Makine Öğrenmesi Yöntemleri Kullanarak Web Uygulama Saldırılarının Tespitinde Genetik Öznitelik Seçimi Yaklaşımı

Genetic Feature Selection Approach in Detection of Web Application Attacks Using Machine Learning Methods

Hüseyin AHMETOĞLU
Mardin Artuklu Üniversitesi,
Bilgisayar Teknolojileri, Mardin, Türkiye
huseyinahmetoglu@artuklu.edu.tr
ORCID: 0000-0002-4320-0198

Resul DAŞ
Fırat Üniversitesi,
Yazılım Mühendisliği Bölümü, Elâzığ, Türkiye
rdas@firat.edu.tr
ORCID: 0000-0002-6113-4649

Öz

İnternet üzerindeki uygulamalar kodlama kaynaklı bir takım güvenlik endişelerini barındırırlar. Zayıflıklar veya güvenlik açıkları, suçluların hassas verileri çalmak için veri tabanlarına doğrudan ve genel erişim elde etmesine olanak tanır. Bu çalışmada, web uygulama saldırılarının hibrit saldırı tespit sistemleri ile daha kolay ve daha doğru tespiti için sezgisel öznitelik seçimi ve makine öğrenmesine dayanan bir yaklaşım önerilmektedir. CIC-IDS2017 ve CSE-CIC-IDS2018 veri setlerindeki web uygulama saldırıları ve normal akış örnekleri bir dizi veri ön işleme aşaması sonrası birleştirilerek ve yeni bir veri seti oluşturuldu. Genetik Algoritma ve Lojistik Regresyon kullanılarak ortalama karesel hata ve öznitelik sayısı optimizasyonu gerçekleştirilip sonuçlar beş farklı makine öğrenmesi algoritması ile test edildi. Elde edilen sonuçlar incelendiğinde, öznitelik sayısının %85 oranında azaltılmasına rağmen sınıflandırmadaki başarım oranlarının %99 seviyesinde kaldığı gözlemlenmiştir.

***Anahtar Sözcükler**: web uygulama saldırısı, makine öğrenmesi, genetik algoritma, öznitelik seçimi, saldırı tespit sistemi*

Abstract

Applications on the Internet have some coding-related security concerns. Weaknesses or vulnerabilities allow criminals to gain direct and public access to databases to steal sensitive data. This study proposes an approach based on heuristic feature selection and machine learning for easier and more accurate detection of web application attacks with hybrid intrusion detection systems. Web application attacks and benign flow examples in CIC-IDS2017 and CSE-CIC-IDS2018 datasets were combined after a series of data preprocessing stages, and a new dataset was created. Using Genetic Algorithm and Logistic Regression, mean square error and feature count optimization were performed, and the results were tested with five different machine learning algorithms. When the results obtained were examined, it was observed that the success rate in classification remained at the level of 99%, although the number of features was reduced by 85%

***Keywords**: web application attack, machine learning, genetic algorithm, feature selection, intrusion detection system.*

Gönderme ve kabul tarihi: 03.11.2021 - 29.11.2021
Makale türü: Araştırma

1. Giriş

İnternet gün geçtikçe insanların sosyal yaşamlarının büyüyen bir parçası haline gelmenin yanı sıra, insanların yaşam tarzlarını şekillendiriyor. Küreselleşen dünyayla internet, toplumdaki yerini kademeli olarak derinleştiriyor. Hükümetlerin hayati altyapıları internet ile bütünleşmiş durumda ve internet, sosyoekonomik büyümenin en önemli kaynaklarından biri haline geliyor. İnternetin bu denli derinleşmesi ve genişleyen yapısı, insanları çeşitliliği sürekli artan yeni tehditlere maruz bırakıyor. Ağ trafiğinde bu tehditlerin nasıl tespit edilebileceği,

bugünün siber güvenliğinin en önemli konuları arasındadır [1]. Siber saldırı, veri veya bilgi sistemlerini çalmak, değiştirmek veya yok etmek için çeşitli yöntemler kullanarak bilgisayar bilgi sistemlerini, altyapıları, bilgisayar ağlarını veya kişisel bilgisayar cihazlarını hedef alan her türlü saldırı eylemidir. Özelde web uygulama saldırıları ise web üzerindeki varlıkları hedef alan saldırılardır. Avantajlarına rağmen, web uygulamaları uygunsuz kodlamadan kaynaklanan bir dizi güvenlik endişesi doğurur. Ciddi zayıflıklar veya güvenlik açıkları, suçluların hassas verileri çalmak için veri tabanlarına doğrudan ve genel erişim elde etmesine olanak tanır. Veri tabanlarının çoğu, onları sık sık saldırıların hedefi haline getiren değerli bilgiler içerir. Kurumsal web sitelerini tahrif etmek gibi eylemlerin yanında, günümüzde saldırganlar, önemli verilerin satılmasındaki muazzam getiriler nedeniyle veri tabanı sunucusunda bulunan hassas verilere veri ihlalleri ile erişmeyi tercih ediyorlar. Yukarıda açıklanan çerçevede, suçlular, veri tabanında bulunan verilere saldırganlık, şans, ihmal veya insan hatası ile hızlı bir şekilde erişebiliyorlar ve web uygulamalarındaki güvenlik açıklarını keşfedebiliyorlar. Web uygulamaları güvenli değilse, yani çeşitli bilgisayar korsanlığı tekniklerinden en az birine karşı savunmasızsa, hassas bilgiler içeren veri tabanının tamamı ciddi bir web uygulaması saldırısı riski altındadır.

Saldırı Tespit ve Saldırı Önleme Sistemleri (IDS/IPS) temel olarak veri paketlerini analiz ederek saldırı olup olmadığını belirler. Ağ hareketlerini analiz ettikten sonra sistem, sonuca göre bazı önlemler alabilmektedir. IDS/IPS'ler operasyonel mantığa dayalı olarak iki ana kategoriye ayrılır; (1) İmzaya Dayalı IDS, (2) Anomaliye Dayalı IDS. İmza Tabanlı IDS, tanınan güvenlik açıklarının bilgileriyle oluşturulan saldırı imzası ile çalışır. İmzalar, saldırılar hakkında ayrıntılı bilgi içerir. Bu tür sistemler tanınan saldırılar için yüksek doğruluk oranına sahiptir, ancak ilk defa karşılaşılmış saldırıları tespit edemezler. Bu nedenle, yeni saldırılar keşfedildiğinde yeni imzalar oluşturulmalı ve bu imza derhal sisteme aktarılmalıdır. Bu sistemler Sıfırıncı Gün Saldırılarına karşı dayanıklı değildirler. Anomali Tabanlı IDS, Sıfırıncı Gün Saldırılarını tespit edebilmektedir ancak bu IDS'ler aynı zamanda yüksek yanlış alarm oranına sahiptir. Anomali Tabanlı IDS'ler, yüksek esnekliğe sahiptir ve üst düzey makine öğrenimi tekniklerini yapılarında barındırabilirler. Bu şekildeki hibrit çözümlerle yanlış alarm oranları düşürülebilir [2].

Bu çalışmada hibrit saldırı tespit sistemlerinin, web uygulama saldırılarının tespitindeki performanslarını artırmak için makine öğrenmesi ve sezgisel öznitelik seçim yöntemlerinden Genetik Algoritma (GA) kullanan yeni bir yaklaşım önerilmektedir. Uygulamada açık erişimle sunulan CIC-IDS2017 [3] ve CSE-CIC-IDS2018 [4] veri setlerindeki web uygulama saldırısı ve normal ağ akış örneklerinin bir dizi ön işlemden geçirilerek birleştirilmesi ile oluşturulmuş bir veri seti kullanılmıştır. Oluşturulan veri setinde SQL Injection, Cross-Site Scripting ve Brute-Force saldırı verileri bulunmaktadır. Bu veri seti kullanılarak GA ve Lojistik Regresyon (LR) kullanılarak bir öznitelik seçim optimizasyonu gerçekleştirilmiştir. Seçilen öznitelikler, Rassal Orman (Random Forest-RF), Destek Vektör Makineleri (Support Vevtor Machine-SVM), Naif Bayes (Naive Bayes-NB), K En Yakın Komşu (k-Nearest Neighbors-KNN) ve Derin Sinir Ağları (Deep Neural Network-DNN) olmak üzere beş farklı sınıflandırıcı model eğitiminde kullanılmıştır. Bu yöntemler belirlenirken hem sığ hem de derin makine öğrenmesi algoritmaları seçilmiştir. Böylelikle seçilen özniteliklerin farklı yöntemlerde nasıl sonuçlar vereceği gözlemlenmek istenmiştir. Modellerin test verileri üzerlerindeki performansları farklı metriklerle izlenmiştir. Bu metriklerin kullanılması ile modellerin sadece saldırıyı tespit etmedeki başarısını ölçmek değil, aynı zamanda farklı saldırı türlerinin analizi hakkında da bilgi sağlanması hedeflenmiştir. Bu çalışma ile çeşitliliği ve etki alanı sürekli artan ağ tehditlerine karşı daha etkili savunma sistemlerinin geliştirilmesine katkı sunulmak amaçlanmıştır. Ağ üzerinde bir anomalinin tespiti ne kadar kolay ve ne kadar hızlı olursa saldırılara karşı tepkisel önlemlerin ve onarım işlemlerinin etkileri de aynı oranda güçlü olmaktadır. Bu noktada makine öğrenmesi temelli siber güvenlik çözümlerinin daha etkin bir şekilde kullanılması için ağ özellikleri ile oluşturulan en etkili özniteliklerin seçilmesi önemlidir. Yukarıdaki amaçlara ek olarak bu çalışma ile özelde web uygulamalarını hedef alan saldırılara karşı geliştirilecek yapay zekâ temelli önlemlerin gücünü artırmayı amaçlayan araştırmacılar için yeni bakış açıları sunulmak istenmiştir.

2. Literatür İncelemesi ve Temel Bilgiler

Veri setindeki örnek sayısının fazla olması, geleneksel makine öğrenmesi algoritmalarına kıyasla

Derin Öğrenme (Deep Learning-DL) tabanlı algoritmalarda hem daha kritik öneme sahiptir hem de örnek sayısı arttıkça eğitim modelinin başarısını artırmaktadır [5]. Buna rağmen hem sığ hem derin makine öğrenmesi yöntemlerinde özniteliklerin bir kısmı belirli sınıflandırma problemleri için bir takım önem derecelerine sahip olabilir. Yüksek öznitelik boyutuna sahip veri setleri ile eğitim ve test işlemleri karmaşık bir hale gelebilir. Bu durumlar düşünüldüğünde daha başarılı sonuçlar elde etmek için çok boyutlu veri setlerinin üzerinde Öznitelik Mühendisliği (Feature Engineering-FE) işlemleri gerçekleştirmek zorunludur. FE, makine öğrenmesi araştırmalarının vazgeçilmez bir parçası haline gelmiştir ve sınıflandırıcıların performanslarını artırmak için bazı özniteliklerin birleştirilmesi ve bazılarının çıkarılması olarak tanımlanabilir. Bu bölümde, çalışmanın arka planında bulunan materyal ve yöntemler açıklanacak ve literatürdeki ilgili çalışmalar incelenecektir.

2.1 Literatürdeki İlgili Çalışmalar

Siber güvenlik problemlerine yapay zekâ çözümleri sunmayı hedefleyen araştırmacılar hesaplama karmaşıklığını azaltmak için öznitelik seçimi odaklı birçok çalışma gerçekleştirmişlerdir. Bu çalışmalar hem öznitelik seçimi hem de boyut azaltma uygulamalarını içermektedir. Khammassi ve Krichen [6] KDDCUP'99 ve UNSW-NB15 veri setlerini kullanarak daha etkin saldırı tespit sistemleri için en iyi öznitelik alt kümelerini bulmayı amaçlamışlardır. Elde ettikleri sonuçlarda öznitelik sayısındaki azalmaya rağmen yüksek algılama oranının sürdüğünü gözlemlemişlerdir. Zhu ve ark. [7] öznitelik seçimi için çok amaçlı optimizasyon yaklaşımını önermişlerdir. Cloud Computing senaryosunda güvenlik duvarına bir uyarı sistemi eklemeyi önerdiler ve modelin başarısını artırmak için NSGA-II algoritmasını kullanmışlardır. Bir başka çalışmamızda [8] oylamaya dayalı öznitelik seçimi yaklaşımını IDS için hem ikili hem de çoklu sınıflandırma için önerdik. Altı farklı öznitelik seçim yönteminde en çok oyu almış öznitelikler DNN kullanılarak sınıflandırılmış ve ikili sınıflandırma için 8 öznitelikle yüzde 92 başarı elde edilmiştir. Wang ve ark. [9] öznitelik mühendisliği yaklaşımları için farklı yönde bir yöntem sunmuşlardır. Logaritmik yoğunluk dönüşümü ile boyut azaltma gerçekleştirmişlerdir. NSL-KDD veri seti üzerindeki deneylerinde daha yüksek performans elde etmek için geliştirilmiş öznitelikler üretmeyi amaçlamışlardır. Xu ve ark. [10] Ağ saldırısı tespiti için geliştirilmiş bir ikili balina optimizasyon algoritmasına (Whale Optimization Algorithm -WOA) dayalı bir özellik seçme yöntemi sunmuşlardır. Önerdikleri yöntemi GA ile karşılaştırmışlar ve sınıflandırma başarısını KDD Cup veri seti üzerinde ölçmüşlerdir. Elde ettikleri sonuçlarda önerilen yönetimin %97'lik bir başarı gösterdiğini açıklamışlardır. Gharaee ve Hosseinvand [11] yenilikçi bir uygunluk fonksiyonu ile öznitelik seçimi için genetik algoritma yaklaşımını önermişlerdir. Gerçekleştirdikleri uygulamada düşük yanlış pozitif oranı korunurken doğruluğu artırmayı başarmışlardır. Önerilen IDS'nin testi için UNSW-NB15 ve KDD Cup veri setini kullanmışlardır. Her sınıfın testi için ayrı ayrı veri setleri oluşturmuşlardır. Thakkar ve Lohiya [12] saldırı tespit sistemleri için üretilen sığ ve derin makine öğrenmesi modellerinin performanslarını artırmak için kullanılan sürü ve evrimsel yöntemleri kullanan çalışmaları incelemişler ve sonuçları karşılaştırmalı olarak sunmuşlardır. Hosseini [13] gerçekleştirdiği çalışmada öznitelik seçimi için hem genetik algoritma hem de parçacık sürü optimizasyonunu kullanmıştır. NSL-KDD veri setinden seçilen öznitelikler ANN ile test etmiş ve ikili sınıflandırma için %94.41 doğru sınıflandırma başarısı elde etmiştir. Onah ve ark. [14] IDS'ler için GA ve NB temelli bir yaklaşım önermişler ve NSL-KDD veri setini kullanmışlardır. %99.73'lük doğru sınıflandırma başarısı ve %0.6'lık yanlık alarm oranı elde etmişlerdir. Z. Halim ve ark. [15] yeni bir GA tabanlı öznitelik seçim yönetimi önermişler ve sonuçları dört farklı sınıflandırma algoritmasında test etmişlerdir. 10 öznitelik ile en başarılı sonucu %99.8 ile XgBoost algoritması vermiştir. Moustafa ve Slay [16] hibrit bir öznitelik seçim yöntemi önermişler ve sonuçları LR ile değerlendirmişlerdir. 11 öznitelik ile eğitilmiş LR modeli %83'lük bir doğruluk elde etmiştir. Tama ve ark. [17] PSO ve GA temelli hibrit bir öznitelik seçim yöntemi öndermişler ve 8 öznitelik ile salıdır tespit sistemleri için %91,27 oranında doğruluk elde etmişlerdir. Kasongo ve Sun [18] önerdikleri sarmalayıcı yaklaşımı RF ile test etmişler ve 22 öznitelikle %77,16 oranında bir doğruluk elde etmişlerdir. Nazir ve Khan [19] optimizasyon temelli saldırı tespit sistemi yaklaşımlarında RR algoritmasını temel almışlar ve 16 öznitelik ile %83.12 oranında bir doğruluk elde etmişlerdir.

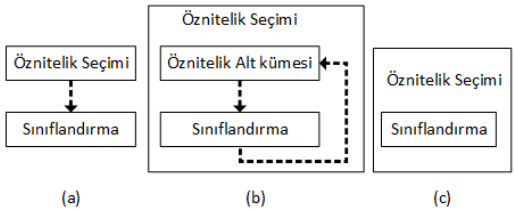
2.2 Web Uygulama Saldırıları

SQL Injection doğrudan veri tabanlarını hedefleyen saldırı türüdür, hala en yaygın ve en tehlikeli güvenlik açıklarından biridir. Kötü niyetli SQL ifadelerini çalıştırmayı mümkün kılan bir enjeksiyon saldırısı

türüdür. SQL ifadeleri ile bir web uygulamasının arkasındaki bir veri tabanı sunucusu kontrol edilir. Saldırganlar, uygulama güvenlik önlemlerini atlatmak için SQL Injection güvenlik açıklarını kullanabilir. Saldırganlar, kullanıcıları kandırmak ve onları kimlik avı sitelerine yönlendirmek için savunmasız web uygulamalarının kullanıcı girişini kullanarak kurban bilgisayarına kötü amaçlı kod enjekte edebilirler [20]. Cross-Site Scripting (XSS) saldırısı web sunucuları ve veri tabanı motorunun kendilerinde herhangi bir güvenlik açığı olmamasına rağmen kullanılabilir. Saldırgan, meşru bir web sayfasına veya web uygulamasına kötü amaçlı kod ekleyerek kurbanın web tarayıcısında kötü amaçlı komut dosyaları çalıştırmayı amaçlar. Asıl saldırı, kurban kötü amaçlı kodu yürüten web sayfasını veya web uygulamasını ziyaret ettiğinde gerçekleşir. Web sayfası veya web uygulaması, kötü amaçlı komut dosyasını kullanıcının tarayıcısına iletmek için bir araç haline gelir. Siteler Arası Komut Dosyası Çalıştırma saldırıları için yaygın olarak kullanılan savunmasız araçlar forumlar, mesaj panoları ve yorumlara izin veren web sayfalarıdır. Web geliştiricilerinin karşılaştığı yaygın bir tehdit de kaba kuvvet saldırısı olarak bilinen bir parola tahmin saldırısıdır [21]. Brute-Force saldırısı, çalışan tek doğru kombinasyonu bulana kadar olası her harf, sayı ve simge kombinasyonunu sistematik olarak deneyerek bir parola keşfetme girişimidir. Bu çalışma kapsamında bu üç saldırı türüne ait veri örnekleri üzerinden uygulama gerçekleştirilecektir [22].

2.3 Öznitelik Seçim Yöntemleri

Veri setlerinin karmaşıklığını azaltmak ve sınıflandırma performansını artırmak amacıyla geliştirilmiş birçok öznitelik seçme yöntemi vardır. Bunlar üç başlık altında incelenebilir. Filtre yöntemleri, sarmalayıcı yöntemler ve gömülü yöntemler [23]. Bu yöntemlerin öznitelik seçiminde izledikleri Şekil-1'de verilmiştir. Filtre tabanlı öznitelik seçim yöntemleri verilerin analizindeki kullanılamayacak önemde olan özniteliklerin ortadan kaldırılmasında kullanılır. Bu yöntemler öznitelikleri değerlendirmek için sınıflandırıcıya ihtiyaç duymazlar. Öznitelikler değerlendirme kriterlerine göre seçilirler [24]. Sarmalayıcı öznitelik seçim yöntemleri arama ve değerlendirme olmak üzere iki bölümden oluşur. Arama işlemi değerlendirme neticesinde aldığı dönütlerle özniteliklerin önemini bulur. Sarmalayıcı teknikler sınıflandırma algoritmasına ihtiyaç duyar. Sınıflandırma değerlendirmesi nedeniyle filtre yöntemlere göre daha yavaş çalışır. Gömülü öznitelik seçimi ise filtre ve sarmalayıcı tekniklerinin oluşturduğu hibrit bir yaklaşımdır. İçerisinde barındırdığı bayes ve karar ağacı sınıflandırıcıları ile model için en etkili öznitelikler seçilebilir [8]. Bu çalışmada ise sezgisel ve sığ makine öğrenmesi tekniklerinden oluşan sarmalayıcı bir yaklaşım önerilmektedir.



Şekil-1: (a) Filtre, (b) Sarmalayıcı ve (c) Gömülü öznitelik seçme yöntemleri

2.4 Veri Setleri

CICIDS2017 ve CSE-CIC-IDS2018 veri setleri ağ akışı özelliklerinin istatiksel özellikleri kullanılarak oluşturulmuştur [25]. CICIDS2017 veri seti, toplam 25 kullanıcının gerçekleştirdiği ağ olaylarını temsil eden gerçekçi ağ arka plan trafiğinden oluşur. Kullanıcıların profilleri, HTTP, HTTPS, FTP, SSH ve E-posta gibi protokolleri içerecek şekilde ayarlanmıştır. Geliştiriciler, ağ olaylarını içeren belirli özellikleri kapsayacak şekilde minimum, maksimum, ortalama ve standart sapma gibi istatistiksel metrikler kullanmışlardır. Bu istatistiksel metrikleri belirlenirken, ağdaki paket boyut dağılımı, akış başına paketlerin sayısı, ağdaki yükün boyutu, protokollerin talep süresi ve yükteki bazı desenler kullanılmıştır. Veri setindeki her bir örnek toplam 79 öznitelik içerir. CICIDS2017'yi oluşturanlar Kanada Siber Güvenlik Enstitüsü'nün trafik verilerini izlemişler ve veri setini beş gün içeren sekiz farklı dosya halinde servis etmişlerdir. CSE-CIC-IDS2018 veri seti İletişim Güvenlik Kuruluşu (CSE) ve Kanada Siber Güvenlik Enstitüsü (CIC) tarafından oluşturulmuştur. Veri setinde Heartbleed, Brute-force, DoS, DDoS, Web attacks, Botnet, and infiltration olmak üzere yedi farklı saldırı senaryosu içerir. CICDS2017 veri seti gibi, ağ trafiğini kullanarak 80 ağ akışı özelliğiyle oluşturulmuştur. 50 makinenin saldırgan ağı kullanılarak 30 sunuculu kurban ağı için 420 istemci makine ile hazırlanmıştır.

2.5 Performans Metrikleri

Performans ölçümlerinin doğru yorumlanması, makine öğrenme modellerinin başarısını doğru bir

şekilde değerlendirmek için büyük önem taşımaktadır. Bu metriklerle, modelin etki seviyesi belirlenir. Saldırı tespiti için geliştirilen makine öğrenme modellerinin analizi için birden fazla metrik değerlendirilmesi gerekir. Çizelge- 1 saldırı ve normal davranışlar için oluşturulan karmaşıklık matrisini göstermektedir.

Çizelge- 1: IDS için Karmaşıklık Matrisi

		Tahmin edilen sınıf	
	Sınıf	**Normal**	**Saldırı**
Gerçek Sınıf	**Normal**	True Positive (TP)	False Positive (FP)
	Saldırı	True Negative (TN)	False Negative (FN)

Hatırlama (Recall, Sensitivity, Detection Rate-DR):

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

(Precision, True Positive Rate-TPR):

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Doğruluk (Accuracy-Acc):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

F1-Skoru (F-Measure, F1-Score):

$$F - Measure = 2 * \frac{precision * recall}{precision + recall} \quad (7)$$

ROC eğrisi ve AUC: False Pozitif Oranı (FPR) ile DR arasında göreceli değişkenlik gösteren 2-B bir grafiktir. ROC eğrisi altında kalan alana AUC denir ve bir saldırı tespit sistemi için modelin saldırıyı normal durumdan ne ölçüde iyi ayırt edebildiğini gösterir.

Cohen'in kappa katsayısı: Diğer birçok değerlendirme ölçütü gibi, Cohen'in kappa'sı da karışıklık matrisine dayalı olarak hesaplanır. Bununla birlikte, genel doğruluğu hesaplamanın aksine, Cohen'in kappa'sı sınıf dağılımındaki dengesizliği hesaba katar. Kappa katsayısı -1 ile +1 arasında değişir. Eğer gözlenen uyum şansa bağlı olarak uyumdan daha büyük veya eşit olma durumu aranır. Kappa katsayısının yorumlanabilir aralığı 0 ile 1 arasındadır ve negatif değerler güvenilirlik açısından anlam taşımamaktadır. Denklem 5 Kappa puanının nasıl hesaplandığını göstermektedir. p_a modelin genel doğruluğunu p_e ise tahminler ile gerçek sınıflar arasındaki uyuşmanın ölçüsünü vermektedir.

$$K = \frac{p_a - p_e}{1 - p_e} \quad (8)$$

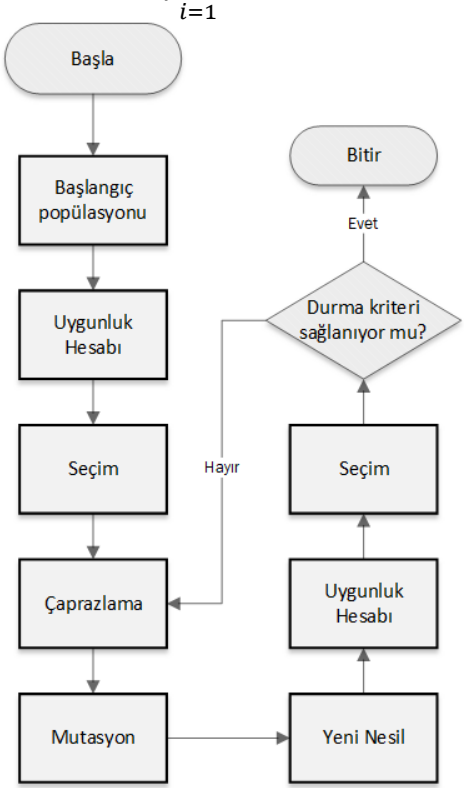
2.6 Genetik Algoritma

Genetik algoritmalar biyolojik doğal genetik kanunlarını ve doğal seleksiyon yasalarının taklit edilerek oluşturulmuş evrimsel bazlı algoritmaların bir parçasıdır. Genetik algoritmalar sezgisel optimizasyonda çokça kullanılırlar. Rastgele arama yöntemleri ile problemlere optimum çözümler üretebilirler. Popülasyon adı verilen arama kümesinden rastgele seçilen bireylerle algoritma başlatılır. Bu bireylerden başarıca üstün olanlar seçilir ve bir sonraki nesil için ebeveyn olarak kabul edilirler. Genetik algoritma ilk popülasyondan rasgele seçimler sonrasında dört adımdaki tekrarlardan oluşur. GA performans değerlendirmeleri için uygunluk fonksiyonu hesaplamaları yapılır. Aday çözümler içerisinden seçimler yapılır. Seçimler üzerinde çaprazlama ve sonrasında mutasyon gibi genetik operatör işlemleri gerçekleşir. Tüm bu prosedürler belirli bir durma kriterinin oluşması durumuna kadar tekrar eder[15]. Şekil-2 genetik algoritma akış diyagramını göstermektedir.

2.7 Ortalama Karesel Hata

Bir makine öğrenmesi modeli eğitilirken yapılan tahminler gerçek değerlerin bir miktar üzerinde ya da altında olabilir. Gerçek değer ve tahmin edilen değer birbirine eşit olmadığı müddetçe hata yapılmış demektir. Amaç beklenen değer ile tahmin edilen değer arasındaki farkın miktarının ölçülmesi ise bu değerin pozitif ya da negatif bir değer olmasının bir anlamının olmaması gerekir. Aksi durumda hatanın ölçülmesi sırasında negatiflik etkisi ortaya çıkacaktır bu durumda hatanın ölçüsünü artıracaktır. Beklenen ve tahmin edilen değerin farkı alındığında sonuç bu etkiden kurtulacaktır. Tüm hata oranlarının karelerinin ortalaması alınarak elde edilen metriğe Ortalama Karesel Hata (Mean Squared Error-MSE) denir. Denklem (9 Ortalama Karesel Hata'nın nasıl hesaplandığını göstermektedir.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad (9)$$



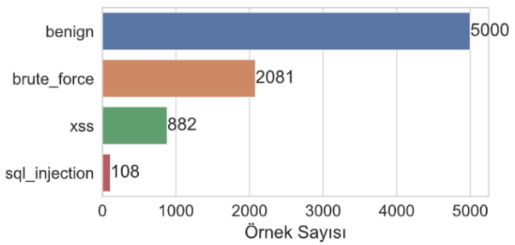
Şekil-2: Genetik Algoritma Akış Şeması

3. Önerilen Yaklaşım

Web uygulama saldırılarının tespiti için oluşturulacak makine öğrenmesi modellerinde kullanılacak veri setlerinin nitelikleri model performansı için kritik öneme sahiptir. Günümüzde açık erişimle sunulan veri setlerindeki dengesizlik farklı çalışmaların konusu olmuştur [26]. Bu çalışmada CICIDS2017 ve CSE-CIC-IDS2018 veri setlerindeki web uygulama saldırı örnekleri bir veri setinde birleştirilmiştir. Her iki veri seti de CICFlowMeter [27] kullanılarak oluşturulmuş özniteliklere sahiptir. Veri setlerinde süre, paket ve veri boyutu sayısını temel alan ağ trafiği özelliklerinden oluşan 80'den fazla öznitelik bulunmaktadır. Bu öznitelikler oluşturulurken kaynaktan hedefe ve hedeften kaynağa doğru olan çift yönlü ağ akışları kullanılmıştır. Veri setlerinin benzer özniteliklere sahip olması her iki veri setinde yer alan web uygulama saldırısı örneklerinin bir veri setinde toplanmasına imkân sağlamıştır. Yeni veri seti oluşturulurken web uygulama saldırı örneklerinin yanında her iki veri setinden de rasgele normal örnekler seçilmiştir. Veri setinin hazırlanması sırasında bir dizi veri ön işleme adımı uygulanmıştır. Web uygulama saldırılarını barındıran farklı CSV dosyaları birleştirilmiştir. Veri setlerinde tekrar eden satırlar, benzersiz değer sayısı bir olan sütunlar ve model eğitimine olumsuz etkiler bırakacak sütunlar kaldırılmıştır (flow_id, src_ip, dst_ip, timestamp vb). Veri setine saldırı ve normal durumu sayısal olarak etiketleyecek iki yeni sütun eklenmiş saldırı durumu ikili ve saldırı türleri çoklu sınıf olacak şekilde kategorilendirilmiştir. Tüm bu işlemler sonrasında veri setinde etiket sütunları hariç 67 öznitelik kalmış ve veri setinin bu hali genetik öznitelik seçiminde kullanılmıştır. Şekil-3 oluşturulan veri seti için etiket dağılımını göstermektedir. Veri seti etiket dağılımında dengesizmiş gibi görünse de temel amaç web uygulama saldırılarının tespitinde öznitelik seçimi olduğu için tüm web uygulama saldırı türleri tek etiket altında birleştirilip modeller saldırı ve normal durum olmak üzere iki etiketli şekilde işleme alınmıştır. Çizelge- 2 ön işleme sonrasında veri setinde kalan 67 özniteliği göstermektedir. Bu öznitelikler ağ akış özelliklerinde elde edilmiş istatistiksel değerlerdir. Özniteliklerin tanımlarına Kanada Siber Güvenlik Enstitüsü web sayfasından ulaşılabilir [28].

Öznitelik seçimi için önerilen yöntem GA ve LR barındıran sezgisel bir sarmalayıcı öznitelik seçimi yaklaşımıdır. En az sayıdaki en etkili özniteliklerin aranması için genetik algoritmanın kullanılması ve sonuçların değerlendirilmesi için de lojistik regresyonun genetik algoritma içerisine entegre edilmesinden oluşturulmuştur. Genetik algoritma için uygunluk fonksiyonu, oluşturulan veri seti için LR'nin doğru sınıflandırma başarısına, ortalama karesel hataya ve öznitelik sayısına bağlıdır. Rastgele seçilmiş bir öznitelik alt kümesi ile başlanarak oluşturulan veriler LR sınıflandırıcısına sunulur. Elde edilen model başarıları ve ortalama karesel hata ile ikili olarak kodlanmış özniteliklerin bir sonraki popülasyona aktarılıp aktarılmayacağı belirlenir. Seçilen öznitelik alt kümelerine çaprazlama ve mutasyon gibi genetik algoritma operatörleri uygulanır. Önerilen yöntemde genetik algoritmanın sonlandırma kriteri jenerasyon sayısıdır, yani tüm bu işlemler belirli bir sayıdaki nesil sayısınca tekrar edecektir. Şekil-4 önerilen yaklaşımın uygulama

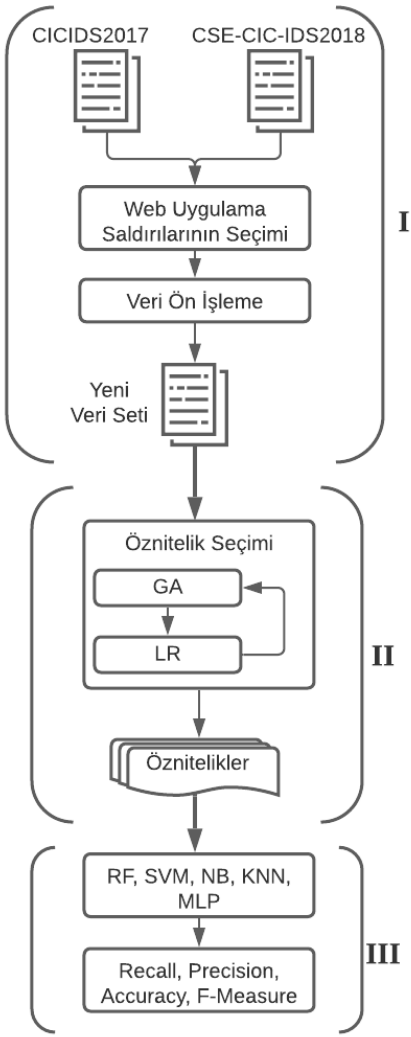
adımlarını göstermektedir. Şekil-4 'te I. bölüm veri ön-işleme ile yeni veri setinin oluşturulmasını, II. bölüm öznitelik mühendisliğini ve III. bölüm de test aşamalarını temsil etmektedir.



Şekil-3: Oluşturulan Veri Seti Örnek Sayıları

Çizelge- 2: Ön İşleme Sonrası Öznitelikler

1	ack_flag_cnt	24	flow_iat_max	47	idle_std
2	active_max	25	flow_iat_mean	48	init_bwd_win_byt
3	active_mean	26	flow_iat_min	49	init_fwd_win_byt
4	active_min	27	flow_iat_std	50	pkt_len_max
5	active_std	28	flow_pkts_s	51	pkt_len_mean
6	bwd_header_len	29	fwd_act_data_pkt	52	pkt_len_min
7	bwd_iat_max	30	fwd_header_len	53	pkt_len_std
8	bwd_iat_mean	31	fwd_iat_max	54	pkt_len_var
9	bwd_iat_min	32	fwd_iat_mean	55	pkt_size_avg
10	bwd_iat_std	33	fwd_iat_min	56	psh_flag_cnt
11	bwd_iat_tot	34	fwd_iat_std	57	rst_flag_cnt
12	bwd_pkt_len_max	35	fwd_iat_tot	58	subflow_bwd_byts
13	bwd_pkt_len_mean	36	fwd_pkt_len_max	59	subflow_bwd_pkts
14	bwd_pkt_len_min	37	fwd_pkt_len_mean	60	subflow_fwd_byts
15	bwd_pkt_len_std	38	fwd_pkt_len_min	61	subflow_fwd_pkts
16	bwd_pkts_s	39	fwd_pkt_len_std	62	syn_flag_cnt
17	bwd_seg_size_avg	40	fwd_pkts_s	63	tot_bwd_pkts
18	down_up_ratio	41	fwd_psh_flags	64	tot_fwd_pkts
19	dst_port	42	fwd_seg_size_avg	65	totlen_bwd_pkts
20	ece_flag_cnt	43	fwd_seg_size_min	66	totlen_fwd_pkts
21	fin_flag_cnt	44	idle_max	67	urg_flag_cnt
22	flow_byts_s	45	idle_mean		
23	flow_duration	46	idle_min		

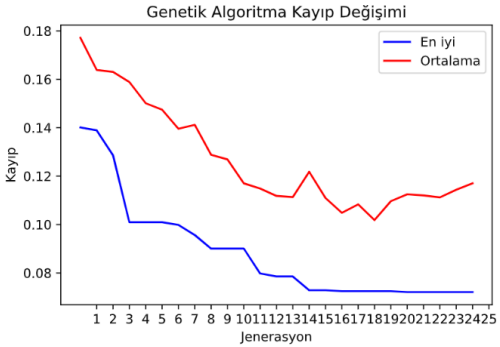


Şekil-4: Önerilen Yaklaşımın Akış Diyagramı

4. Uygulama Sonuçları

Genetik algoritma ile öznitelik seçimi en az öznitelik sayısına karşın en iyi sınıflandırma sonucuna ulaşmayı hedefleyen bir optimizasyon işlemidir. Gerçekleştirilen uygulamada genetik operatörlerden mutasyon ve çaprazlama için oranlar sırasıyla 0.05 ve 0.5 olarak seçilmiştir. Her iterasyon için 4 ebeveyn ve 40 çocuk, popülasyonu belirleyen birey sayılarıdır. Seçim işlemi rütbeleme ile gerçekleştirilmiş ve uygunluk fonksiyonu lojistik regresyon için Ortalama

Karesel Hata ile öznitelik sayısının minimuma yaklaşması ile hesaplanmıştır. Bitirme kriteri olarak iterasyon sayısı 25 olarak seçilmiştir. Şekil-5 yukarıda sayılan parametreler ile gerçekleştirilen GA uygulaması sırasında oluşan kayıp değerinin iterasyona göre değişimini göstermektedir.

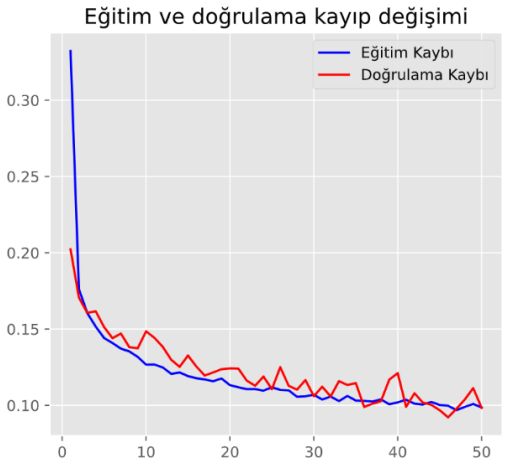


Şekil-5: Genetik Algoritma Kayıp Değişimi

Orijinal veri setleri bir dizi veri ön işleme adımında geçirilmiş, 80 olan öznitelik sayısı 67'ye düşürülmüştür. GA uygulaması sonrası da özniteliklerin sayısı 10'a düşürülmüştür. **Error! Reference source not found.** seçilen öznitelikleri göstermektedir. Bu seçilen öznitelikler ile veri setinden rastgele seçilmiş normal davranış örnekleri ve web uygulama saldırıları ile oluşturulmuş yeni veri seti RF, SVM, NB, KNN ve DNN olmak üzere beş farklı makine öğrenme algoritmasının model eğitiminde kullanılmıştır. Seçilen öznitelikler hem sığ hem de derin öğrenme algoritmaları üzerinde test edilmiştir. Şekil - **6** DNN modelinin eğitimi sırasında oluşmuş eğitim ve doğrulama verilerine ait kayıp değişimini göstermektedir. Model eğitiminde veri setinin %75'i kullanılmış geriye kalan kısım test işlemi için ayrılmıştır. Şekil-7 test verileri üzerindeki model eğitim başarılarını vermektedir. Elde edilen sonuçlara göre en başarılı yöntemler RF ve DNN algoritmaları olmuştur. Her iki algoritmada da tüm performans metriklerinde yaklaşık %99'luk bir ölçüm gözlemlenmiştir. RF'yi ve DNN'i sırasıyla KNN ve SVM yöntemleri izlemektedir. En başarısız yöntem olarak da NB belirlenmiştir.

Çizelge- 3: Seçilen Öznitelikler

1	active_max	6	fwd_iat_max
2	bwd_iat_tot	7	fwd_iat_min
3	dst_port	8	pkt_len_min
4	flow_iat_mean	9	totlen_bwd_pkts
5	fwd_act_data_pkts	10	totlen_fwd_pkts

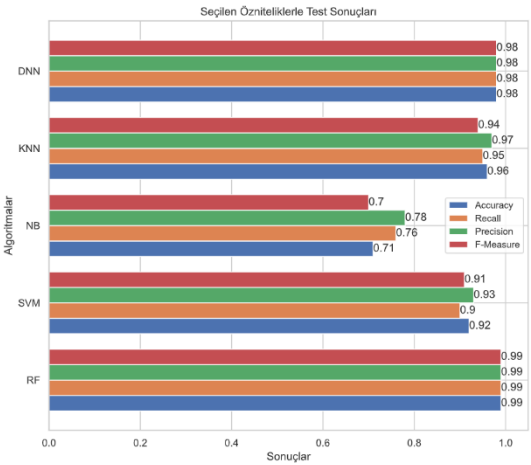


Şekil - 6: DNN eğitim ve doğrulama verileri kayıp değişimi.

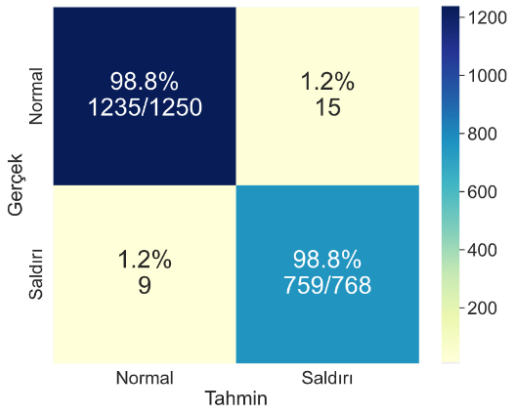
Genetik algoritmanın öznitelik seçimi sırasında veri setinde yer alan tüm web uygulama saldırı türleri tek sınıf olarak kodlanmıştır. Bu durumda Şekil-3 incelendiğinde 5000 normal ağ davranışı örneğine karşılık veri setinde 3071 saldırı örneği yer almaktadır. Şekil-8 en başarılı sonuçları veren RF algoritmasının ikili sınıflandırma ölçümleri sırasında elde edilen karmaşıklık matrisini göstermektedir. Şekil - **9** ise DNN ile elde edilen karmaşıklık matrisini göstermektedir. Önerilen yaklaşımın, çoklu sınıflandırma için gösterdiği performansı izlemek için veri setindeki saldırı türleri etiketlenerek kodlanmıştır. Veri seti bu haliyle RF yöntemine sunulmuş ve çoklu sınıflandırma için model eğitimi gerçekleştirilmiştir. Şekil-10 her saldırı türündeki örnekler ve normal ağ akış örnekleri için dört sınıflı karmaşıklık matrisini vermektedir. Buna göre modelin normal ağ akışını ayırt edilmedeki başarısı yüksektir. Bu anomali tespitindeki False Positive oranının düşük olduğunu gösterir. En iyi tespit edilen saldırı türü Brute Force saldırısıdır. XSS ve SQL Injection saldırılarının tespitindeki düşük oran saldırı örneklerinin sayısının yetersizliğinden kaynaklanmaktadır. Çizelge- 4 eğitilen tüm modelleri

test verileri üzerindeki Cohen'in Kappa puanlarını göstermektedir. Buna göre sınıflandırma da en başarılı yöntemler sırasıyla RF, k-NN ve DNN olmuştur.

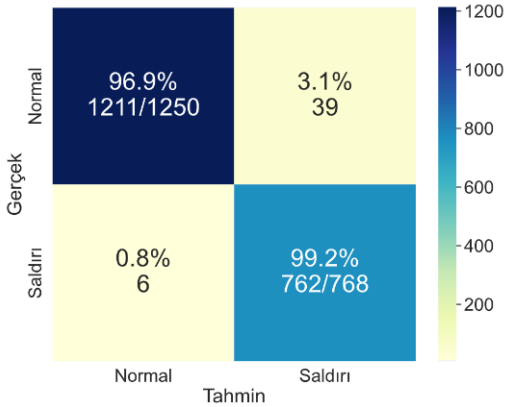
Çizelge- 5 benzer çalışmalar ile yapılan karşılaştırmaları göstermektedir.



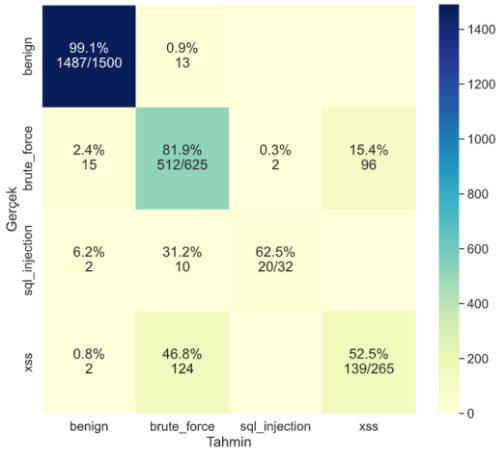
Şekil-7: Seçilen Özniteliklerin Test Sonuçları



Şekil-8: RF ile İkili Sınıflandırma Karmaşıklık Matrisi



Şekil - 9: DNN ile ikili Sınıflandırma Karmaşıklık Matrisi



Şekil-10: Çoklu Sınıflandırma Karmaşıklık Matrisi

Çizelge- 4: 10 Öznitelikle Eğitilen İkili Sınıflandırıcı Modellerinin Kappa Puanları

Sınıflandırma Algoritması	Kappa Puanı
Random Forest	0,97
Support Vevtor Machine	0,87
Naive Bayes	0,30
k-Nearest Neighbors	0,95
Deep Neural Network	0,96

Çizelge- 5: İkili Sınıflandırmada Benzer Çalışmalarla Karşılaştırmalar

Ref.	Yıl	Yöntem	Sınıflandırıcı	Öznitelik Sayısı	Başarım (Acc)
[16]	2017	Hibrit	LR	11	%83
[6]	2017	GA-LR	DT	21	%81.42
[17]	2019	PSO ve GA	TSE-IDS	19	%91.27
[18]	2020	WFEU	RF	22	%77.16
[13]	2020	GSPSO	ANN	5	%94.41
[14]	2021	GA	NB	19	%99.73
[15]	2021	GA	XgBoost	10	%99.80
[19]	2021	TS-RF	RF	16	%83.12
Bu Çalışma	2021	GA-LR-MSE	RF, DNN	10	%98.71

Sonuç

Saldırı tespit sistemlerinin etkinliğini artırmak için birçok yöntem önerilmiştir. Makine öğrenmesi çözümleri önerilen yöntemler arasında en başta gelenlerden biridir. Bu çalışmada hibrit saldırı tespit sistemlerinin, web uygulama saldırılarının tespitindeki performanslarını artırmak için makine öğrenmesi ve sezgisel öznitelik seçim yöntemlerinden GA kullanan yeni bir yaklaşım önerilmektedir. Önerilen öznitelik seçimi yaklaşımı için benzer özniteliklere sahip CICIDS2017 ve CSE-CIC-IDS2018 veri setlerinde bulunan saldırı örnekleri yeni bir veri setinde toplanmıştır. Genetik algoritma ile bu veri setinin lojistik regresyona sınıflandırma için sunulması sonrası oluşan ortalama karesel hata ve veri setindeki öznitelik sayısı minimuma yaklaştırılmıştır. Önerilen sarmalayıcı öznitelik seçim yaklaşım ile düşürülen öznitelik sayısının sınıflandırma performansına etkileri ölçülmüştür. Bu işlem için beş farklı sınıflandırma yöntemi kullanılmış ve en başarılı yöntemin hem ikili hem de çoklu sınıflandırma sonuçları incelenmiştir. Elde edilen sonuçlara göre veri setindeki öznitelikler 67'den 10'a düşürülmüştür. Seçilen öznitelikler ile oluşturulmuş veri seti ikili sınıflandırma için RF, SVM, NB, KNN ve DNN algoritmalarına sunulmuş ve bu algoritmalar içinde en başarılı sonuçları tüm performans metriklerinde %99'luk bir başarım ile RF ve DNN algoritmaları göstermiştir. Aynı veri seti çoklu sınıflandırma için de model eğitiminde kullanılmış ve elde edile karmaşıklık matrisine göre normal ağ akış örneklerinin %99,1 ile tespit edilebildiği görülmüştür. Brute Force, SQL Injection ve XSS saldırılarının tespit başarımları sırasıyla %81,9, %62.5 ve %52.5'tir.

Çizelge- 5'e göre Bu çalışmayı benzer çalışmalardan ayıran en önemli özellik web uygulama saldırılarına odaklanılması ve bu amaçla oluşturulan yeni veri setidir. Benzer çalışmalarda önerilen genetik algoritma yaklaşımları ile bu çalışmada önerilen yaklaşım arasındaki en önemli fark uygunluk fonksiyonunda kullanılan ortalama karesel hata ve öznitelik sayısının minimuma yaklaştırılmasıdır.

Bu çalışma ve bu çalışmayla benzer nitelikteki gelecek çalışmalar siber güvenlik problemlerine makine öğrenmesi temelli çözümler üretmek için motivasyon verebilir. Makine öğrenmesi yöntemleri çok boyutlu veri setlerine ihtiyaç duymaktadırlar ancak veri setlerindeki yüksek boyutluluk eğitim ve test işlemlerini karmaşık hale getirebilmektedir. Siber güvenlik alanındaki büyük ölçekli veri setlerini konu alan öznitelik mühendisliği çalışmaları bu alandaki araştırmacılar için farklı perspektifler sunacaktır.

Kaynakça

[1] K. Seyhan, T. N. Nguyen, S. Akleylek, K. Cengiz, and S. K. H. Islam, "Bi-GISIS KE: Modified key exchange protocol with reusable keys for IoT security," Journal of Information Security and Applications, vol. 58, p. 102788, May 2021, doi: 10.1016/J.JISA.2021.102788.

[2] H. Ahmetoglu and R. Das, "Derin Öğrenme ile Büyük Veri Kumelerinden Saldiri Türlerinin Siniflandirilmasi," 2019. doi: 10.1109/IDAP.2019.8875872.

[3] "IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/ids-2017.html (accessed Oct. 27, 2021).

[4] "IDS 2018 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/ids-2018.html (accessed Oct. 27, 2021).

[5] S. M. Kasongo, "Genetic Algorithm Based Feature Selection Technique for Optimal Intrusion Detection," no. June, pp. 1–22, 2021, doi: 10.20944/preprints202106.0710.v1.

[6] C. Khammassi and S. Krichen, "A GA-LR wrapper approach for feature selection in network intrusion detection," Computers & Security, vol. 70, pp. 255–277, Sep. 2017, doi: 10.1016/J.COSE.2017.06.005.

[7] Y. Zhu, J. Liang, J. Chen, and Z. Ming, "An improved NSGA-III algorithm for feature selection used in

intrusion detection," Knowledge-Based Systems, vol. 116, pp. 74–85, Jan. 2017, doi: 10.1016/J.KNOSYS.2016.10.030.

[8] H. Ahmetoglu and R. Das, "Analysis of Feature Selection Approaches in Large Scale Cyber Intelligence Data with Deep Learning," 2021. doi: 10.1109/siu49456.2020.9302200.

[9] H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," Knowledge-Based Systems, vol. 136, pp. 130–139, Nov. 2017, doi: 10.1016/J.KNOSYS.2017.09.014.

[10] H. Xu, Y. Fu, C. Fang, Q. Cao, J. Su, and S. Wei, "An improved binary whale optimization algorithm for feature selection of network intrusion detection," Proceedings of the 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems, IDAACS-SWS 2018, pp. 10–15, Nov. 2018, doi: 10.1109/IDAACS-SWS.2018.8525539.

[11] H. Gharaee and H. Hosseinvand, "A new feature selection IDS based on genetic algorithm and SVM," 2016 8th International Symposium on Telecommunications, IST 2016, pp. 139–144, Mar. 2017, doi: 10.1109/ISTEL.2016.7881798.

[12] A. Thakkar and R. Lohiya, "Role of swarm and evolutionary algorithms for intrusion detection system: A survey," Swarm and Evolutionary Computation, vol. 53, p. 100631, Mar. 2020, doi: 10.1016/J.SWEVO.2019.100631.

[13] S. Hosseini, "A new machine learning method consisting of GA-LR and ANN for attack detection," Wireless Networks, vol. 26, no. 6, pp. 4149–4162, 2020, doi: 10.1007/s11276-020-02321-3.

[14] J. O. Onah, S. M. Abdulhamid, M. Abdullahi, I. H. Hassan, and A. Al-Ghusham, "Genetic Algorithm based feature selection and Naïve Bayes for anomaly detection in fog computing environment," Machine Learning with Applications, vol. 6, no. April, p. 100156, 2021, doi: 10.1016/j.mlwa.2021.100156.

[15] Z. Halim et al., "An effective genetic algorithm-based feature selection method for intrusion detection systems," Computers and Security, vol. 110, p. 102448, 2021, doi: 10.1016/j.cose.2021.102448.

[16] N. Moustafa and J. Slay, "A hybrid feature selection for network intrusion detection systems: Central points," pp. 5–13, Jul. 2017, doi: 10.4225/75/57a84d4fbefbb.

[17] B. A. Tama, M. Comuzzi, and K. H. Rhee, "TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System," IEEE Access, vol. 7, pp. 94497–94507, 2019, doi: 10.1109/ACCESS.2019.2928048.

[18] S. M. Kasongo and Y. Sun, "A deep learning method with wrapper based feature extraction for wireless intrusion detection system," Computers & Security, vol. 92, p. 101752, May 2020, doi: 10.1016/J.COSE.2020.101752.

[19] A. Nazir and R. A. Khan, "A novel combinatorial optimization based feature selection method for network intrusion detection," Computers and Security, vol. 102, p. 102164, 2021, doi: 10.1016/j.cose.2020.102164.

[20] Ö. Kasim, "An ensemble classification-based approach to detect attack level of SQL injections," Journal of Information Security and Applications, vol. 59, p. 102852, Jun. 2021, doi: 10.1016/J.JISA.2021.102852.

[21] I. Tariq, M. A. Sindhu, R. A. Abbasi, A. S. Khattak, O. Maqbool, and G. F. Siddiqui, "Resolving cross-site scripting attacks through genetic algorithm and reinforcement learning," Expert Systems with Applications, vol. 168, p. 114386, Apr. 2021, doi: 10.1016/J.ESWA.2020.114386.

[22] A. B. Puthuparambil and J. J. Thomas, "Freestyle, a randomized version of ChaCha for resisting offline brute-force and dictionary attacks," Journal of Information Security and Applications, vol. 49, p. 102396, Dec. 2019, doi: 10.1016/J.JISA.2019.102396.

[23] D. Ö. Şahin, O. E. Kural, S. Akleylek, and E. Kılıç, "A novel Android malware detection system: adaption of filter-based feature selection methods," Journal of Ambient Intelligence and Humanized Computing 2021, vol. 1, pp. 1–15, Jul. 2021, doi: 10.1007/S12652-021-03376-6.

[24] M. DASH and H. LIU, "Feature selection for classification," Intelligent Data Analysis, vol. 1, no. 1–4, pp. 131–156, Jan. 1997, doi: 10.1016/S1088-467X(97)00008-5.

[25] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy, 2018, vol. 2018-Janua, pp. 108–116. doi: 10.5220/0006639801080116.

[26] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, "Detecting web attacks using random undersampling and ensemble learners," Journal of Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00460-8.

[27] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of tor traffic using time based features," ICISSP 2017 - Proceedings of the 3rd International Conference on Information Systems Security and Privacy, vol. 2017-Janua, pp. 253–262, 2017, doi: 10.5220/0006105602530262.

[28] "Applications | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/research/applications.html#CICFlowMeter (accessed Oct. 28, 2021).