Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 8(2) (2025) 130-140
DOI: 10.38016/jista.1635809

Nesnelerin İnterneti Saldırılarının Tespitinde Makine Öğrenmesi Algoritmalarının Performanslarının Karşılaştırılması

İsa Yıltırak[1*], Ali Öztürk[2,3]

[1]KTO Karatay Üniversitesi Lisansüstü Eğitim Enstitüsü, Konya, Türkiye

[2]KTO Karatay Üniversitesi Mühendislik ve Doğa Bilimleri Fak. Bilgisayar Müh. Bölümü, Konya, Türkiye

[3]Havelsan A.Ş., Konya, Türkiye

isayilt332@gmail.com, ali.ozturk@karatay.edu.tr, aliozturk@havelsan.com.tr

Öz

Bu çalışmada IoT cihazlarındaki çeşitli saldırı türleri için farklı makine öğrenmesi algoritmalarının sınıflandırma performansları karşılaştırılmıştır. Bu amaçla kullanılan algoritmalar AdaBoost, CatBoost, XGBoost, Karar Ağaçları, K En Yakın Komşu, Rastgele Orman, Light GBM, Lojistik Regresyon ve Gaussian Naïve Bayes'tir. CICIoT2023 veri seti kullanılarak 33 farklı saldırı tipi ve 7 farklı saldırı grubu için saldırı sınıflandırması yapılmıştır. Deneysel sonuçlara göre, makine öğrenmesi algoritmaları arasında Rastgele Orman (RF), 33 saldırı türünü sınıflandırmada %94,90 doğruluk, %94,90 kesinlik, %94,89 hatırlatma, %94,84 F1 puanı, 7 saldırı grubunu sınıflandırmada %94,33 doğruluk, %94,35 kesinlik, %94,33 hatırlatma, %94,31 F1 puanı ve ikili sınıflandırmada %96,81 doğruluk, %97,15 kesinlik, %96,81 hatırlatma, %96,79 F1 puanı oranını elde etti. RF, 33 saldırı türünü sınıflandırmada en iyi performansı gösterdi.

Anahtar kelimeler: Makine Öğrenmesi Algoritmaları, Nesnelerin İnterneti, Saldırı Tespiti, Saldırı Türleri.

Comparison of Performances of Machine Learning Algorithms in Detectiion of Internet of Things Attacks

Abstract

In this study, the classification performances of different machine learning algorithms were compared for various types of attacks in IoT devices. The algorithms used for this purpose are AdaBoost, CatBoost, XGBoost, Decision Trees, K Nearest Neighbour, Random Forest, Light GBM, Logistic Regression and Gaussian Naïve Bayes. By using the CICIoT2023 data set, attack classification was carried out for 33 different attack types and 7 different attack groups. According to the experimental results, among the machine learning algorithms, Random Forest (RF) achieved 94.90% accuracy, 94.90% precision, 94.89% recall, 94.84% F1 score in classifying 33 attack types, 94.33% accuracy, 94.35% precision, 94.33% recall, 94.31% F1 score in classifying 7 attack groups and 96.81% accuracy, 97.15% precision, 96.81% recall, 96.79% F1 score in binary classification. RF performedbest in classifying 33 attack types.

Keywords: Machine Learning Algorithms, Internet of Things, Attack Detection, Attack Types.

1. Giriş (Introduction)

Teknolojinin hızlı ilerlemesi kablosuz sensor teknolojisini kullanan, kendi verilerini herhangi bir ağ ya da internet üzerinden gönderen nesneler ve cihazları ortaya çıkarmıştır. 1991 yılında Cambridge Üniversitesi'nde çalışan akademisyenler, üst katta bulunan kahve makinesinin içerisindeki kahvenin bitip bitmediğini kontrol etmek için bir kamera sistemi

* Sorumlu yazar.
E-posta adresi: isayilt332@gmail.com

Alındı : 9 Şubat 2025
Revizyon : 9 Nisan 2025
Kabul : 23 Mayıs 2025

geliştirmişlerdir. 1999 yılında Kevin Ashton çalıştığı firma için geliştirdiği RFID (Radyo Frekansı ile Kimliklendirme) sistemi sunumu ile "Nesnelerin İnterneti" (Internet of Things- IoT) kavramını ortaya koymuştur (Bozdogan ve Kara, 2015). Gelişen ve dönüşüm geçiren teknolojiyle birlikte; günümüzde bilgiler elektronik ortama aktarılmış ve bu bilgilere cep telefonu, tablet ve çeşitli IoT cihazlarıyla ulaşılabilmesi sağlanmıştır. Böylece herhangi bir yerden internete bağlantı sağlayan IoT cihazlarının sayısı gün geçtikçe artmıştır (Thereza ve Ramli, 2023).

Günümüzde dünya genelinde 2023 yılında IoT'yi kullanan cihaz sayısının 15 milyara ulaştığı belirtilmekte ve bu sayının 2025 yılına kadar 30 milyar olması beklenmektedir (Saiyed ve Al Anbagi, 2023). Kullanılan cihaz sayısının artması ile birlikte kaydedilen, işletilen, işlenen ya da değiştirilen bilgi de artmıştır. Bunun yanında, bilgilerin korunması ve güvenliğinin sağlanması da önem kazanmıştır. Elektronik ortamlardaki bilgilerin güvenlikleri çeşitli yollarla sağlanmaktadır. Bunlara antivirüs yazılımları, güvenlik duvarları ve saldırı tespit sistemleri (Intrusion Detection System- IDS) örnek olarak verilebilir (Saiyed ve Al Anbagi, 2023).

IoT ağlarındaki anormallikleri tespit için yaygın olarak kullanılan yöntemlerin başında IDS'ler gelmektedir. IDS'ler, ağ trafiğini izler ve topladığı verileri ön işlemeden geçirerek analiz yapacağı özellikleri seçer. Seçilen özellikler ile analiz aşamasında normal ağ trafiği modeli oluşturur. Bu model kullanılarak ağdaki anormallikler tespit edilir. (Satılmış ve Akleylek, 2021). IDS'lerde kullanılan tespit tabanlı yöntemlerden birisi Anomali Tabanlı (Anomaly-Based) yöntemdir. Bu yöntem, istatistiksel modelleri, bilgi tabanlı yöntemleri veya makine öğrenmesi tekniklerini kullanıp bilgisayar sisteminin ağdaki normal ve anormal davranışlarını inceleyerek standart bir model oluşturmayı sağlar (Rahim vd., 2024).

IoT cihazlarının sayısının artması çok büyük miktarda veri üretimine neden olmuştur. Düşük hesaplama kapasitesine sahip geleneksel yöntemler büyük veriyi işlemede ve sızmaları gerçek zamanlı olarak tespit etmede yetersiz kalmaktadırlar. Makine Öğrenmesi (Machine Learning-ML) yöntemleri geçmiş verileri kullanıp doğru tahminler yapmayı sağlayarak saldırıların tespitinde ve saldırıları hafifletmede başarılı olmaktadırlar. ML yöntemleri; IoT cihazları tarafından üretilen yüksek yoğunluklu verilerden dinamik olarak öğrenebilmekte ve insan müdahalesi gerektirmeden sonuca ulaşabilmektedirler. (Yemmanuru vd., 2023).

Bu çalışmada, CICIoT2023 veri seti birçok saldırı türü ve saldırı grubu bilgilerini içerdiği için tercih edilmiştir. CICIoT2023 veri setinde veri dengesizliği sorununu aşmak için saldırı türlerine ve saldırı gruplarına göre veri sayıları eşitlenmiştir. CICIoT2023 veri setinden 33 farklı saldırı türü, 7 farklı saldırı grubu ve ikili sınıf için (saldırı var veya yok) eşit sayıda veriler alınarak yeni veri setleri oluşturulmuştur. Veri sayılarını eşitlemek için, SMOTE aşırı örnekleme algoritması kullanılmıştır. Ayrıca, veri setlerindeki özelliklerden sabit değerli olanlar korelasyon matrisi yöntemiyle elenmiştir. Daha sonra, elde edilen veri setleri ile AdaBoost, CatBoost, XGBoost, Karar Ağaçları, K En Yakın Komşu, Rastgele Orman, Light GBM, Lojistik Regresyon ve Gaussian Naïve Bayes yöntemleri üzerinde 5 katlı Çapraz Doğrulama yapılarak Doğruluk, Kesinlik, Hatırlatma ve F1-Puanı ölçütleri elde edilmiştir.

Bu çalışmanın geri kalanı şu şekilde düzenlenmiştir. Bölüm 2 'de daha önce yapılan çalışmalara yer verilmiştir. Bölüm 3' te, çalışmada kullanılan ML yöntemleri, özellik azaltma yöntemi ve CICIoT2023 veri seti hakkında bilgi verilmiştir. Bölüm 4'te, elde edilen deneysel sonuçlar açıklanmış ve daha önce yapılan çalışmalarla karşılaştırma yapılmıştır. Bölüm 5'te ise çalışmaya dair yorumlar yapılmış ve ileride yapılacak çalışmalar hakkında bilgiler verilmiştir.

2. Literatür (Literature)

ML algoritmalarıyla IoT ataklarını sınıflandırmak için çeşitli yöntemler ve çeşitli veri setleri kullanılmıştır. Kasongo (2021), büyük ölçekli verilere sahip Endüstriyel IoT (IIoT)' de Genetik Algoritma (GA) ve Ağaç Tabanlı Algoritmalara dayalı gelişmiş bir saldırı algılama sistemi önermiştir. Özellik seçimi için GA kullanarak bir IDS önermiş ve GA uygunluk puanlarını üretmek için Rastgele Orman (RF) kullanmıştır. UNSW-NB15 veri seti üzerinde özellik seçiminden sonra sırasıyla RF, Doğrusal Regresyon (LR), Karar Ağacı (DT), Extra Ağaçlar (ET) ve Aşırı Gradient Boosting (XGB) sınıflandırma algoritmalarını uygulamıştır. GA-RF yönteminde %98 doğruluk (Accuracy-AUC) sonucuna ulaşmıştır.

Samdekar vd. (2021), IoT 'de saldırı tespiti verimliliğinin artırılması konusuna yoğunlaşmışlardır. Bot-IoT veri seti üzerinde özellik seçimi için Ki-Kare (Chi Square), Ekstra Ağaç Sınıflandırıcı (Extra Tree Classifier), Temel Bileşen Analizi (PCA) ve Ateş Böceği (FA) algoritmaları kullanmışlardır. Sınıflandırıcı olarak Destek Vektör Makinelerini (SVM) uygulamışlar ve SVM+FA'nın %99.38 ile en yüksek doğruluk verdiğini bulmuşlardır.

Guo (2022), 45 özelliği bulunan TON-IoT veri setini kullanmıştır. Bu veri seti üzerinde DT, RF, ET, XGB, K En Yakın Komşu (KNN), Kategorik Artırma (CB), Gradyan Artırma (GB), Uyarlamalı Güçlendirme (AB) ML algoritmaları ve iki topluluk modelini karşılaştırmıştır. Bunun için Doğruluk, Hassasiyet, Hatırlatma, F1- puanı, Çok Merkezli Hesap (MCC) ve Alıcı İşletim Karakteristik Eğrisi Altındaki Alan (AUC) ve zaman tüketimi değerlerini kullanmıştır. Dengesiz veri kümeleri için tercih edilen MCC ölçü birimini dikkate almıştır. XGB algoritmasının hem doğruluk hem de zaman tüketimi açısından çoklu sınıflandırmada %99.17 MCC değerini ve ikili

sınıflandırmada %99.84 MCC değerini alarak en yüksek performansı gösterdiğini tespit etmiştir.

Mittal vd. (2023), CICIDS2017 ve Bot-IoT veri setlerini kullanmışlardır. Veri setlerindeki eksik ve tekrar eden değerleri kaldırmışlar, veriyi önişleme tabi tutmuşlardır. Veri setlerindeki özellik seçme işlemi için Yinelemeli Özellik Eleme (RFE) kullanmışlardır. Veri setinin %20' sini test verisine ayırdıktan sonra farklı sınıf saldırı kategorilerini dengelemek için Near Mist alt örnekleme ve SMOTE aşırı örnekleme algoritmasını kullanmışlardır. Sınıflandırma sürecinde Gaussion Naïve Bayes, DT, RF, KNN ML algoritmalarını denemişlerdir. Sonuç olarak CICIDS2017 veri setinde RF %99.3 doğruluk ve Bot-IoT veri setinde DT %99 doğruluk oranına ulaşmıştır.

Abdulkareem vd. (2022), akıllı evler, akıllı şehirler ve akıllı ulaşım bilgilerine sahip IoT ağlarında Toplu Öğrenme Algoritmalarının (ELC) performanslarını karşılaştırmışlar ve BoT-IoT veri setini kullanmışlardır. Bu işlemi gerçekleştirmek için, BoT-IoT veri setinde bulunan özelliklerin veri önişleme sırasında aynı özellikleri taşıdığını ve bazı özelliklerin aynı bilgileri depoladıklarını tespit etmişler ve bu özellikleri kaldırmışlardır. Veri setini %20 test verisi için ayırmışlar ve Min-Max ölçeğini kullanarak değerlerin 0-1 arasında normalizasyonunu sağlamışlardır. Elde ettikleri veri setinin son haline Adaboost (AB), LightGBM (LGBM), RF, Catboost (CB) ve XGB ML algoritmalarını uygulamışlardır. Karşılaştırma işlemini gerçekleştirmek için Doğruluk, Hatırlatma, Hassasiyet, F1-puanı, eğitim süresini ve test süresini kullanmışlardır. Sonuçta CB %99,99 doğruluk oranına, en kısa test ve eğitim süresi ile ulaşmıştır.

Kumar vd. (2024), IoT cihazlarına yapılan siber saldırıları sınıflandırmak ve saldırıları tespit etmede RF, SVM, AB, LR ve DT algoritmalarının etkinliğini karşılaştırmışlar ve yeni veri seti CICIoT2023' ü tercih etmişlerdir. Veri setini 34 farklı sınıfa (33 farklı saldırı türü ve normal trafik) ayırmışlar, saldırı grubunu 8 farklı sınıfa (7 farklı saldırı grubu ve normal trafik) ayırmışlar ve ikili gruba (saldırı var veya yok) ayırmışlardır. Sonuçta RF algoritması 34 farklı sınıfta %99.19 doğruluk, 8 farklı sınıfta %99.49 ve ikili sınıflandırmada %99.69 doğruluk değerine ulaşmıştır.

Neto vd. (2023); IoT ağında veri trafiğini ikili (saldırı var veya yok) sınıflandırmak için CICIoT2023 veri setini kullanarak ML ve Derin Öğrenme (DL) algoritmalarının performanslarını değerlendirmişlerdir. Bu amaçla literatürde yapılan çalışmaları inceleyerek, bu çalışmalar için deneysel ortamdan elde edilen PCAP dosyalarını inceleyerek CICIoT2023 veri setini oluşturmuşlardır. Veri setini veri önişleme sürecine sokarak boş kayıtları ve tekrar eden kayıtları temizlemişlerdir. Veri setini 34 farklı sınıf, 8 faklı sınıf ve ikili sınıflandırdıktan sonra Standard Scaler kullanarak verilere normalizasyon işlemini uygulamışlardır. Sınıflandırma işlemi için LR, Perceptron, AB, RF ve Derin Sinir Ağları (DNN) ML algoritmalarını kullanmışlardır. 34 farklı sınıflandırmada RF %99.16 doğruluk, 8 farklı sınıflandırmada RF %99.43 doğruluk ve ikili sınıflandırmada RF %99.68 doğruluk değerlerini bulmuşlardır.

Thereza ve Ramli, (2023), DDoS saldırılarını tespit etmek için CICIoT2023 veri setini kullanmışlardır. Bunun için; öncelikle veri setindeki hesaplanamayan özellikleri kaldırarak temizlemişlerdir. İkinci aşamada aralarında en az ilişki olan özellikleri kaldırmak için Ki-Kare algoritmasını kullanmışlardır. En uygun 30 özelliği seçmişlerdir. Üçüncü aşamada eğitim ve test verilerine ayrılan veri kümesine DT, RF, KNN ve NB ML algoritmalarını uygulamışlardır. DT için %100 doğruluk, KNN için %99.95 doğruluk, RF için %100 doğruluk ve NB için %93.23 doğruluk bulmuşlardır.

Kouekam ve Khennou (2024), topluluk öğrenme yöntemlerini kullanarak IoT saldırılarının tespiti için en iyi sınıflandırıcıyı bulmuşlar ve veri merkezli inceleme yaparak sınıf dengesizliği konusunu incelemişlerdir. CICIoT2023 veri seti dengesizdir. Bu sorunu çözmek için veri temizliği yapmışlar, veri ön işleme aşamasında verileri sayısallaştırarak standart hale getirmişler, verileri eğitim için %80, test için %20 olarak ayırmışlardır. Verileri dengelemek için SMOTE yöntemini kullanmışlardır. Her bir saldırı grubu için 12325 veri, normal trafik için 40000' den fazla veri seçmişlerdir. Özellik seçimi işlemi için Korelasyon Katsayısı ve Bilgi Kazanımını uygulamışlardır. AB, LGBM ve RF algoritmalarını karşılaştırmışlar ve RF'in %96 ile en yüksek doğruluğa sahip olduğunu bulmuşlardır.

Literatürdeki çalışmalara bakıldığında; Mittal vd. (2023) CICIDS2017 ve Bot-IoT veri setlerinindeki farklı sınıflara ait örnekleri eşit hale getirerek dengelemek için Near Mist ve SMOTE algoritmalarını kullanmışlardır. Kouekam ve Khennou (2024) ise, CICIoT2023 veri setinde dengeleme yapmak için SMOTE algoritmasını kullanarak her saldırı grubu için eşit sayıda örneklem elde etmişlerdir.

Bu çalışmada ise, SMOTE algoritması kullanılarak CICIoT2023 veri seti 33 farklı saldırı türüne ve 7 farklı saldırı grubuna ayrı ayrı dengelenerek yeni veri setleri oluşturulmuştur. Korelasyon matrisi kullanılarak, yeni veri setleri içerisindeki sabit değerler çıkarılmıştır. Daha sonra, veri setlerindeki özelliklere standart ölçeklendirme uygulanarak değerler standart hale getirilmiştir. Bu şekilde oluşturulan veri setleri ile aşırı uyumu engellemek için K-Katlı çapraz doğrulama yöntemi kullanılarak AdaBoost, CatBoost, XGBoost, Karar Ağaçları, K En Yakın Komşu, Rastgele Orman, Light GBM, Lojistik Regresyon ve Gaussian Naïve Bayes algoritmalarının performansları karşılaştırılmıştır.

3. Metot (Method)

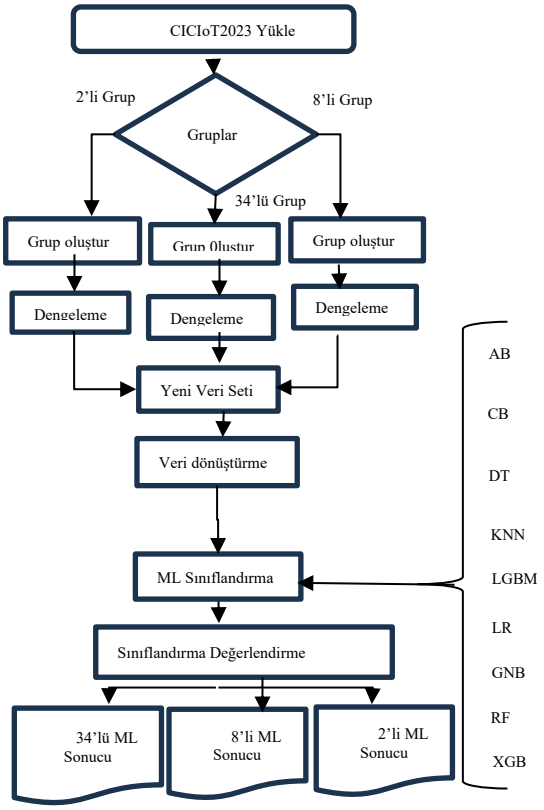
Bu çalışmada CICIoT2023 veri setini kullanmamızın etken sebepleri; 105 farklı cihazdan elde edilmesi, 33 farklı saldırı türü ve 7 farklı saldırı

grubunu içermesi, veri setinin PCAP dosyalarından oluşturulurken literatürdeki farklı kaynaklardan yararlanılarak oluşturulmuş olmasıdır. Önerdiğimiz metodoloji ile, veri setindeki saldırı türleri ve saldırı grupları için daha dengeli veri setleri oluşturarak ML algoritmalarının performanslarını artırmak amaçlanmıştır.

3.1. Metodoloji (Methodology)

Önerilen modelin mimari yapısı Şekil 1' de gösterilmiştir. CICIoT2023 veri setinden 2'li sınıflandırma, 8'li sınıflandırma ve 34'lü sınıflandırma için veriler alınarak gruplar oluşturulur. Daha sonra veri önişleme gerçekleştirilir.

Veri ön işleme sürecinde; orijinal veri setinden, saldırı türleri ve saldırı gruplarına göre yeni veri setleri oluşturulmuştur. Bunun için, SMOTE aşırı örnekleme algoritması ve korelasyon matrisi yöntemleri kullanılmıştır. SMOTE aşırı örnekleme yöntemi veri kümesindeki saldırı türlerinin ve saldırı gruplarının sayılarını eşitler. Veri setleri içerisindeki sabit değerleri çıkarmak için ise korelasyon matrisi kullanılmıştır. Böylece, saldırı türlerine ve saldırı gruplarına göre veriler dengelenmiştir. Daha sonra, veri seti özelliklerine standart ölçeklendirme uygulanarak standart hale getirilmiştir. 5 katlı çapraz doğrulama yapılarak ML algoritmalarıyla sınıflandırılır.



Şekil 1. Önerilen saldırı algılama mimarisi (Proposed attack detection architecture)

3.2. CICIoT2023 veri seti (CICIoT2023 data set)

Kanada Siber Güvenlik Enstitüsü (CIC) tarafından oluşturulmuş gerçek zamanlı bir veri setidir ve bu veri setindeki veriler 105 IoT cihazından elde edilmiştir. Burada 33 farklı saldırı türü ve normal trafik olmak üzere 34 farklı sınıf bulunmaktadır. Veri seti, toplam otuz üç saldırı türü, yedi farklı saldırı grubu ve normal trafik (Benign Traffic) olarak birleştirilmiştir. PCAP ve CSV formatındaki CICIoT2023 veri seti dosyalarına https://www.unb.ca/cic/datasets/iotdataset-2023.html adresinden erişilebilmektedir (Neto vd., 2023).

Tablo 1. CICIoT2023 veri seti (CICIot2023 data set)

Saldırı Grubu	Saldırı Türü	Satır sayısı
DDoS	ACK Fragmentation	285 104
	UDP Flood	5412287
	SlowLoris	23 426
	ICMP Flood	7200504
	RSTFIN Flood	4045285
	PSHACH Flood	4094755
	HTPP Flood	28 790
	UDP Fragmentation	286 925
	ICMP Fragmentation	452 489
	TCP Flood	4 497 667
	SYN Flood	4 059 190
	SynonymousIP Flood	3 598 138
DoS	TCP Flood	2 671 455
	HTPP Flood	71 864
	SYN Flood	2 028 834
	UDP Flood	3 318 595
Recon	Ping Sweep	2 262
	OS Scan	98 259
	Vulnerability Scan	37 382
	Port Scan	82 284
	Host Discovery	134 378
Web-Based	Sql Injection	5 245
	Command Injection	5 409
	Backdor Malware	3 218
	Uplaoding Attack	1 252
	XSS	3 846
	Browser Hijacking	5859
Brute Force	DictionaryBruteForce	13 064
Spoofing	Arp Spoofing	307 593
	DNS Spoofing	178 911
Mirai	GREIP Flood	751 682
	Greeth Flood	991 866
	UDPPlain	890 576
BenignTraffic	BenignTraffic	1 098 195
Total number of rows		46 686 579

3.3. Makine öğrenmesi algoritmaları (Machine learning algorithms)

Karar ağaçları (Decision Trees –DT) , makine öğrenmesi yöntemleri arasında sınıflandırma işlemlerinde ilk akla gelen algoritmalardandır ve yaygın olarak kullanılmıştır. DT algoritması, verileri çeşitli şekilde parçalara ayırır. Veri kümesindeki çeşitli özelliklerin varyansını bularak sınıflandırma işlemlerini gerçekleştirir. En yüksek varyansa sahip veriyi kök düğüme atar. Verileri kök düğümden başlayarak ağaç yapısına benzer şekilde belli bir kurala göre sıralanmış şekilde düzenler (Thereza ve Ramli, 2023).

Adaboost algoritması, DT algoritmalarının performansını artırmak için geliştirilmişitir. Zayıf öğrenenleri bir kümeye alarak sırasıyla ekler. Eğitim aşamasında bu zayıf kümeleri iyileştirme yaparak performansını artırır. Eğer yine de zayıf öğrenen varsa tekrar ve tekrar eğitime alarak en iyi performansı sağlar (Sibai vd., 2023).

GPU desteği ve görselleştirme özelliklerine sahip olan CatBoost; ayrı verileri yeterli ve verimli şekilde analiz eder. Bu şekilde gradyan önyargısı ve tahmin kayması sorunlarını çözerek tahmin doğruluğunu daha iyi bir duruma getirir. Aşırı uyum sorununa çözüm getirir(Abdulkareem vd., 2022).

K En Yakın Komşu algoritması, benzer örnek verileri bir araya toplayan bir algoritmadır. Öklid, Manhattan ve Minkowski mesafe ölçütlerini kullanarak mesafeleri hesaplar. Dezavantajı ise hesaplamada tüm durumları değerlendirir ve güç tüketimini artırır (Guo, 2022).

Ligth GBM (LGBM), ağaç tabanlı algoritmaların bir çeşididir. Özellikle büyük veri kümeleriyle sınıflandırma yaparken verimli bir şekilde muhteşem bir performans gösterir. LGBM bir önceki yaptığı hataları düzeltmek amacıyla bir grup karar ağacı oluşturur. LGBM yaprak bazında ağaç büyümesi adı verilen bir yöntem kullanarak büyümeyi hedefler (Kouekam ve Khennou, 2024).

Lojistik Regresyon algoritması, belirli bir girdinin hangi sınıfa ait olma olasılığını tahmin eder. İstatiksel bir model olarak hizmet eder ve sınıflandırma işlemlerinde özellikle ikili sınıflandırmalarda daha çok kullanılırlar(Kumar vd., 2024) .

Gaussian Naive Bayes, olasılık tabanlı bir algoritmadır. NB olasılıkları, önceki sınıf olasılıklarını ve tahmin edici önceki olasılıkları hesaplar ve bunları sözlükte depolar. Algoritma test verilerini önceden işler. Olasılık sözlüğünü, sınıfların önceki olasılıklarını ve eğitim modülünün tahmin edicilerini kullanarak en yüksek olasılığa sahip sınıfları işaretler (Mittal vd., 2023).

Rasgele Ağaçlar (Random Forests – RF), regresyon ve sınıflandırma modellemelerinde diğer ağaç algoritmalarından daha iyi bir performans ortaya koymaktadır. Çeşitli veri setlerini mükemmel bir şekilde işleyen RF, ölçeklenebilirliği ve çok yönlü olması tahmin hassasiyetini artırırken, aşırı uyum sorununu en aza indirir (Happy vd., 2024).

XGBoost, özellikle tablolu verilerle çalışırken verimliliği ve etkinliğiyle tanınan, eğitim performansını artırmak için geliştirilmiş bir algoritmadır. Veri bilimi ve pratik uygulamalarda hızlı olması, taleplere uygun şekilde cevap vermesi ve performansı nedeniyle tercih edilmektedir. XGBoost'un temelinde güçlü ve güvenilir bir tahmin modeli oluşturmak için genellikle karar ağaçları olmak üzere birden fazla zayıf tahmin modelini bir araya getirerek eğitim artırma tekniklerini kullanır. Modeli sistematik oluştururken her yeni model üzerinde bunu uygulayarak hata ve eksiklikleri giderir (Manchala vd., 2023).

3.4. Değerlendirme metrikleri (Evaluation metrics)

ML algoritmalarının modellemelerini göstermek, bilgi sahibi olmak amacıyla verileri görselleştirerek fikir vermek için kullanılan matris Karmaşıklık Matrisi olarak adlandırılmaktadır. Tablo 2 Karmaşıklık Matrisini açıklamaktadır (Sahu, ve Mukherjee, 2020).

Tablo 2. Karmaşıklık matrisi (Confusion matrix)

		Gerçek değerler	
		Pozitif sayısı	Negatif sayısı
Tahmin edilen değerler	Pozitif değerler	Doğru pozitif (TP)	Yanlış pozitif (FP)
	Negatif değerler	Yanlış negatif (FN)	Doğru negatif (TN)

Doğru Pozitif (TP): Değerlerin içerisinden doğru sınıflandırılanların sayısını vermektedir.
Yanlış Pozitif (FP): Değerlerin içerisinden yanlış olarak doğru diye sınıflandırılanların sayısıdır.
Yanlış Negatif (FN): Değerlerin içerisinde yanlışlıkla negatif olarak sınıflandırılanların sayısıdır.
Doğru Negatif (TN): Değerlerin içesinden doğru sınıflandırılan negatif olanların sayısıdır.
Bu bilgiler doğrultusunda aşağıdaki doğruluk, kesinlik, hatırlatma ve F1 puanı hesaplamaları yapılmaktadır.

Doğruluk, kesinlik, hatırlatma ve F1_puanı ölçütlerinin nasıl hesaplandığı sırasıyla Denklem 1-4'te verilmiştir.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Kesinlik = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Hatırlatma} = \frac{TP}{TP+FN} \quad (3)$$

$$F1\_Puanı = 2 * \frac{Kesinlik*Hatırlatma}{Kesinlik+Hatırlatma} \quad (4)$$

Doğruluk, model tarafından yapılan ölçüm sonuçlarının gerçek değere ne kadar yakın olduğunu gösterir. Kesinlik, toplam pozitif sayıları içerisinden saldırıları doğru tanımlama kapasitesidir. Hatırlatma,

sınıflandırma yaparken gerçek saldırılardan saldırıları tahmin etmek için kullanılır. F1-Puanı, hatırlatma ve kesinlik değerlerinin harmonik ortalamasıdır. (Thereza ve Ramli, 2023).

3.5. Veri önişleme (Data preprocessing)

CICIoT2023 veri seti CSV formatında 01.02.2024 ile 01.05.2024 tarihleri arasında https://www.unb.ca/cic/datasets/iotdataset-2023.html adresinden indirilmiştir. Veri setini, Pcap dosyasından 46 özelliğini seçerek oluşturmuşlardır. Bu özellikleri, yapılan araştırmalar sonucunda elde edilen bilgilere göre çıkarmışlardır (Neto vd., 2023). Veri seti içerisindeki eksik ve boş kayıtları temizleyerek kullanılması için son şeklini vermişlerdir. Tablo 1'de veri setindeki 7 saldırı grubu altındaki toplam 33 farklı saldırı türünün satır sayıları verilmiştir. Bazı saldırı türlerinin satır sayıları arasında çok büyük farklar vardır. Hatta saldırı gruplarının içerisindeki saldırı türleri arasında farklar görülmektedir. Mesela DDoS saldırı grubunda 12 tane saldırı türü varken Brute Force saldırı grubunda bir tane saldırı türü vardır. Bu nedenlerden dolayı CICIoT2023 veri seti dengesiz bir veri setidir. Dengesiz bir veri seti ML algoritmalarının sınıflandırma işlemlerinde modelin hatalı sonuç vermesine neden olabilmektedir. CICIoT2023 veri setindeki bilgiler 34'lü (33+1 iyi huylu) sınıflandırma, 8'li (7+1 iyi huylu) sınıflandırma ve ikili (1+1) sınıflandırma olarak üç faklı kategoriye ayrılmıştır. Ikili sınıflandırma normal trafik (Benign Traffic) 1098195 tane satıra sahiptir. Normal trafik satır sayısını 33 farklı saldırıya paylaştırırsak her bir saldırı türü için yaklaşık 33280 tane satıra ihtiyaç vardır. Her bir saldırı türünden 33300 tane örnek seçilmiştir. 33300 örnekten az sayıda olan saldırı türleri vardır. Bu saldırı türlerinin sayısını eşitleyebilmek için Smote Aşırı Örnekleme Algoritması kullanılmıştır (Mittal vd., 2023).

Tablo 3. CICIoT2023 dengeli veri seti (CICIoT2023 balanced dataset)

Saldırı Grubu	Saldırı Türü	Satır sayısı
DDoS	ACK Fragmentation	33300
	UDP Flood	33300
	SlowLoris	33300
	ICMP Flood	33300
	RSTFIN Flood	33300
	PSHACH Flood	33300
	HTPP Flood	33300
	UDP Fragmentation	33300
	ICMP Fragmentation	33300
	TCP Flood	33300
	SYN Flood	33300
	SynonymousIP Flood	33300
DoS	TCP Flood	33300
	HTPP Flood	33300
	SYN Flood	33300
	UDP Flood	33300
Recon	Ping Sweep	33300
	OS Scan	33300
	Vulnerability Scan	33300
	Port Scan	33300
	Host Discovery	33300
Web-Based	Sql Injection	33300
	Command Injection	33300
	Backdor Malware	33300
	Uplaoding Attack	33300
	XSS	33300
	Browser Hijacking	33300
Brute Force	DictionaryBruteForce	33300
Spoofing	Arp Spoofing	33300
	DNS Spoofing	33300
Mirai	GREIP Flood	33300
	Greeth Flood	33300
	UDPPlain	33300
BenignTraffic	BenignTraffic	33300
Total number of rows		1132200

Tablo 3 ile 33 faklı saldırı türü ve normal trafik ile oluşan 34 sınıfı sınıflandırmak için hazırlanmış veri seti bilgileri verilmiştir.

Tablo 4. CICIoT2023 8 farklı sınıf dengelenmiş veri seti (CICIoT2023 8 different class balanced dataset)

Saldırı Grubu	Saldırı Türü	Satır Sayısı
DDoS	ACK Fragmentation	5453
	UDP Flood	5618
	SlowLoris	5460
	ICMP Flood	5408
	RSTFIN Flood	5610
	PSHACH Flood	5717
	HTPP Flood	5500
	UDP Fragmentation	5629
	ICMP Fragmentation	5559
	TCP Flood	5527
	SYN Flood	5569
	SynonymousIP Flood	5550
DoS	TCP Flood	16588
	HTPP Flood	16780
	SYN Flood	16562
	UDP Flood	16670
Recon	Ping Sweep	13287
	OS Scan	13158
	Vulnerability Scan	13486
	Port Scan	13327
	Host Discovery	13342
Web-Based	Sql Injection	10995
	Command Injection	11141
	Backdor Malware	11177
	Uplaoding Attack	11149
	XSS	11085
	Browser Hijacking	11053
Brute Force	DictionaryBruteForce	66600
Spoofing	Arp Spoofing	33248
	DNS Spoofing	33352
Mirai	GREIP Flood	22469
	Greeth Flood	22014
	UDPPlain	22117

BenignTraffic	BenignTraffic	66600
Total number of rows		532800

Tablo 4'te her bir saldırı grubundan 66600 tane veri seçilerek 8 farklı sınıf için oluşturulmuş dengeli veri setine ait satır sayıları verilmiştir. Örneğin DDoS saldırı grubuna ait 12 farklı saldırı türü vardır. Bu saldırı türlerine ait örnek sayıların toplamı 66600 tane bilgiden oluşmaktadır. Burada Brute Force saldırı grubundan bir tane saldırı çeşidi (Dictionary Brute Force) olduğu için 66600 tane örnek alınmıştır. Diğer saldırı grupları ve normal trafik sayıları bu şekilde oluşturulmuştur. Bu yöntemle saldırı grupları içerisinde bulunan saldırı türleri arasında daha dengeli bir veri dağılımı oluşturulmuştur.

Tablo 5. CICIoT2023 2 farklı sınıf dengelenmiş veri seti (CICIoT2023 2 different class balanced dataset)

Saldırı Grubu	Saldırı Türü	Satır Sayısı
DDoS	ACK Fragmentation	33300
	UDP Flood	33300
	SlowLoris	33300
	ICMP Flood	33300
	RSTFIN Flood	33300
	PSHACH Flood	33300
	HTPP Flood	33300
	UDP Fragmentation	33300
	ICMP Fragmentation	33300
	TCP Flood	33300
	SYN Flood	33300
	SynonymousIP Flood	33300
DoS	TCP Flood	33300
	HTPP Flood	33300
	SYN Flood	33300
	UDP Flood	33300
Recon	Ping Sweep	33300
	OS Scan	33300
	Vulnerability Scan	33300
	Port Scan	33300
	Host Discovery	33300
Web-Based	Sql Injection	33300
	Command Injection	33300
	Backdor Malware	33300
	Uplaoding Attack	33300
	XSS	33300
	Browser Hijacking	33300
Brute Force	DictionaryBrute Force	33300
Spoofing	Arp Spoofing	33300
	DNS Spoofing	33300
Mirai	GREIP Flood	33300
	Greeth Flood	33300
	UDPPlain	33300
Benign Traffic	BenignTraffic	1098195
Total number of rows		2 197 095

Tablo 5'te 33 farklı saldırı türünün her birinden 33300 tane veri alınarak toplam 1098900 saldırı veri ile CICIoT2023 veri setinde bulunan tüm normal trafik (Benign Traffic) birleştirilerek ikili sınıflandırma için dengelenmiş bir veri seti oluşturulmuştur.

Oluşturulan veri setlerine Korelasyon Matrisi uygulanmıştır (Thereza ve Ramli, 2023). Buna göre, sabit değerler; 34 faklı sınıfta "Telnet ve DHCP", 8 farklı grupta "Telnet, SMTP, IRC ve DHCP" ve ikili sınıflandırmada "Telnet ve DHCP" olarak tespit edilmiş, bu değerler kaldırılmıştır. Veri setlerinde bulunan verilerin daha iyi işlenebilmesi için kategorik sayısal tür dönüşümleri Label Encoder ile yapılmıştır (Bala ve Behal, 2024). Veri normalizasyonu işlemi için Standart Scaler ölçeği kullanılmıştır (Elaziz vd., 2024). Yapılan değişiklikler ile veriler daha tutarlı hale getirilmiş, ML algoritmalarının sınıflandırma işlemlerini daha hızlı gerçekleştirmesi sağlanmıştır. Standart Scaler'ı hesaplamak için kullanılan formül Denklem (5)'te verilmiştir (Elaziz vd., 2024).

$$S' = \frac{S-\mu}{\sigma} \quad (5)$$

S : Anlık değer
μ : Ortalama değer
σ : Standart sapma
S' : Ölçeklendirilmiş değer

K katlı çapraz doğrulama ML modelini değerlendirmede, kullanılan modelin veri kümesi üzerinde aşırı uyumu azaltan ve genellemeyi yaygınlaştırarak veri değerlendirmesine katkıda bulunan bir yöntemdir. Model birden fazla ve farklı test kümesine bölündüğünden K Katlı Çapraz Doğrulama, uygulanan modelin performansını daha objektif şekilde değerlendirmesine katkıda bulunur (Sadhwani vd., 2024). ML algoritmalarıyla sınıflandırma işlemini gerçekleştirmeden önce oluşturduğumuz veri setlerine 5 katlı çapraz doğrulama yapılmıştır.

4. Uygulama Sonuçları (Application Results)

Bu bölümde, deneylerin gerçekleştirildiği çalışma ortamından, ML algoritma sonuçları ve diğer çalışmalardan bahsedilmektedir.

4.1. Çalışma ortamı (Working environment)

Deneylerimizi gerçekleştirdiğimiz çalışma ortamında Python programlama dili kullanılmıştır. Python programlama dili; Anaconda çalışma platformunun Spider IDE arayüzünde gerçekleştirilmiştir. Python ile çalışmalarımızı gerçekleştirirken Scikit-learn, Pandas ve MatPlotLib kütüphaneleri kullanılmıştır. Bilgisayar özellikleri, 11. nesil corei5 2.42 Ghz işlemci, DDR4 3200 Mhz 20 Gbyte ram, Windows 11 Home 64-bit işletim sistemi ve 1Gbyte İntel iRIS paylaşımlı ekran kartı kullanılmıştır.

4.2. ML algoritmalarında kullanılan parametreler (Parameters used in ML algorithms)

Bu çalışmada kullanılan ML algoritmalarına ait parametreler ve değerleri Tablo 6'da verilmiştir.

Tablo 6. ML algoritmalarının kullanılan parametreleri (Parameters used in ML algorithms)

Algoritmalar	Parametreler
AB	estimator=DecisionTreeClassifier(max_depth=2), n_estimators=100, learning_rate=1.0, random_state=42, algorithm='SAMME'
CB	task type="CPU", learning_rate: 0.03, iterations=1000, depth=6, l2_leaf_reg=4, border_count=254, objective:None,depth:6,num_leaves:31,custom_metric:None,reval_metrics:'Logloss', random_state:None, l2_leaf_reg:3, subsample:1.0, use-best-model :True, nan-mode :'Min',verbose:100, scale_pos_weight:1.0
DT	criterion="entropy", splitter="best", max_depth=None, min_samples_split=2, min_samples_leaf=5, max_features=None,random_state=42,class_weight="balanced", min_density=None, compute_importances=None, max_leaf_nodes=None
KNN	n_neighbors=7, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski',metric_params=None, n_jobs= -1
LGBM	lambda_l1 :0.1, lambda_l2:0.1, n_estimators:150,learning_rate:0.05, boosting_type:'gbdt', colsample:0.7, bytree:1.0, max_depth=-1, subsample:0.7, min_child_weight=1, min_split_gain=0.0, n_jobs=-1, num_leaves=33, subsample_for_bin=1000, num_iterations :100, num_threads :-1, device_type :'CPU', min_data_in_leaf :25, bagging_fraction :1.0, bagging_freq :0, feature_fraction :1.0, max_delta_step:0.0, max_cat_to_onehot :4, cat_smooth :10.0, cat_l2 :10.0, verbosity :1, max_bin :255, zero_as_missing :false, boost_from_average :true.
LR	penalty='l2', dual=False, tol=0.0001, C=0.5, fit_intercept=True, intercept_scaling=1, class_weight="balanced", random_state=42, solver='lbfgs', max_iter=1000, multi_class='deprecated', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None
GNB	priors=None, var_smoothing=1e-08
RF	n_estimators=200, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None, monotonic_cst=None
XGB	learning_rate:0.3, n_estimators:200, objective:"multi:softmax", max_depth:6, subsample:1, booster:'gbtree', enable_categorical=False,gamma:0,verbosity:1,,min_child_weight:1,validate_parameters:false,eval_metric:'logloss', updater:'all' ,nthread:none,max_delta_step:0 , colsample_bytree:1, tree_method':auto', scale_pos_weight:1, grow_policy:' depthwise', max_leaves:0, max_bin:256, num_parallel_tree:1, seed:0,

4.3. Deneysel Sonuçlar (Experimental Results)

Tablo 3, Tablo 4 ve Tablo 5 ile gösterilen veri setlerindeki veriler Denklem (5)' e göre standartlaştırılmıştır. Daha sonra, 5-Katlı çapraz doğrulama ile ML algoritmaları kullanılarak sınıflandırma yapılmıştır.

Tablo 7' deki sonuçları elde etmek için Tablo 3' teki CICIoT2023 34 sınıf için dengelenmiş veri seti kullanılmıştır. Tablo 7'de IoT cihazlarına yapılan 34 sınıf çoklu sınıflandırma için 9 farklı ML algoritmalarının doğruluk, kesinlik, hatırlatma ve F1 puanı değerlendirme metrikler verilmiştir. 34 sınıf için çoklu sınıflandırma sonuçlarını incelediğimizde, RF algoritması 9 farklı ML algoritması arasında en iyi %94.90 doğruluk, %94.90 kesinlik, %94.89 hatırlatma ve %94.84 F1-puanına ulaşmıştır.

Tablo 8' deki sonuçları elde etmek için Tablo 4' teki CICIoT2023 8 farklı sınıf için dengelenmiş veri seti kullanılmıştır. Tablo 8'de, 8 farklı sınıf grubu sınıflandırmada 9 farklı ML algoritmalarının doğruluk, kesinlik, hatırlatma ve F1 puanı değerlendirme metrikleri verilmiştir. Tablo 8'deki sonuçlara göre, RF ML algoritması %94.33 doğruluk, %94.35 kesinlik, %94.33 hatırlatma ve %94.31 F1 puanına sahip olduğu görülmüştür.

Tablo 9'da ki sonuçları elde etmek için Tablo 5' teki CICIoT2023 2 sınıf için dengelenmiş veri seti kullanılmıştır. Tablo 9'deki sonuçlara göre, RF ML algoritmasının %96.81 doğruluk, %97.15 kesinlik, %96.81 hatırlatma ve %96.79 F1 puanına sahip olduğu görülmüştür. Ayrıca Tablo 8'ye göre AB ML algoritmasının Tablo 9'daki sonuç performansının belirgin bir şekilde arttığı görülmüştür.

Tablo 7. 34 sınıf çoklu sınıflandırma sonuçları (34 class multi-class classification results)

Değerlendirme Metrikleri	AB	CB	DT	KNN	LGBM	LR	GNB	RF	XGB
Doğruluk	%72.06	%89.82	%91.62	%81.17	%92.17	%60.06	%47.91	**%94.90**	%93.82
Kesinlik	%77.26	%90.38	%91.67	%81.01	%92.38	%61.06	%58.45	**%94.90**	%93.99
Hatırlatma	%72.06	%89.82	%91.62	%81.17	%92.17	%60.06	%47.91	**%94.89**	%93.82
F1-Puanı	%71.64	%89.84	%91.62	%80.84	%92.15	%58.76	%43.19	**%94.84**	%93.82

Tablo 8. 8 farklı sınıf grubunu sınıflandırma sonuçları (Classification results of 8 different class groups)

Değerlendirme Metrikleri	AB	CB	DT	KNN	LGBM	LR	GNB	RF	XGB
Doğruluk	%79.58	%91.96	%91.69	%80.65	%93.09	%66.34	%32.70	%94.33	%94.34
Kesinlik	%80.62	%92.05	%91.69	%80.81	%93.13	%67.72	%41.12	%94.35	%94.35
Hatırlatma	%79.58	%91,96	%91.69	%80.65	%93.09	%66.34	%32.70	%94.33	%94.34
F1-Puanı	%70.60	%91.94	%91.68	%80.61	%93.07	%65.94	%25.89	%94.31	%94.32

Tablo 9. İkili sınıf sınıflandırma sonuçları (binary class classification results)

Değerlendirme Metrikleri	AB	CB	DT	KNN	LGBM	LR	GNB	RF	XGB
Doğruluk	%93.67	%96.04	%95.31	%91.00	%95.72	%84.39	%66.75	%96.81	%96.98
Kesinlik	%94.34	%96.48	%95.70	%92.31	%96.14	%85.76	%79.42	%97.15	%97.28
Hatırlatma	%93.67	%96.04	%95.31	%91.00	%95.72	%84.39	%66.76	%96.81	%96.98
F1-Puanı	%93.62	%93.62	%95.28	%90.81	%95.70	%84.12	%62.00	%96.79	%96.96

4.4. CICIoT2023 veri seti ile diğer çalışmaların karşılaştırılması (Comparison of CICIoT2023 dataset with other studies)

CICIoT2023 veri setinin bu çalışmada seçilmesinin nedeni; 105 farklı cihazdan elde edilmesi, 33 farklı saldırı türü ve 7 farklı saldırı grubunu içermesi, veri setinin PCAP dosyalarından oluşturulurken literatürdeki farklı kaynaklardan yararlanılarak oluşturulmuş olmasıdır.

Neto vd. (2023), çalışmalarında RF 34 sınıf çoklu sınıflandırmada %99.16 doğruluk, 8 farklı sınıf grubu sınıflandırmada %99.43 doğruluk ve ikili sınıflandırmada %99.68 doğruluk bulmuşlardır. Fakat çalışmada verileri dengelememişlerdir.

Thereza ve Ramli, (2023), çalışmalarında DDoS saldırılarını tespit etmek için yaklaşık 23412707 tane örnek kullanmışlardır. RF ve DT %100 doğruluk oranına ulaşmışlardır. Diğer saldırı çeşitleriyle ilgili bir çalışma yapmamışlardır.

Kumar vd. (2024), çalışmalarında 34 farklı sınıf için DT %99.19 doğruluk, 8 farklı sınıf için RF %99.45 doğruluk ve ikili sınıflandırma için RF %99.68 doğruluk oranına ulaşmışlardır. Fakat verileri dengelemek için yaptıkları çalışmalardan ve kullandıkları veri sayısından bahsetmemişlerdir.

Kouekam ve Khennou. (2024), çalışmalarında AB, LGBM ve RF ML algoritmalarını çeşitli özellik seçme teknikleriyle kullanmışlardır. İkili sınıflandırmada veri dengesizliği çalışmaları yapmışlardır. AB %94 doğruluk, LGBM %96 doğruluk ve RF %96 doğruluk sonuçlarına ulaşmışlardır. Saldırı gruplarını dengelemeye çalışmışlardır. Fakat saldırı grupları içerisinde bulunan saldırı türleri için bir dengelemeden bahsetmemişlerdir.

Saif vd. (2024), çalışmalarında 34 farklı sınıf için RF %99.42 doğruluk, 8 farklı sınıf için %99.60 doğruluk ve ikili sınıflandırma için %99.73 doğruluk oranına ulaşmışlardır. Toplam 19598221 örnek kullanmışlardır. Fakat verilerin dengelenmesi için gerekli çalışmalardan bahsetmemişlerdir.

Çalışmamızda veri dengesi sorununu çözmek için saldırı türü ve saldırı gruplarına göre veri sayısı ayarlanmıştır. Böylece hem veri setinin bütününde hem de veri seti saldırı grupları arasındaki veri paylaşımı dikkate alınmıştır. Daha dengeli bir veri seti ortaya çıkarılmıştır. Bu veri setleri Tablo 3, Tablo 4 ve Tablo 5' te verilmiştir. Diğer çalışmalarda sadece saldırı türüne ya da sadece saldırı grubuna göre genel veri dengelenmesi yapılmıştır. Ama saldırı gruplarının kendi içerisinde özel bir veri dengeleme işlemi yapılmamıştır. Örneğin Tablo 1'i incelediğimizde DDoS saldırı grubunda 12 tane saldırı türü vardır. Bu saldırı türleri arasında ICMP Flood 7200504 tane veriye sahiptir ve SlowLorish 23406 tane veriye sahiptir. Saldırı türleri sınıflandırılırken, her bir saldırı türünden kaçar tane örnek olduğu, seçilen örnek sayılarına göre saldırı gruplarının kendi içerisindeki saldırı türleri arasında nasıl bir dağılım olduğu sorularına, yapılan diğer çalışmalarda çözüm sunulmamıştır. Çalışmamızda ise, bu sorulara Tablo 3, Tablo 4 ve Tablo 5' te saldırı türü ve saldırı gruplarına ait satır sayıları verilerek açıklık getirilmiştir.

5. Sonuçlar (Conclusions)

Bu makalede IoT ağlarında yapılan siber saldırılara karşı ML algoritmalarının performansları karşılaştırılmış ve değerlendirilmiştir. Birçok veri kümesi incelenmiş ve CICIoT2023 veri seti, en geniş

saldırı türlerini içerdiği için seçilmiştir. CICIoT2023 veri seti dengesiz bir veri setidir. CICIoT2023 veri seti içerisindeki saldırı gruplarına ait saldırı türlerinde de dengesizlik vardır. Bu dengesizlikleri gidermek ve veri setini dengeli hale getirmek için bir yöntem izlenmiştir. Bu yöntem ile, veri seti içerisinde bulunan saldırı türleri veri sayılarından belirli sayıda örnekler alınmış ve yeni veri setleri oluşturulmuştur. Bu veri setleri içerisindeki saldırı türleri örnek sayıları SMOTE ile çoğaltılarak eşitlenmiştir. Bu yöntem veri setindeki ML algoritmalarının sınıflandırmadaki başarılarına katkısını göstermektedir. RF ML algoritması genel olarak saldırı tespitinde her üç sınıflandırmada %94.90-%96.79 arasında doğruluk ile çok iyi bir performans gerçekleştirmiştir. XGB ML algoritmasının her üç sınıflandırma türünde RF ML algoritmasına çok yakın performans göstermiştir. LR ve GNB ML algoritmalarının iyi performans göstermediği ve siber saldırılara karşı zayıf olduğu gözlemlenmiştir. LGBM, DT ve CB ML algoritmaları ise ikili sınıflandırmada iyi bir performans ortaya koymuştur. AB ML algoritması 34 sınıf çoklu sınıflandırmada iyi performans gösterememiştir. Çalışmamızda kullanılan veri sayısının artması ML algoritmalarının performanslarını artırdığını göstermiştir. Bu çalışma, veri dengesizliğini giderebilmek için bir yöntem geliştirmiş ve IoT cihazlarında saldırı tespiti ve güvenliğini sağlamak için ML algoritmalarının etkinliğini artırmıştır. Gelecekte yapılacak olan çalışmalara farklı bir bakış açısı getirmiştir. Brute Force, Recon, Spoofing ve Web-Based saldırı türlerini daha iyi tespit edebilmek için gelecekteki çalışmalarda derin öğrenme modelleri uygulanması planlanmaktadır.

Kaynaklar (References)

Abdulkareem, S. A., Foh, C. H., Lee, H., Carrez, F., Moessner, K., 2022. IoT Network Intrusion Detection with Ensemble Learners. International Conference on ICT Convergence, 2022-October, 510-514.

Bala, B., Behal, S., 2024. A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques. 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2024- Proceedings, 1755-1762.

Bozdogan, Z., Kara, R., 2015. Layered model architecture for internet of things. Journeal of Engineering Reearch and Applied Science , 4(1), pp 260-264.

Elaziz, B., Ommane, Y., Eddabbah, M., Laaziz, Y., 2024. Enhancing Diabetes Detection Through Data Preprocessing: A Comparative Analysis of Machine Learning Algorithms. 1st International Conference on Computing, Internet of Things and Microwave Systems, ICCIMS 2024, pp. 1-6

Guo, G., 2022. An Intrusion Detection System for the Internet of Things Using Machine Learning Models. 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2022, 332-335.

Happy, Chhikara, R., Kashyap, N., 2024. A Comparative Analysis of Machine Learning Prediction Algorithms for Detecting IoT Botnet Activities. 2024 International Conference on Intelligent Systems for Cybersecurity, ISCS 2024., pp. 1-6.

Kasongo, S., M., 2021.An advanced intrusion detection system for IIoT Based on GA and tree based algorithms', IEEE Access, 9, pp. 113199–113212

Kouekam, M., Khennou, F., 2024. Securing the Internet of Things: Exploring Ensemble Learning for Attacks Classification. Proceedings- 2024 5th International Conference on Industrial Engineering and Artificial Intelligence, IEAI 2024, 78-84.

Kumar, A. G., Rastogi, A., Ranga, V., 2024. Evaluation of Different Machine Learning Classifiers on New IoT Dataset CICIoT2023. 2024 International Conference on Intelligent Systems for Cybersecurity, ISCS 2024, pp. 1-6.

Manchala, Y., Nayak, J., Behera, H. S. 2023. Detection of Malicious Traffic in IoMT Environment Using Intelligent XGboost Approach. 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development, OTCON 2022., pp. 1-6.

Mittal, S., Mishra, A. K., Tripathi, V., Singh, P., Pandey, P. 2023. A Comparative Analysis of Supervised Machine Learning Models for Smart Intrusion Detection in IoT Network. 2023 3rd Asian Conference on Innovation in Technology, ASIANCON 2023., pp. 1-6 .

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., Ghorbani, A. A. 2023. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. Sensors, 23(13).

Rahim, R., Chishti, M. A., Raheem, M. M., 2024. Improving the security of Internet of Things (IoT) using Intrusion Detection System(IDS). 21st International Learning and Technology Conference: Reality and Science Fiction in Education, L and T 2024, 290-295.

Sadhwani, S., Modi, U. K., Muthalagu, R., Pawar, P. M. 2024. SmartSentry: Cyber Threat Intelligence in Industrial IoT. IEEE Access, 12, 34720-34740.

Sahu, N., K., Mukherjee, I., 2020. Machine Learning based anomaly detection for Network. Proceedings of the 4th International Conference on Trends in Electronics and Informatics (ICOEI 2020): 15-17, June 2020. IEEE., pp. 787-794, 2020.

Saif, S., Hossain, M. A., Islam, M. S., 2024. IoT Security Fortification: Enhancing Cyber Threat Detection Through Feature Selection and Advanced Machine Learning. 1st International Conference on Innovative Engineering Sciences and Technological Research, ICIESTR 2024- Proceedings., pp. 1-6.

Saiyed, M., Al Anbagi, I., 2023. Entropy and Divergence-based DDoS Attack Detection System in IoT Networks. International Conference on Wireless and Mobile Computing, Networking and Communications, 2023-June, 224-230.

Samdekar, R., Ghosh, S. M., Srinivas, K., 2021. Efficiency enhancement of intrusion detection in iot based on machine learning through bioinspire. Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021, 383-387.

Satılmış H., Akleylek S., 2021. IoT Güvenliği için Kullanılan Makine Öğrenimi ve Derin Öğrenme Modelleri Üzerine Bir Derleme. Bilişim Teknolojileri Dergisi, 14(4), 457-481.

Sibai, F. N., Asaduzzaman, A., Sibai, A., 2023. A Comparative Study of Machine Learning Methods for Intrusion Detection. Proceedings- 2023 10th

International Conference on Electrical and Electronics Engineering, ICEEE 2023, 184-188.

Thereza, N., Ramli, K., 2023. Development of Intrusion Detection Models for IoT Networks Utilizing CICIoT2023 Dataset. Proceedings of the 3rd 2023 International Conference on Smart Cities, Automation and Intelligent Computing Systems, ICON-SONICS 2023, 66-72.

Yemmanuru, P. K., Yeboah, J., Khakata Esther, N. G., 2023. Systematic Literature Review of Machine Learning for IoT Security. Proceedings- 2023 International Conference on Computational Science and Computational Intelligence, CSCI 2023, 227-233.