All Sciences Proceedings

5th International Conference on Applied Engineering and Natural Sciences

July 10-12, 2023 : Konya, Turkey

All Sciences Proceedings
http://as-proceeding.com/

https://www.icaens.com/

Kurumsal Ağlara Yapılan Saldırılarda Açıklanabilir Yapay Zekânın Yeri

Özer Yaz[1] , Ahmet Ali Süzen[2*]

[1]Bilgisayar Mühendisliği Bölümü /Lisansüstü Eğitim Enstitüsü, Isparta Uygulamalı Bilimler Üniversitesi, Türkiye
[2*]Bilgisayar Mühendisliği Bölümü /Teknoloji Fakültesi, Isparta Uygulamalı Bilimler Üniversitesi, Türkiye

[*](ahmetsuzen@isparta.edu.tr)

Özet **–** Bilişim sistemlerinin her alanda kullanılması nedeniyle siber güvenlik tehditleri artmaktadır. Bilindiği üzere günümüzde bilgiye erişim ihtiyacı her an doğabilmektedir. Artan kişisel erişim ve İnternet kullanımı milyarlarca cihaza yayılmakta, zettabayt boyutunda verileri ortaya çıkarmakta ve siber güvenlik tehditleri oluşturmaktadır. Siber güvenlik tehditleri, kişi ve kurumların bilgi ve iletişim sistemlerindeki bilgi güvenliği açıklarını her geçen gün artırmaktadır. Bu durum sistem çalışmazlığına, finansal kayıplara ve siber güvenlik riskine neden olmaktadır. Yeni teknolojilerin siber savunma sistemlerine adapte edilmesi, tehditlerdeki yeni teknolojik artış göz önüne alındığında çok kritik olduğu görülmektedir. Bu kapsamda yapay zekâ uygulamaları, siber güvenlik alanında kullanıcı davranışlarını analiz ederek iyi ve kötü aktiviteleri ayırmak için kullanılabilmekte, birbirinden bağımsız gibi görünen saldırı göstergelerini yorumlayıp korelasyon kurallarına göre alarm üretebilmektedir. Ancak yapay zekâ sistemleri açıklanabilirlik konusunda yetersiz kalabilir ve sınırları olabilir. Açıklanabilir yapay zekâ (AYZ) burada devreye giriyor. AYZ sonuçların geliştirilmesine, anlaşılmasına ve yorumlanmasına yardımcı olur. Bu araştırma kapsamında kurumsal ağlara yapılan saldırılarda, saldırıların tespitinde başarı oranının artırılması, yeni saldırı tiplerinin tespit edilmesi ve saldırı tiplerinin açıklanması amacıyla açıklanabilir yapay zekâ sistemi oluşturma amaçlanmıştır.

Anahtar Kelimeler – Açıklanabilir Yapay Zekâ, Kurumsal Ağlar, Siber Saldırılar, Veri Sızıntısı

I. GİRİŞ

Bilgi teknolojisi geliştikçe bilgiye ulaşma olanakları artmakta bu da bilgi güvenliğinin sağlanmasını zorlaştırmaktadır. Bilişim sistemlerinin her alanda kullanılması nedeniyle siber güvenlik tehditleri artmaktadır. Bilindiği üzere günümüzde bilgiye erişim ihtiyacı her an doğabilmektedir. [1] Aklımıza takılan herhangi bir soruyu akıllı telefonlar kullanarak yanıtlamak mümkün olabilmektedir. Bu soruyu yanıtlamak için akıllı telefonlarda yüklü bulunan uygulamalar kullanılmakta ve bu uygulamalar birçok veriye erişim için izin istemekte bu izinler verilmektedir. Evimizde bulunan beyaz eşyalarda da ağ bağlantısının olduğu bir dönemde bulunuyoruz. Artan kişisel erişim ve İnternet kullanımı milyarlarca cihaza yayılmakta, zettabayt boyutunda verileri ortaya çıkarmakta ve siber güvenlik tehditleri oluşturmaktadır. Siber güvenlik tehditleri, kişi ve kurumların bilgi ve iletişim sistemlerindeki bilgi güvenliği açıklarını her geçen gün artırmaktadır. Bu durum sistem çalışmazlığına, finansal kayıplara ve siber güvenlik riskine neden olmaktadır. Kurumlar ve bireyler ancak gelişmiş risk yönetimi ve kapsamlı güvenlik stratejileri ile siber risklerin önüne geçebilir[2]. Siber güvenlik uzmanları yeni tehdit ve zayıf noktaları her zaman göz önünde bulundurmalı ve siber savaş alanındaki teknolojik gelişme ve ilerlemeler ile karmaşık tehditlere daha akılcıl çözümler üretilmelidir.

Yeni teknolojilerin siber savunma sistemlerine adapte edilmesi, tehditlerdeki yeni teknolojik artış göz önüne alındığında çok kritik olduğu görülmektedir. Bu kapsamda yapay zekâ uygulamaları, siber güvenlik alanında kullanıcı davranışlarını analiz ederek sınıflandırıp iyi ve kötü

aktiviteler arasında ayrım yapabilme, birbirinden bağımsız gibi görünen saldırı göstergelerini yorumlanarak korelasyon kurallarına göre alarm üretme gibi çeşitli kritik işlevlere sahip olabilmektedir. Özellikle siber güvenliğe odaklanan yapay zekâ uygulamaları, siber savunma ekiplerinin işini kolaylaştırması nedeniyle ilerleyen yıllarda daha da önem kazanacak güvenlik trendlerinden biri olarak sıralanıyor. Ancak yapay zekâ sistemleri, açıklanabilirlik konusunda yetersiz kalabilmekte ve sistemlerin sınırları bulunabilmektedir. Yapay zekânın çıkışı ve benimsenmesinde, doğru sonuçlar alındığı sürece, modelin neyi öngördüğü anlaşılamasa da sorun olmamaktaydı. Şimdilerde ise, odak noktası insan tarafından anlaşılabilir modellerin oluşturulmasıdır. Açıklanabilir yapay zekâ (AYZ) burada devreye giriyor. AYZ sonuçların geliştirilmesine, anlaşılmasına ve yorumlanmasına yardımcı olur. Bu çalışma kapsamında kurumsal ağlara yapılan saldırılarda, saldırıların tespitinde başarı oranının artırılması, yeni saldırı tiplerinin tespit edilmesi ve saldırı tiplerinin açıklanması amacıyla açıklanabilir yapay zekâ sistemi oluşturma amaçlanmıştır.

II. SİBER GÜVENLİK ALANINDA YAPAY ZEKÂ KULLANIMI LİTERATÜR TARAMASI

Aytan ve ark. tarafından yapılan çalışmada, saldırı tespit sistemleri ile ilgili uygulamalarda en yaygın kullanılan veri setlerinden biri olan "KDD Cup'99" veri seti kullanılarak, hizmet engelleme saldırılarını ve veri tarama saldırılarını tespit etmek için makine öğrenmesi algoritmaları test edilmiş olup Weka aracı ile yüzde doksan dokuz başarı sağlanmıştır[3].

Remzi ATAY ve ark.[4] tarafından yapılan çalışmada CSE-CIC-IDS2018 veri kümesi üzerinde saldırı tespiti amaçlanmıştır. Çalışmada Rastgele Orman, Evrişimsel Sinir Ağı ve Hafif Gradyan Artırma makine öğrenmesi yöntemleri iki olarak hibrit yöntem kullanılarak veri kümesi ele alınmıştır. 0.86 macro F-skoru ve %98 doğruluk oranı ile Evrişimsel Sinir Ağı + Rastgele Orman hibrit modelinin en iyi saldırı tespiti yapan model olduğu görülmüştür.

Serdar ASARKAYA ve ark.[5] tarafından yapılan çalışmanın amacı, DDoS saldırılarını tespit etmek için makine öğrenmesi yöntemlerini kullanarak saldırıları sınıflandırmaktır. Araştırmada seçilen veri setinin materyali optimize edilerek K-Yakınlık Alanları, Çok Katmanlı Perceptron, Destek Vektör Makinesi ve Random Forest sınıflandırıcı modelleri geliştirilmiştir. Değerlendirmede ROC eğrileri ve kesinlik, geri çağırma, F1 puanı ve kesinlik ölçüleri kullanıldı ve çok katmanlı Perceptron modeli için yüksek düzeyde kesinlik elde edildi.

Mujibullah SHAMS[6] tarafından yapılan bu çalışmada, anormallik tespitinde iyi sonuçlar gösteren yedi denetimli makine öğrenimi algoritması seçildi. Bu algoritmalar JK-En yakın Komşu, AdaBoost, Multilayer Perceptron, 48, Random Forest, Support Vector Machines ve Naive Bayes'dir. Bu algoritmaların performansı, en son iki veri seti olan CIC-IDS-2017 ve CSE-CIC-IDS-2018 kullanılarak doğruluk, kesinlik, geri çağırma ve işleme süresi, F-ölçüsü açısından değerlendirildi. Özellik seçimi ve sınıflandırma yöntemlerinin rolünü değerlendirmek için iki tür sınıflandırma oluşturulmuştur. Sonuçlar J48, RF, KNN ve NB'nin başarılı sonuçlar elde edebildiğini ve en etkili sınıflandırıcılar olarak belirlendiğini göstermektedir.

III. VERİ SIZINTISI

Veri, kurumlar için ulusal güvenliği ilgilendiren bilgilerden, kimlik bilgilerine kadar geniş bir yelpazede yer alırken aynı zamanda rekabetin devamlılığını sağlamak için gerekli fikri mülkiyetlerden kuruluş içi gizli bilgilere kadar genişletilebilir[7]. Kişisel veri, kimliği belirli veya belirlenebilir gerçek kişiye ilişkin her türlü bilgiyi ifade etmektedir (KVKK, 2016). Hassas veri, sadece yetkili kişilerin erişebildiği ve kanuni dayanak olmadan ifşa edilmeye karşı korunan bilgilerdir. Özel nitelikli kişisel veriler ise ırk, etnik köken, din, mezhep, siyasi düşünce, sağlık vb. bilgileri içerir. Veri Sızıntısı, bahsi geçen bu verilerin kasıtlı olarak veya yapılan bir yanlışlıkla kurum dışına aktarılmasına, hukuka aykırı olarak, farklı yollardan yetkisiz kişilerin eline geçmesine denir.

IV. SİBER SALDIRI

Siber terimi, Sibernetik biliminden gelir ve 1958'de Louis Couffignal tarafından kurulan makine-canlı iletişim araştırmalarını temel alır.

Siber saldırı, kişilik haklarına yapılan saldırı türüdür. Dijital ortamda şahsi bilgilere saldırı yaparak normal hayattaki taciz, tehdit ve şantaj gibi durumları gerçekleştirir. Sanal ortamdaki saldırılar, çocuklar ve gençler için daha riskli durumdadır. Siber saldırı, bilgisayarlar veya ağlar aracılığıyla veri kopyalama, değiştirme ve silme gibi kötü amaçlı işlemler için yapılan bilinçli eylemlerdir. İnternet sitelerine erişimi engellemek ve yavaşlatmak gibi amaçları da olabilir.

Siber tehditler, gelişen bilgi sistemleriyle birlikte hem devlet hem de devlet dışı aktörler için yeni tehditler oluşturuyor. Bu tehditler soyut alandan geliyor ve tespit edilebilirlikleri düşük olduğu için sonuçları öngörülemez olabiliyor. Ayrıca tehditlerin merkezi bir yapıya sahip olmaması da belirsizliği arttırıyor. Tehdit kaynağı bireyler, gruplar, terör örgütleri veya devletler olabilir. Siber saldırılar devletlerarası boyuta ulaştığında "Siber Savaş" olarak adlandırılır. Devlet kurumlarına saldırarak iş göremez hale getirme ya da gizli bilgi sızdırma amaçlarına odaklanılır. Siber saldırılar büyük kayıplara neden olabilir.

2020 sonrasında Covid-19 salgınının da etkisiyle iş modellerinde değişim gerçekleşmiş ve insanların uzaktan çalışmaya geçmesi hızlanmıştır. Fakat günümüzde daha yaygın hâle gelen uzaktan çalışma yönteminin siber saldırı sayısında ciddi bir artışa neden olduğunu farklı ülkelerde yapılan araştırma raporları ortaya koyuyor. Araştırmalar herhangi bir siber tehdit karşısında kuruluşların geçmişe nazaran daha fazla olumsuz etkilendiğini gösteriyor.

A. Kurumsal Ağlarda En Çok Karşılaşılan Siber Saldırı Türleri

1) Kötü Amaçlı Yazılımlar

Zararlı yazılımlar internet kullanıcısına zarar verirken saldırgana da maddi kazanç sağlayabiliyor. Kötü amaçlı yazılımlar kişisel bilgileri çalıp tehdit olarak kullanabilir, büyük ölçekli şirketlerden devlete bağlı kurumlara kadar ulusal ya da uluslararası öneme sahip verileri sızdırarak, silerek ya da şifreleyerek önemli zararlara ve bunun sonucunda haksız maddi kazanca neden olabilir [19].

2) Sosyal Mühendislik

Kişiler arası iletişim ve insan hareketlerindeki kalıpları zayıflık olarak kabul edip bunlardan faydalanarak güvenlik aşamalarını atlatma yöntemine dayalı müdahaleleri içerir. Karşı tarafın güvenilir bir kaynak olduğuna inanmak, Hedef sistemin atıklarını karıştırmak, Ortak tanıdıklar aracılığıyla yakınlık kurmak, Başkasının kimliğine bürünmek, Gizlice zor bir durum yaratmak ve yardım ediyor görüntüsü vermek en sık kullanılan sosyal mühendislik yöntemleridir [20].

3) Dağıtık Hizmet Engelleme (Distributed Denial of Service- DDoS)

DDoS, sistemleri belirli kapasite sınırlarının üstünde veriye maruz tutarak kullanıcıların sisteme veya siteye girişini engelleyen saldırıdır. DoS türü ise yalnızca tek bir kaynaktan hedefe saldırı yapılmasıdır. DDoS, çok sayıda kaynaktan tek hedefe doğru yapılarak şiddetini artırır. Sistem kurulurken kullanıcı sayıları, hat kapasitesi, istek sayısı gibi unsurlar için değerler öngörülür ve bu değerlerin üstündeki yükü kaldırabilecek şekilde tasarım yapılır. DDoS, sistem yükünü çok üzerinde kullanıcı sayısı veya istek ile yorarak sistemi cevap veremez hale getirerek veya hattı doldurarak erişilebilirliği engelleyen bir saldırı türüdür. DDoS saldırıları kolaylıkla yapılabilir ancak kullanım amacına ve stratejik plana göre daha karmaşık düzeyde de olabilir [18].

4) Fidye Saldırıları

Fidye yazılımı, dosyaları şifreleyerek para talep eden kötü amaçlı yazılımlardır. Siber saldırganlar, bu yöntemle dijital şantaj yaparak para kazanmaktadır. Fidye yazılımı verileri yok etmek yerine şifreler ve bu nedenle kurtarma seçenekleri sınırlıdır. Fidye yazılımına yakalanma durumunda yedek yoksa verileri kurtarmanın tek yolu fidyeyi ödemektir. Ancak, işletmeler talep edilen fidyeyi ödese bile, siber saldırganlar anahtarı bazen göndermezler [21].

5) Flooding

Bilgisayar ağları, iletişim protokolleri veya diğer bilgi işlem sistemleri açısından, flooding genellikle aşırı yüklenmeye, hizmet kesintilerine veya kaynak tükenmesine neden olan bir saldırı türünü ifade eder.[8] Flooding saldırıları, ağa veya

sistem kaynaklarına sürekli olarak büyük miktarda trafiğin gönderilmesiyle gerçekleştirilir. Bu trafiğin yoğunluğu, ağın veya sistem kaynaklarının normal işlevselliğini bozacak düzeyde olabilir. Saldırganlar, ağdaki iletişimi engellemek, kaynakları tüketmek, hizmetleri kesintiye uğratmak veya sistemleri çökertmek gibi amaçlarla flooding saldırılarını gerçekleştirebilir. Flooding saldırıları farklı şekillerde gerçekleştirilebilir. Örneğin, ağ katmanında yapılan bir saldırıda, ağa çok sayıda paket gönderilerek ağ bant genişliği tüketilir. ICMP (Internet Control Message Protocol) flooding, SYN flooding, UDP (User Datagram Protocol) flooding gibi saldırı teknikleri de başlıca flooding örneklerindendir [17].

ICMP, IP ağ protokolü üzerinden iletişim kurmaya yardımcı olan bir iletişim protokolüdür. ICMP flooding saldırısı, ağa büyük miktarda ICMP trafiği göndererek ağ kaynaklarını tüketmeyi amaçlar. Bu saldırı türünde, saldırganlar genellikle ICMP Echo Request (ping) mesajları gönderirler. Bu mesajlar normalde ağdaki cihazların erişilebilirlik testlerinde kullanılır. Ancak, saldırganlar büyük miktarda ICMP Echo Request mesajı göndererek ağa yoğun bir trafiğin yığılmasına neden olurlar. Bu, ağ cihazlarının kaynaklarını tüketir ve ağ performansını olumsuz etkiler.[8]

UDP, IP tabanlı ağlarda veri iletimini sağlayan bir iletişim protokolüdür. UDP flooding saldırısı, saldırganların hedef ağa veya sistemlere yoğun UDP trafiği göndererek ağ kaynaklarını tüketmeyi amaçlar. Bu tür bir saldırıda, saldırganlar genellikle rastgele kaynak IP adreslerinden UDP paketleri oluşturarak hedef ağa gönderirler. Saldırganlar, hedefe doğrudan veya yansıma saldırısı olarak bilinen bir yöntemle, başka sistemlerin yanıtlarını hedefe ileterek saldırı yoğunluğunu artırabilirler. Bu süreçte, hedef ağ veya sistemlerde aşırı yüklenme, kaynak tükenmesi veya hizmet kesintileri meydana gelebilir.

DNS Flood, bir tür hizmet reddi saldırısıdır. Bir ağ kaynağı veya makinedeki trafiğin bir süre durdurulması işlemidir. Suçlu, kaynağa veya makineye çok sayıda istek gönderir, böylece kaynağa ulaşmaya çalışanlar tarafından kullanılamaz hale gelir. Bir DNS Flood saldırısında, saldırgan, belirli bir DNS sunucusunu görünürde geçerli bir trafikle ve çok fazla sunucu kaynaklarıyla sunucuların meşru istekleri bölge kaynaklarına yönlendirme yeteneğini engellemeye çalışır.

MAC flooding, ağdaki Ethernet switchlerine yönelik bir saldırı türüdür. Ethernet switchleri, ağdaki cihazları birbirine bağlayan ve veri iletimini yönlendiren cihazlardır. Her switchin bir MAC tablosu bulunur ve bu tablo, hangi MAC adresinin hangi port üzerinden ulaşabileceğini belirler. Normalde, switchler bu tabloyu öğrenerek günceller ve veri paketlerini doğru portlara yönlendirirler. MAC flooding saldırısında, saldırganlar ağdaki switchleri yanıltmak amacıyla çok sayıda sahte MAC adresi ve bağlantı talepleri gönderirler. Switch, bu sahte MAC adreslerini tablosuna eklemeye çalışırken aşırı yüklenir ve sonunda bu tablo dolup taşar. Bu durumda switch, MAC tablosunu doldurduğu için veri paketlerini yönlendiremez ve normal ağ trafiği engellenir.[9]

V. YAPAY ZEKÂ

Yapay Zekâ; insan benzeri düşünme ve öğrenme yeteneklerine sahip bilgisayar sistemlerinin tasarımı ve geliştirilmesiyle ilgilenen bir disiplindir. Yapay zekâ, karmaşık problemleri çözebilen, desenleri tanıyabilen, verileri analiz edebilen ve kararlar verebilen bir bilgisayar sistemi oluşturma amacını taşır. Bu sistemler, genellikle algoritma ve veri tabanlı yöntemler kullanarak büyük miktarda veriyi işleyebilir ve bu verilerden bilgi çıkarabilir [10]. Yapay zekâ, çeşitli uygulama alanlarında kullanılmaktadır, örneğin otomasyon, sağlık hizmetleri, finans, oyunlar, görüntü işleme, siber güvenlik ve robotik gibi birçok alanda kullanılmaktadır. Yapay zekânın başlıca özellikleri arasında öğrenme, tahmin etme, karar verme, ses ve görüntü tanıma gibi yetenekler bulunur. Makine öğrenmesi ve derin öğrenme gibi teknikler kullanarak, yapay zekâ sistemleri kendi kendine gelişebilir ve optimize edilebilir hale gelebilir. Yapay zekâ, teknolojik ilerlemelerle birlikte hızla gelişen ve büyük bir potansiyele sahip olan bir alandır [11].

Yapay zekâ, siber güvenlik alanında çeşitli şekillerde kullanılabilir ve güvenlik uzmanlarına önemli bir yardımcı olabilir. Yapay zekâ algoritmaları, ağ trafiği analizi ve davranışsal modelleme gibi teknikler kullanarak, anormal aktiviteleri veya potansiyel tehditleri tespit etmek için kullanılabilir. Yapay zekâ tabanlı sistemler, normal davranış modellerini öğrenerek, saldırıları tespit etmek için anormallikleri belirleyebilir. Güvenlik duvarları ve saldırı önleme sistemleri gibi savunma mekanizmalarını güçlendirmek için kullanılabilir. Yapay zekâ, uygulama ve sistemlerde güvenlik açıklarını tespit etmek için kullanılabilir. Zayıf noktaları ve açıkları otomatik olarak tarayabilir ve bu bilgileri güvenlik uzmanlarına raporlayabilir. Geçmiş saldırılar ve siber saldırılarla ilgili verileri analiz ederek saldırganların yöntemlerini ve davranışlarını öğrenebilir. Bu bilgiler, savunma sistemlerini güncellemek ve gelecekteki saldırıları tespit etmek için kullanılabilir [12].

VI. AÇIKLANABİLİR YAPAY ZEKÂ

Açıklanabilir yapay zekâ, normal yapay zekâdan farklı olarak karar süreçlerini ve sonuçlarını insanlar tarafından anlaşılabilir ve izlenebilir bir şekilde açıklayabilen bir yapay zekâ türüdür. Normal yapay zekâ modelleri genellikle karmaşık yapıları ve içsel işleyişleri nedeniyle kararlarını verirken kullanılan faktörleri net bir şekilde açıklayamazlar. Açıklanabilir yapay zekâ, karar verme sürecinin adımlarını ve sonuçlarını insanlara anlatma veya gösterme kabiliyetine sahiptir. Bu kararların nedenlerini, veriye dayalı kanıtları ve kullanılan algoritma veya modelin nasıl çalıştığını açıklayabilmeleri anlamına gelir. Bu sayede insanlar, yapay zekânın nasıl kararlar verdiğini anlayabilir, doğruluğunu değerlendirebilir. Aynı şekilde gerekirse kararları sorgulayabilir [14].

Açıklanabilir yapay zekâ, özellikle etik, hukuk ve güvenlik gibi alanlarda önemli bir rol oynar. Kararlarını anlamak ve izlemek insanların güvenini artırırken, yanlış veya hatalı kararlar üzerinde denetim sağlar. Ayrıca, açıklanabilirlik, adalet, ayrımcılık ve yanlılık gibi endişeleri ele almayı ve yapay zekânın etkisini izlemeyi kolaylaştırır.

A. Açıklanabilir Yapay Zekânın Avantajları

Güvenilirlik: Açıklanabilir yapay zekâ, karar süreçlerini ve sonuçlarını insanlar tarafından anlaşılabilir bir şekilde açıklayabildiği için güvenilirlik sağlar. Bu, insanların yapay zekânın neden belirli kararlar verdiğini anlamalarını ve doğruluğunu değerlendirmelerini sağlar. Ayrıca, hatalı veya yanlış kararları tespit etmek ve düzeltmek için denetim mekanizmalarının oluşturulmasını kolaylaştırır [15].

Şeffaflık: Açıklanabilir yapay zekâ, karar verme sürecinin adımlarını ve kullanılan faktörleri anlatarak şeffaflık sağlar. Bu, yapay zekânın nasıl çalıştığını anlamak isteyen insanlar için önemlidir. Kararların şeffaf olması, adalet, etik ve ayrımcılık gibi endişelerin ele alınmasına yardımcı olur [16].

Sorumluluk: Açıklanabilir yapay zekâ, yapay zekânın kararlarının ardındaki sorumluluğu belirlemeyi kolaylaştırır. Kararların nedenleri ve veriye dayalı kanıtları açıklanabilir olduğu için, hatalı veya zararlı kararlar durumunda sorumluları belirlemek ve gerekli önlemleri almak daha kolay olabilir[15].

Eğitim ve İyileştirme: Açıklanabilir yapay zekâ, insanların yapay zekâyı eğitmek ve iyileştirmek için daha fazla bilgiye sahip olmalarını sağlar. Açıklanabilirlik, yapay zekânın eksikliklerini ve hatalarını anlamak ve düzeltmek için geri bildirim sağlamayı kolaylaştırır [16].

Hukuki Uyumluluk: Bazı durumlarda, yapay zekânın kararları hukuki veya düzenleyici gerekliliklere uyumlu olmalıdır. Açıklanabilir yapay zekâ, bu gereklilikleri karşılamayı daha kolay hale getirir. Kararların açıklanabilir olması, yapılan işlemlerin ve kullanılan verilerin hukuki standartlara uygun olduğunu doğrulamayı sağlar.

Kabul Edilebilirlik: İnsanların yapay zekâ teknolojilerini kabul etmeleri için güven ve anlayış önemlidir. Açıklanabilir yapay zekâ, insanlara yapay zekânın nasıl çalıştığını ve kararlarını nasıl verdiğini anlatarak bu kabul edilebilirlik sürecini kolaylaştırır.

Bu avantajlar, açıklanabilir yapay zekânın daha etkili, güvenilir ve insan odaklı bir şekilde kullanılmasını sağlar. Aynı zamanda, yapay zekâ teknolojilerinin toplum tarafından daha iyi anlaşılmasını ve benimsenmesini destekler.

B. Açıklanabilir Yapay Zekânın Siber Güvenlikte Kullanımı

Siber güvenlikte açıklanabilir yapay zekâ (AYZ) yöntemleri, yapay zekâ algoritmalarının karar süreçlerini daha anlaşılır ve izlenebilir hale getirmeyi hedefler. Özellikle siber saldırıları tespit etme, tehdit analizi yapma ve olay yanıtı süreçlerini destekleme gibi alanlarda kullanılabilir.

Açıklanabilir yapay zekâ algoritmaları, ağlardaki anormal aktiviteleri veya potansiyel tehditleri tespit etmek için kullanılabilir. Bu algoritmaların çalışma prensipleri açıklanabilir ve izlenebilir olduğunda, saldırgan faaliyetleri daha etkin bir şekilde tespit etmek ve yanlış pozitif veya yanlış negatif sonuçları azaltmak mümkün olur.

Saldırı analizi: Siber saldırıların analizinde kullanılabilir. Saldırganların kullanabileceği farklı teknikleri ve saldırı modellerini öğrenerek, saldırıları tanımlayabilir ve belirli bir saldırının niyetini veya etkilerini tahmin edebilir. Bu analizleri nasıl gerçekleştirdiği açıklanabilir bir şekilde sunulabilir.

Güvenlik açığı tespiti: Yazılım veya ağlarda güvenlik açıklarını tespit etmek için kullanılabilir. Güvenlik açıklarını belirlemek için AYZ algoritmaları, uygulamaların veya sistemlerin yapısını, kodunu veya yapılandırmasını analiz edebilir. AYZ'nin bu tespit süreci ve sonuçları açıklanabilir bir şekilde sunularak, güvenlik açıklarının daha hızlı ve etkin bir şekilde giderilmesi sağlanabilir.

Karar destek sistemi: Siber güvenlik operasyonlarında karar destek sistemleri olarak kullanılabilir. Örneğin, AYZ algoritmaları, saldırıların önceliğini belirlemek, zararlı trafikleri filtrelemek veya olaylara öncelik atamak gibi görevlerde kullanılabilir.

Açıklanabilir yapay zekâ, siber güvenlik alanında kullanılan yapay zekâ algoritmalarının güvenilirliğini artırır ve karar süreçlerini daha şeffaf hale getirir. Bu, siber güvenlik uzmanlarına, algoritmaların neden belirli kararları verdiğini anlama ve güvenlik önlemlerini geliştirme konusunda daha fazla kontrol ve anlayış sağlar.

VII. SONUÇ

Açıklanabilir Yapay Zekâ'nın siber saldırıları tespit etme, tehdit analizi yapma ve olay yanıtı süreçlerini destekleme gibi alanlarda, saldırganların kullanabileceği farklı teknikleri ve saldırı modellerini öğrenerek, saldırıları tanımlamada, saldırıların önceliğini belirlemek, zararlı trafikleri filtrelemek veya olaylara öncelik atamak gibi görevlerde kullanılabilir. Siber güvenlik tehditlerinin karmaşıklığı ve çeşitliliği göz önüne alındığında, açıklanabilirlik, güvenlik uzmanlarının karşılaştıkları zorlukları çözmelerine yardımcı olabilir.

KAYNAKÇA

[1] Yıldırım, E. Y. (2018). Bilişim Sistemlerine Yönelik Siber Saldırılar ve Siber Güvenliğin Sağlanması., (s. 24-33).

[2] Şeker, E. (2015). Yapay Zekâ Tekniklerinin / Uygulamalarının Siber. *Uluslararası Bilgi Güvenliği Mühendisliği Dergisi*, 108-115.

[3] Aytan, B. N. B. (2018). Siber Savunma Alanında Yapay Zekâ Tabanlı Saldırı Tespiti ve Analizi. Samsun.

[4] Atay,R. D. E. (2019). İki Seviyeli Hibrit Makine Öğrenmesi Yöntemi ile Saldırı Tespiti . *Gazi Mühendislik Bilimleri Dergisi* , 258-272.

[5] Asarkaya, S. O. K. (2021). Ddos Saldırılarının Makine Öğrenimi Algoritmalarıyla Tespiti. *Tasarım Mimarlık ve Mühendislik Dergisi* , 221-232.

[6] Shams, M. (2020). Ağ Anomalisi Tespitinde Makine Öğrenmesi Algoritmalarının Kullanımı Ve Karşılaştırmalı Analizi.

[7] Burak Oğuz, H. K. (2010). BT Yönetiminde Bilgi Sızıntısı ve Ağ Tabanlı Çoklu Protokol.

[8] Tandoğan, E. (2020). *FLOOD SALDIRILARI*. https://www.siberdinc.com/siber/flood-saldirilari/.html adresinden alındı

[9] Çelik, S. (2021). *MAC FLOODİNG SALDIRISI NEDİR?* https://www.siberguvenlik.web.tr/index.php/2021/01/13/mac-flooding-saldirisi-ve-analizi/ adresinden alındı

[10] Winston, P. H. (1984). Artificial intelligence. Addison-Wesley Longman Publishing Co., Inc..

[11] Ramesh, A. N., Kambhampati, C., Monson, J. R., & Drew, P. J. (2004). Artificial intelligence in medicine. Annals of the Royal College of Surgeons of England, 86(5), 334.

[12] Li, J. H. (2018). Cyber security meets artificial intelligence: a survey. Frontiers of Information Technology & Electronic Engineering, 19(12), 1462-1474.

[13]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence

[14]. (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

[15]. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

[16]. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), 4793-4813.

[17]. Rizal, R., Riadi, I., & Prayudi, Y. (2018). Network forensics for detecting flooding attack on internet of things (IoT) device. Int. J. Cyber-Security Digit. Forensics, 7(4), 382-390.

[18]. Chadd, A. (2018). DDoS attacks: past, present and future. Network Security, 2018(7), 13-15.

[19]. Kumar, S. (2020). An emerging threat Fileless malware: a survey and research challenges. Cybersecurity, 3(1), 1-12.

[20]. Ansari, M. F., Sharma, P. K., & Dash, B. (2022). Prevention of phishing attacks using AI-based Cybersecurity Awareness Training. Prevention.

[21]. Kansagra, D., Kumhar, M., & Jha, D. (2016). Ransomware: a threat to cyber security. CS Journals, 7(1).