, 5, 6, 9**	**10, 11, 12, 13, 14, 15**
Timestamp, processId, threadId, parentProcessId, userId, mountNamespace, processName, hostName, eventId, eventName, stackAddresses, argsNum, returnValue, args, sus, evil	Timestamp, processId, parentProcessId, userId, processName, hostName, eventId, eventName, argsNum, returnValue, args, sus, evil	Timestamp, SourceIP, DestinationIP, DnsQuery, DnsAnswer, DnsAnswerTTL, DnsQueryNames, DnsQueryClass, DnsQueryType, NumberOfAnswers, DnsResponseCode, DnsOpCode, SensorId, sus, evil

BETH veri seti içerisindeki Şekil 1'deki veri seti dosyalarının yukarıdan aşağı sırasıyla içerdiği satır sayıları 763.146, 713.869, 485.243, 479.435, 409.933, 378.426, 188.969, 188.969, 199.224, 271, 271, 271, 271, 271 ve 271 olarak gözlemlenmiştir. Çıkarılan özellikler incelendiğinde "sus" özelliğinin değeri 1 ise işlemin şüpheli olduğunu ve "evil" özelliğinin değeri 1 ise işlemin saldırı olduğunu göstermektedir. 0 değeri tam tersi anlamındadır. Tablo 5'teki özellikler ile satır sayıları birleştirildiğinde ortaya çıkan milyonlarca kayıttan veri setinin büyük veri tanımına uyduğu görülmektedir.

4. BETH, KDD99, NSL-KDD ve ISCX Karşılaştırılması

Literatürde KDD99, NSL-KDD ve ISCX veri setlerinin karşılaştırılmasının bütünüyle ya da ikili gruplar halinde yapıldığı birçok çalışma mevcuttur. Bütünüyle karşılaştırmanın yapıldığı bir araştırma [28] sonuçları ile bu çalışmadaki incelenen BETH veri setinin özellik sayıları, saldırı tipleri ve olay sayıları açısından karşılaştırılması Tablo 6'da gösterilmiştir.

Tablo 6. BETH, KDD99, NSL-KDD ve ISCX karşılaştırması

Veri Seti	Özellik Sayısı	Olay Sayısı	Atak Tipleri
BETH	28	8,004,918	Kullanıcı tarafından oluşturulmuş gürültü veriden arındırılmış gerçek OS ve bulut trafik atakları [10]
KDD99	42	4,898,431	Dos, Probe, R2L, and U2R
NSL-KDD	41	125,973	Dos, Probe, R2L, and U2R
ISCX	14	2,545,935	HTTP denial of service (DoS), Brute force SSH, and Distributed Denial of service using an IRC botnet (DDoS)

Tablo 6 yorumlandığında BETH veri setinin güncel saldırı ve sızma girişimlerine ait ve kullanıcı verisinden arındırılmış saf atak verilerini içerdiği, bununla birlikte diğer veri setlerinden daha fazla olay sayısına sahip olduğu, KDD99 veri setinin tekrar eden verilerden temizlenerek NSL-KDD olarak sunulmasıyla olay sayısının yaklaşık %97,5 oranında azaldığı ve bunun KDD99 veri setinin oldukça yüksek gürültü ve tekrarlanan veri sayısına sahip olduğu anlamına geldiği görülmektedir.

5. BETH Veri Setinin KDD99 Veri Seti ile Karşılaştırmalı Görselleştirilmesi

Literatürde veri görselleştirmesinde kullanılan birçok araç mevcuttur. Bu araçlar genellikle verinin yapısallığına, hacmine ve akıp akmadığına göre uygunluk değerlendirilerek seçilmektedir. Power BI, Apache Spark ve Tableau [29] birbirlerinden farklı özelliklere sahip görselleştirme araçlarıdır. Akan verinin görselleştirilmesi açısından genellikle Spark aracı tavsiye edilmektedir. Microsoft'un ürünü olan Power BI ise kullanımı oldukça kolay olan ve akan veride kısıtlı özelliklerle de olsa API sayesinde kullanılabilen bir araçtır [31].

İncelenen bir araştırmada [30] KDD99 veri setinin siber güvenlik amaçlı olarak görselleştirilmesi konusu çalışılmıştır. Bahse konu çalışmada KDD99 veri setinde birçok fazlalık olduğu, veri setindeki fazlalıkların atılınca normal trafik yüzdesinin oldukça arttığı ve bu sorunu vurgulamak için yeni bir örnekleme modeli önerdiklerini vurgulamışlardır. Örnekleme metotlarında KDD99 veri setinden 10.000 kayıt alarak bunları sınıflara ayırdıkları, sınıflar içerisinden seçilen kayıtların HASH bilgilerinin alınarak kaydedildiği, HASH'i alınan kayıtlar temizlenerek bu algoritmanın tüm veriye uygulanmasıyla tekrarlayan veri ve gürültülü veriden arındırdıklarını belirmişlerdir. BETH veri seti ise araştırmacılara göre bu gürültü ve artıklıklara sahip değildir.

Araştırmada [30] insan gözüyle incelenebilecek ve değerlendirilebilecek veri sayısının görselleştirme çıktısında rahat görünebilmesi için en ideal olarak 800 örneklem veri olduğu sonucu çıkarılmıştır. Aynı çalışmada bu 800 örneklem veri Multi-Dimensional Scaling (MDS) ve Principal Component Analysis (PCA) yöntemleri kullanılarak görselleştirilmiş ve 800 örneklem kullanımının en ideal karar olacağı tespitlerine dayanak olarak gösterilmiştir.

Bu çalışmada BETH veri setinde test veri seti olarak sunulmuş olan lablled_testing_data.csv isimli veri seti içerisindeki eventid, sus, evil ve processName özelliklerinden sus ve evil değerleri sıfır olan 200 dpkg, sus ve evil değerleri 1 olan 300 sshd ve sus değeri 1 olup evil değeri 0 olan 300 systemd özelliği alınarak karşılaştırma yapılan çalışmadaki [30] şekilde 800 örnek alınmıştır. Bu veri seti test1.csv olarak adlandırılmıştır. Test1.csv veri setinde bulunan özelliklerin kendi içerisindeki bütün veri setine oranı sırasıyla %25 (sus ve evil=0), %37,5 (sus ve evil=1) ve % 37.5 (sus=1 ve evil=0) olarak belirlenmiştir. Çalışmada [30] KDD99 veri seti üzerinde PCA ve MDS kullanılarak görselleştirme yapılmış olması ile bu çalışmada PCA kullanılarak görselleştirme yapılmıştır.

PCA kullanılarak gerçekleştirilen görselleştirme algoritması Python dilinde oluşturulmuştur. Google Colabs yeteneklerinden faydalanılarak oluşturulan ipynb dosyası [32] ise açık erişim olarak sunulmuştur. Oluşturulan test1.csv veri seti algoritmada URL olarak okunmaktadır. Bu nedenle internete yüklenmiştir [33].

Algoritmanın devamı için başlangıçta pandas, numpy, matplotlib.pyplot, PCA (sklearn.decomposition'dan), StandardScaler (sklearn.preprocessing'den) kütüphaneleri yüklenmiştir. Pd.read fonksiyonu ile csv dosyası URL'den okunmuş ve başlıkları okutulduğunda Şekil 2'deki sonuç elde edilmiştir.

	eventid	sus	evil	processname
0	21	0	0	dpkg
1	1005	0	0	dpkg
2	257	0	0	dpkg
3	5	0	0	dpkg
4	3	0	0	dpkg

Şekil 2. Test1.csv veri seti ilk 5 satır

PCA, eksenler boyunca varyansı en üst düzeye çıkaran bir özellik alt uzayı sağladığından, özellikle farklı ölçeklerde ölçülmüşse, verileri standartlaştırmak mantıklıdır. Test1 veri setindeki sayısal veriler, birçok makine öğrenme algoritmasının optimal performansı için bir gereklilik olan verileri birim ölçeğe (ortalama = 0 ve varyans = 1) dönüştürülmüştür. StandardScaler kütüphanesi kullanılarak yapılan dönüştürme neticesinde Şekil 3'teki çıktı elde edilmiştir.

	eventid	sus	evil
0	-0.590166	-1.732051	-0.774597
1	1.975999	-1.732051	-0.774597
2	0.025297	-1.732051	-0.774597
3	-0.631892	-1.732051	-0.774597
4	-0.637108	-1.732051	-0.774597

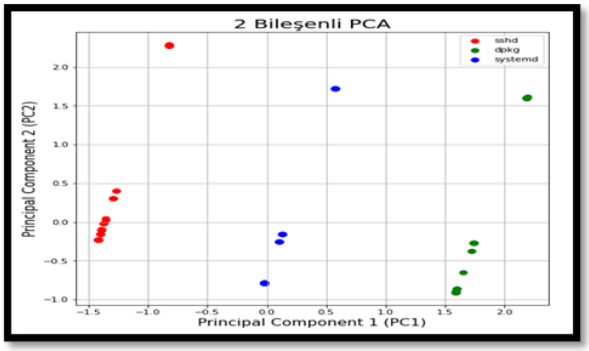
Şekil 3. Test1.csv veri seti standardizasyonu neticesinde okunan ilk 5 satırı

PCA işleminin iki boyutlu olarak sunumu için Principle Component 1 (PC1) ve Principle Component 2 (PC2) bileşenleri tanımlanmıştır. PC1 ve PC2 bileşenlerinin aldığı değerler iki boyutlu olarak sunumuna ait ilk 5 satırın görselleştirilmiş hali Şekil 4 olarak sunulmuştur.

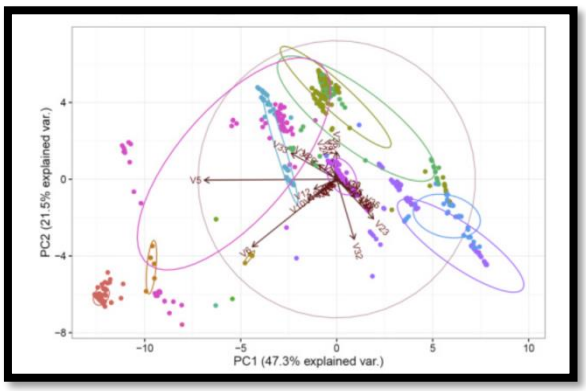
	principal component 1	principal component 2
0	1.602502	-0.866438
1	2.188025	1.600170
2	1.742932	-0.274853
3	1.592981	-0.906546
4	1.591791	-0.911559

Şekil 4. PC1 ve PC2 olarak PCA iki boyutlu sunum

PCA sunumunu 2 boyutlu olarak görselleştirirken farklı renkler ve şekiller kullanılmıştır. Buradaki amaç sınıfların birbirinden en iyi şekilde ayrılıp ayrılmadığının görülebilmesidir.Test1.csv veri setinin iki bileşenli PCA yapılarak görselleştirilmesi neticesinde elde edilen grafik Şekil 5 olarak sunulmuştur. Üç sınıfın da birbirinden iyi şekilde ayrıldığı net olarak görülmektedir.



Şekil 5. Test1.csv veri setinin PCA görselleştirilmesi

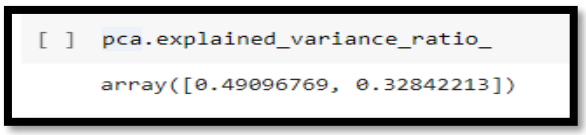


Şekil 6. Çalışmada [30] KDD99 veri setindeki PCA görselleştirmesi

Çalışmada [30] KDD99 veri setinden alınan 800 örnek veri üzerinden gerçekleştirilen PCA görselleştirmesi (Şekil 6) bu çalışmadaki BETH veri setinden alınan 800 örnek veri seti üzerinden gerçekleştirilen PCA görselleştirmesi (Şekil 5) karşılaştırıldığında PCA çalışmasından beklendiği gibi bu çalışmada elde edilen görsel bir insan tarafından net bir şekilde okunabilecek halde olup veri seti özellikleri ayrımları net bir şekilde görülebilmektedir.

Açıklanan varyans, temel bileşenlerin her birine ne kadar bilgi (varyans) atfedilebileceğini söyler. Diğer çalışmada [30] ilk temel bileşen (PC1), varyasyonun %47,3'ünü temsil etmektedir ve ikincisi

(PC2) varyasyonun yaklaşık %70'inde toplanan, varyasyonun %21,5'ini temsil etmektedir. Bu çalışmada ise ilk iki temel bileşen birlikte bilgilerin %81,9'unu içerir (Şekil7). Birinci temel bileşen varyansın %49,09'unu, ikinci temel bileşen ise varyansın %32,84'ünü içermektedir. Üçüncü ve dördüncü ana bileşen, veri kümesinin geri kalan varyansını içermektedir.



Şekil 7. Bu çalışmadaki PCA açıklamalı varyans oranları

Siber güvenlik loglarının oluşturduğu büyük verinin artık verilerden temizlenerek elde edilecek sonuç veri üzerinde bu çalışmada değinilen UMAP ve uygulanan PCA yöntemlerinin yanında Linear Discriminant Analysis (LDA) gibi yöntemlerin uygulanması da benzer sonuçlar verecektir. LDA, gruplar arasındaki ayrılabilirliği maksimize eden bir özellik alt uzayı bulmaya odaklanır. LDA PCA'den farklı olarak veri görselleştirmesini denetimli olarak yapmaktadır. Çalışmamızda başka bir akademik çalışmadaki uygulamanın güncel bir veri seti ile test edilmesi ve UMAP ve LDA gibi yöntemlerin karşılaştırma yapılan çalışmada bulunmaması nedeniyle sadece PCA yöntemi uygulanmıştır.

6. Sonuç

PCA'deki verileri ölçeklendirme zorunluluğu, en büyük varyansa sahip yönlerin en ilgi çekici değer olduğunun var sayılması, sadece orijinal değişkenlerin ortogonal dönüşümlerini dikkate alması, yalnızca ortalama vektör ve kovaryans matrisini temel alması gibi kısıtlamaları bulunmasına rağmen artık veriden arındırılmış büyük veri üzerinden veri seti uygun olduğu durumlarda kullanıcı tarafından net okunabilecek görselleştirme gerçekleştirebilmektedir.

Siber güvenlik amaçlı olarak büyük verinin görselleştirilmesi hakkındaki çalışmada [30] KDD99 veri setinden alınan 800 özelliğin PCA ile görselleştirmesi sonucunda siber saldırıların tespit edilmesi için kullanıcı (insan) tarafından net şekilde rahatça okunabilecek bir grafik elde edildiği belirtilmiştir. Bu çalışmada 2021 yılında oluşturulduğu görülen BETH veri seti üzerinde aynı çalışma uygulandığında PCA grafiğinin çok daha net ayrıldığı ve daha rahat okunabildiği tespit edilmiştir. Çalışma neticesinde siber güvenliğin sağlanması için gerçekleştirilen büyük veri görselleştirilmesi çalışmalarının, yeni saldırı türlerini içeren yeni veri setleri kullanılarak gerçekleştirilebileceği ortaya konmuştur. Ayrıca eski tarihli olan KDD99, ISCX gibi veri setlerinin sırf ağ loglarından oluşmasından dolayı oldukça fazla artık ve tekrar eden veri sayısına sahip olmaları araştırmacıları zorlayan bir durumdur. Bu özellik büyük veride artık verinin ve tekrar eden verilerin temizlenmesi çalışmalarında kullanılabilir ancak zamanın kritik olduğu siber güvenlik amaçlı saldırı ve girişimlerin tespiti açısından olumsuz bir durumdur. Bu nedenle siber güvenlik görselleştirmesi için kullanılacak veri setlerinin, girişim ve saldırı türüne göre etkilediği işlem ve zararlı olup olmadığı bilgileri sadeliğinde tutulması ayrıca faydalı olacaktır ve kayıt tutan sistemler için oldukça yük hafiflemesine sebep olacaktır. Siber güvenlik hem bilgi güvenliği hem de kişisel veri güvenliği açısından çok önemli bir kavramdır. Bu çalışma siber saldırıların engellenmesi için siber güvenlik büyük verisinin görselleştirilmesinin önemini ortaya koyması ve bu alanda gerçekleştirilen çalışmalarda güncel tarihli veri setlerinin kullanılmasının önemi ortaya koyarak hem siber güvenliğe hem de akademik çalışmalara olumlu katkı sunmaktadır.

Çıkar Çatışması

Yazarlar, çıkar çatışması olmadığını beyan eder.

Kaynaklar

[1]. Demir, H., Güllü, A., Taş Dokusunun Yüzey Pürüzlülüğü ve Taşlama Kuvvetlerine Etkilerinin İncelenmesi, Gazi Üniv. Müh. Mim. Fak. Dergisi, 2008, 23 (1).

[2]. Wang, Y., Kung, L., Byrd, TA. Big data analytics: "Understanding its capabilities and potential benefits for healthcare organizations", Technological Forecasting and Social Change, Volume 126, P. 3-13, 2018.

[3]. Oussous, Ahmed, Benjelloun, Fatima-Zahra, Lahcen, Ayoub Ait, Belfkih, Samir, "Big data technologies: a survey", J. King Saud Univ.-Comput. Inf. Sci. 30 (4), 431–448, 2018.

[4]. Yaqoob, I., Ahmed, E., Gani, A., Mokhtar, S., Imran, M., Guizani, S., "Mobile adhoc cloud: a survey", Wireless Commun. Mobile Comput. 16 (16), 2572–2589, 2016.

[5]. Jianzheng Liu, Jie Li, Weifeng Li, Jiansheng Wu, "Rethinking big data: A review on the data quality and usage issues", ISPRS Journal of Photogrammetry and Remote Sensing, Volume 115, Pages 134-142, 2016.

[6]. Javaid, N. "Integration of context awareness in Internet of Agricultural Things", ICT Express, 2021.

[7]. Najada, H. A., Mahgoub, I. and Mohammed I., "Cyber Intrusion Prediction and Taxonomy System Using Deep Learning And Distributed Big Data Processing," 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 631-638, doi: 10.1109/SSCI.2018.8628685, 2018.

[8]. Johnsan, J., "Global number of web attacks blocked per day from 2015 to 2018(in 1,000s)", statista.com/statistics/494961/web-attacks-blocked-per-day-worldwide/, 23 Aralık, 2021.

[9]. Lynch, K., "How Big Data Aids Cybersecurity", 2019, https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2019/how-big-data-aids-cybersecurity, 23 Aralık, 2021.

[10]. Jagpreet, K., Ramkumar K. R, "The recent trends in cyber security: A review", Journal of King Saud University - Computer and Information Sciences, 2021.

[11]. Highnam, K., Arulkumaran, K., Hanif, Z.D., and Jennings, N.R, "BETH Dataset: Real Cybersecurity Data for Anomaly Detection Research", ICML In workshop on Uncertainty and Robustness in Deep Learning 2021 and Conference on Applied Machine Learning for Information Security (CAMLIS 2021), 2021.

[12]. Ibrahim, A., Targio, H., Ibrar, Y., Nor Badrul, A., Salimah, M., Abdullah, G., Samee Ullah, K., "The rise of "big data" on cloud computing: Review and open research issues", Information Systems, Volume 47, Pages 98-115, 2015.

[13]. Rin, H., Hironori, U., Asato, N., Keitaro, T., Yoshihisa, N., "Large scale log anomaly detection via spatial pooling", Cognitive Robotics, Volume 1, Pages 188-196, 2021.

[14]. Amir, F. T. and Aaron, G., "Unsupervised log message anomaly detection", ICT Express, Volume 6, Issue 3, Pages 229-237, 2020.

[15]. Tuncer, O., Ates, E., Zhang, Y., Turk, A., Brandt, J., Leung, V. J., and Coskun, A. K., "Diagnosing Performance Variations in HPC Applications Using Machine Learning". In: Kunkel J.M., Yokota R., Balaji P., Keyes D. (eds) High Performance Computing. ISC High Performance 2017. Lecture Notes in Computer Science, vol 10266. Springer, 2017.

[16]. Wang, B., Shi, Y., Cheng, G., Wang, R., Yang, Z., and Dong, B., "Log-Based Anomaly Detection with the Improved K-Nearest Neighbor", International Journal of Software Engineering and Knowledge Engineering, 2020.

[17]. Wittkopp, T., Wiesner, P., Scheinert, D. and Kao, O., "A Taxonomy of Anomalies in Log Data". In AIOPS workshop 2021 co-located with ICSOC 2021, 2021.

[18]. Oliner, A. and Stearley, J., "What Supercomputers Say: A Study of Five System Logs," 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07), pp. 575-584, 2007.

[19]. Muhammad, A., Sagheer, A., Khan, M.A., Areej, F., Muhammad Adnan, K. and Sang-Woong, L., "MapReduce Based Intelligent Model for Intrusion Detection Using Machine

Learning Technique", Journal of King Saud University - Computer and Information Sciences, 2021.
[20]. Lee, H. and Lee, S., "A Study on Security Event Detection in ESM Using Big Data and Deep Learning", International Journal of Internet, Broadcasting and Communication Vol.13 No.3 42-49, 2021.
[21]. Tommaso, Z., Andrea, C., Lorenzo, S. and Andrea, B., "On the educated selection of unsupervised algorithms via attacks and anomaly classes", Journal of Information Security and Applications, Volume 52, 2020.
[22]. Ma, Q., Huang, W., Jin Y. and Mao J., "Encrypted Traffic Classification Based on Traffic Reconstruction," 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 572-576, 2021.
[23]. Sahu, S. K., Mohapatra, D. P., Rout, J. K., Sahoo, K. S., & Luhach, A. K., "An ensemble-based scalable approach for intrusion detection using big data framework," Big Data, 9(4), 303-321, 2021.
[24]. Pande, S., Aditya K., and Deepak G., "Recommendations for DDOS Attack-Based Intrusion Detection System Through Data Analysis." Proceedings of Second Doctoral Symposium on Computational Intelligence. Springer, Singapore, 2022.
[25]. KDD-99 Veri Seti. The Fifth International Conference on Knowledge Discovery and Data Mining Konferansında Sunulmuştur, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html adresinden erişilmiştir, 23 Aralık, 2021.
[26]. Sinha, A. and Rastogi, S. and Kaur, G., "Mining Anomalies in Large ISCX Dataset Using Machine Learning Algorithms in KNIME (April 28, 2018)". Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018 held at Malaviya National Institute of Technology, Jaipur (India) on March 26-27, 2018.
[27]. Tavallaee, M., Bagheri, E., Lu, W. and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
[28]. BETH Veri Seti Erişim. https://www.kaggle.com/katehighnam/beth-dataset adresinden erişilmiştir. Erişim Tarihi: 23/12/2021
[29]. Ghurab, M., Al-gaphari, G., Alshami, F., Alshamy, R. & Othman, S., "A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System", 2021.
[30]. Shetty, SD. "Sentiment Analysis, Tweet Analysis and Visualization on Big Data Using Apache Spark and Hadoop". IOP Conf. Ser.: Mater. Sci. Eng., 2021.
[31]. Zichan, R., Yuantian, M., Lei, P., Nicholas, P., Jun, Z., "Visualization of big data security: a case study on the KDD99 cup data set", Digital Communications and Networks, Volume 3, Issue 4, Pages 250-259, 2017.
[32]. Microsoft. "Real-Time Streaming in Power BI", https://docs.microsoft.com/en-us/power-bi/connect-data/service-real-time-streaming, 24 Aralık, 2021.
[33]. BETH_Veri_Seti_Örneğinde_PCA_Görselleştirmesi.ipynb, https://colab.research.google.com/drive/1Ll8riSCBEUhWleWVEPcKMexBNmYZIMQQ?usp=sharin, 24 Aralık, 2021.
[34]. Test1.csv, https://www.dset.com.tr/wp-content/uploads/test1.csv, 24 Aralık, 2021.