BİLİŞİM DÜNYASINDA VE SİBER GÜVENLİK ALANINDA AÇIKLANABİLİR YAPAY ZEKA (XAI) KULLANIMI

Dr. Esra SÖĞÜT
ORCID: 0000-0002-0051-2271
esrasogut@gazi.edu.tr
Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye

Özet

Bilişim dünyasında ilgilenilen alanların genişlemesi ve kullanılan yöntemlerin veya araçların çeşitlilik kazanması sonucunda her an yeni gelişmeler yaşanabilmektedir. Bilişim sistemleri ile alakalı yapılan çalışmaların ve hazırlanan ürünlerin sayısı ve çeşitliliği de arttıkça bilişim dünyası kaçınılmaz şekilde siber âlemin tehlikeleriyle karşı karşıya gelmektedir. Sistemlerde en önemli değer olan veri anonim şekilde, kişisel, kurumsal veya ülkeye ait özel veri şeklinde de olabilir. Üretilen, kullanılan ve düzenlenen veri çeşitliliği ve büyüklüğü düşüldüğünde siber güvenliğin sağlanmasının kolay olmadığı; mutlaka güvenlik anlamında özel çalışmaların yapılması gerektiği anlaşılmaktadır. Ayrıca kullanılan araçların veya sistemlerin de kendilerine özgü güvenlik açıklıkları da olabilmekte ve kötü niyetli kişiler tarafından bunlar sömürülebilmektedir. Bu sebeplerden dolayı bilişim sistemlerinde siber güvenliğin sağlanması zorunlu hale gelmiştir. Siber güvenlik alanında da gelişmeler yaşanmakta ve daha verimli yöntemler üzerinde çalışmalar yapılmaktadır. Geleneksel makine öğrenmesi, veri madenciliği, kural tabanlı gibi sistemlerin kullanılmasına ek olarak yapay zekâ sistemleri veya yöntemleri de kullanılmaya başlanmıştır. Yapay zeka sistemlerinin yeteneklerinin geliştirilerek sistemin ve elde edilen sonuçların yorumlanmasını kolaylaştırmak için Açıklanabilir Yapay Zeka (XAI: Explainable Artificial Intelligence-AYZ) kavramı ortaya konulmuştur. Bir saldırı tespit sistemi hazırlamak için AYZ kullanımı ile elde edilen sonuçların; geleneksel yöntemlere göre elde edilen sonuçlardan daha anlaşılabilir ve geliştirilebilir duruma getirilmesi sağlanacaktır. Ayrıca, elde edilen analiz sonuçlarının yorumlanabilmesine de imkân sunulmaktadır. Bu çalışma kapsamında bilişim dünyasının siber açıklıklarına dikkat çekilmesi, siber saldırılar hakkında bilgi verilmesi, siber saldırıların tespit edilmesi için kullanılan yöntemlerin açıklanması ve AYZ kullanımının bu sistemlere sunabileceği katkıların anlatılması amaçlanmıştır. Bilişim dünyasına ve siber güvenlik alanına AYZ kullanımı için yol gösterici olması hedeflenmiştir.

Anahtar Kelimeler: Açıklanabilir Yapay Zeka, AYZ, siber güvenlik, saldırılar, tespit sistemleri

USE OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN THE WORLD OF INFORMATICS AND CYBER SECURITY

Abstract

In the world of informatics, new developments can occur at any time due to the expansion of the fields of interest and the diversification of the methods or tools used. As the number and diversity of studies and prepared products related to information systems increase, the informatics world inevitably faces the dangers of the cyber world. Data, the most critical value in the systems, can be anonymous, personal, corporate, or country-specific data. Considering the variety and size of data produced, used, and organized, it takes work to ensure cyber security. Therefore, special studies must be carried out in terms of security. In addition, the tools or systems used may have security vulnerabilities, and malicious people can exploit them. For these reasons, ensuring cyber security in information systems has become mandatory. There are also developments in cyber security, and studies are being carried out on more efficient methods. In addition to using traditional systems such as machine learning, data mining, and rule-based, artificial intelligence systems or methods have also begun to be used. The concept of Explainable Artificial Intelligence (XAI) has been introduced to facilitate the development of the capabilities of artificial intelligence systems, the interpretation of the system, and the interpretation of the results obtained. The results obtained by using XAI to prepare an attack detection system will be made more understandable and improvable than the results obtained according to traditional methods. In addition, it is possible to interpret the analysis results obtained. Within the scope of this study, it is aimed to draw attention to the cyber vulnerabilities of the IT world, provide information about cyber attacks, explain the methods used to detect cyber attacks and explain the contributions that the use of XAI can make to these systems. It aims to be a guide for using XAI in the IT world and cyber security field.

Keywords: Explainable Artificial Intelligence, XAI, cyber security, attacks, detection systems

1-GİRİŞ

Verinin kullanıldığı, işlendiği ve kaydedildiği alanlar giderek artmakta ve çeşitlilik kazanmaktadır. Buna ek olarak, değerlendirilen verinin türü, büyüklüğü, içeriği veya önemi de farklılık göstermektedir. Bilişim dünyasında ele alınan her veri değerlidir fakat anonim olmayan, önemli görülen veya gizli olan verinin korunması zorunlu bir durumdur. Bu durum da siber güvenlik çalışmalarının ilgi alanına girmektedir.

Siber güvenlik dünyasında; güvenliğin sağlanması verinin gizliliği, bütünlüğü ve erişebilirliği ana temellerine dayanmaktadır. Bunlara ek olarak süreklilik, yetkilendirme, kayıt altına alma ve sınama gibi güvenlik unsurları da yer almaktadır. Evlerde kişisel eşyalar tarafından üretilen veri için de siber güvenlik sağlanmalıdır. Büyük fabrikalarda üretilen veya kurumlarda kullanılan veri için de güvenlik sağlanmalıdır. Kullanılan cihaz sayısının artması, internet kullanımının yaygınlaşması ve farklı uygulamaların veya yazılımların kullanılması hayatı kolaylaştırmakta ve işlemlerin yapılmasını hızlandırmaktadır. Üretilen, kullanılan ve kayıt altına alınan verinin türü ve boyutu da farklılık göstermektedir. Dolayısıyla tek bir siber güvenlik çözümü tüm sistemler için uygulanabilir olamaz [1]. Akıllı telefondaki bir uygulamada sorun olması çok büyük problemlere yol açmazken; belediyeye ait bir kurumda gerçekleştirilen işlemlerin sürekliliğinin bozulması büyük sorunlara yol açabilir. Örneğin emlak işlemleri ile ilgilenen bir kurumun siber saldırıya uğraması sonucunda ilgili veriye ulaşılamaması veya ilgili verinin yetkisiz kişilerce değiştirilmesi gibi sorunlar yerel halkı tedirgin edecek ve yapılmak istenen işlemleri çıkmaza sokacak bir durumdur. Hatta emlak işlemleriyle bağlantılı olan diğer işlemleri de aksatacak ve böylece birçok sistem kullanılamaz hale getirilecektir. Farklı bir örnek de kritik sistemlerin saldırıya maruz kalması üzerine olabilir. Örneğin su dağıtım sistemlerine yönelik meydana gelebilecek saldırılar yüzünden halkın temiz suya erişmesi günlerce engellenebilir veya şebeke suyundaki kimyasal maddelerin kullanım oranları değiştirilerek halkın sağlığı tehlikeye atılabilir [2].

Siber saldırılar geliştikçe saldırılarla mücadele etme konusunun da güncellenmesi zorunlu hale gelmiştir. Yeni ve gelişmiş teknikler ile hazırlanan saldırılar, geleneksel saldırı tespit sistemlerini kolay ve hızlı şekilde atlatabilmektedir. Bu yüzden saldırı tespit sistemlerinde güncel, gelişmiş ve yenilikçi yaklaşımlar kullanılmalıdır. Geliştirilen çözümler arasında imza, kural, veri madenciliği, makine öğrenmesi, derin öğrenme ve yapay zeka tabanlı gibi çeşitli saldırı tespit çalışmaları yapılmaktadır. Son zamanlarda yapay zeka alanında yapılan çalışmalar ivme kazanmıştır.

Bilişim dünyasında ve siber güvenlik alanında yapay zeka kullanımı ile birçok probleme çözüm üretilmesi amaçlanmaktadır. Karmaşık, sistematik ve kapsamlı siber saldırıların tespit edilmesi için özellikle yapay zekanın kullanılması sistemlerin güvenliğinin sağlanması için önemli katkılar sunmaktadır. Sunulan katkıların daha anlaşılabilir ve geliştirilebilir duruma getirilmesi ihtiyacı duyulmuş ve bunun sonucunda Açıklanabilir Yapay Zeka (AYZ - Explainable Artificial Intelligence - XAI) kavramı ortaya çıkmıştır [3]. Saldırıların tespit edilmesinde AYZ kullanımı; elde edilen analiz sonuçlarının daha açık şekilde yorumlanması ve geleneksel yöntemlere göre daha anlamlı çıkarımlar elde edilmesi gibi avantajlar sunmaktadır.

Bu çalışma kapsamında bilişim dünyasında meydana gelen siber saldırılara, var olan güvenlik açıklıklarına, siber saldırı tespit yöntemlerine ve bu alanda AYZ kullanımına yer verilmektedir. Siber güvenliğin sağlanması için AYZ kullanımının ne gibi faydalar sağlayacağına dikkat çekilmesi amaçlanmıştır.

2- SİBER GÜVENLİK SORUNLARI

Verinin kullanıldığı her sektöre, alana, kuruma, şirkete, fabrikaya veya şahsa yönelik siber saldırılar gerçekleştirilebilir. Siber saldırılar ile kötü niyetli kişiler veri çalmayı, değiştirmeyi veya silmeyi hedefleyebilir. Verinin kullanıldığı, işlendiği ve kaydedildiği sistemlerin çalışmasını engellemeyi ve daha da ileri giderek sistemleri bozmayı amaçlayabilirler. Bu gibi durumları gerçekleştirmek için kötü niyetli kişi veya kişilerin motivasyonları hedeftekilere şantaj yapmak veya hedef şirketlerin imajına zarar vermek olabilir. Hatta ülkeler söz konusu olduğunda; kritik altyapılara siber saldırılar düzenlenerek ülke bazında veri kayıplarının olması (kişisel verilerin çalınması gibi), veri ihlallerinin yaşanması (çalınan verilerin internette paylaşılması gibi) veya günlük yaşamın durdurulması (elektrik iletiminin kesilmesi gibi) gibi istenmeyen sonuçlar yaşanabilir. Bu örnekler daha da artırılabilir. Bunun gibi siber güvenlik sorunlarının temelinde siber güvenlik açıklıkları ve siber saldırılar yer almaktadır.

2.1. Siber Güvenlik Açıklıkları

Sistemlerde kullanılan yazılımlara, donanımlara, servislere ve hatta yeni gelen güncelleme işlemlerine göre farklı türde güvenlik açıklıkları bulunabilmektedir. Bir sistemin güvenlik açıklığının olup olmadığının belirli aralıklarla kontrol edilmesi ve alarm durumunda hızlıca müdahale edilmesi gereklidir. Çok eski bir açıklık sistemde hala bulunuyor olabilir veya daha yeni üretilmiş bir açıklık sisteme yerleştirilmiş olabilir. Bu yüzden açıklıklar hakkında bilgi sahibi olunması siber güvenliğin sağlanması için büyük öneme sahiptir [4]. Siber açıklıklar çok çeşitli olabilmektedir (Şekil 1). Bunlardan en sık karşılaşılanlar bu bölümde incelenmektedir.



Şekil 1. Sıklıkla karşılaşılan güvenlik açıklık türleri

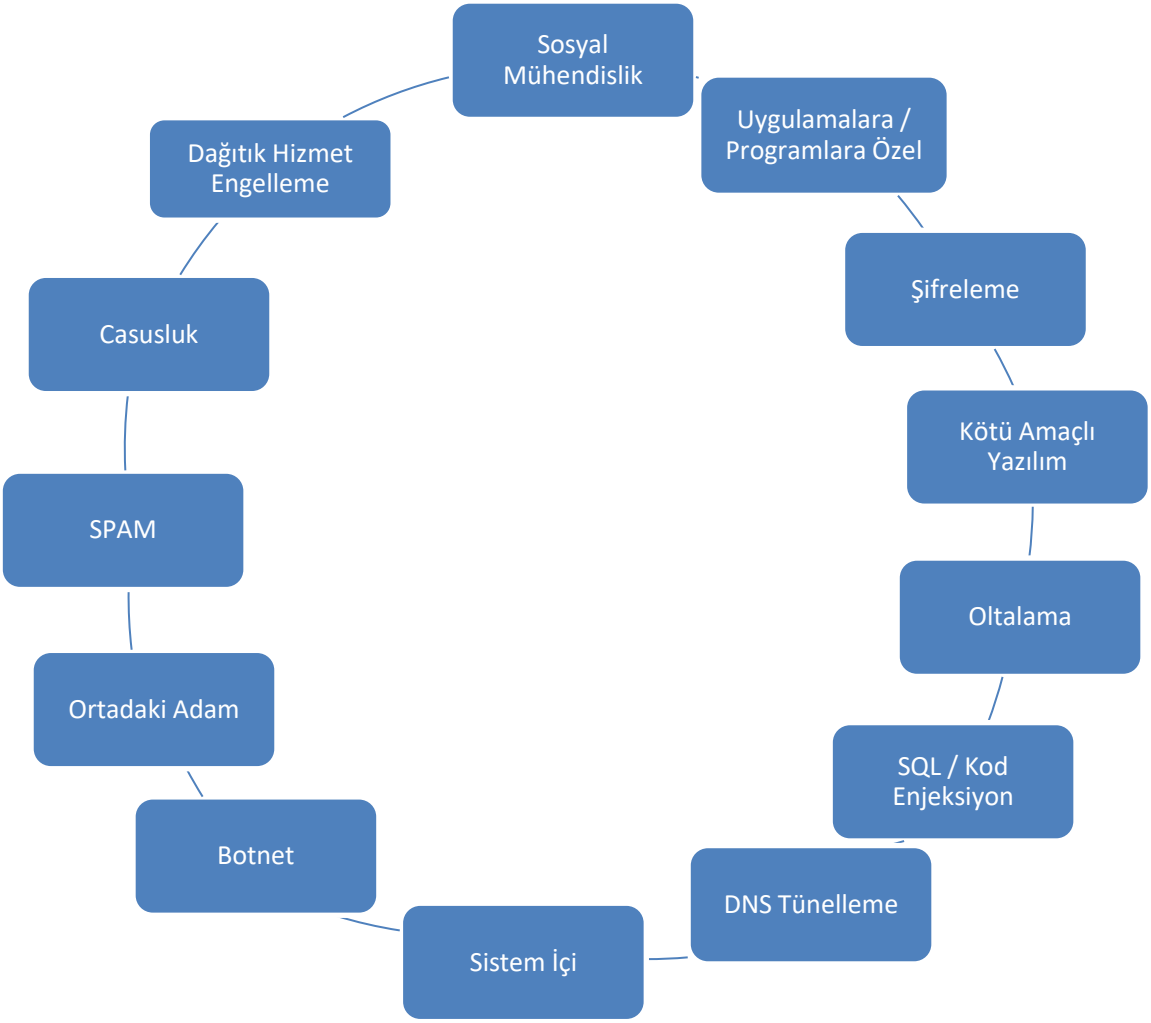
- İlk olarak, yazılımdan kaynaklı açıklıklar incelenmektedir. Sistemlerde yer alan uygulamalarda, programlarda veya cihazlarda bulunan hatalar, yanlış kodlamalar ve gömülü özel kod parçacıkları kötü niyetli kişiler tarafından kullanılmaktadır. Bu şekilde sistemler tehlike altında kalabilmektedir. Bunlara ek olarak, işletim sistemlerinde yer alan güvenlik açıklıkları da görülebilmektedir. İşletim sistemlerinin güncel tutulmaması, güvenlik yamalarının zamanında yapılmaması ve uygun şekilde kullanılmaması ciddi güvenlik sorunlarına yol açmaktadır.
- İkinci olarak güvenlik duvarı açıklıkları ele alınmaktadır. Bir ağa bağlı tüm sistemler için önemli bir güvenlik adımıdır. Yazılımları ve donanımları saldırıya karşı korur. Güvenlik duvarında bir açıklık olması; korumanın yönünü tersine çevirerek sistemin içinde bulunduğu ağa saldırılar uygulanmasının önünü açabilir. Güvenlik duvarının mimarisi, yapılandırılması, kurulması ve çalıştırılması aşamalarından birinde yapılan bir hata veya eksiklik büyük bir açıklığa sebep olabilmektedir.
- Bir başka açıklık ise iletişim protokollerini ilgilendirmektedir. Kullanılan ağın farklı katmanlarına ait farklı açıklıklar bulunabilir. Güvenli olmayan bu ağda yer alan protokoller de açıklıklara sahip olabilir. Kötü niyetli kişiler kullanılan protokol bilgisinden yola çıkarak var olan açıklıklara ulaşabilir ve saldırılarını bu yollardan gerçekleştirmeyi deneyebilir. Aynı şekilde kablosuz ağ üzerinden de saldırılar

gerçekleştirilebilir. Özellikle, güvenli olmayan kablosuz ağ erişim yerleri kötü niyetli kişilerin saldırı uygulamaları için cazip alanlardır.

- Ağda aktif olarak kullanılan veya kullanılmayan donanımlara ait açıklıklar da olabilmektedir. Donanımların içine gömülü şekilde açıklıklar bulunabilir veya daha sonradan ele geçirilmiş donanımlar da olabilir. Ayrıca sisteme yeni eklene donanımlar doğru yapılandırılmadıysa ve tüm donanımların kontrolleri sıklıkla yapılmıyorsa bu durum da büyük sorunlara yol açabilir.
- Bir diğer açıklık da sunuculardan kaynaklı olabilmektedir. Sistemlerde kullanılan sunucuların güncel olmaması, uygun şekilde yapılandırılmamaları ve rutin kontrollerinin gerçekleştirilmemesi güvenlik sorunlarına davetiye çıkarmaktadır. Sunucuların siber güvenliği sağlanmazsa tüm sistem işleyişi ve güvenliği tehlike altına girmektedir.

2.2. Siber Saldırılar

Bilişim dünyasının gelişmesiyle farklı siber saldırı türleri de gerçekleşmeye başlamıştır (Şekil 2). Bu saldırılar kişisel, kurumsal veya ülkesel tabanlı olabilmekte ve saldırıların etkileri de farklılık göstermektedir. Genel olarak gerçekleşen saldırı türlerinden bazılarına bu bölümde yer verilmektedir [5].



Şekil 2. Sıklıkla karşılaşılan siber saldırı türleri

- En yaygın şekilde görülen saldırılardan birisi sosyal mühendislik saldırılarıdır. Bireysel olarak veya kurumsal olarak çalışan kişileri kandırarak sisteme sızmak amaçlanmaktadır. Bu saldırı yöntemi ile kişiler farkında olmadan sahte adreslere tıklayabiliyor veya virüs içeren belgeleri indirebiliyor. Bundan sonrası ise kötü niyetli kişilerin becerisine kalmaktadır. Özellikle bu konuda ara ara eğitimler verilmeli, kontrol amaçlı uygulamalar yapılmalı ve ilgili kişiler uyarılmalıdır.
- Bir diğer saldırı türü de uygulamalara veya programlara özel saldırılarıdır. Kötü niyetli kişiler hedef sistemde kullanılan uygulamaları tespit etmeyi amaçlamaktadır. Daha sonra bu uygulamaların güvenlik açıklıkları araştırılmaktadır. Açık kaynak, özel, ticari, büyük kapsamlı, küçük kapsamlı, eski üretim veya yeni üretim uygulamalar olabileceği gibi farklı yazılım dilleri ile kodlanmış uygulamalar da olabilir.

Hedef uygulamaya ait belirlenen güvenlik açıklıkları kullanılarak sistemlere sızılmakta ve saldırılar düzenlenmektedir.

- Şifreleme saldırıları ise veriyi gönderen ve alan arasındaki şifreli iletişime sızmayı ve bu iletişimi bozmayı hedefleyen bir saldırı türüdür. Burada, kötü niyetli kişi saldırı kod içerisindeki, şifrelemedeki, şifreleme protokolündeki veya anahtar yönetim şemasındaki açıklıklardan faydalanmayı amaçlamaktadır. Saldırıyı gerçekleştirecek bir yol bulunursa şifreli iletişime dâhil olunarak verinin ve iletişimin güvenliği bozulmuş olur.
- Kötü amaçlı yazılım saldırıları da sıklıkla karşımıza çıkmaktadır. Sistemlere sızmak, sistemden veri çalmak veya kişileri rahatsız etmek amaçlanmaktadır. Virüsler, solucanlar ve Truva atları en bilinen örneklerdir. Yazılım dilinden ve platformdan bağımsız şekilde hazırlanabilmektedir. Kişileri uyarmadan sistemde saklanabilmekte ve hedeflerini gerçekleştirmektedir. Bunlara ek olarak, şantaj amaçlı kullanımlar için kişilerden para ödemeleri talep edilebilmektedir.
- Dağıtık Hizmet Engelleme saldırıları olarak bilinen ve hedefteki sistemi kullanılmaz hale getirmeyi amaçlayan bir saldırı türü de bulunmaktadır. Bu saldırı ile hedef sistemin kapasitesinin aşılması ve sistemin gelen isteklere cevap veremez hale getirilmesi hedeflenmektedir. Böylece kısa veya uzun bir süre sistem devre dışı kalmaktadır. Sistemde yer alan donanımlar, yazılımlar veya ağ bileşenleri aşırı tüketilerek hizmet veremez duruma getirilmektedir. Bu bölümde bahsedilen saldırı örnekleri dışında farklı saldırı türleri de bulunmaktadır.

3-SİBER SALDIRILARIN TESPİT EDİLMESİ

İlk bilgisayar sistemlerinin ortaya çıktığı 1940'lardan bu yana bilişim teknolojileri sürekli gelişme göstermiştir. Bilişim dünyasındaki gelişmelere paralel olarak saldırı fikirleri ortaya çıkmış ve bunun için yeni teknikler geliştirilmiştir. Bu gelişmelere bağlı olarak; siber güvenliğin sağlanmasına yönelik saldırı tespit sistemleri kullanılmaya başlanmıştır. Bunların kapsamlı olması, verimli kullanılabilmesi, uygulanabilir olması ve farklı teknolojileri içermesi için çalışmalar yapılmakta ve yeni yöntemler geliştirilmektedir. Saldırı tespit sistemlerinde sıkılıkla kullanılan yöntemlere, metotlara ve tekniklere bu bölümde yer verilmektedir.

- İlk olarak verinin anlaşılması, tanımlanması ve üzerinde analizler yapılması gerekmektedir. Veri içerisinde yer alan kalıpların veya desenlerin belirlenmesi, bunların kullanım oranlarının hesaplanması ve analiz işlemlerinin yapılması gerekmektedir. Bu işlemler için istatistik ve olasılık tekniklerinden faydalanılmaktadır. İstatistik teknikleri saldırı verileri üzerinde analiz yapılmasına ve elde edilen

sonuçların yorumlanmasına olanak sağlar. Olasılık teknikleri ise veri üzerinde bir olayın veya saldırının gerçekleşme olasılığının hesaplanmasını ve bunun üzerine yorumlar yapılmasını sağlamaktadır [5].

- Veri boyutunun artması sonucunda, istatistik ve olasılık tekniklerini destekleyecek farklı yöntemler de kullanılmaya başlanmıştır. Bu aşamada karşımıza veri madenciliği yöntemleri çıkmaktadır. Veriler arasındaki benzerlikleri, ilişkileri tanır ve tüm veri için ortak bir model oluşturur. Bu modelleri de benzerliklerine göre gruplandırır. Yeni bir saldırıyı tespit etmek için üzerinde çalışılan ve analiz edilmiş veriye ihtiyaç duyar [6].
- Bilişim alanında uzun zamandır kullanılan bulanık mantık sistemleri de saldırı tespitinde kullanılmaktadır. Bu sistemlerde kesinlik olmadığı için veride yer alan sınıfların üyelik derecelendirmeleri değerlendirilir. Analiz yaparken üyelik fonksiyonları kullanılır, derecelendirmeler 0-1 arasında yapılır ve bunlara göre veri sınıflandırması gerçekleştirilir [7].
- Sistemde ağ akışının kontrol edildiği bir cihaz üzerinden saldırılı verinin geçip geçmediğini belirlemek için kural tabanlı ve imza tabanlı olmak üzere farklı saldırı tespit sistemleri de kullanılmaktadır. Saldırılı veri belirlendikten sonra yeni gelen veri buna göre kıyaslanır, analiz edilir ve kayıt altına alınır. Sistem işleyişini sürdürürken eğer daha önceden kural olarak belirlenmiş veya imzası alınmış saldırı verisi ile saldırıya maruz kalırsa, saldırı tespit sistemi alarm haline geçer. Ağ yöneticisine durumu gösterir. Bu işlemlerin yapılabilmesi için ağın belirli aralıklarla taranması ve kontrollerin yapılması gerekmektedir [8].
- Kural tabanlı ve imza tabanlı saldırı tespit sistemlerine benzer şekilde çalışma yapısına sahip olan bir diğer sistem de anomali tabanlı saldırı tespit sistemidir. Sisteme ait ağ akış verisi toplandıktan sonra ağdaki paketlerin hareketleri veya davranışları kaydedilir ve incelenir. Saldırı olduğunda ve normal durumda paketlerin özellikleri kaydedilir. Sistem yeni bir saldırıya maruz kaldığında anomali tabanlı saldırı tespit sistemi devreye girer ve anormal bir tespit ettiğinde alarm üretir [9].
- İmza tabanlı ve kural tabanlı saldırı tespit sistemleri farklı veya yeni bir saldırı ile karşılaşınca tespit işleminde başarılı olamayabilir. Anomali tabanlı tespit sistemi ise saldırıları veya saldırı olmayan veriyi yanlış şekilde tespit edebilir. Bu gibi sorunlar veri üzerinde daha derin analizler yapılmasının gerekli olduğunu göstermiştir. Bu aşamada karşımıza makine öğrenmesi teknolojisi çıkmaktadır [10]. Makine öğrenmesi, verileri belirli algoritmalar çerçevesinde tanımlar ve karmaşık yapıdaki desenleri bilgisayar sistemine öğretir. İstatistik, olasılık gibi yöntemleri ve kurulu olduğu bilgisayar sisteminin kaynaklarını kullanarak gerekli olan hesaplamaları yapar. Bu yeteneklerini kullanarak saldırıların tespiti için deneyimlerden faydalanır ve karar verme süreçlerini işler [11].

- Makine öğrenmesinin özel bir uygulaması olarak karşımıza derin öğrenme uygulamaları çıkmaktadır. Derin öğrenme tabanlı saldırı tespit sistemleri veri boyutunun ve çeşitliliğinin artması sonucu kullanılmaya başlanmıştır. Bu sistemde derin sinir ağları bulunmakta ve bunların içerisinde yer alan katmanlar giriş verilerini işlemektedir. İşlenen veri çıkış katmanına iletilmekte ve burada sınıflandırma gibi belirli işlemler gerçekleştirilmektedir. Veride yer alan özniteliklerin ağırlıklarının veya kullanım oranlarının hesaplanması ve analizde bunlara yer verilmesi gibi imkânlar da sunulmaktadır [12].
- Büyük boyutlu ve karmaşık yapılı verileri analiz etmek için kullanılan bir diğer sistem de yapay zeka sistemleridir. Derin öğrenme ve makine öğrenmesi uygulamaları yapay zeka sistemlerinin alt dallarını oluşturmaktadır. Yapay zeka tabanlı saldırı tespit sistemleri pratik, esnek ve sezgiye dayalı yöntemler içermektedir. Algoritmaları ve veritabanı kullanarak kararlar verebilen bir yapıya sahiptir. Kapsamlı bir teknoloji olduğu için üzerinde çalıştığı bilgisayarın kaynaklarını hızlıca tüketmektedir [3].
- Tek başına kullanılan tespit sistemleri farklı sistemler ile beraber kullanılarak daha verimli çalışmalar elde edilebilmektedir. Örneğin hibrit saldırı tespit sistemi elde etmek için imza tabanlı tespit sistemi ile anomali tabanlı tespit sistemi beraber kullanılabilir.

4-AÇIKLANABİLİR YAPAY ZEKA

Literatürde kullanılan yapay zeka yöntemlerine yeni özelliklerin eklenmesi ve kullanım kolaylığının sunulması ile Açıklanabilir Yapay Zeka (AYZ) çalışmaları ön plana çıkmıştır. AYZ analiz için kullanılan karar süreçlerinin ve elde edilen sonuçların anlaşılmasını kolaylaştırmaktadır. Bu sistemi kullanan kişiler, uzmanlar veya araştırmacılar daha anlaşılabilir, izlenebilir ve yorumlanabilir sonuçlar elde etmektedir.

Genel olarak, saldırı tespit sistemlerinin işleyişinde arka planda neler olduğu çok net şekilde anlaşılamamaktadır. Her sistemin kendine özgü metotları ve yöntemleri bulunmakta ve işleyişler farklılık göstermektedir. Bu durumlar da kullanıcıların saldırı tespitinin nasıl yapıldığı ve kararların nasıl alındığı konularında bilgi sahibi olmasını, çıkarımlarda bulunmasını ve sistemlere doğrudan müdahale etmesini zorlaştırmaktadır. AYZ kullanımı ile bu sorunlar için yeni çözümler geliştirilmektedir [13].

AYZ saldırı tespit sisteminde, saldırı olup olmadığına karar vermek için arka planda işlenen adımlar ve bu adımların sonuçları kullanıcılara gösterilmektedir. Sistem içinde verilen kararların hangi temellere dayandığının gösterilmesine imkân sunulmaktadır. Böylece kullanılan yöntemlerin, metotların veya modellerin çalışma prensipleri hakkında bilgi sahibi olmanın önü açılmıştır. Elde edilen analiz sonuçlarının irdelenmesi ve yorumlanabilmesi özellikleri sayesinde kullanıcılar bu sistemin doğruluğunu da eleştirebilir hale gelmiştir [3].

Tüm yapay zeka sistemlerinin açıklanabilir olması gerekliliği yoktur. Yapay zeka sisteminin açıklanabilir olması amaçlandığında dört temel prensibe uyulması beklenir. Bunlar açıklanabilirlik, anlamlı olma, açıklamanın doğruluğu ve bilgi sınırı prensipleridir. Bir sistemin ürettiği sonuçları veya gerçekleştirdiği işlemleri destekleyen kanıtların veya nedenlerin sunulması açıklanabilirlik olarak ifade edilmektedir. İkinci prensip olan anlam veya anlamlı olma ise bir sistemin hedef sistem, tüketici veya alıcı için anlaşılabilir açıklamalar sağlamasıdır. Bir açıklama, çıktının üretilme nedenini ve işlemlerin sürecini doğru bir şekilde yansıtırsa açıklama doğruluğu prensibi uygulanmış olur. Bilgi sınırı ise bir sistemin sadece tasarlandığı koşullar altında ve yeteri kadar güvenilir sonuca ulaşması durumunda çalışmasını ifade etmektedir. Yapay zekanın aldığı veya kullandığı veri ile uzmanla veya kullanıcıyla olan ilişkisi ele alındığında bu dört prensip önemli etkiye sahip olur. Bu nedenle bir açıklamanın gerekliliği ve nasıl iletildiği yapay zeka kullanıcılarına, uzmanlarına veya tüketicilere göre farklılık gösterebilir [14].

4.1. Avantajları ve Siber Güvenlik Alanında Kullanımı

AYZ geniş kullanım alanına hitap etmektedir. Bilişim dünyasında ve siber güvenlik alanında kullanımı ile saldırı tespit sistemlerine şeffaflık ve açıklanabilirlik gibi özellikler sağlamıştır. Saldırıların hedefindeki sistemin güvenlik sorumlusu veya güvenlik üzerine araştırmalar yapan kişiler tarafından saldırı tespit aşamalarının izlenebilir olması büyük avantajlar sağlamaktadır. Sistemde toplanan verinin saldırı içerip içermediğini anlayabilmek, hangi özelliklerin analizde kullanıldığını bilmek gibi olanaklar ilerisi için kullanılabilecek ve iyileştirme imkanları sunabilecek değerli bilgiler haline gelmiştir.

AYZ kullanımı ile güvenlik araştırmacıları saldırı tahmin ve tespit modellerini daha iyi kavrayabilir ve arka planda verilen kararları daha kolay anlayabilir. Bu durumlar, AYZ tespit sistemlerinin daha güvenilir ve uygulanabilir olmasını sağlamaktadır. Ayrıca güvenlik uzmanlarının saldırı tespit konusunda kendilerini geliştirmelerini, daha bilgili olmalarını ve daha hızlı işlemler yapmasını da teşvik etmektedir. Saldırılara karşı anında ve uygun tepkiler verilmesi de kolaylaşmaktadır. Uzmanların sorumlu oldukları sistemin güvenliğine daha hakim olmalarını ve gelebilecek saldırılara karşı bilinçli olmalarını sağlamaktadır. Bunlara ek olarak saldırıların analizinde, farklı saldırı tekniklerinin öğrenilmesinde ve saldırıların tanımlanmasında AYZ kullanılabilir. Sistemlerde yer alan donanımların ve yazılımların güvenlik açıklıklarının tespitinde de kullanılabilir. Tüm bu durumlar kullanıcıların, sistem sorumlularının veya güvenlik araştırmacılarının yapay zeka teknolojilerine daha hızlı adapte olmalarını, bu teknolojileri daha çok kullanmalarını ve bu alanda daha çok çalışma yapılmasını desteklemekte ve teşvik etmektedir.

5- SONUÇ VE ÖNERİLER

Bilişim dünyasında ve siber güvenlik alanında yaşanan gelişmeler olabilecek saldırıların çeşitliliğini, kapsamını ve boyutlarını da etkilemektedir. Güvenliğin sağlanmasındaki en önemli adım olan saldırı tespit sistemleri de gelişmelere ayak uydurmak için kendini güncellemek zorunda kalmıştır. Makine öğrenmesi, veri madenciliği gibi var olan tespit sistemlerine yeni özellikler, metotlar veya yöntemler eklenmektedir. AYZ teknolojisi de bu şekilde geliştirilmiştir.

AYZ teknolojisinin kullanımı ile yapay zeka tabanlı saldırı tespit sistemlerine açıklanabilirlik özelliği eklenmiş ve farklı bakış açıları sunulması sağlanmıştır. Saldırıların anlaşılması, analiz edilmesi ve tespit edilmesi süreçlerinde AYZ kullanılmaktadır. Bu kullanım ile saldırıların karmaşıklığı, büyüklüğü ve türleri daha kolay analiz edilmektedir. Uzmanlara veya kullanıcılara saldırı tespit sisteminin çıktıları hakkında daha derin anlayışlar sunarak yapay zekanın daha güvenilir şekilde kullanılmasını sağlar. AYZ anlamlı ve doğru açıklamalar sunmasından dolayı; uzmanların veya kullanıcıların saldırı tespiti için sunulan bilgileri başka sistemlere uyarlamasına, sistem kararını eleştirmesine, sistemi geliştirmesine veya iyileştirmesine yönelik yetkiler verir.

AYZ tabanlı tespit sistemleri işlemlerini yaparken şeffaflık, açıklanabilirlik ve kolaylık sunduğu için güvenlik uzmanları bu sistemleri giderek daha verimli şekilde kullanabilecektir. Ayrıca, karşılaşılan sorunlara daha hızlı ve uygun tepkiler verilebilmesi de kolaylaşacaktır. Saldırı tespit sistemlerine daha fazla güven duyulması ve bu sistemlerin daha etkili kullanılması sağlanacaktır.

KAYNAKLAR

1. Çıtlak, O., Dörterler, M., & Doğru, İ. (2023). "A hybrid spam detection framework for social networks". *Politeknik Dergisi*, 26(2), 823-837. https://doi.org/10.2339/politeknik.933785
2. Söğüt, E., & Erdem, O. A. (2023). "A Multi-Model Proposal for Classification and Detection of DDoS Attacks on SCADA Systems". *Applied Sciences*, *13*(10), 5993. https://doi.org/10.3390/app13105993
3. Yaz, Ö., & Süzen, A. A. (2023). "Kurumsal Ağlara Yapılan Saldırılarda Açıklanabilir Yapay Zekânın Yeri". *The 5 th International Conference on Applied Engineering and Natural Sciences.* 528-534. Konya.
4. Humayun, M., Niazi, M., Jhanjhi, N., Alshayeb, M., & Mahmood, S. (2020). "Cyber Security Threats and Vulnerabilities: A Systematic Mapping Study". *Arab. J. Sci. Eng.* 45, 3171–3189. https://doi.org/10.1007/s13369-019-04319-2
5. Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions". *Electronics*, *12*(6), 1333. https://doi.org/10.3390/electronics12061333

6. Bakir, C., Hakkoymaz, V., Diri, B., & Güçlü, M. (2020). "Dağıtık veritabanlarında saldırı önleme metotları". *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, *10*(2), 425-441. https://doi.org/10.17714/gumusfenbil.612774
7. Atacak, İ., Çıtlak, O., & Doğru, İ. A. (2023). "Application of interval type-2 fuzzy logic and type-1 fuzzy logic-based approaches to social networks for spam detection with combined feature capabilities". *PeerJ Computer Science*, *9*, e1316.
8. Kılınç, F., & Eyüpoğlu, C. (2023). "AĞ ORTAMINDAKİ SALDIRI TÜRLERİ: SALDIRI SENARYO ÖRNEKLERİ". *İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi*, *6*(1), 99-109. https://doi.org/10.56809/icujtas.1282687
9. Erol, D., & Bakır, H. (2023). "Hiperparemetreleri Ayarlanmış Makine Öğrenmesi Yöntemleri Kullanılarak Ağdaki Saldırıların Tespiti". In the *International Conference on Pioneer and Innovative Studies,* Vol. 1, pp. 274-286.
10. Çelik, B. D., & Arıcı, N. (2021). "Air Quality Forecast in the Covid-19 Outbreak Process: Zonguldak Case", *Gazi Journal of Engineering Sciences,* 7(3), 222-232. https://doi.org/10.30855/gmbd.2021.03.05
11. Orakcı, M., Ciylan, B., Kök, İ., & Sevri, M. (2016). "Suç Analizinde Veri Madenciliği Teknikleri Ve Makine Öğrenmesi Algoritmalarının Kullanılması". *Akademik Bilisim Konferansi 2016,* Aydın.
12. M. Güllü, & N. Barışçı. (2022). "Android Malware Classification with Gray Wolf Optimization Algorithm and Deep Neural Network Hybrid Approach". *The 30th Signal Processing and Communications Applications Conference (SIU),* Safranbolu, Turkey. 1-4, https://doi.org/10.1109/SIU55565.2022.9864822
13. Erhan, B. A. Ş., & Süzen, A. A. (2023). "Web Uygulama Zafiyetlerinin Keşfinde Açıklanabilir Yapay Zekânın Yeri". *The 5 th International Conference on Applied Engineering and Natural Sciences.* 517-527. Konya.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). "Four principles of explainable artificial intelligence". *National Institute of Standards and Technology Interagency or Internal Report 8312*, 43. Gaithersburg, Maryland, 18.

View publication stats