

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

İSTATİKSEL YAYILIM YÖNTEMLERİ İLE ATAKLARA DAYANIKLI
FEDERE ÖĞRENME

YÜKSEK LİSANS TEZİ

Fatma Zehra SOLAK

AĞUSTOS 2024

TRABZON



KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**İSTATİKSEL YAYILIM YÖNTEMLERİ İLE ATAKLARA DAYANIKLI
FEDERE ÖĞRENME**

Fatma Zehra SOLAK

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"BİLGİSAYAR YÜKSEK MÜHENDİSİ"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 08 / 08 / 2024

Tezin Savunma Tarihi : 20 / 08 / 2024

Tez Danışmanı : Prof. Dr. Güzin ULUTAŞ

Trabzon 2024

ÖNSÖZ

“İstatiksel Yayılım Yöntemleri ile Ataklara Dayanıklı Federe Öğrenme” isimli bu tez Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Programı’nda yapılan bir çalışmadır.

Tez çalışmam boyunca hiçbir desteğini esirgemeyen ve bilim insanı olma yolundaki ilk adımlarımı atarken kıymetli bilgileriyle ve tecrübesiyle yolumu aydınlatan danışman hocam Sayın Prof. Dr. Güzin ULUTAŞ’a teşekkürü bir borç bilirim. Ayrıca, yapıcı eleştirileri ve önerileri ile tezime büyük katkıları bulunan saygıdeğer hocalarım Sayın Dr. Öğr. Üyesi Beste ÜSTÜBİOĞLU’na, Arş. Gör. Hakan ERDÖL’e ve Eda Sena Erdöl’e eğitim öğretim hayatımda katkısı olan tüm hocalarıma ve ortak olarak çalışmalar yürüttüğümüz ve tez çalışması süresinde de hiçbir yardımdan kaçınmayan arkadaşlarım Arş. Gör. Mervenur ÇAKIR, Süheyla ŞİMŞEK ve Huriye BAYRAKÇI’ya teşekkür ederim.

Tez çalışması sürecinde bana desteklerini esirgemeyen Prof. Dr. Servet KIRAN hocama teşekkürlerimi sunarım. Son olarak, doğduğum günden beri elimi hiç bırakmayan, hayatım boyunca aldığım tüm kararlarda bilgi ve tecrübelerinden faydalandığım ve her kararında arkamda olan, maddi ve manevi desteklerini hiçbir zaman esirgemeyen ve bugünlere gelmemde en büyük paya sahip olan annem Ayşedudu SOLAK ve babam Kazim SOLAK’a, kardeşlerim Emine TUNÇEZ, Mesut Yesevi SOLAK, Mualla SOLAK, Ümmügülsüm SOLAK, Abdullah SOLAK ve Muhammed Şevket SOLAK’a teşekkür eder, minnettarlığımı sunarım.

Bu tez çalışmasının bundan sonraki çalışmalara katkı sağlamasını temenni ederim.

Fatma Zehra SOLAK
Trabzon 2024

TEZ ETİK BEYANNAMESİ

Doktora Tezi olarak sunduğum “İstatistiksel Yayılım Yöntemleri ile Ataklara Dayanıklı Federe Öğrenme” başlıklı bu çalışmayı baştan sona kadar danışmanım Prof. Dr. Güzin ULUTAŞ’ın sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 20/08/2024

Fatma Zehra SOLAK

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ.....	III
TEZ ETİK BEYANNAMESİ.....	IV
İÇİNDEKİLER.....	V
ÖZET	VII
SUMMARY	VIII
ŞEKİLLER DİZİNİ.....	IX
TABLolar DİZİNİ.....	XII
SEMBOLLER VE KISALTMALAR DİZİNİ	XIV
1. GENEL BİLGİLER	1
1.1. Federe Öğrenme Ağ Topolojileri	3
1.1.1. Merkezi Federe Öğrenme.....	3
1.1.2. Dağıtık Federe Öğrenme	4
1.2. Federe Öğrenmede Veri Paylaşım Yaklaşımları	6
1.3. Federe Öğrenmede Güvenlik ve Gizlilik.....	8
1.3.1. Veri Zehirlenmesi Saldırıları	10
1.3.2. Model Zehirlenmesi Saldırıları	13
1.3.3. Bedavacılık Saldırıları.....	14
1.3.4. Sybil Saldırısı.....	15
1.3.5. Arka Kapı Saldırıları.....	16
1.3.6. Bizans Saldırıları.....	17
1.4. Güvenli Federe Öğrenme Yöntemleri	18
1.4.1. Anomali Tespiti ve İzleme	18
1.4.2. Güçlü Birleştirme Teknikleri	22
2. YAPILAN ÇALIŞMALAR.....	25
2.1. Önerilen Yöntem	25
2.1.1. Son Katmanlar Arası Öklid Mesafesinin Hesaplanması.....	27
2.1.2. Öklid Mesafelerinin İstatiksel Analizi	29
2.1.3. QMAD Yöntemi ile Kötü Niyetli Katılımcı Tespiti	29
2.1.4. QAD Yöntemi ile Kötü Niyetli Katılımcı Tespiti.....	31
2.1.5. ÇİFT SINIR Yöntemi ile Kötü Niyetli Katılımcı Tespiti	34

2.1.6. Kötü Niyetli Katılımcıların Sistemden Elimine Edilmesi Ve Ana Model Güncellenmesi	35
2.2. Veri Setleri	37
2.3. Savunma Yapılan Saldırı ve Saldırgan Türleri.....	40
2.4. Model Mimarileri ve Hiperparametreler	42
2.4.1. Model Mimarileri	42
2.4.2. Hiperparametreler	44
2.5. Değerlendirme Metrikleri.....	45
3. BULGULAR VE TARTIŞMA	48
3.1. MNIST Veri Setinde Yapılan Deneysel Çalışmalar.....	48
3.1.1. Etiket Çevirme Saldırısı	48
3.1.2. Bizans Saldırısı	56
3.2. Fashion MNIST Veri Setinde Yapılan Deneysel Çalışmalar	61
3.2.1. Etiket Çevirme Saldırısı	61
3.2.2. Bizans Saldırısı	68
3.3. CIFAR Veri Setinde Yapılan Deneysel Çalışmalar	74
3.3.1. Etiket Çevirme Saldırısı	74
3.3.2. Bizans Saldırısı	81
4. SONUÇLAR.....	87
5. ÖNERİLER VE TARTIŞMA	89
6. KAYNAKLAR	90
ÖZGEÇMİŞ	

Yüksek Lisans Tezi

ÖZET

İSTATİKSEL YAYILIM YÖNTEMLERİ İLE ATAKLARA DAYANIKLI FEDERE ÖĞRENME

Fatma Zehra SOLAK

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Güzin ULUTAŞ
2024, 92 Sayfa

Yapay zekâ alanında önemli bir yöntem olan federe öğrenme, verilerin katılımcılardan doğrudan alınmadan eğitim yapılmasını sağlayarak, kişisel veri güvenliği açısından önemli avantajlar sunar. Ancak, federe öğrenmede katılımcılardan herhangi bir veri alınmaması, sistemi çeşitli saldırılara karşı savunmasız bırakabilmektedir. Literatürde, federe öğrenmedeki güvenlik açıklarına karşı geliştirilen yöntemler genellikle düşük yoğunlukta kötü niyetli katılımcılara odaklanmıştır. Ancak, yüksek yoğunlukta kötü niyetli katılımcıların bulunduğu durumlarda, bu yöntemler yetersiz kalabilir ve güvenlik zafiyetlerine yol açabilir. Tez çalışmasında, saldırılara karşı dayanıklı bir federe öğrenme yöntemi geliştirilmiştir. Önerilen istatistiksel yayılım yöntemleri kullanılarak, özellikle kötü niyetli katılımcı oranlarının yüksek olduğu senaryolarda yöntemimizin gücü vurgulanmıştır. Çalışmada, Bizans saldırıları ve Etiket Çevirme saldırılarında yöntemin performansı test edilmiştir. Yapılan çeşitli deneysel çalışmalar, önerilen yöntemin etkinliğini ve mevcut yöntemlere kıyasla üstünlüğünü net bir şekilde ortaya koymuştur. Ayrıca, önerilen yöntem literatürdeki yöntemlerle karşılaştırılarak, sağlanan iyileştirmeler detaylı bir şekilde analiz edilmiştir.

Anahtar Kelimeler: Federe öğrenme, Etiket çevirme saldırıları, Bizans saldırıları, Veri zehirlenmesi

Master. Thesis

SUMMARY

ATTACK-RESISTANT FEDERATED LEARNING WITH STATISTICAL PROPAGATION METHODS

Fatma Zehra SOLAK

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Department of Computer Engineering Graduate Program
Supervisor: Prof. Dr. Güzin ULUTAŞ
2024, 92 Pages

Federated learning, an important method in the field of artificial intelligence, offers significant advantages in terms of personal data security by enabling training without directly obtaining data from participants. However, not receiving any data from participants in federated learning can leave the system vulnerable to various attacks. In the literature, methods developed against security vulnerabilities in federated learning have generally focused on low-density malicious participants. However, in cases where there is a high density of malicious participants, these methods may be insufficient and lead to security vulnerabilities. In the thesis study, a federated learning method resistant to attacks has been developed. Using the proposed statistical propagation methods, the power of our method has been emphasized, especially in scenarios where the rate of malicious participants is high. In the study, the performance of the method has been tested in Byzantine attacks and Label Translation attacks. Various experimental studies have clearly demonstrated the effectiveness of the proposed method and its superiority over existing methods. In addition, the proposed method has been compared with the methods in the literature and the improvements provided have been analyzed in detail.

Key Words: Federated learning, Label flipping, Byzantine attacks, Data poisoning

ŞEKİLLER DİZİNİ

	<u>Sayfa No</u>
Şekil 1 Merkezi FL mimarisi.....	4
Şekil 2 Dağıtık FL mimarisi	5
Şekil 3 FL yaklaşımları	7
Şekil 4 FL’de güvenlik saldırıları.....	10
Şekil 5 Veri Zehirlenmesi.....	12
Şekil 6 Model Zehirlenmesi	14
Şekil 7 Sybil Saldırısı	16
Şekil 8 Önerilen Yöntemlerin Akış Şeması.....	27
Şekil 9 QAD Trimodal dağılım	32
Şekil 10 Çift sınır sözde kodu	36
Şekil 11 MNIST veri seti.....	38
Şekil 12 CIFAR veri seti	39
Şekil 13 Fashion MNIST veri seti	40
Şekil 14 MNIST veri setinde 2NN ve CNN mimarisine ait modellerde organize olan ve organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.	55
Şekil 15 MNIST veri setinde 2NN ve CNN mimarisine ait modellerde organize olan ve organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.....	60
Şekil 16 Fashion MNIST veri setinde organize olan ve organize olmayan Etiket Çevirme saldırılarında savunma yöntemlerinin test doğruluk değerleri, (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan	

saldırılarıdaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.....	65
Şekil 17 Fashion MNIST veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden ceket etiketinin kaynak doğruluğu.....	66
Şekil 18 Fashion MNIST veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden kazak etiketinin kaynak doğruluğu	67
Şekil 19 Fashion MNIST veri setinde organize olan ve organize olmayan Bizans saldırılarında savunma yöntemlerinin test doğruluk değerleri, (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.	71
Şekil 20 Fashion MNIST veri setindeki organize ve organize olmayan Bizans saldırılarındaki yöntemlerin F1, Doğruluk ve Keskinlik değerleri. (a), organize saldırılardaki F1 skor, (b) organize olmayan saldırılardaki F1 skor, (c), organize saldırılardaki Doğruluk, (d) organize olmayan saldırılardaki Doğruluk , (e), organize saldırılardaki Keskinlik, (f) organize olmayan saldırılardaki Keskinlik değerleri.	73
Şekil 21 CIFAR veri setinde organize olan ve organize olmayan Etiket Çevirme saldırılarında savunma yöntemlerinin test doğruluk değerleri, (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu.	77
Şekil 22 CIFAR veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden köpek etiketinin kaynak doğruluğu.....	79
Şekil 23 CIFAR veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden kedi etiketinin kaynak doğruluğu	80
Şekil 24 CIFAR veri setinde organize olan ve organize olmayan Bizans saldırılarında savunma yöntemlerinin test doğruluk değerleri (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu.	84
Şekil 25 CIFAR veri setindeki organize ve organize olmayan Bizans saldırılarındaki yöntemlerin F1, Doğruluk ve Keskinlik değerleri. (a),	

organize saldırılardaki F1 skor, (b) organize olmayan saldırılardaki F1 skor, (c), organize saldırılardaki Doğruluk, (d) organize olmayan saldırılardaki Doğruluk , (e), organize saldırılardaki Keskinlik, (f) organize olmayan saldırılardaki Keskinlik değerleri..... 85



TABLÖLAR DİZİNİ

	<u>Sayfa No</u>
Tablo 1 Organize Etiket Çevirme saldırılarında değiştirilen sınıf etiketleri	41
Tablo 2 MNIST-2NN model mimarisi	43
Tablo 3 MNIST-CNN model mimarisi	43
Tablo 4 CIFAR model mimarisi.....	43
Tablo 5 Fashion-MNIST model mimarisi	44
Tablo 6 Eğitim Hiperparametreleri.....	45
Tablo 7 MNIST veri seti 2NN mimarisi organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri	49
Tablo 8 MNIST veri seti 2NN mimarisi organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri.....	49
Tablo 9 MNIST veri seti CNN mimarisi organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri	52
Tablo 10 MNIST veri seti CNN mimarisi organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri.....	52
Tablo 11 MNIST veri seti 2NN mimarisi organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri	56
Tablo 12 MNIST veri seti 2NN mimarisi organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri.....	57
Tablo 13 MNIST veri seti CNN mimarisi organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri.....	58
Tablo 14 MNIST veri seti CNN mimarisi organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri	59
Tablo 15 Fashion MNIST veri setinde organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri	62
Tablo 16 Fashion MNIST veri setinde organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri.....	62
Tablo 17 Fashion MNIST veri setinde organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri	68
Tablo 18 Fashion MNIST veri setinde organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri.....	69
Tablo 19 CIFAR veri setinde organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri	75
Tablo 20 CIFAR veri setinde organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri.....	75

Tablo 21 CIFAR veri setinde organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri	81
Tablo 22 CIFAR veri setinde organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri.....	82



SEMBOLLER VE KISALTMALAR DİZİNİ

CNN	:Konvolüsyonel Sinir Ağları (Convolutional Neural Networks)
FL	: Federe Öğrenme
F1	: F skoru
FN	: Yanlış Negatif
FP	:Yanlış Pozitif
GAN	: Çekişmeli Üretici Ağ(Generative Adversarial Networks)
IOT	: Nesnelerin İnterneti(Internet of Things)
QAD	: Çeyreklik Mutlak Sapma
QMAD	: Çeyreklik Ortalama Mutlak Sapma
RAM	: Rastgele Erişimli Hafıza (Random Access Memory)
TN	:Doğru Negatif
TP	:Doğru Pozitif

1. GENEL BİLGİLER

Yapay zekâ ve özellikle makine öğrenmesinde genellikle büyük miktarda verinin tek bir merkezde toplanması ve verilerin analiz edilmesi ile gerçekleştirilir. Merkezi veri toplama süreci, yapay zekâ algoritmalarının etkin bir şekilde öğrenebilmesi için kritik öneme sahiptir. Sağlık alanındaki tahliller ve radyolojik görüntüler, kişiselleştirilmiş reklam üretimi için kullanılan veriler gibi birçok örnek, sürecin önemli parçalarını oluşturur. Veriler kişisel bilgileri içerdiği için, yapay zekâ modellerinin eğitimi sırasında doğrudan kullanımı önemli gizlilik endişelerine yol açmaktadır. Ülkemizde Kişisel Verilerin Korunması Kanunu, kişisel verilerin işlenmesi ve saklanması konusunda ciddi kısıtlamalar getirmektedir. Kişisel Verilerin Korunması Kanunu'nun getirdiği kısıtlamalar, yapay zekâ eğitim süreçlerini doğrudan etkileyerek, veri işleme yöntemlerinde daha fazla dikkat ve hassasiyet gerektiren bir ortam yaratır. Bu durum, yapay zekâ teknolojisinin gelişimi için hem bir zorluk hem de yeni fırsatlar doğurur, çünkü verilerin güvenli bir şekilde işlenmesi ve mahremiyetin korunması, teknolojinin toplum tarafından kabul edilmesi açısından büyük önem taşır. Bu konu, günümüzde dijital ortamda kişisel verilerin toplanmasıyla ilgili artan endişelerin bir yansımasıdır. Basit bir uygulamadan gezinilen internet sitelerine kadar birçok dijital kullanım alanı kişisel verileri bazen bir çerez ile bazen de verilen basit bir izinle toplamaktadır. Kişinin bazen bu durumdan haberi bile olmamaktadır. Öte yandan akademik çalışmalarda kullanılmak üzere yeni veri setleri oluşturulurken de kişilerin onaylarının gerekmesi, etik kurullardan geçmesi gibi katı kurallar bulunmaktadır. Veri seti oluşumundaki kısıtlamalar, geniş ve heterojen veri setlerinin hazırlanmasını zorlaştırmaktadır. Verilerin tek merkezde toplanmasında sorun yaşanan durumlarda, dağıtılmış öğrenme yöntemlerinden biri olan Federe Öğrenme (FL) ilk olarak 2016 yılında kullanılmaya başlanmıştır.

FL veri gizliliğini koruma ve merkezi veri depolama gereksinimini ortadan kaldırma amacıyla ortaya çıkan yenilikçi bir makine öğrenimi paradigmasıdır. FL, merkezi bir sunucuya veri göndermek yerine, yerel cihazlarda model eğitimi gerçekleştirilmesine olanak tanır. FL, özellikle mahremiyetin kritik olduğu sağlık, finans ve mobil cihazlar gibi alanlarda giderek daha fazla önem kazanmaktadır (McMahan, 2017). Kullanıcıların kişisel verileri gizli tutulurken, yapay zekâ modelleri etkili bir şekilde eğitilebilir. FL'in en büyük avantajlarından biri, veri gizliliğini korumanın yanı sıra, geniş çapta dağılmış ve çeşitli veri

kaynaklarından yararlanabilmesidir. Örneğin, bir sağlık uygulamasında, farklı hastanelerden toplanan verilerle bir model eğitilebilir. Hastaneler, verilerini paylaşmak zorunda kalmadan merkezi bir modele katkıda bulunabilirler. Aynı şekilde, bir mobil uygulama, kullanıcıların cihazlarındaki verileri kullanarak kişiselleştirilmiş hizmetler sunabilir. Bu, kullanıcı gizliliğini korurken uygulamanın etkinliğini artırır. FL'nin bir diğer önemli avantajı, merkezi veri depolama maliyetlerini azaltmasıdır. Verilerin merkezi bir sunucuda toplanması ve depolanması maliyetli olabilir. Ancak FL ile maliyetler minimize edilir çünkü veriler yerel cihazlarda kalır. Ayrıca, FL ağ bant genişliği gereksinimlerini de azaltır. Çünkü büyük veri setlerinin merkezi bir sunucuya gönderilmesi yerine, yalnızca model güncellemeleri gönderilir. Bu, özellikle düşük bant genişliğine sahip ağlarda önemli bir avantajdır. Bunun yanı sıra, FL'nin uygulama alanları oldukça geniştir. Sağlık, finans, mobil uygulamalar, IoT cihazları gibi birçok alanda kullanılabilir. Farklı veri kaynaklarından gelen verileri bir araya getirerek daha güçlü ve genelleştirilebilir modeller oluşturulabilir. Örneğin, bir bankacılık uygulaması, farklı şubelerden gelen verilerle bir kredi risk modeli eğitebilir. Merkezi bir veri deposuna ihtiyaç duymadan daha geniş ve çeşitli bir veri seti ile modelin performansını artırabilir. Sonuç olarak, FL, veri gizliliği ve güvenliği konularında getirdiği avantajlarla birlikte, geniş ve çeşitli veri kaynaklarından yararlanma imkânı sunarak veri toplama ve depolama maliyetlerini azaltırken, yapay zekâ modellerinin etkinliğini ve genelleştirilebilirliğini artırır. Kişisel verilerin korunmasının önemli olduğu alanlarda, FL'nin kullanımı giderek yaygınlaşmaktadır. FL, veri paylaşımı olmadan iş birliği yapma imkânı sunduğu için, özellikle hassas verilerin bulunduğu sektörlerde büyük ilgi görmektedir. Sağlık sektörü, bankacılık, mobil uygulamalar ve IoT cihazları, FL'nin potansiyel kullanım alanlarından sadece birkaçıdır. FL, kullanıcı gizliliğini korurken geniş ve heterojen veri setlerinden yararlanarak güçlü yapay zekâ modelleri oluşturma imkânı sağlar. Veri gizliliği ve güvenliği konusundaki artan endişelerle birlikte FL'nin popülaritesi yükselmektedir.

FL'nin gerçek dünya problemlerinde uygulanmasında verilebilecek en bilinen örneği, otomatik konuşma tanıma için Apple'ın Siri'sidir (Paulik, 2021). Kişisel verileri son derece iyi koruyan FL tek merkezli eğitimden farklı olarak çeşitli zayıf yönleri bulunmaktadır. Bunlar, kötü niyetli katılımcı ve saldırılara karşı açık olmak, yerel katılımcı ve merkez arasındaki iletişim maliyeti ve birçok katılımcının veri kümelerindeki heterojenlik ve tutarsızlıklardır. Merkez sunucu, katılımcılardan gelen modellerin doğru verilerle eğitilip eğitilmediğine dair elinde bir veri bulundurmadığından savunmanın bulunmadığı bir federe

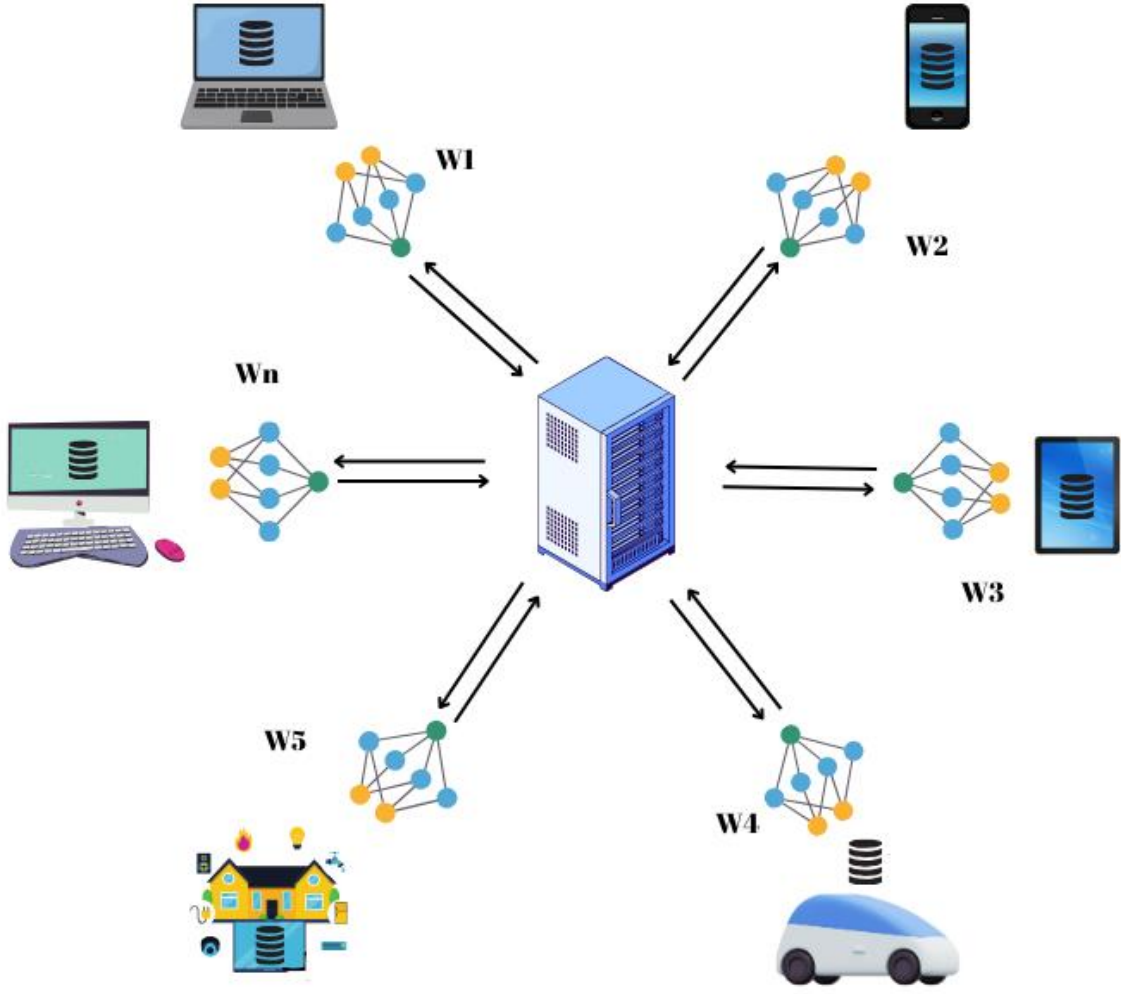
yapısında kötü niyetli katılımcılardan gelen güncellemeler merkezdeki modelin eğitiminde tahribata sebep olur. Gelecek bölümde, FL'deki ağ topolojileri ve veri dağılımları gibi FL'yi oluşturan teknik ve yaklaşımlar incelenecektir.

1.1. Federe Öğrenme Ağ Topolojileri

FL'de cihazların veya düğümlerin nasıl organize edildiği ve birbirleriyle nasıl iletişim kurduğu, ağ topolojisine göre iki kategoriye ayrılabilir: merkezi ve dağıtık. (Lian, 2017).

1.1.1. Merkezi Federe Öğrenme

FL, eğitimin tek merkezde yapılmamasına dayanmasına rağmen, merkezi topolojide katılımcılar eğitilmiş modelleri toplar, ana modeli oluşturur ve tüm istemcilere dağıtırken merkezi bir sunucuya bağlıdır. İstemciler arasında güven sağlamak için genellikle üçüncü taraf bir sistemin devreye girmesi gerekir. Bu topoloji, tek bir sunucu ile çok sayıda katılımcıyı içerir (Kairouz, 2021). Genel çalışma mantığında katılımcılar yerel verilerini üçüncü kişilerle paylaşmadan sunucudaki ana bir modeli toplu olarak eğitirler. Şekil 1'de örnek mimarisi verilen FL'de eğitim işlemi katılımcıların kendi cihazlarında gerçekleştirilir. Sunucunun işlevi ise yerel modellerden ana bir model oluşturarak katılımcılara ana modeli yeniden yollamaktır. Genellikle, FL yinelemeli bir yaklaşım kullanır. Her yinelemede, sunucu ana modelin o anki halini katılımcılara yollar (adım 1), katılımcılar sunucudan gelen modeli kendi verileri ile eğiterek geliştirir (adım 2) ve güncellenmiş modeli sunucuya gönderir (adım 3). Sunucu, güncellenmiş modelleri bir araya getirerek ana modelin bir sonraki versiyonunu oluşturur (adım 4) ve adım 1'e geri dönlür.

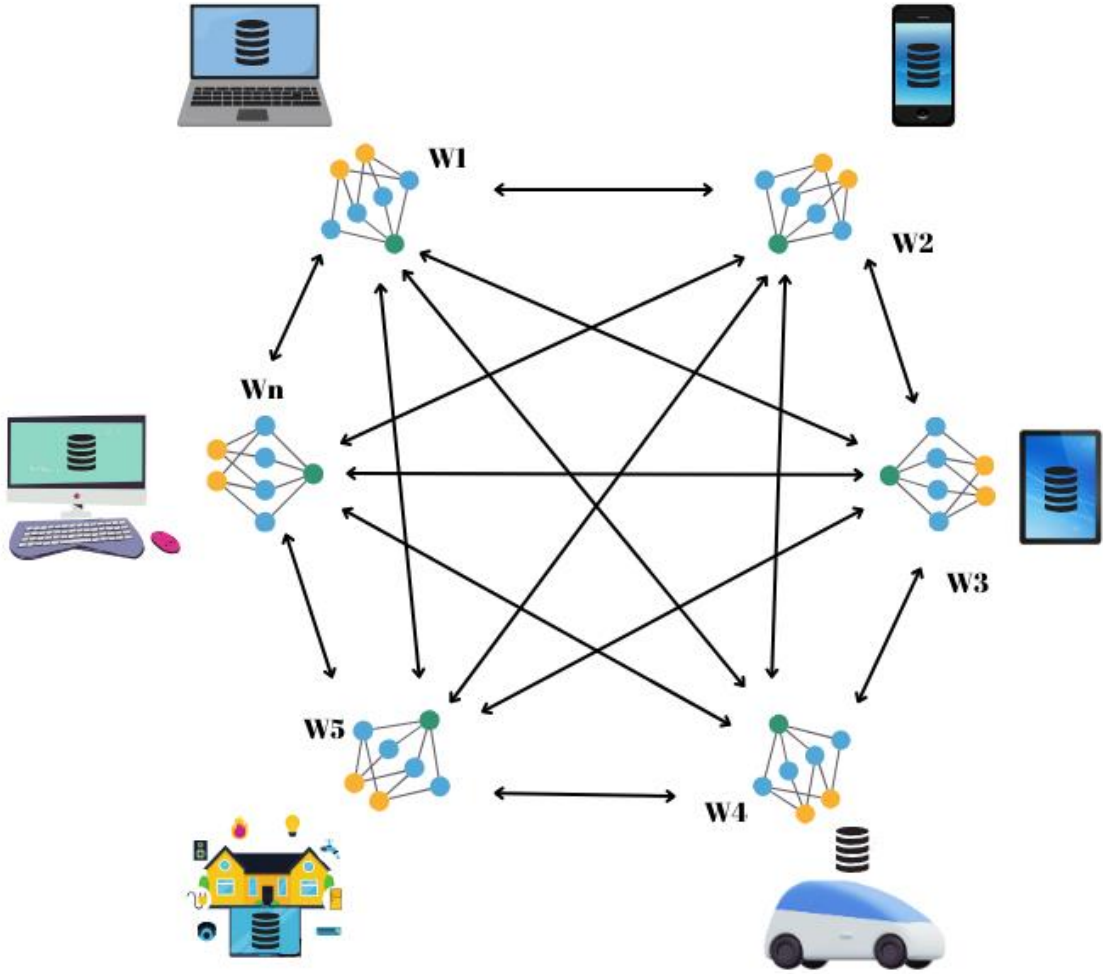


Şekil 1 Merkezi FL mimarisi

1.1.2. Dağıtık Federe Öğrenme

Merkezi topolojideki merkez sunucuya olan bağımlılığı ortadan kaldıran yöntem olarak bilinen dağıtık FL'de, model toplama sürecinde merkezi bir sunucuya ihtiyaç duyulmaz. Bunun yerine, güvenilirliği sağlamak amacıyla çeşitli özel algoritmalar devreye girer. Şekil 2'de mimarisi verilmiş olan yaklaşımda, katılımcılar arasında ortak bir küresel model bulunmaz; her bir katılımcı, yalnızca komşularıyla bilgi paylaşarak kendi modelini günceller ve gelen bilgileri kendi modeli üzerinde uygulayarak ilerler. (Vanhaesebrouck, 2017) Eşler arası topoloji, dağıtık yapılar için ideal bir çözüm olarak öne çıkar ve ağda izlenecek protokolü belirlemek amacıyla merkez sunucu yalnızca eğitim turları sırasında ve sadece bir kez devreye girer. Tamamen dağıtık öğrenme süreçlerini desteklemek için,

günümüzde birçok gelişmiş teknoloji ve algoritma geliştirilmiştir. Geliştirilen teknolojiler, merkezi bir otoritenin eksikliğini telafi ederek, katılımcılar arasındaki iş birliğini en üst düzeye çıkarmayı hedefler.



Şekil 2 Dağıtık FL mimarisi

Gelecek bölümde, FL çerçevesinde kullanılan farklı veri paylaşım yaklaşımları üzerinde durulacak. Yaklaşımların hangi durumlarda tercih edildiği, nasıl uygulandığı ve her birinin avantajları ile dezavantajları incelenecek. Bu kapsamda, her bir yaklaşımın farklı veri yapılarına ve gizlilik gereksinimlerine nasıl uyum sağladığı değerlendirilecek. Çeşitli uygulama alanlarındaki performansları ve güvenlik açısından karşılaştıkları zorluklar da tartışılacaktır.

1.2. Federe Öğrenmede Veri Paylaşım Yaklaşımları

FL’de üç ana yaklaşım bulunmaktadır: yatay federe öğrenme (horizontal federated learning), dikey federe öğrenme (vertical federated learning) ve transfer federe öğrenme (transfer federated learning). Yaklaşımlar, veri dağılımı ve veri sahipleri arasındaki ilişkilerin niteliğine göre farklılık göstermektedir (Yang, 2019).

Yatay federe öğrenme, aynı özellik setine sahip ancak farklı veri örneklerine sahip cihazlar veya kurumlar arasında gerçekleştirilir. Özellikle aynı türde veriye sahip olan farklı kuruluşların iş birliği yapması durumunda kullanılır. Örneğin, farklı hastanelerin benzer sağlık verilerini kullanarak ortak bir model eğitmesi, yatay federe öğrenmenin bir örneğidir (McMahan, 2017).

Yatay federe öğrenme yaklaşımında, her kurum kendi verilerini kullanarak yerel modellerini günceller ve merkezi sunucuya sadece model parametreleri gönderilir. Verilerin gizliliği korunur ve veri paylaşımı gereksinimi ortadan kalkar. Yatay federe öğrenmenin avantajları arasında gizlilik koruma, veri çeşitliliği ve bant genişliği tasarrufu bulunmaktadır. Ancak, model konsensüsü zorluğu, heterojenlik ve koordinasyon gibi dezavantajlar da söz konusudur (Konečný, 2016).

Dikey federe öğrenme ise farklı özellik setlerine sahip ancak aynı veri örneklerine sahip olan cihazlar veya kurumlar arasında gerçekleştirilir. Örneğin, bir banka ve bir sigorta şirketinin müşterileri hakkında sahip oldukları farklı veri türlerini birleştirerek daha kapsamlı bir model oluşturması, dikey federe öğrenmenin bir örneğidir (Yang, 2019). Dikey federe öğrenme yaklaşımında, veri sahipleri farklı özelliklerdeki verilerini birleştirerek daha kapsamlı ve zengin modeller oluştururlar. Dikey federe öğrenmenin avantajları arasında zengin veri setleri, gelişmiş analitik ve gizlilik koruma bulunmaktadır. Bununla birlikte, veri uyumu, gizlilik riskleri ve koordinasyon ve yönetim gibi dezavantajlar da bulunmaktadır (Bonawitz, 2019).

Transfer federe öğrenme, veri sahipleri arasında sınırlı sayıda ortak özellik seti olduğu durumlarda uygulanır. Farklı veri setlerinin birbirine transfer edilerek, ortak bir modelin oluşturulmasını sağlar. Transfer federe öğrenmede, bir veri kümesinden öğrenilen bilgiler, başka bir veri kümesine transfer edilir ve bilgi transferi sayesinde daha geniş ve çeşitli veri setleriyle modeller eğitilir. Örneğin, bir sağlık kuruluşunun genetik verileri ile bir diğer sağlık kuruluşunun klinik verileri arasında transfer federe öğrenme uygulanabilir. Transfer federe öğrenme yaklaşımında, veri sahiplerinin birbirinden öğrenmesini ve böylece daha

güçlü modeller oluşturmalarını sağlar. Transfer federe öğrenmenin avantajları arasında veri çeşitliliği, bilgi transferi ve model performansının artırılması bulunmaktadır. Ancak, veri uyumu ve gizlilik riskleri gibi zorluklar içermektedir. Şekil 3'te FL'deki yaklaşımların çalışma şekilleri verilmiştir.



Şekil 3 FL yaklaşımları

FL yaklaşımının seçiminde dikkate alınması gereken faktörler şunlardır:

- **Veri Yapısı:** Verileriniz aynı özellik setine sahipse yatay federe öğrenme; farklı özellik setlerine sahipse dikey federe öğrenme tercih edilir.
- **Gizlilik Gereksinimleri:** Gizlilik ve güvenlik seviyeniz yüksekse, dikey federe öğrenme tercih edilir. Bu, farklı veri sahipleri arasında veri paylaşımını en aza indirir ve ek gizlilik artırıcı tekniklerin kullanımını destekler.
- **Uygulama Alanı:** Sağlık, finans gibi hassas sektörlerde, genellikle dikey federe öğrenme tercih edilir, çünkü bu alanlarda farklı veri türleri ve özellik setleri bir araya

getirilebilir. Ancak, aynı tür verilerin paylaşımı söz konusuysa yatay federe öğrenme de kullanılabilir.

- **Veri Sahipleri Arasındaki İlişkiler:** Veri sahipleri arasındaki iş birliği düşükse, dikey federe öğrenme daha uygun olabilir çünkü veri paylaşımını en aza indirir. İş birliği yüksekse, yatay federe öğrenme tercih edilebilir çünkü daha verimli bir model eğitimi sağlar. Transfer federe öğrenme ise, farklı veri setleri arasında bilgi aktarımı gerektiğinde kullanılabilir.

Genel olarak, yatay federe öğrenme, aynı türde veriye sahip farklı kuruluşların iş birliği yapması gerektiğinde tercih edilir. Dikey federe öğrenme, farklı veri türlerine sahip kuruluşların ortak bir model oluşturması için uygundur. Transfer federe öğrenme ise, veri setleri arasında bilgi transferi yaparak model performansını artırmak için kullanılır.

Sonuç olarak, en uygun FL yaklaşımını seçerken veri yapısı, gizlilik gereksinimleri, uygulama alanı ve veri sahipleri arasındaki ilişkiler dikkate alınmalıdır. Bu faktörler doğrultusunda, yatay, dikey veya transfer federe öğrenme yaklaşımlarından biri seçilebilir.

Gelecek bölümde, FL'de karşılaşılan güvenlik sorunları ayrıntılı bir şekilde ele alınacak. Özellikle literatürdeki FL'de zehirlenme saldırıları ve diğer potansiyel tehditler incelenecek. FL'nin güvenliğini sağlamak için geliştirilen savunma mekanizmaları, mekanizmaların etkinliği ve alanda karşılaşılan zorluklar üzerine odaklanılacak. Ayrıca, mevcut literatürdeki güvenlik açıkları ve açıkların FL sistemleri üzerindeki etkileri de detaylandırılacak.

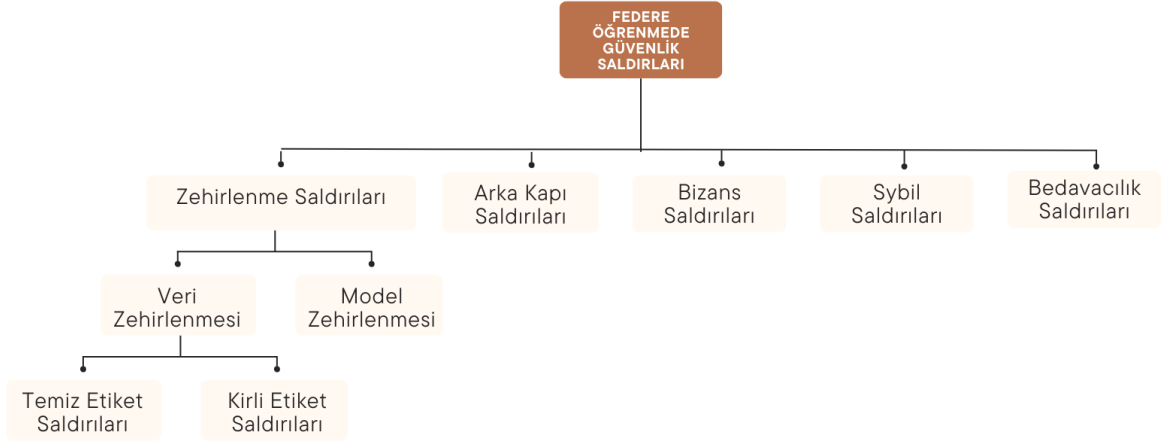
1.3. Federe Öğrenmede Güvenlik ve Gizlilik

FL, kullanıcı verilerini korumayı hedeflerken, çeşitli saldırı türleri gizliliği tehdit edebilir. Kasıtsız Veri Sızıntısı ve Model Yeniden Yapılandırma Saldırıları, model ağırlıklarının analiz edilmesiyle verilerin elde edilebileceği riskine dayanır. Meraklı ancak dürüst saldırganlar, istemci ve sunucu arasındaki veri akışını izleyerek istemci modelini yeniden yapılandırabilirler. Kötü Niyetli Sunucu Saldırıları, merkezi sunucunun yetkisiz erişimle kötüye kullanılması durumunda, istemcilerden alınan verilerin analiz edilerek orijinal verilere dair bilgi sızdırma tehlikesini taşır. GAN Tabanlı Çıkarım Saldırıları, sınır ağlarının hesaplanan gradyanlarının incelenmesi yoluyla eğitim verilerini çıkarmayı hedefler; saldırılar, pasif veya aktif şekilde gerçekleştirilebilir ve gradyanların küçük bir kısmı bile başarılı sonuçlar verebilir. Üyelik Çıkarım Saldırıları, modelin daha önce belirli verilerle eğitilip eğitilmediğini saptamaya çalışır ve saldırganın modelin evrimini izleyerek

bilgi çıkarmasına olanak tanır (Gosselin, 2022). FL'deki gizlilik saldırıları hem verilerin hem de katılımcıların gizliliğini ciddi şekilde tehlikeye atabilir; ancak saldırıların amacı, genellikle eğitimin başarısını olumsuz etkilemek değil, gizli bilgilere erişim sağlamaktır.

FL süreçlerinde, merkezi sunucu, kullanıcıların modellerinin doğru verilerle eğitilip eğitilmediğine dair herhangi bir bilgiye sahip değildir. Bilgi eksikliği, kötü niyetli kullanıcıların sunduğu yanıltıcı güncellemelerin modelin doğruluğunu bozmasına neden olabilir. Özellikle kötü niyetli katılımcıların gerçekleştirebileceği saldırılar, FL modellerinin güvenilirliğini ciddi şekilde tehdit eder ve FL'nin güvenlik açısından daha kırılgan bir hale gelmesine yol açar. Örneğin, bir katılımcı bilinçli olarak yanlış veriler kullanarak model güncellemeleri gönderdiğinde, merkezi sunucu güncellemelerin güvenilirliğini test edeceği yerel verilere sahip olmadığından yanlış ile doğru verileri ayırt etmekte yetersiz kalabilir. Bu durum modelin genel doğruluğunu ve güvenilirliğini azaltır. Dolayısıyla veri gizliliği sağlanırken güvenlik önlemlerinin de alınması zorunludur.

Kötü niyetli katılımcılar, modelin doğruluğunu ve güvenilirliğini hedef alabilir. Saldırıları, modelin performansını düşürmek veya hatalı sonuçlar üretmek amacıyla gerçekleştirilebilir. Özellikle dikkat çeken saldırılardan biri veri zehirlenme (data poisoning) saldırılarıdır. Kötü niyetli katılımcılar, modelin eğitilmesi sırasında zararlı veriler ekleyerek modelin performansını düşürmeyi hedefler. Saldırganlar, modelin yanlış sonuçlar üretmesine veya belirli bir sınıfa karşı yanlış olmasına neden olabilir. Model zehirlenme (model poisoning) saldırılarında ise kötü niyetli katılımcılar, model güncellemelerini manipüle ederek merkezi modele zarar vermeyi amaçlar. Bu, modelin genel doğruluğunu ve güvenilirliğini azaltabilir (Yang, 2019). Şekil 4'te Literatürde yer alan FL'deki saldırı türleri gösterilmiştir.



Şekil 4 FL’de güvenlik saldırıları

Sonuç olarak, FL, veri gizliliği ve güvenliğini ön planda tutarak, merkezi veri depolama gereksinimini ortadan kaldıran yenilikçi bir makine öğrenimi yöntemidir. Ancak, güvenlik tehditleri ve saldırılara karşı savunma mekanizmalarının geliştirilmesi, FL’nin başarılı bir şekilde uygulanması için kritik bir gerekliliktir. FL sistemlerinin güvenilirliği ve etkinliği, yalnızca veri gizliliğinin sağlanmasıyla değil, aynı zamanda güçlü güvenlik önlemlerinin de alınmasıyla mümkün olabilir.

Bir sonraki bölümde, FL’deki güvenlik saldırılarını ayrı başlıklar altında ayrıntılı olarak inceleyeceğiz. Bu kapsamda, her iki saldırı türünün detayları, etkileri ve saldırılara karşı alınabilecek önlemler ele alınacaktır.

1.3.1. Veri Zehirlenmesi Saldırıları

FL’de veri zehirlenmesi saldırıları, kötü niyetli katılımcıların, yerel veri setlerinde kasıtlı değişiklikler yaparak modelin doğruluğunu ve güvenilirliğini bozmayı amaçladığı bir tehdit türüdür. Saldırıları, özellikle FL’nin merkezi olmayan doğasından dolayı büyük bir risk teşkil eder. Veri zehirlenmesi saldırılarının temel amacı, modelin çıktılarının güvenilmez hale gelmesini sağlamaktır (Fung, 2018).

Veri Zehirlenmesi saldırıları genellikle iki ana kategoriye ayrılır: Temiz etiket saldırıları ve Kirli etiket saldırıları.

Temiz Etiket veri zehirlleme saldırıları, makine öğrenimi modellerine yönelik sofistike bir saldırı türüdür. Saldırgan, eğitim setine doğru etiketlenmiş ancak özellikleri manipüle edilmiş veriler ekleyerek modelin test aşamasında belirli bir örneği yanlış sınıflandırmasını amaçlar. Saldırgan, hedef sınıfa ait bir veri seti içinden bir örnek seçer ve örnek üzerinde küçük değişiklikler yaparak hedeflenen diğer sınıfa benzeyecek şekilde manipüle eder. Manipüle edilmiş veri, doğru etiketiyle eğitim setine eklenir ve model manipüle verilerle eğitilir. Eğitim sürecinde model, manipüle edilmiş verileri doğru etiketiyle öğrenir; ancak süreçte sınıflandırma sınırları kaymaya başlar. Kayma, test aşamasında saldırırganın hedeflediği örneğin yanlış sınıflandırılmasına yol açar (Shafahi, 2018).

Örneğin, saldırırgan bir modelin köpek fotoğraflarını yanlışlıkla balık olarak sınıflandırmasını sağlamak isteyebilir. Saldırgan, balık sınıfına ait bir eğitim verisi seçer ve veriyi küçük ama fark edilmesi zor değişikliklerle köpek fotoğrafına benzetir. Manipüle edilen veri, hala balık olarak görüldüğü ve doğru bir şekilde balık olarak etiketlendiği için modelin eğitim setine eklenir. Model, manipüle edilmiş verilerle eğitildiğinde zehirlenmiş balık fotoğraflarını öğrenir ve sınıflandırma sınırlarını manipüle edilmiş verilere göre ayarlar. Sonuç olarak, test aşamasında model hedeflenen köpek fotoğrafını gördüğünde bu fotoğrafı balık olarak sınıflandırma eğiliminde olur. Saldırgan, stratejisiyle modelin genel performansını bozmadan sadece belirli bir hedef üzerinde istenilen yanlış sınıflandırmayı elde eder.

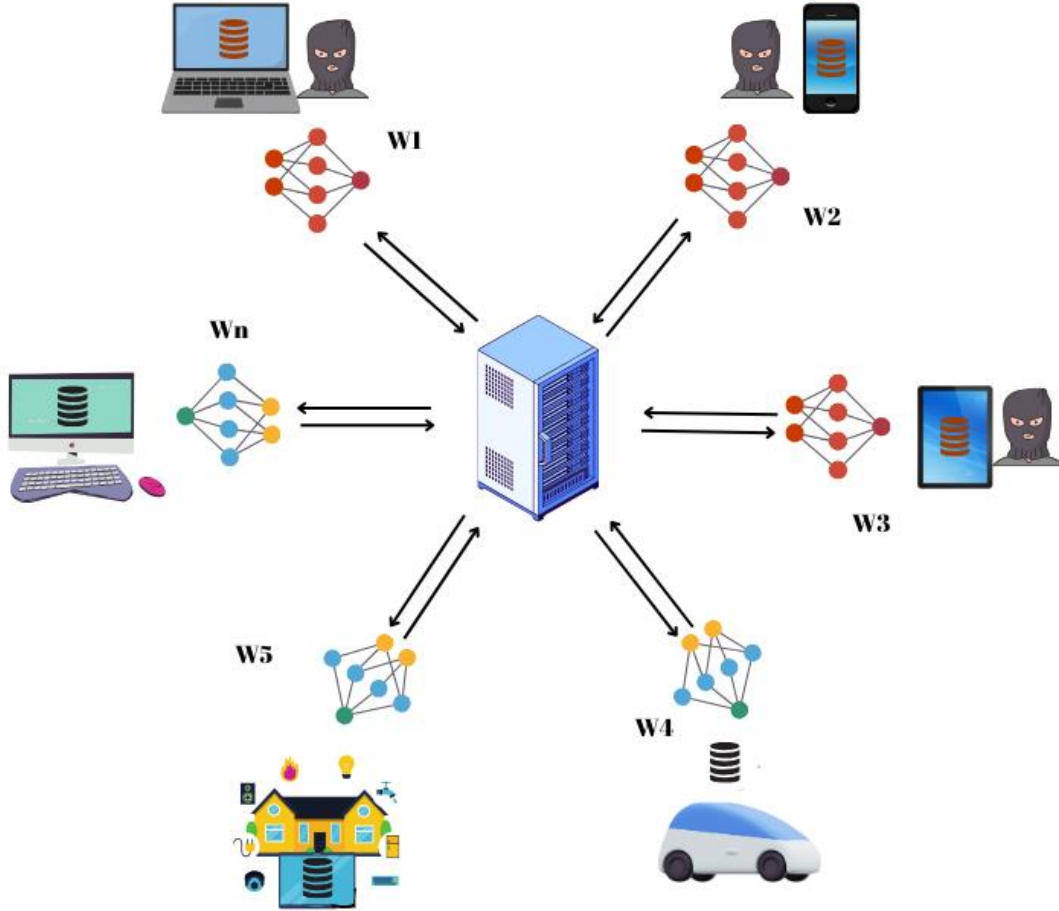
Başka bir saldırı türü ise, veri setlerine gürültü ekleyerek yapılan saldırılardır. Kötü niyetli katılımcılar, verilerine rastgele veya belirli bir desenle gürültü ekleyerek verilerin bozulmasına neden olurlar. Gürültülü verilerle eğitilen modeller, gerçek dünya verilerine karşı duyarlılığını kaybeder ve tahminlerde ciddi hatalar yaparak ana modelin güvenilirliğini ve doğruluğunu düşürür. Gürültü ekleme saldırıları, özellikle regresyon problemlerinde modelin tahmin performansını büyük ölçüde etkileyebilir. Veri zehirlenmesi saldırılarının amacı, ana modelin performansını ve güvenilirliğini düşürerek, modelin çıktılarının güvenilirmez hale gelmesini sağlamaktır. Saldırırlara karşı savunma stratejileri geliştirmek, FL sistemlerinin güvenliği için kritik öneme sahiptir. Temiz etiket saldırıları, etiketler doğru olduğu için tespit edilmesi zor, ancak makine öğrenimi sistemleri için ciddi bir güvenlik riski oluşturan bir saldırı türüdür.

Temiz etiketli saldırılarda, saldırırganın eğitim verilerinin etiketlerini değiştiremeyeceği varsayılır çünkü verilerin doğru etiketlerle sınıflandırıldığını garanti eden bir süreç mevcut olduğu için verilerin sınıfları değiştirilemez. Ancak kirli etiketli zehirlleme saldırılarında

saldırgan, yanlış sınıflandırmayı amaçladığı veri örneklerini, hedeflenen etiketlerle eğitim setine ekleyebilir. Kirli etiket saldırıları sınıfların doğru etiketlendiğini garanti eden bir algoritmanın olmadığı sistemler üzerinde gerçekleştirilir (Lyu, 2020).

Kirli Etiket saldırılarından da en bilinen saldırı türü Etiket Çevirme (label flipping) saldırıdır. Etiket çevirme saldırılarında, kötü niyetli katılımcılar belirli bir sınıfın veya tüm veri setinin etiket bilgilerini kasıtlı olarak birbirleriyle değiştirir. Yanlış etiketli verilerle eğitilen yerel modeller, merkezde birleştiğinde ana modelin doğruluğunu önemli ölçüde azaltır. Yanlış etiketleme modelin belirli sınıfları yanlış tanımasına neden olur ve modelin genel performansını olumsuz etkiler. Etiket Çevirme saldırıları, özellikle sınıflandırma problemlerinde büyük bir tehdit oluşturur ve modelin belirli sınıflara karşı yanlı hale gelmesine yol açar.

Veri zehirlenmesi saldırılarının işleyişi Şekil 5'te gösterilmiştir.



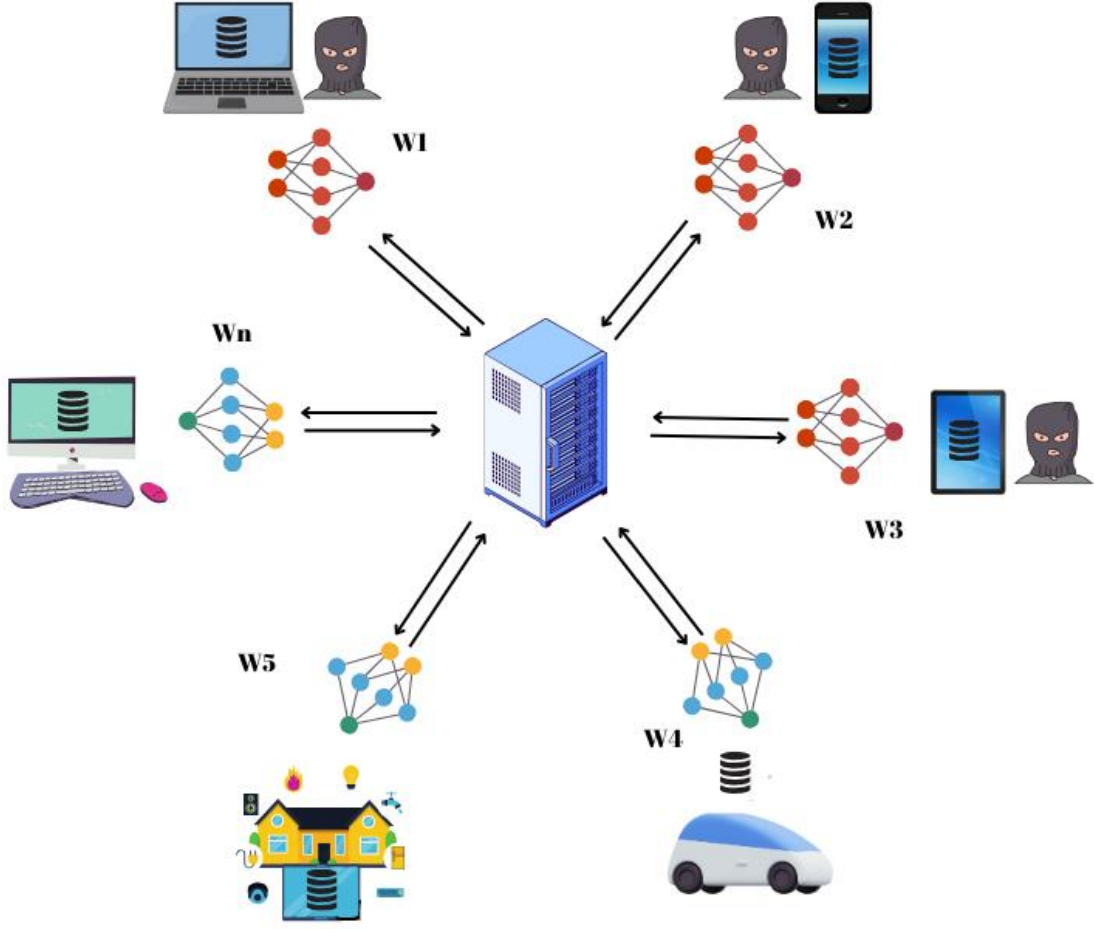
Şekil 5 Veri Zehirlenmesi

Veri zehirlenmesi saldırılarının engellenmesi ve saldırılara karşı dirençli sistemler geliştirilmesi, FL'nin güvenli ve etkili bir şekilde uygulanabilmesi için hayati öneme sahiptir. Bu nedenle, veri zehirlenmesi saldırılarının mekanizmalarının anlaşılması ve saldırılara karşı uygun savunma stratejilerinin geliştirilmesi üzerine daha fazla araştırma yapılması gerekmektedir.

1.3.2. Model Zehirlenmesi Saldırıları

Model zehirlenmesi saldırılarında, kötü niyetli katılımcılar yerel eğitim sürecinden bağımsız olarak doğrudan model gradyanlarında değişiklik yaparlar. Saldırganlar modelin öğrenme sürecini manipüle ederek genel performansını düşürmeyi amaçlar. Örneğin, işaret çevirme (sign flipping) saldırılarında kötü niyetli katılımcılar model güncellemelerindeki gradyan işaretlerini tersine çevirirler. Modelin öğrenme süreci bozularak ana modelin doğruluğu ve güvenilirliği ciddi şekilde zedelenir (Blanchard, 2017). Hedefsiz ataklarda, kötü niyetli katılımcılar rastgele zararlı güncellemeler göndererek modelin tamamına yönelik bir saldırı yaptıkları için bütün model çıktısını bozmaktadır.

Sonuç olarak, model zehirlenmesi FL sistemlerinde önemli güvenlik açıklarına yol açar. Saldırganlar modelin doğruluğunu, güvenilirliğini ve genel performansını olumsuz etkiler. Saldırılara karşı etkili savunma stratejileri geliştirmek FL'nin güvenliği için kritik öneme sahiptir. Savunma stratejileri arasında anomali tespiti, güvenli model güncellemeleri ve sağlam toplama teknikleri bulunmaktadır. Şekil 6'da model zehirlenmesi saldırı türlerinin çalışma prensipleri açıklanmıştır



Şekil 6 Model Zehirlenmesi

1.3.3. Bedavacılık Saldırıları

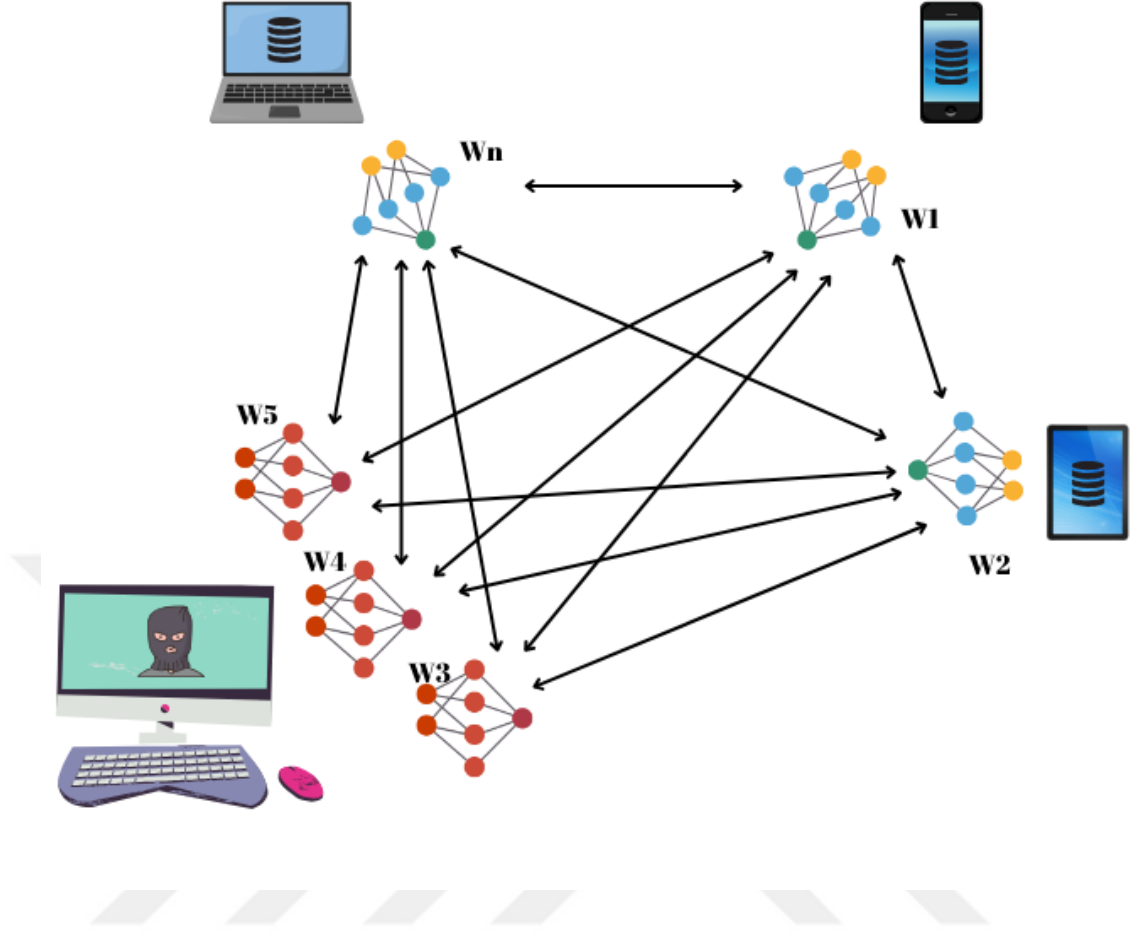
Bedavacılık saldırıları, FL sistemlerinde, hiçbir katkı yapmadan sistemden haksız kazanç sağlamayı amaçlayan katılımcılar tarafından gerçekleştirilen saldırılardır. (Lin, 2019). Katılımcıların yerel verilerini kullanarak model eğitimi yapmaları ve güncellemeleri merkezi sunucuya göndermeleri beklenir. Ancak bedavacılar gerçek veri kullanmadan rastgele veya sahte güncellemeler gönderir. Amaç, diğer katılımcıların çabalarından yararlanarak ödüller veya modelin nihai kullanım hakkını elde etmektir. Bedavacılık saldırıları FL'in ortak model oluşturma sürecini bozar ve modelin performansını düşürür ve FL'in gizlilik ve veri güvenliği avantajlarını suistimal eder. Bedavacılar sistemin sunucusunu kandırarak sanki gerçekten model eğitimi yapmış gibi görünerek adil katılımcıların çabalarını boşa çıkarabilir ve sistemin güvenilirliğini zedeler. Bedavacılık saldırılarını tespit

etmek zor olabilir çünkü sahte güncellemeler genellikle diğer katılımcıların güncellemeleriyle benzer görünmeye çalışır.

Bedavacılık saldırılarını önlemek için FL sistemleri kötü niyetli veya fırsatçı hareketleri caydırarak dürüst katılımı desteklemelidir. Bunun bir yolu, yüksek kaliteli güncellemeler yapan müşterileri ödüllendirirken fırsatçılık yapanları cezalandıracak teşvik sistemleri oluşturmaktır. Ayrıca FL çerçeveleri, saldırıların etkisini en aza indirmek ve iş birliğini artırmak amacıyla model doğrulama, ödül çeşitliliği ve müşteri seçimi stratejileri gibi yöntemler kullanabilir. Yaklaşımlar katkı sürecinde adaleti sağlayarak ücretsiz saldırıları etkili bir şekilde engelleyebilir ve katılımcılar arasında iş birliğini teşvik edebilir (Zhang, 2024).

1.3.4. Sybil Saldırısı

Sybil saldırısı merkezi olmayan ağlarda, bir saldırganın birden çok sayıda sahte kimlik oluşturmasıyla gerçekleşir(Douceur, 2002). Sahte kimlikler, ağdaki güvenilir katılımcıları oy çoğunluğuyla yenerek ağın güvenliğini ve dayanıklılığını zayıflatır. Sonuç olarak, işlemlerin sıralaması değiştirilebilir ve onaylanmaları engellenebilir. Kötü niyetli bir katılımcı, FL’de birden fazla katılımcı gibi davranarak ana modelin hesaplanma sürecine birden fazla kez katılabilir. Sybil katılımcı, saldırıları çeşitli amaçlarla gerçekleştirebilir; örneğin, eğer Sybil katılımcı ana modelin belirli bir sınıfın etiketleme başarısını düşürmek istiyorsa, sadece o sınıfın verilerini manipüle ederek oluşturduğu modeli çoğaltarak birden fazla kez ana model hesaplamasına katılabilir. Alternatif olarak, tüm modelin başarısını düşürmek istiyorsa, birden fazla zararlı model üreterek modelleri sahte katılımcı kimlikleriyle ana model hesaplanma sürecine dahil edebilir. Saldırının şeması Şekil 7’de verilmiştir. Şekilde bir Sybil katılımcının 3 model oluşturarak ana model oluşum sürecine etki ettiği gösterilmektedir.



Şekil 7 Sybil Saldırısı

1.3.5. Arka Kapı Saldırıları

Arka Kapı saldırıları, FL’de sinsi ve tehlikeli bir güvenlik tehdidi oluşturur. Arka kapı saldırılarında, kötü niyetli katılımcılar modeli eğitirken gizli arka kapılar eklerler. Arka kapılar, modelin belirli koşullar altında beklenmedik ve istenmeyen davranışlar sergilemesine neden olur. (Bagdasaryan, 2020)

Örneğin, bir yüz tanıma modeline eklenen bir arka kapı, belirli bir aksesuar veya işaret kullanıldığında modelin bir kişiyi yanlış tanımlamasına neden olabilir. Modelin genel performansını olumsuz etkilemeden sadece spesifik koşullar altında devreye girer. Saldırganlar, arka kapıyı yerleştirirken modelin genel doğruluğunu ve güvenilirliğini bozmak istemezler. Bunun yerine, arka kapı yalnızca hedeflenen spesifik durumlar ortaya çıktığında etkinleşir. Böylece model, normal kullanımda beklenen performansını sürdürürken, saldırganın belirlediği özel durumlarda gizli işlevini yerine getirir. Arka kapı ekleme yaklaşımı, saldırının fark edilmesini zorlaştırır çünkü modelin genel davranışında

belirgin bir bozulma gözlemlenmez. Örneğin, saldırgan bir görüntü sınıflandırma modelini manipüle ederek modelin belirli renklerdeki nesneleri yanlış sınıflandırmasını sağlayabilir, ancak yalnızca belirli bir renk kombinasyonu veya belirli bir şekil ortaya çıktığında devreye girer.

1.3.6. Bizans Saldırıları

Literatürde Bizans saldırıları olarak bilinen saldırılar, katılımcıların eğitim sürecinde ya da eğitim sonrasında model gradyanlarını rastgele veya stratejik olarak değiştirerek gerçekleştirilir. Kötü niyetli katılımcılar, doğru öğrenme sürecini sabote etmek amacıyla model güncellemelerine kasıtlı olarak yanlış gradyanlar eklerler. Modelin belirli görevlerde başarısız olmasına neden olabilirler ve genel güvenilirliğini azaltırlar.

Bizans saldırılarında, saldırganlar modelin çıktılarının yanlış olmasına neden olacak hatalı bilgileri bilinçli olarak sağlarlar. Rastgele veya stratejik olarak değiştirilmiş gradyanlarla, modelin belirli görevlerde tamamen başarısız olmasına yol açabilirler. Saldırganlar, modelin kullanıcıların gerçek ihtiyaçlarını karşılamasını zorlaştırır ve genel güvenilirliğini sarsar. Bizans saldırılarının temel amacı, modelin doğruluğunu ve güvenilirliğini azaltmaktır. Merkezi sunucu, kullanıcıların modellerinin doğru verilerle eğitilip eğitilmediğine dair bir garantiye sahip olmadığından, Bizans türü saldırılar modelin performansını ciddi şekilde etkiler. Bizans saldırıları, modelin güvenliğini ve performansını tehdit eder, çünkü modelin doğru verilerle eğitildiğine dair merkezi sunucunun bir kontrolü yoktur.

Örneğin, kötü niyetli bir katılımcı, yerel model güncellemeleri sırasında bilinçli olarak yanlış gradyanlar hesaplayabilir ve merkezi sunucuya iletebilir. Merkezi sunucu, yanlış gradyanları kullanarak güncellenmiş modeli oluşturduğunda, modelin doğruluğu ve güvenilirliği ciddi şekilde zedelenir. Bizans türü saldırılar, modelin belirli görevlerde tamamen başarısız olmasına veya yanlış sonuçlar üretmesine neden olabilir.

Bizans saldırıları, modelin doğruluğunu ve güvenilirliğini hedef alarak modelin çıktılarının güvenilmez hale gelmesini sağlar. Saldırganlar, modelin öğrenme sürecini manipüle ederek modelin genel performansını düşürmeyi amaçlar.

Bir sonraki bölümde, saldırılara karşı geliştirilen savunma yöntemleri kapsamlı bir şekilde incelenecektir. Literatürde öne çıkan güvenli model güncellemeleri ve birleştirme yöntemleri, katılımcıların güncellemelerini güvenli bir şekilde toplarken merkezi sunucunun

bu güncellemelerin doğruluğunu ve güvenilirliğini kontrol etmesini sağlar. Anomali tespit sistemleri ve sağlam toplama teknikleri, zararlı güncellemelerin etkilerini en aza indirmek için önemli savunma mekanizmaları olarak değerlendirilmektedir (Bonawitz, 2019). gelecek bölümde, yöntemlerin etkinliği ve uygulanabilirliği üzerinde durulacaktır.

1.4. Güvenli Federe Öğrenme Yöntemleri

Güvenli FL yöntemleri, genellikle iki ana başlık altında incelenir: anomali tespiti ve izleme yöntemleri ile güçlü birleştirme teknikleri. Bu başlıklar, FL süreçlerinde güvenliği ve veri bütünlüğünü sağlamak için kritik öneme sahiptir. Sonraki bölümde her iki yaklaşıma ilişkin detaylar sunulacaktır.

1.4.1. Anomali Tespiti ve İzleme

Anomali tespiti ve izleme, FL süreçlerinde katılımcılardan gelen verilerdeki olağandışı veya kötü niyetli davranışları tespit ederek sistemin güvenilirliğini ve bütünlüğünü korumak için kullanılır. Literatürde yapılan çalışmalar, farklı yöntem ve yaklaşımlar çerçevesinde incelenmiş ve tekniklerin FL süreçlerine katkıları detaylandırılmıştır.

Ece ve arkadaşları (Işık Polat, 2023) tarafından tanıtılan ARFED (Attack-Resistant Federated Averaging based on Outlier Elimination) yöntemi, FL ortamlarında güvenliğini artırmayı amaçlar. Veri ve model zehirlenme saldırılarına karşı dayanıklılığı sağlamak için çeyrekler açıklığı (IQR) tabanlı bir koordinat düzleminde çalışır. Çeyrekler açıklığı, veri setini %25, %50 ve %75 dilimlerine böler. %25'lik dilim, verilerin alt çeyreğini, %50 medyanı ve %75'lik dilim ise verilerin üst çeyreğini temsil eder. IQR, Q1 ve Q3 arasındaki fark olarak tanımlanır ve verilerin merkezi eğilimlerini belirlemede kullanılır. ARFED, heterojen veri setlerine karşı etkinlik göstermiştir ve çeşitli çalışmalarda heterojen veri setleri üzerinde test edilmiştir. Yöntem, organize ve organize olmayan saldırılara karşı test edilmiştir. Organize saldırılarda, belirli bir koordinasyon içerisinde hareket eden saldırganlar, yöntemin zayıflıklarını daha etkin bir şekilde kullanabilmektedir. Organize olmayan saldırılarda ise, bireysel saldırganların rastgele hareketleri dikkate alınmıştır. Bizans saldırıları durumunda, kötü niyetli kullanıcı oranı %25'i aştığında ARFED'in performansında belirgin bir düşüş gözlemlenmiştir. Düşüş, çeyrekler açıklığı tabanlı yaklaşımın %25'lik dilimi aşan saldırgan oranlarında etkisiz kalmasından

kaynaklanmaktadır. Özellikle, %25'lik dilim aşıldığında, saldırganlar güvenli alan olarak değerlendirilen bölgeye girmiş olur ve ARFED'in saldırganları tespit etme kapasitesini azaltır. Kötü niyetli kullanıcı oranı %25'i geçtiğinde, özellikle organize saldırılar sırasında, yöntemin performansında ciddi kayıplar yaşanır.

Eda ve arkadaşları (Erdol, 2024) tarafından önerilen yöntemde ise, kullanıcıların model güncellemelerinden en çok değişen nöronları belirleyerek çalışır. İlk adımda nöronların boyutları Principal Component Analysis (PCA) kullanılarak küçültülür. Daha sonra, boyutları küçültülen vektörler One-Class Support Vector Machines (OC-SVM) yöntemi ile analiz edilerek kötü niyetli kullanıcılar tespit edilir. Belirlenen kötü niyetli kullanıcılar eğitim döngüsünden çıkarılarak modelin zehirlenme saldırılarına karşı daha dayanıklı olmasını sağlar. Önerilen yöntem, üç farklı veri kümesi üzerinde test edilmiş ve etkinliği gösterilmiştir.

Li ve arkadaşları (Li, 2019) RSA yöntemini önerdi. RSA, her işçinin hesapladığı gradyanları bir araya getirirken, olası hatalı gradyanları düzenler. Her gradyan, genel gradyanla olan farkına göre bir düzenleme terimine tabi tutulur. Eğer bir işçinin gradyanı genel gradyandan farklıysa, farkın işareti kullanılarak gradyana bir terim eklenir. Yapılan düzenleme ile gradyanın genel modelle daha uyumlu hale gelmesini sağlar ve hatalı veya kasıtlı olarak yanlış gönderilen gradyanların etkisini dengeler.

Ancak, RSA yöntemi, gradyan yönüyle ilgili bozucular ekleyerek belirli saldırı türlerini, özellikle yön tersine çevirme gibi saldırıları önlemek için tasarlanmıştır. Bununla birlikte, Min-Max saldırıları gibi bazı saldırı türleri, gradyan yönüyle düşük bir korelasyon gösterir. RSA'nın Min-Max gibi saldırılar karşısında etkisiz kalabileceği göz önünde bulundurulmalıdır. Sonuç olarak, RSA yöntemi genel olarak hatalı verilerin olumsuz etkilerini azaltmada etkili olsa da, belirli saldırı senaryolarında savunma gücü sınırlı olabilir.

Shejwalkar ve arkadaşları (Shejwalkar, 2021) DnC (Divide-and-Conquer) algoritmasını önerdi. DNC, her iterasyonda rastgele bir koordinat alt kümesi seçerek güncellemeleri analiz eder. Alt kümedeki güncellemelerin ortalaması alınır ve ardından güncellemeler, ana bileşen analizi kullanılarak en yüksek varyansa sahip bileşen boyunca projekte edilir. Projeksiyonlar kullanılarak her güncelleme için bir "aykırı değer" puanı hesaplanır. Belirli bir parametreye göre en yüksek aykırı değer puanlarına sahip güncellemeler hariç tutulur ve geri kalan güncellemeler "iyi" olarak kabul edilir. Belirli sayıda iterasyon boyunca tekrarlanır ve her seferinde farklı koordinat alt kümeleri seçilir.

Son olarak, iyi olarak kabul edilen güncellemelerin ortalaması alınarak güvenilir bir model güncellemesi elde edilir.

Ancak, DnC'nin verimli çalışabilmesi için her iterasyonda kaç güncelleme için aykırı değer olarak kabul edilip hariç tutulacağını belirlemek için kullanıcıdan bir parametre girişi gereklidir. Bu, FL'nin dinamik ve heterojen doğası göz önüne alındığında optimal olmayabilir. FL'de katılımcılar arasında veri dağılımları ve hesaplama yetenekleri büyük ölçüde değişebilir, bu da sabit bir parametrenin her duruma uyum sağlayamamasına neden olabilir. Ayrıca, sabit bir parametre belirlemek, kötü niyetli güncellemelerin doğru bir şekilde filtrelenmesini zorlaştırabilir ve sistemin güvenilirliğini riske atabilir. FL ortamlarında aykırı değerlerin adaptif ve veri odaklı tekniklerle dinamik olarak belirlenmesi, daha esnek ve güvenli bir çözüm sağlayabilir.

Nguyen ve arkadaşlarının önerdiği (Nguyen, 2022) FLAME arka kapı saldırılarını engellemeye yönelik geliştirilen bir savunma framework'üdür. Çalışma prensibi üç temel bileşene dayanır: dinamik model filtreleme, adaptif kesme ve gürültü ekleme. Dinamik model filtreleme, farklı model güncellemeleri arasındaki açısal farklılıkları analiz eden ve büyük sapmalar gösteren modelleri tespit eden bir kümeleme algoritması kullanır. Sapmalar tespit edildiğinde, ilgili model güncellemeleri küresel modelden çıkarılır, böylece saldırının etkisi minimize edilir. Adaptif kesme, model güncellemelerinin ağırlıklarını belirli bir sınırdan tutarak saldırının manipülasyon etkisini azaltır. Kesme sınırı, modelin zararsız güncellemelerini koruyacak şekilde dinamik olarak belirlenir. Son olarak, FLAME, küresel modele dikkatlice ayarlanmış bir miktarda Gauss gürültüsü ekler. Eklenen gürültü, arka kapı saldırılarının kalıntı etkilerini ortadan kaldırır ve modelin ana görevindeki performansını büyük ölçüde korur.

FLAME, üç bileşeni bir araya getirerek, FL süreçlerinde güvenliği artırır ve saldırıların manipülasyonlarına karşı sistemi dayanıklı hale getirir. Ancak, FLAME'in önemli bir eksiği, istemcilerin gizliliğini korumak için kullanılan Güvenli Çok Partili Hesaplama protokollerinin ciddi bir performans yükü getirmesidir. Protokol birden fazla tarafın özel verilerini güvenli bir şekilde paylaşmasını sağlar, ancak süreç yüksek hesaplama maliyetleri doğurur ve sistemin ölçeklenebilirliğini sınırlayabilir. Bu nedenle, FLAME arka kapı saldırılarına karşı etkili bir savunma sağlarken, gizlilik koruması ve performans optimizasyonu açısından bazı zorluklarla karşı karşıya kalır.

Cao ve arkadaşları (Cao D. a., 2019) Sniper algoritmasını önerdiler. Sniper zehirlenmiş yerel modelleri tespit edip filtrelemeyi amaçlayan etkili bir yöntemdir.

Katılımcılardan toplanan yerel modeller arasındaki Öklid mesafelerini hesaplayarak, birbirine yakın modelleri belirler. Elde edilen benzer modeller, bir grafik üzerinde maksimum klik probleminin çözümüyle tanımlanır ve güvenilir modeller küresel modelin güncellenmesinde kullanılır. Sonuç olarak, saldırganlar tarafından zehirlenmiş modellerin küresel modeli olumsuz yönde etkilemesi engellenir.

Yöntemin güçlü yönlerinden biri, saldırganlar tarafından gönderilen zehirlenmiş modellerin etkili bir şekilde filtrelenmesini sağlamasıdır. Bu sayede, küresel modelin doğruluğu korunur ve sistemin güvenliği artırılır. Ayrıca, Sniper mekanizması, farklı FL senaryolarında esnek bir şekilde uygulanabilir.

Ancak Sniper yönteminin bazı eksiklikleri de bulunmaktadır. En önemli eksikliklerden biri, eşik değeri (γ) belirleme sürecidir. Eşik değeri, modeller arasındaki mesafenin ne kadar yakın olması gerektiğini belirler ve eşik değeri deneysel olarak optimize edilmelidir. Yanlış bir eşik değeri seçimi, güvenilir modellerin göz ardı edilmesine veya zehirlenmiş modellerin küresel modele dahil edilmesine yol açabilir. Eğer eşik değeri çok düşük seçilirse, güvenilir modellerin bazıları filtrelenebilir ve küresel modelin doğruluğu olumsuz etkilenebilir. Öte yandan, eşik değeri çok yüksek seçilirse, saldırganlar tarafından zehirlenmiş modeller de küresel model güncellemesine dahil olarak saldırıların etkisini artırabilir.

Bu eksiklik, Sniper mekanizmasının her senaryoda optimal sonuçlar vermesini zorlaştırabilir ve yöntem, farklı veri setleri ve saldırı türleri için ayrı ayrı ayarlanmayı gerektirebilir. Eşik değerinin belirlenmesi, Sniper yönteminin en kritik adımlarından biri olup, dikkatli bir şekilde ele alınması gereken bir zayıflıktır.

Li ve arkadaşları (Li S. a., 2020) FL sistemlerinde kötü niyetli istemcileri tespit etmek için spektral anomali tespitine dayalı bir yöntemi tanıttılar. Yöntem model güncellemelerindeki anormallikleri belirlemek için spektral anomali tespiti tekniğini kullanmaktadır. Spektral anomali tespiti, FL sistemindeki istemcilerin merkezi sunucuya gönderdiği model güncellemelerini analiz eder. Analiz sonucunda güncellemelerin düşük boyutlu temsilleri çıkarılır. Düşük boyutlu temsiller, güncellemelerdeki önemli özellikleri korurken gereksiz ve gürültülü verileri filtreler. Kötü niyetli istemcilerden gelen zararlı model güncellemeleri, normal güncellemelerden farklı özelliklere sahiptir. Farklılıklar, spektral anomali tespiti ile belirginleşir ve anomali tespit edilerek anormal davranışlar olarak sınıflandırılır. Çalışmada, anomali tespiti gerçekleştirmek için Varyasyonel Otomatik Kodlayıcı (VAE) modeli kullanılmıştır. VAE, normal model güncellemelerinin özelliklerini öğrenir. Öğrenme sonucunda, anormal güncellemeler yüksek yeniden yapılandırma hataları

ile tespit edilir. Her güncelleme turunda, istemcilerden gelen güncellemelerin yeniden yapılandırma hatalarına göre dinamik olarak bir eşik değeri belirlenir. Eşik değerinin üzerinde hata oranına sahip olan güncellemeler kötü niyetli olarak kabul edilir. Kötü niyetli güncellemeler model birleştirme sürecinden hariç tutulur. Savunma mekanizması, kötü niyetli güncellemelerin etkisini ortadan kaldırmayı hedefler.

Chen ve arkadaşları, (Chen H. a., 2023) FL sistemlerinde Bizans, işaret çevirme, Fang, LIE ve Min-Max gibi saldırılara karşı dayanıklı bir savunma algoritması olan FLRAM'ı (Federated Learning Robust Aggregation Method) geliştirmiştir. FLRAM yerel gradyanların genlik ve sembollerini analiz ederek anomali tespiti yapar. İlk adımda, gelen yerel gradyanların genliklerinde anormallikler belirlenir ve anomali tespiti, İzolasyon Ormanı (Isolation Forest) algoritması kullanılarak gerçekleştirilir. İzolasyon Ormanı, veri noktalarının izole edilme kolaylığına dayanan bir anomali tespit algoritmasıdır ve anormal veri noktalarını tespit etmekte etkilidir. Genlik analizinden sonra, gradyanların sembollerinde (yönlerinde) anormallikler tespit edilir. Bu adımda IDBSCAN (Incremental Density-Based Spatial Clustering of Applications with Noise) algoritması kullanılır. IDBSCAN, yoğunluk temelli bir kümeleme algoritmasıdır ve zaman içinde güncellenen veri kümelerinde anormal sembol davranışlarını belirler. İzolasyon Ormanı ve IDBSCAN algoritmaları ile belirlenen anormal gradyanlar toplama sürecinden çıkarılır, böylece kötü niyetli katılımcıların model performansını bozması engellenir. Anomaliler çıkarıldıktan sonra, güvenilirlik matrisi oluşturulur. Güvenilirlik matrisi, her kullanıcının yerel gradyanlarının güvenilirliğini değerlendirir. Güvenilirlik puanları, kullanıcıların geçmiş performansları ve mevcut gradyanlarının güvenilirlik analizleri baz alınarak hesaplanır. Güvenilirlik matrisine göre en yüksek güvenilirlik puanına sahip gradyanlar seçilir. Güvenilir gradyanlar, küresel modelin güncellenmesi için kullanılır. Böylece, FL modelinin doğruluğu ve güvenliği korumayı amaçlar.

1.4.2. Güçlü Birleştirme Teknikleri

Güçlü birleştirme teknikleri, FL süreçlerinde merkezi sunucuya gönderilen model güncellemelerinin güvenilir bir şekilde birleştirilmesini sağlayarak kötü niyetli veya hatalı katılımcılardan gelen güncellemelerin etkisini azaltmak için kullanılır. Başlıca güçlü birleştirme teknikleri şunlardır:

Yin ve arkadaşları (Yin, 2018) model parametrelerinden maksimum ve minimum değerleri çıkartarak geri kalanların ortalamasını hesaplayan Trimmed Mean ve model parametrelerine dayalı medyanı hesaplayan CwMedian yöntemleri tanıtılmıştır. TRIMMED yöntemi, her bir model parametresi için gelen değerlerin en küçüğünden en büyüğüne doğru sıralanmasını ve ardından en küçük ve en büyük belirli sayıda değerlerin çıkarılmasını içerir. Geri kalan değerlerin ortalaması alınarak güvenli bir birleştirme sağlanır. Yöntem, uç değerlerin (outliers) etkisini azaltarak daha güvenilir bir ortalama sağlar. Ancak, kötü niyetli katılımcı oranının önceden bilinmesi gerektiğinden bazı kısıtlamaları vardır. Ancak, kötü niyetli katılımcı oranını önceden bilmeyi gerektirdiği için FL yapısına tam uyumlu değildir. CwMedian yöntemi ise, her bir model parametresi için gelen değerlerin medyanını hesaplar. Yerel modellerden gelen her bir parametre değeri sıralanır ve ortanca (medyan) değer küresel model parametresi olarak belirlenir. Sonuç olarak aşırı uç değerlerin etkisini minimize eder ve küresel modelin daha dayanıklı olmasını sağlar

Chang Zhang ve arkadaşları (Blanchard P. E., 2017) tarafından tanıtılan Krum yöntemi, her katılımcının model güncellemesini diğer tüm katılımcıların güncellemeleri ile olan mesafesini hesaplayarak başlar. Genellikle Öklid mesafesi olarak hesaplanır. Her güncelleme için, diğer güncellemeler arasından en yakın komşular belirlenir. Daha sonra, her güncelleme için seçilen en yakın komşuların mesafeleri toplanır ve toplam mesafe, güncellemenin güvenilirliğini temsil eder. En düşük toplam mesafe skoruna sahip güncelleme, merkezi modele dahil edilir. Ancak, tüm güncellemeler arasındaki mesafelerin hesaplanması ve en yakın komşuların belirlenmesi, özellikle büyük ölçekli sistemlerde yüksek hesaplama maliyetine yol açabilir. Ayrıca, Krum yöntemi, verilerin bağımsız ve aynı dağılıma sahip olduğu varsayımına dayanır, bağımsız ve aynı dağılıma sahip olmayan veri setlerinde performansının düşmesine neden olabilir.

Cao ve arkadaşları (Cao X. a., 2020) FLTrust yöntemini önermiştir. FLTrust'ta sunucu, istemcilerden gelen yerel gradyanları değerlendirirken kendi topladığı küçük, temiz bir doğrulama veri kümesine dayanarak bir güven puanı hesaplar. Doğrulama veri kümesi, sunucunun güvenilir bir model oluşturmak için başvurduğu temel dayanak noktasıdır. İstemcilerden gelen yerel gradyanlar, doğrulama veri kümesiyle karşılaştırılır; ardından gradyanların güvenilirliğini belirlemek için kosinüs benzerliği kullanılır ve güven puanları ReLU fonksiyonu ile işlenir. Sonuç olarak, güven puanlarına göre ağırlıklandırılmış gradyanlar birleştirilerek küresel model güncellenir.

Ancak, yöntemin etkinliği, sunucunun elindeki doğrulama veri kümesinin, genel eğitim verisi dağılımından çok fazla sapmamasına bağlıdır. Sunucu, böyle bir doğrulama veri kümesine erişemezse, özellikle çok sayıda kötü niyetli istemci bulunduğu anda, küresel model bozulabilir. Tipik bir federatif öğrenme senaryosunda, sunucunun doğrulanma veri kümesine sahip olmaması durumunda, FLTrust yöntemi kötü niyetli girişimleri engellemekte yetersiz kalabilir. Bu nedenle, FLTrust yönteminin tam anlamıyla başarılı olabilmesi için, sunucunun güvenilir ve temsili bir doğrulama veri kümesine sahip olması gereklidir.

Chen ve arkadaşları (Chen Y. a., 2017) Bizans gradyan inişi yöntemini tanıttılar (BGD). Geleneksel gradyan iniş algoritmasında, tüm gradyanlar toplanır ve ortalama alınarak model güncellenir, ancak yaklaşım hatalı verilerle karşılaştığında güvenilir hale gelebilir. BGD, bu sorunu aşmak için daha dayanıklı bir yaklaşım benimser. Öncelikle, tüm gradyanlar belirli gruplara ayrılır ve her grup için ortalama gradyan hesaplanır. Ardından, grup ortalamalarının geometrik medyanı alınır. Geometrik medyan, aykırı değerlere karşı ortalamadan daha dirençlidir ve hatalı gradyanların etkisini sınırlandırır. Böylece, merkezi sunucu, daha güvenilir bir model güncellemesi yapabilir. Özellikle güvenliğin kritik olduğu büyük ölçekli ve dağıtık makine öğrenimi uygulamalarında etkili bir şekilde kullanılır. BGD, sistemin hızlı ve doğru bir şekilde öğrenmesini sağlarken, hatalı makinelerin olumsuz etkilerini en aza indirir. Yöntem, belirli sayıda kötü niyetli katılımcıyı tolere edebilir, ancak belirli bir sayı aşıldığında başarısız olma riski taşır.

2. YAPILAN ÇALIŞMALAR

Bu çalışmada, Bizans ve Etiket Çevirme saldırılarına karşı dayanıklı bir FL yöntemi önerilmiştir. Literatürde yer alan saldırılara dayanıklı FL yöntemleri analiz edildiğinde, tez çalışmasında aşağıda belirtilen unsurların iyileştirilmesi hedeflenmiştir:

Saldırılarına dayanıklı FL yöntemlerinde genellikle düşük orandaki kötü niyetli katılımcı oranlarındaki yöntem başarısı incelenmiştir. Ancak tez kapsamında %40'a kadar olan kötü niyetli katılımcı oranlarında bile saldırılara dayanıklı bir sistem geliştirilmiştir.

Literatürdeki saldırılara dayanıklı algoritmalarda çoğunlukla tek tip saldırı türünde savunma gerçekleştirilmiştir. Örneğin, model zehirlenmesi saldırılarına odaklanan çalışmalar, Bizans ve işaret değiştirme gibi benzer amaç ve yöntemlere sahip saldırı yöntemlerine yönelmiştir. Bu çalışmada ise Etiket Çevirme ve Bizans saldırıları gibi çeşitli saldırı türlerinde etkinlik deneysel olarak gösterilerek literatüre katkı sağlanmıştır.

Literatürde yeni bir konu olan saldırganların organize ya da organize olmayan davranışları da deneysel olarak incelenmiş, saldırganların koordinasyon seviyelerinin sistem üzerindeki etkileri analiz edilmiştir.

2.1. Önerilen Yöntem

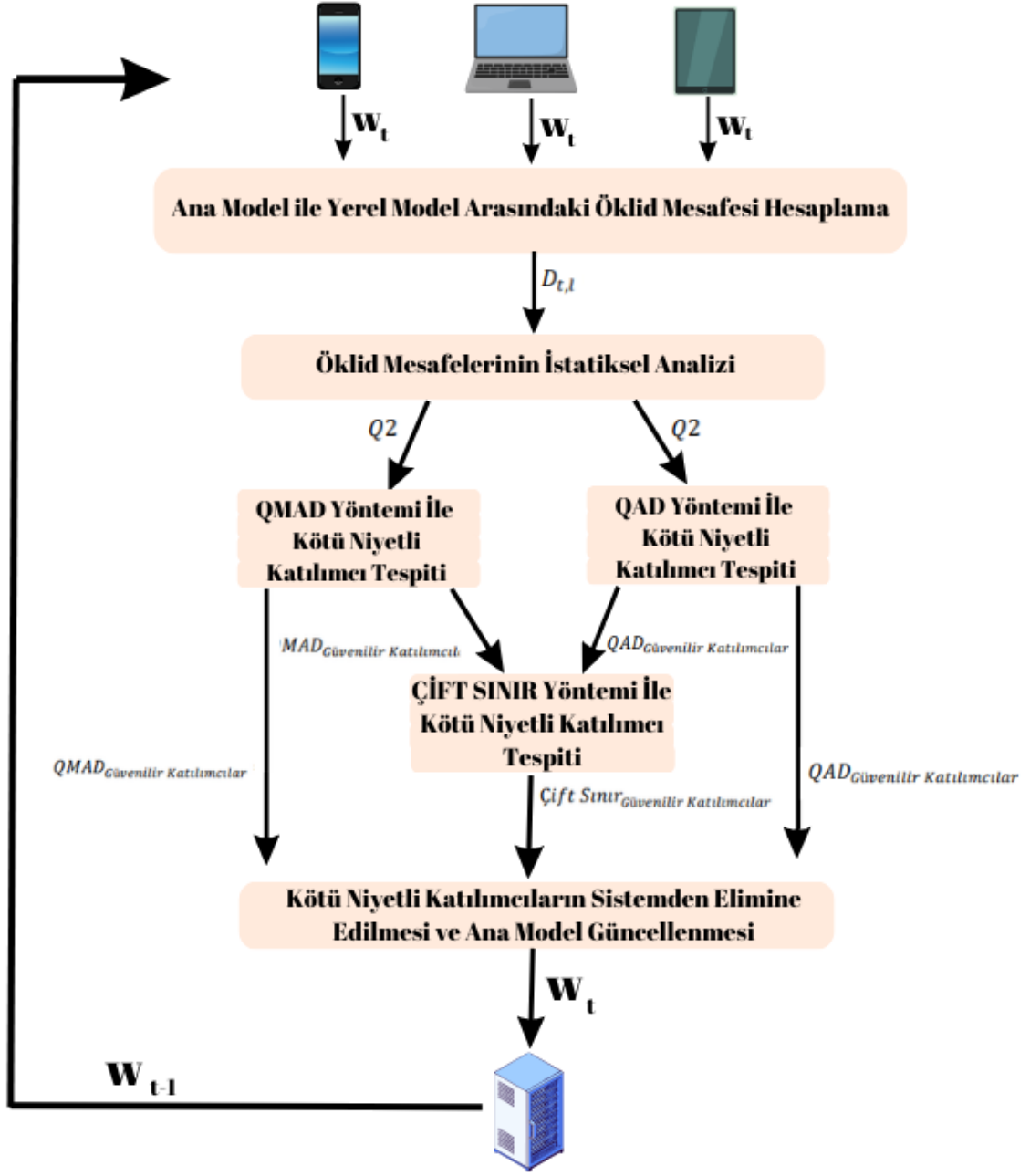
Tez kapsamında, FL modelinin son katmanında aykırı değerlerin tespit edilmesi için QAD, QMAD ve Çift Sınır kötü niyetli katılımcı tespit yöntemleri sunulmaktadır. Çalışmanın amacı, modelin son katmanında bulunan anormal verileri belirleyerek, verilerin sahibi olan kötü niyetli katılımcıların eğitime zarar vermesini engellemektir.

Önerilen yöntemlerde, Federated Averaging (FedAvg) (McMahan, 2017) algoritmasının güvenli bir versiyonu ele alınmaktadır. FedAvg, FL uygulamalarında yaygın olarak kullanılan bir algoritmadır (Işık Polat, 2023). Çalışma mantığı ise şu şekildedir: Güvenilir bir sunucu, her FL turunun başında ana modeli katılımcılara gönderir. Katılımcılar, modeli kendi yerel verileriyle eğitir ve yalnızca modelin gradyanlarını sunucuya geri gönderirler. Güvenilir sunucu, yerel modellerin gradyanlarını toplar ve ortalamalarını alarak ana modelin güncellenmiş bir versiyonunu oluşturur.

FedAvg'nin tercih edilme nedeni, literatürde FL güvenlik çalışmalarının büyük bir çoğunluğunun FedAvg algoritmasını kullanmasıdır. Ortak bir algoritmanın kullanılması

savunma mekanizmalarının etkinliğini açıkça gözlemlememizi sağlar. Tez kapsamında, literatürle uyumlu olarak FedAvg algoritmasının güvenli versiyonları sunulmuştur.

Kötü niyetli katılımcıların olmadığı bir FL ortamında, bir tur eğitim sonucunda katılımcıların yerel modellerinin ana modelden aşırı uzaklaşması beklenmez. Ancak, kötü niyetli katılımcıların yerel modelleri, güvenilir olanlardan ve ana modelden uzaklaşırlar. (Reddi, 2020). Bu sebeple önerilen yöntemlerin, FedAvg'nin toplama adımından hemen önce bir savunma mekanizması ortaya koyarak ana modelin doğruluğunu korumayı amaçlar. Şekil 8'da önerilen QAD, QMAD ve ÇİFT SINIR Yöntemlerinin akış şeması verilmiştir. Yöntemlerin üçü de ana modelin son katmanı ile katılımcıların son katmanı arasındaki uzaklığı hesaplayarak katılımcıların bir iterasyondaki model güncellemelerinin ana modelden ne kadar uzaklaştığına odaklanır. Güvenli katılımcıların model güncellemeleri belirli bir alan içerisinde kalacaktır. Alanı belirleyebilmek için ise üst ve alt sınırların konumunu bilmek gerekir. Alt ve üst sınır belirlenmesi için QAD, QMAD ve Çift Sınır yöntemleri önerilerek bütün güvenilir katılımcıları kapsarken, kötü niyetli katılımcıların güvenli belirlenen alan dışında kalacağı optimum değerler hesaplanır. Hesaplanan optimum değerler ile güvenilir katılımcı ile kötü niyetli katılımcıları ayırt eder. Güvenilir olarak işaretlenen katılımcılar toplama adımına katılırken, kötü niyetli katılımcı olarak işaretlenen katılımcılar hesaplama dışı çıkarılır. Sözel olarak açıklanan yöntemler sırasıyla şu adımları içermektedir: Son katmanlar arasındaki Öklid mesafesinin hesaplanması, mesafelerin istatistiksel analizi, önerilen savunma yöntemleri ile güvenilir katılımcılar ve ana model arasındaki optimum uzaklık değerlerinin hesaplanması, ardından kötü niyetli katılımcıların tespit edilmesi. Kötü niyetli katılımcıların belirlenmesinde QAD, QMAD ve Çift Sınır yöntemleri kullanılmıştır. Son olarak, kötü niyetli katılımcıların sistemden elenmesi ve ana modelin güncellenmesi işlemleri gerçekleştirilmiştir. Gelecek bölümde son katmanlar arasındaki Öklid mesafesinin hesaplamasından başlayarak sırası ile yöntemlerin aşamaları detaylı olarak anlatılmıştır.



Şekil 8 Önerilen Yöntemlerin Akış Şeması

2.1.1. Son Katmanlar Arası Öklid Mesafesinin Hesaplanması

FL sürecinde, her bir kullanıcının modeli $W_{t,p}$ ile ana modelin t-1 iterasyonundaki ağırlıkları W_{t-1} arasındaki Öklid mesafesi hesaplanır ve her katmandaki ağırlıkların tek bir vektöre dönüştürülmesi (flattening) yapılır. Flattening, matris veya çok boyutlu dizilerin tek boyutlu bir vektöre dönüştürülmesi işlemidir.

Örneğin, bir ağırlık matrisi $W_{t-1}[l]$, (1) ile gösterildiği gibi üç boyutlu bir matristir

$$W_{t-1}[l] = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \quad (1)$$

$W_{t-1}[l]$ matrisinin yassılaştırılmış hali (2)'de tek boyutlu bir vektör olarak ifade edilmiştir. Bu işlem, matrisin elemanlarının tek bir vektörde sıralanmasını sağlar.

$$\text{flatten}(W_{t-1}[l]) = [w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33}] \quad (2)$$

(3)'de yerel katılımcı modellerinin, ana modele olan Öklid mesafeleri hesaplanılırken (2)'deki yassılaştırma yapılarak matristen tek boyutlu vektöre geçiş yapılır.

$$D_{t,p,l} = \|\text{flatten}(W_{t-1}[l]) - \text{flatten}(W_{t,p}[l])\|_2 \quad (3)$$

(3)'deki $D_{t,p,l}$ sembolü, i iterasyonundaki p yerel katılımcı modelinin l katmanı ile ana modelin $t-1$ iterasyonundaki l katmanı arasındaki Öklid mesafesini ifade eder. $(W_{t-1}[l])$ Ana modelin $t-1$ iterasyonundaki l katmanının yassılaştırılmış hali olarak tanımlanırken, $(W_{t,p}[l])$ ise yerel katılımcının t iterasyonundaki l katmanını temsil eder.

(4)'de her bir yerel katılımcı için hesaplanan Öklid mesafeleri bir matriste arka arkaya eklenir. Matris, her bir katılımcının tüm katmanlar için Öklid mesafelerini içerir ve tek bir liste halinde tutulur.

$$D_{t,l} = [D_{(t,p_1,l)}, D_{(t,p_2,l)}, \dots, D_{(t,p_n,l)}] \quad (4)$$

(4)'de $D_{t,l}$ ile sembolize edilen değer t iterasyonundaki tüm katılımcıların l katmanı için Öklid mesafelerini içeren listedir.

2.1.2. Öklid Mesafelerinin İstatiksel Analizi

Önerilen yöntemin ilk adımında, modelin son katmanındaki Öklid mesafelerinin istatistiksel analizi yapılır. Analiz sürecinde, veri setinin çeyrek değerleri(Quantile) hesaplanır. Çeyrek değerler sıralanmış veri setinde belirli yüzde dilimlerine karşılık gelen 25, 50 ve 75’lik dilimlerinde yer alan değerlerdir. Bu değerler, verilerin yayılımını anlamak için kullanılır.

X veri setindeki çeyrek değerler (5) ile gösterildiği gibi hesaplanılmaktadır.

$$quantile(X, k) = X_{\left(\frac{k(n+1)}{100}\right)} \quad (5)$$

(5)’de k istenilen çeyrek değerini ifade ederken n değeri X listesindeki eleman sayısını ifade eder. Önerilen yöntemde çeyrek değerlerini bulmak istediğimiz liste, tüm katılımcılardan gelen gradyanlar ile ana modelin t-1 iterasyonundaki ağırlıkları arasındaki Öklid uzaklıklarının yer aldığı $D_{t,l}$ listesidir. $D_{t,l}$ listesinin çeyrek değerleri (6), (7) ve (8)’de verilmiştir.

$$Q1 = quantile(D_{t,l}, 0.25) \quad (6)$$

$$Q2 = quantile(D_{t,l}, 0.50) \quad (7)$$

$$Q3 = quantile(D_{t,l}, 0.75) \quad (8)$$

Q1, verilerin yüzde 25’lik kısmının altındaki değeri; Q2, verilerin ortasındaki değeri; Q3 ise verilerin yüzde 75’lik kısmının altındaki değeri temsil eder.

2.1.3. QMAD Yöntemi ile Kötü Niyetli Katılımcı Tespiti

QMAD (Quantile Mean Absolute Deviation), verilerin ortalama etrafındaki ortalama mutlak sapmasını temsil eder. Ortalama, veri setinin merkezini temsil ederken, sapmalar, verilerin merkez etrafında ne kadar yayıldığını gösterir.

QMAD yönteminde yerel katılımcı modeli ile ana model arasındaki Öklid mesafelerini ifade bir liste $D_{(t,l)} = \{D_{(t,p_1,l)}, D_{(t,p_2,l)}, \dots, D_{(t,p_n,l)}\}$ kullanılır. $D_{t,l}$ listesinden (7) kullanılarak Q2 ile sembolize edilen %25 çeyreklik hesaplanır. (9)'da $D_{t,l}$ ile sembolize edilen yerel katılımcıların modelleri ile ana model arasındaki Öklid uzaklarından Q2 değeri çıkarılarak mutlak sapma (d_i) hesaplanılır

$$d_i = |D_{(t,l)} - Q_2| \quad (9)$$

(9)'da hesaplanan mutlak sapma değerinin ortalaması (10)'da alınarak, Çeyreklik Mutlak Sapma (QMAD) değeri elde edilir.

$$QMAD = \frac{1}{n} \sum_{i=1}^n |D_{(t,l)} - Q_p| \quad (10)$$

(11)'de, kullanılmak üzere (7)' de hesaplanan ve Q2 ile sembolize edilen %25 çeyreklik değerinden QMAD değeri çıkarılarak yerel katılımcıların modeli ile ana model arasındaki olabilecek minimum uzaklık değeri hesaplanır.

$$QMAD_{Alt\ Sınır} = Q2 - QMAD \quad (11)$$

(12)'de ise (7)' de hesaplanan ve Q2 ile sembolize edilen %25 çeyreklik değeri ile QMAD değeri toplanarak yerel katılımcıların modeli ile ana model arasındaki olabilecek maksimum uzaklık değeri hesaplanır.

$$QMAD_{Üst\ Sınır} = Q2 + QMAD \quad (12)$$

(11) ve (12)'de belirlenen alan aralığında kalan güvenilir katılımcılar (13)'te hesaplanarak $QMAD_{Güvenilir\ Katılımcılar}$ listesinde tutulur

$$QMAD_{Güvenilir Katılımcılar} = (D_{t,l} > QMAD_{Alt Sınır}) \wedge (D_{t,l} < QMAD_{Üst Sınır}) \quad (13)$$

Tek sınır olarak QMAD yöntemi ile savunulan bir federe öğrenme için bu aşamadan sonra $QMAD_{Güvenilir Katılımcılar}$ listesi, (27)'de W_t hesaplanmasında kullanılır ve işlemler 100 iterasyon boyunca devam eder.

2.1.4. QAD Yöntemi ile Kötü Niyetli Katılımcı Tespiti

QAD (Quantile Absolute Deviation) (Akinshin, 2022) MAD (Medyan Mutlak Sapma)'nin daha genel bir formudur ve medyan etrafındaki mutlak sapmanın belirli bir çeyrek değerini ölçer.

MAD, bir veri setinin yayılımını ölçen sağlam bir istatistiksel ölçüdür. Özellikle uç değerlere karşı duyarlılığı azdır ve normal dağılımlardan sapma durumlarında bile sağlam kalabilen bir yayılım ölçüsüdür. MAD, veri setindeki medyan etrafındaki mutlak sapmaların medyanını hesaplar. Medyan, sıralanmış bir veri setinin ortasındaki değerdir.

Liste $X = \{x_1, x_2, \dots, x_n\}$ için MAD değeri (15)'de gösterilmiştir.

$$MAD(X) = \text{medyan}(|X - \text{medyan}(X)|) \quad (14)$$

(15)'de, X listesinin medyanı etrafındaki mutlak sapma hesaplanır. Bulunan medyan etrafındaki mutlak sapmanın tekrar medyanı hesaplanılır. Bulunan değer medyan mutlak sapmadır (MAD).

Medyan, bir listenin ortanca değeridir. Medyan hesaplamasında, listenin eleman sayısı n ise; n tek sayı ise medyan (16) ile, eğer n çift sayı ise (17) ile hesaplanır.

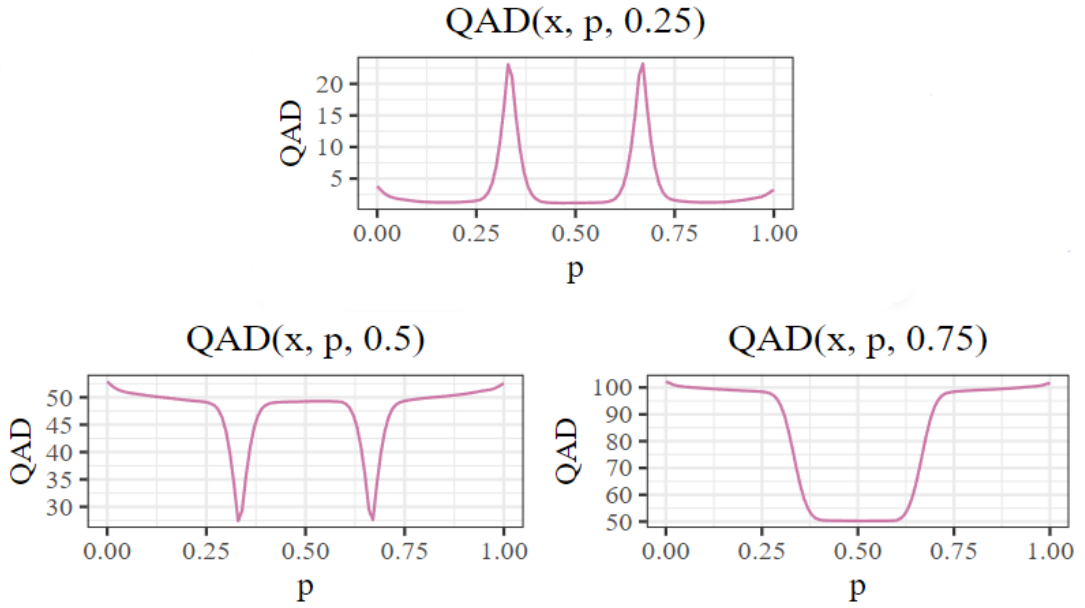
$$\text{med}(X) = X_{(\frac{n+1}{2})} \quad (15)$$

$$\text{med}(X) = \frac{X_{\frac{n}{2}} + X_{(\frac{n}{2}+1)}}{2} \quad (16)$$

Liste $X = \{x_1, x_2, \dots, x_n\}$ için QAD'ın hesaplanması ise (18)'de anlatılmıştır. K , bir ölçek sabitidir. Ve yöntemde 1 olarak alınmıştır

$$QAD(X, p, q) = K \cdot Q(|X - \text{quantile}(X, p)|, q) \quad (17)$$

Şekil 9 farklı p değerleri (0.25, 0.5, 0.75) için QAD değişimlerini göstermektedir. Her bir grafik, belirli bir p değeri için QAD'ın dağılımını ve sapmalarını ortaya koymaktadır.



Şekil 9 QAD Trimodal dağılım

Şekil 9'da da görülebileceği gibi QAD yönteminde seçilen çeyrekliğe göre dağılımlar değişmektedir. Model eğitiminde öğrenme katsayısı olarak 0.05 seçilmiştir dolayısıyla bir iterasyonda bile model güncellemesi fazla olacağı anlamına gelmektedir. Buradan yola çıkarak çalışmamızda geniş dağılıma sahip $QAD(D_{(t,l)}, 0.50, 0.75)$ kullanılmıştır. QAD alt ve üst sınırının bulunabilmesi için önce QADdata bulunmalıdır. (19)'da $D_{t,l}$ ile sembolize edilen yerel katılımcıların modelleri ile ana model arasındaki Öklid uzaklarından, (7)'de hesaplanan Q_2 ile sembolize edilen %25 çeyreklik değeri çıkarılarak (20)'de çeyreklik değerleri hesaplanacak veri olan, yerel katılımcıların Öklid mesafelerinin Q_2 'den mutlak sapmaları hesaplanır.

$$QADdata = |D_{t,l} - Q2| \quad (18)$$

(20)'de yerel katılımcıların Öklid mesafelerinin $Q2$ 'den mutlak sapmalarının %75 çeyrek değeri $QAD(D_{(t,l)}, 50,75)$ bulunur. Bu değer önerilen savunma mekanizmalarından olan QAD yöntemi için sınır değerlerini (21) ve (22) hesaplarırken kullanılacaktır.

$$QAD(D_{(t,l)}, 50,75) = \text{quantile}(QADdata, 0.75) \quad (19)$$

$QAD_{Alt Sınır}$ değeri güvenilir katılımcılar ile ana model arasındaki minimum Öklid uzaklığını ifade eder.

$$QAD_{Alt Sınır} = Q2 - QAD(D_{(t,l)}, 0.50, 0.75) \quad (20)$$

$QAD_{Üst Sınır}$ ise güvenilir katılımcılar ile ana model arasındaki maksimum Öklid uzaklığını ifade eder.

$$QAD_{Üst Sınır} = Q2 + QAD(D_{(t,l)}, 0.50, 0.75) \quad (21)$$

(21) ve (22)'de belirlenen alan aralığında kalan güvenilir katılımcılar (23)' de bulunarak $QAD_{Güvenilir Katılımcılar}$ listesinde tutulur

$$QAD_{Güvenilir Katılımcılar} = (D_{t,l} > QAD_{Alt Sınır}) \wedge (D_{t,l} < QAD_{Üst Sınır}) \quad (22)$$

Tek sınır olarak QAD yöntemi ile savunulan bir federe öğrenme için bu aşamadan sonra (27)'de W_t hesaplanmasında $QAD_{Güvenilir Katılımcılar}$ listesi kullanılır ve bu işlemler 100 iterasyon boyunca devam eder

2.1.5. ÇİFT SINIR Yöntemi ile Kötü Niyetli Katılımcı Tespiti

QAD ve QMAD yöntemleri ile 2 farklı Güvenli alan bulunmuştur. Çift sınır yönteminin hibrit bir kullanımı gerçekleştirilerek daha yüksek doğrulukta bir tespit yapılması amaçlanmıştır. Çift sınır yönteminde bir katılımcının kötü niyetli olarak etiketlenebilmesi için hem QAD hem de QMAD yöntemlerine göre aykırı değer olarak belirlenmiş olması gerekmektedir. Bu sayede yöntemlerin güvenilir katılımcıları yanlış etiketleyerek kötü niyetli katılımcı olarak sistemden elimine edilmesinin önüne geçilmek istenilmiştir.

(23)'de yerel modellerinin ana modele olan Öklid uzaklıkları QMAD yöntemi ile belirlenen alt ve üst sınır arasında kalan yerel katılımcılar $QMAD_{Güvenilir\ Katılımcılar}$ listesine eklenmektedir.

$$QMAD_{Güvenilir\ Katılımcılar} = (D_{t,l} > QMAD_{Alt\ Sınır}) \wedge (D_{t,l} < QMAD_{Üst\ Sınır}) \quad (23)$$

(24)'de, yerel modellerinin ana modele olan Öklid uzaklıkları QAD yöntemi ile belirlenen alt ve üst sınır arasında kalan yerel katılımcılar $QAD_{Güvenilir\ Katılımcılar}$ listesine eklenmektedir.

$$QAD_{Güvenilir\ Katılımcılar} = (D_{t,l} > QAD_{Alt\ Sınır}) \wedge (D_{t,l} < QAD_{Üst\ Sınır}) \quad (24)$$

(24) ve (25)'de bulunan güvenli katılımcı listeleri (26)'da "veya" işlemine tabii tutularak her iki yöntemde de ortak olan Kötü niyetli katılımcılar belirlenir.

$$\begin{aligned} \text{Çift Sınır}_{Güvenilir\ Katılımcılar} \\ = QAD_{Güvenilir\ Katılımcılar} \vee QMAD_{Güvenilir\ Katılımcılar} \end{aligned} \quad (25)$$

Bir sonraki bölümde sıra önerilen savunma yöntemlerinden olan Çift sınır yöntemi ile belirlenen kötü niyetli katılımcıların sistemden elimine edilmesi gelmektedir.

2.1.6. Kötü Niyetli Katılımcıların Sistemden Elimine Edilmesi Ve Ana Model Güncellenmesi

FL süreçlerinde kötü niyetli katılımcıların tespit edilerek verilerinin model güncellemelerinden çıkarılması, modelin güvenilirliğini artırmak ve performansını optimize etmek için önemli adımlardır. Kötü niyetli katılımcıların tespit edilmesi ve çıkarılmasının ardından, FL süreçlerinde yaygın olarak kullanılan FedAvg yöntemi ile güncellenmiş ana model hesaplanır ve katılımcılara geri gönderilir. FedAvg’de güncellenmiş ana model ağırlıkları şu şekilde hesaplanır:

$$W_t = \frac{1}{N} \sum_{i=1}^N W_{t,i} \quad (26)$$

(26)’ da W_t t iterasyonundaki güncellenmiş ana modelin ağırlıklarını temsil eder. N, aykırı değerler çıkarıldıktan sonra kalan yerel model sayısını gösterir. W_t , t iterasyonundaki her bir yerel modelin ağırlıklarının ortalamasını alarak güncellenmiş ana modeli oluşturur.

Güncellenmiş ana model hesaplandıktan sonra, modelin ağırlıkları tüm katılımcılara geri gönderilir süreç 100 iterasyon tamamlanana kadar devam eder. Şekil 10’ de Çift sınır yöntemine ait sözde kod verilmiştir. Kod ile teorik anlatım arasında bağlantılar kurularak ayrıntılı bir şekilde anlatılmıştır.

Algorithm 1 Çift sınır QARFL

```

1: Prosedür QARFL
2:  $W^0 \leftarrow$  başlangıç ağırlıkları (rastgele)
3: for  $t = 1, 2, \dots, T$  do
4:   for  $p = 1, 2, \dots, P$  do
5:     for  $e = 1, 2, \dots, E$  do
6:        $W_p^t \leftarrow W^{t-1} - \eta \frac{1}{n_p} \sum_{i=1}^{n_p} \nabla W(L(x_{p,i}, y_{p,i}, W^{t-1}))$  //katılımcıların
        yerel eğitimi
7:     end for
8:   end for
9:   // Son katman için uç değer tespiti
10:  her bir katılımcı  $p \in P$  için // katılımcıların l katmanını ile ana modelin
    t-1 iterasyonundaki l katmanını arasındaki Öklid mesafesi
11:     $D_{t,p,l} \leftarrow \|W_p^t[\text{katman\_indeksi}] - W^t[\text{katman\_indeksi}]\|_2$ 
12:     $D_{t,p,l}$ 'yi  $D_{t,l}$  ekle
13:  end for
14:   $Q1, Q2, Q3 \leftarrow \text{quantiles}(D_{t,l})$  // 25., 50., 75. yüzdeler
15:   $mad \leftarrow \frac{1}{n} \sum_{i=1}^n |D_{t,l} - Q2|$ 
16:   $QMAD_{alt} \leftarrow Q2 - mad$ 
17:   $QMAD_{ust} \leftarrow Q2 + mad$ 
18:   $data2 \leftarrow |D_{t,l} - Q2|$ 
19:   $qad1, qad2, qad3 \leftarrow \text{quantiles}(data2)$  // Mutlak farkların 25., 50., 75.
    yüzdeleri
20:   $QAD_{alt} \leftarrow Q2 - qad3$ 
21:   $QAD_{ust} \leftarrow Q2 + qad3$ 
22:   $QMAD_{icinde} \leftarrow (D_{t,p,l} > QMAD_{alt}) \& (D_{t,p,l} < QMAD_{ust})$  //QMAD
    sınırına uygun katılımcılar
23:   $QAD_{icinde} \leftarrow (D_{t,p,l} > QAD_{alt}) \& (D_{t,p,l} < QAD_{ust})$  //QAD sınırına
    uygun katılımcılar
24:   $guvenilir\_katilimci \leftarrow QMAD_{icinde} \& QAD_{icinde}$  //çift sınıra uygun
    nihai güvenilir katılımcılar
25:  her bir katılımcı  $p \in P$  için
26:    eğer  $[p]guvenilir\_katilimci$  değilse
27:      bu tur için  $p$ 'yi model güncellemesinden hariç tut
28:    end if
29:  end for
30:   $W^t = \frac{\sum_{p \in P} n_p \cdot W_p^t}{\sum_{p \in P} n_p}$  // Kalan katılımcıların ağırlıklı ortalaması
31: end for
32: return  $W^T$ 
33: end prosedür

```

Şekil 10 Çift sınır sözde kodu

Şekil (10)'da gösterilen Çift Sınır algoritması, belirli bir sayıda katılımcının (Satırlar 8-26) iş birliğiyle, rastgele başlatılmış bir küresel modelin (Satır 2) T tur boyunca eğitilmesini sağlar. Her turda t , katılımcılar önceki turun W_{t-1} parametreleriyle temsil edilen ana modelini alır ve W_{t-1} modeli üzerinde yerel eğitim yaparlar (Satırlar 4-7). Yerel eğitim sonrasında, her katılımcının l katmanını ile ana modelin l katmanını arasındaki Öklid mesafesi

hesaplanır (Satır 11). $D_{t,l}$ listesine eklenir eklenir(Satır 12) Öklid mesafesi, yerel katılımcı modeli ile ana model arasındaki uzaklığı ifade eder.

Tüm katılımcılar için Öklid mesafesi hesaplandıktan sonra, uzaklık değerlerinin çeyrek değerleri hesaplanır (Satır 14). Algoritma, çeyrek değerleri kullanarak QMAD tabanlı alt ve üst sınırları hesaplar (Satırlar 15-17). QMAD, ana model ile yerel katılımcı modelleri arasındaki mesafenin kabul edilebilir alt ve üst sınırlarını belirler. QMAD sınırları, ana model ile yerel katılımcı modeli arasındaki uzaklığın optimum seviyesini belirlemek amacıyla kullanılır. Bir sonraki aşamada, QAD tabanlı alt ve üst sınırlar da belirlenir (Satırlar 18-21).

Önce QMAD sınırlarına göre her katılımcının modeli değerlendirilir. Eğer bir katılımcının modeli, hesaplanan QMAD alt ve üst sınırları içinde kalıyorsa, QMAD yöntemine göre güvenilir olarak kabul edilir. (Satır 22).

Sonrasında QAD sınırlarına göre her katılımcının modeli değerlendirilir. Eğer bir katılımcının modeli, hesaplanan QAD alt ve üst sınırları içinde kalıyorsa, QAD yöntemine göre güvenilir olarak kabul edilir. (Satır 22).

Değerlendirmeler sonucunda, QMAD ve QAD sınırlarından birine göre güvenilir kabul edilen katılımcılar sistemde kalmaya devam eder (Satırlar 22-24). Ancak hem QMAD hem de QAD sınırlarını aşan, yani her iki kritere göre de güvenilir olmayan katılımcılar sistemden çıkarılır.

Son olarak, ana model ile olan uzaklıkları belirlenen sınırlar içinde kalan güvenilir katılımcıların ağırlıkları kullanılarak ana model güncellenir (Satırlar 25-28). Algoritma, T tur boyunca devam eder ve her turun sonunda modelin güncellenmiş ağırlıkları döndürülür (Satır 30).

2.2. Veri Setleri

Tez kapsamında yapılan deneylerde 3 farklı veri seti kullanılmıştır. Bunlar sırası ile MNIST (LeCun, 1998) ve Fashion-MNIST (Xiao, 2017) ve CIFAR10 (Krizhevsky, 2009)

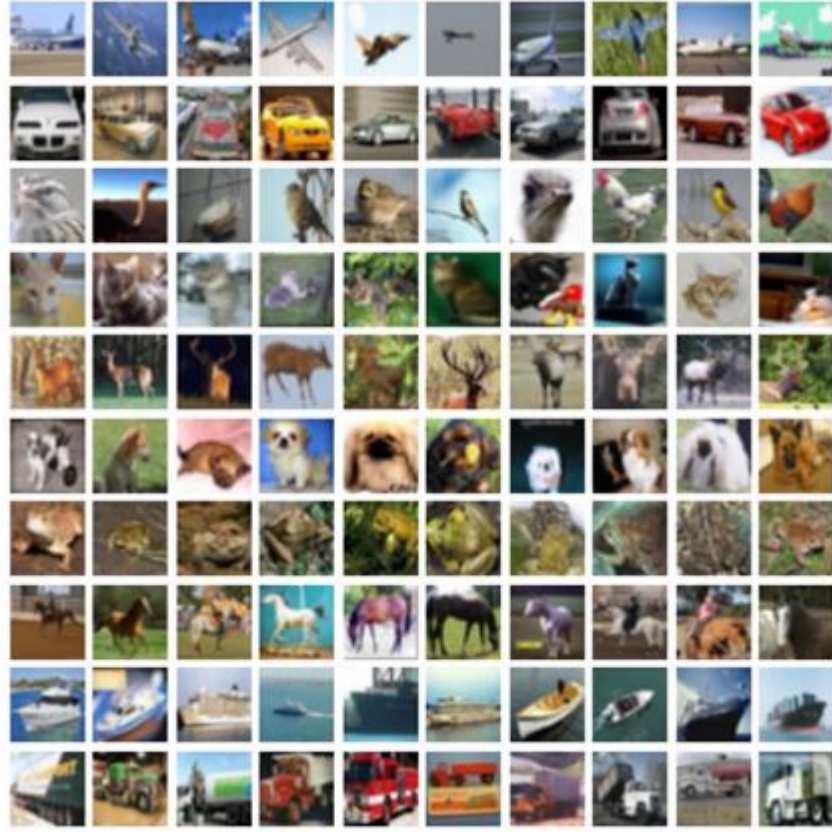
MNIST veri kümesi, Tablo 8'de örneklendiği gibi 28x28 gri tonlamalı el yazısıyla yazılmış rakam görüntüleri içerir. MNIST veri kümesi 50.000 eğitim, 10.000 doğrulama ve 10.000 test görüntüsünden oluşur. MNIST veri seti, makine öğrenimi modellerinin temel bir değerlendirme aracı olarak yaygın şekilde kullanılmaktadır. Her rakam 0-9 arasında

etiketlenmiştir ve veri setinin popülaritesi, literatürdeki diğer çalışmalarla karşılaştırmak açısından kolaylık sağlamaktadır.



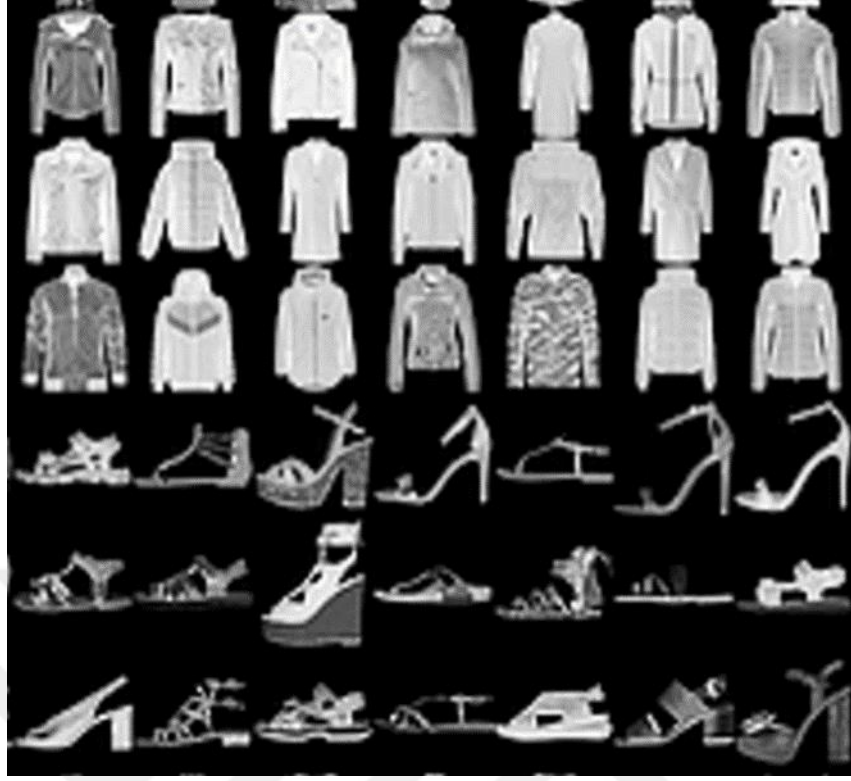
Şekil 11 MNIST veri seti

CIFAR-10 veri seti, 32x32 renkli görüntülerden oluşur ve 10 sınıfta toplam 60.000 görüntü içerir. Veri setin, 50.000 eğitim ve 10.000 test görüntüsüne sahiptir. CIFAR-10 veri seti, doğal görüntülerden oluşması ve renk bilgisi içermesi nedeniyle, diğer 2 veri setinden daha karmaşıktır ve genel olarak test başarısı daha düşüktür. CIFAR-10 veri setindeki sınıflar arasında uçak, otomobil, kuş, kedi, geyik, köpek, kurbağa, at, gemi ve kamyon bulunmaktadır. Veri setinden alınan örnekler Şekil 12 de verilmiştir.



Şekil 12 CIFAR veri seti

Fashion-MNIST veri seti, moda ürünlerinden oluşan 70.000 adet 28x28 gri tonlamalı görüntü içermektedir. Veri seti, 10 farklı kategoriye ayrılmış olup, her bir kategoride 7.000 görüntü bulunmaktadır. Eğitim seti 60.000, test seti ise 10.000 görüntüden oluşmaktadır. Fashion-MNIST, orijinal MNIST veri seti ile uyumlu şekilde tasarlanmıştır bu sayede MNIST veri setine uygun hale getirilmiş eğitimlerde uygulama sorunu yaşanmamaktadır. Veri setinden alınan örnekler Şekil 13 de verilmiştir



Şekil 13 Fashion MNIST veri seti

2.3. Savunma Yapılan Saldırı ve Saldırgan Türleri

Bu tez kapsamında Etiket Çevirme ve Bizans saldırıları bağlamında, %40'a kadar kötü niyetli katılımcı oranlarında organize ve organize olmayan saldırgan tiplerine karşı önerilen algoritmaların performanslarını değerlendirmektedir.

Etiket Çevirme saldırıları, FL sistemlerinde yaygın bir veri zehirleme saldırısı türüdür. Etiket Çevirme saldırılarında, kötü niyetli katılımcılar modelin belli bir sınıfa ait doğruluğunu düşürmek amacıyla verilerin etiketlerini manipüle ederler. Etiket Çevirme saldırısında, verilerin özellikleri değiştirilmeden sadece etiketler yanlış bir şekilde değiştirilir. Bu, modelin o etiketi yanlış sınıflandırmalar yapmasına yol açar ve genel performansını olumsuz etkiler (Jebreel, 2022)

Kötü niyetli katılımcılar, Etiket Çevirme saldırısını eğitim verilerinin etiketlerinin sistematik olarak değiştirilmesiyle gerçekleştirir. Kötü niyetli katılımcılar kendi içinde ikiye ayrılmaktadır: organize kötü niyetli katılımcılar ve organize olmayan kötü niyetli katılımcılar. Örneğin, organize olan kötü niyetli katılımcılar CIFAR-10 veri setindeki kendi verilerinde "kedi" fotoğraflarını "köpek" ve "köpek" fotoğraflarını "kedi" olarak etiketlemiş ve eğitimi yanlış etiketlenmiş verilerle gerçekleştirmiştir. Kötü niyetli katılımcılar, yanlış

verilerle eğitilen model güncellemelerini ana modele yollayarak ana modelin genel doğruluğunu düşürürken belirli bir etiketin yanlış etiketlenmesini hedefler. Organize olmayan kötü niyetli katılımcılar ise birbirlerinden farklı sınıf etiketlerinde değişiklik yaparlar; bir kötü niyetli katılımcı kedi ile köpek etiketini değiştirirken, diğer bir kötü niyetli katılımcı “koyun” ile “at” etiketlerini değiştirerek modellerini eğitirler. Organize saldırganlar için Etiket Çevirme saldırılarında değiştirilen etiketler Tablo 1’de verilmiştir.

Tablo 1 Organize Etiket Çevirme saldırılarında değiştirilen sınıf etiketleri

MINST		FASHION-MNIST		CIFAR	
Orjinal	Değiştirilen	Orjinal	Değiştirilen	Orjinal	Değiştirilen
0	9	T-shirt/üst	Gömlek	Uçak	Kuş
1	7	Pantolon	Elbise	Araba	Kamyon
2	5	Kazak	Mont	Kuş	Uçak
3	8	Elbise	Kazak	Kedi	Köpek
4	3	Mont	Pantolon	Geyik	At
5	2	Sandalet	Sneaker	Köpek	Kedi
6	4	Gömlek	T-shirt/üst	Kurbağa	Gemi
7	6	Sneaker	Bot	At	Geyik
8	1	Çanta	Sandalet	Gemi	Kamyon
9	0	Bot	Sneaker	Kamyon	Araba

Bizans saldırıları, FL modelinin tamamını bozmaya yönelik saldırılar kapsamında üzerine çokça çalışma yapılmış bir saldırı türüdür. Bizans saldırılarında, kötü niyetli katılımcılar modelin tamamının doğruluğunu düşürmek amacıyla modellerinin gradyanlarında rastgele ya da örtülü değişiklikler gerçekleştirirler. Rastgele güncellemelerde, kötü niyetli katılımcılar modelin parametrelerini rastgele değiştirerek ana modele yollarlar. Kötü niyetli katılımcılardan gelen model güncellemeleri ana modelde büyük hatalara yol açar. Örtülü güncellemelerde ise kötü niyetli katılımcılar modellerinin gradyanlarında küçük değişiklikler yaparak ana modele yollarlar.

Bu tez kapsamında, Bizans saldırısında rastgele gradyan değerleri ile saldırı gerçekleştiren kötü niyetli katılımcılar incelenmiştir. Bunun sebebi, örtülü saldırılarda kötü

niyetli katılımcıların tespit edilememesi durumunda bile gradyan değerlerinde belli bir oranda değişiklik yapıldığı için model başarısına etkisinin daha az olmasıdır.

Bizans saldırılarında da Etiket Çevirme saldırılarında olduğu gibi iki tip kötü niyetli katılımcının etkileri incelenmiştir. Bizans saldırılarında organize saldırı gerçekleştirilirken kötü niyetli katılımcılar, kendi aralarında belirledikleri rastgele gradyan değerlerini yollarlar. Organize olmayan saldırılarda ise her kötü niyetli katılımcı, kendisi rastgele gradyan değerleri belirler ve onları ana modele yollar. Hem organize hem de organize olmayan Bizans saldırılarının amacı, modelin tamamında bir tahribata yol açarak bütün sınıfların tespitini engellemektir.

2.4. Model Mimarileri ve Hiperparametreler

Bu bölümde, FL eğitiminde kullandığımız model mimarilerini ve modellerin performansını optimize etmek amacıyla belirlediğimiz hiperparametre ayarlarını ayrıntılı olarak ele alacağız.

2.4.1. Model Mimarileri

Her veri kümesi için kullanılan model mimarileri Tablo 2, Tablo 3, Tablo 4 ve Tablo 5 gösterilmiştir.

Seçilen model mimarilerinin tercih edilme nedeni, savunma yönteminin başarısının açık bir şekilde ortaya konmasını sağlamaktır. Literatürde yer alan birçok çalışma ile aynı mimarileri kullanarak, eğitimin başarısında savunma yönteminin etkisini net bir şekilde görebilmeyi amaçladık. Savunma yöntemimizin literatürdeki diğer yöntemlerle karşılaştırılabilirliğini artırarak, yöntemin etkinliğini ve geçerliliğini sağlam bir zemine oturtmaktadır. Aynı mimarileri kullanarak elde edilen sonuçlar, savunma yöntemimizin diğer çalışmalarda kullanılan yöntemlerle nasıl bir performans sergilediğini doğrudan karşılaştırmamıza olanak tanımaktadır.

Tablo 2 MNIST-2NN model mimarisi

Katman	Boyut
Tam Bağıntılı	(784.200)
Tam Bağıntılı	(200.200)
Tam Bağıntılı	(200.10)

Tablo 3 MNIST-CNN model mimarisi

Katman	Boyut
Konvolüsyonel	32@5x5
Maksimum Havuzlama	2x2
Konvolüsyonel	64@5x5
Maksimum Havuzlama	2x2
Tam bağıntılı	(1024, 512)
Tam bağıntılı	(512, 10)

Tablo 4 CIFAR model mimarisi

Katman	Boyut
Konvolüsyonel	32@3x3, Pad=1
Konvolüsyonel	32@3x3, Pad=1
Maksimum Havuzlama	2x2
Konvolüsyonel	64@3x3, Pad =1
Konvolüsyonel	64@3x3, Pad =1
Maksimum Havuzlama	2x2
Konvolüsyonel	128@3x3, Pad=1
Konvolüsyonel	128@3x3, Pad=1
Maksimum Havuzlama	2x2
Tam Bağıntılı	(2048, 128)
Tam Bağıntılı	(128, 10)

Tablo 5 Fashion-MNIST model mimarisi

Katman	Boyut
Konvolüsyon el	32@5x5, pad=2
Maksimum Havuzlama	2x2
Konvolüsyon el	64@5x5, pad=2
Maksimum Havuzlama	2x2
Tam bağıntılı	(3136, 500)
Tam bağıntılı	(500 10)

2.4.2. Hiperparametreler

Her veri kümesi için kullanılan FL ayar parametreleri Tablo 6’da gösterilmiştir. Her veri seti için öğrenme katsayısı 0.5 olarak belirlenmiş ve iletişim turu sabit 100 olarak seçilmiştir. Bu tercih, yüksek iletişim turu gerçekleştirerek daha düşük bir öğrenme katsayısı kullanmanın eğitim süresini aşırı derecede uzatması nedeniyle. FL’de, katılımcı cihazların yüksek işlem gücüne sahip olması zorunlu değildir; bu cihazlar telefon, tablet veya IoT cihazları olabilir. Bu bağlamda, aşırı uzun eğitim süreleri FL prensiplerine ters düşmektedir.

İletişim turunun 100 olarak sabitlenmesi, daha kısa sürede öğrenme katsayısının 0.5 olarak seçilmesiyle uzun eğitim sürelerine duyulan ihtiyacı ortadan kaldırmıştır. Ancak, öğrenme katsayısının yüksek olması, kötü niyetli katılımcıların sisteme sızması durumunda sistem doğruluğunu önemli ölçüde etkileyebilir. Bu durum göz önüne alındığında, savunma yönteminin kötü niyetli katılımcıları tespit etmede yüksek doğruluk oranına sahip olması kritik bir gerekliliktir. Tablo 6’da gösterilmektedir

Tablo 6 Eğitim Hiperparametreleri

Parametreler	MNIST-2NN	MNIST-CNN	Fashion MNIST	CIFAR
Katılımcı sayısı (n)	100	100	100	100
İletişim turu (t)	100	100	100	100
Batch boyutu	32	32	25	32
Epoch sayısı	0.9	0.9	0.9	0.9
momentum	0.9	0.9	0.9	0.9
Öğrenme kat sayısı	0.05	0.05	0.05	0.05

2.5. Değerlendirme Metrikleri

Deneylerde, üç veri setinde ortak olarak modelin performansını değerlendirmek için doğruluk (accuracy) metriği kullanılmıştır. Doğruluk, modelin doğru tahmin ettiği örneklerin toplam tahminlere oranını gösteren temel bir performans ölçütüdür. Doğruluk metrik, modelin genel performansını hızlı ve etkili bir şekilde değerlendirmek için uygun olup hem pozitif hem de negatif örneklerin doğru sınıflandırılma oranını yansıtır. Doğruluk

$$\text{Doğruluk} = \frac{\text{Doğru Tahmin Sayısı}}{\text{Toplam Tahmin Sayısı}} = \frac{TP + TN}{TP + TN + FP + FN} \quad (27)$$

biçiminde hesaplanır. TP doğru pozitifleri, TN doğru negatifleri, FP yanlış pozitifleri ve FN yanlış negatifleri ifade eder. Çalışmamızda doğruluk metriği, farklı saldırı senaryolarında modelin dayanıklılığını ve güvenilirliğini objektif bir şekilde kıyaslamak için kullanılmıştır.

Fashion MNIST ve CIFAR veri setlerinde Bizans saldırıları için doğruluk metriğine ek olarak F1, Duyarlılık ve Keskinlik metrikleri kullanılmıştır. Keskinlik bir modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu ölçen bir metriktir. Keskinlik sayesinde savunma yönteminin sistemden elimine ettiği katılımcıların gerçek kötü niyetli katılımcılara oranına bakılarak yöntemin güvenilir katılımcıların ne oranda yanlış etiketlendiğini ölçer. Yani, modelin doğru pozitif tahminlerinin toplam pozitif tahminlerine oranıdır Keskinlik

$$Keskinlik = \frac{Gerçek\ Zararlı\ Katılımcılar\ (TP)}{Gerçek\ Zararlı\ Katılımcılar\ (TP) + Yanlış\ Zararlı\ Katılımcılar\ (FP)} \quad (28)$$

biçiminde hesaplanır. Gerçek Zararlı (True Positives, TP): Savunma yönteminin zararlı olarak doğru bir şekilde tespit ettiği katılımcılar. Yani, hem tespit edilmesi gerekenler arasında yer alan hem de savunma yöntemi tarafından zararlı olarak işaretlenen katılımcılar.

Yanlış Zararlı (False Positives, FP): Savunma yönteminin zararlı olarak işaretlediği, ancak aslında zararlı olmayan katılımcılar. Yani, savunma yöntemi tarafından zararlı olarak işaretlediği fakat gerçekte zararsız olan katılımcılar. Duyarlılık ise savunma yönteminin zararlı katılımcıları ne kadar iyi yakaladığını ölçer. Yani, tespit edilmesi gereken zararlı katılımcıların ne kadarının savunma yöntemi tarafından doğru bir şekilde zararlı olarak işaretlendiğini gösterir ve

$$Duyarlılık = \frac{Gerçek\ Zararlı\ Katılımcılar\ (TP)}{Gerçek\ Zararlı\ Katılımcılar\ (TP) + Kaçırılan\ Zararlı\ Katılımcılar\ (FN)} \quad (29)$$

Biçiminde hesaplanır. (29)'da Kaçırılan Zararlı (False Negatives, FN): Savunma yönteminin zararlı olarak işaretlemediği, fakat gerçekte zararlı olan katılımcılar. Yani, savunma yöntemi tarafından zararlı olarak tespit edilemeyen katılımcıları ifade eder. Duyarlılık ve Keskinlik metrikleri kullanılarak hesaplanan F1 Skoru, savunma yöntemlerinizin zararlı katılımcıları tespit etme performansını değerlendirirken keskinlik ve duyarlılığı dengeleyen bir metriktir. (30)'da F1 skoru, hem zararlı katılımcıları doğru bir şekilde tespit etmenin (keskinlik) hem de tespit edilmesi gereken tüm zararlı katılımcıları yakalamanın (duyarlılık) önemini eşit olarak göz önünde bulundurarak hesaplanır.

$$F1 = \frac{Keskinlik * Duyarlılık}{Keskinlik + Duyarlılık} \quad (30)$$

F1 skoru, kesinlik ve duyarlılığın harmonik ortalamasıdır. Bu, savunma yöntemlerinizin hem ne kadar isabetli tespitler yaptığını (kesinlik) hem de zararlı katılımcıları ne kadar iyi yakaladığını (duyarlılık) birleştirir.

Fashion MNIST ve CIFAR veri setlerinde Etiket Çevirme saldırılarına karşı savunmanın etkinliğini ölçmek için ise Kaynak Doğruluğu metriği kullanılmıştır. Kaynak Doğruluğu bir savunma yöntemi uygulana FL sisteminin hedeflenen (saldırı yapılan) etiketin ne kadar doğru bir şekilde etiketlendiğini değerlendiren bir metriktir. Bu metrik, savunma yönteminizin saldırıya uğrayan etiketlerin orijinal doğru etiketlerini ne kadar iyi koruyabildiğini ölçer. (31)'de hesaplanan Kaynak doğruluk değerinin yüksek olması savunma yönteminin saldırıya uğrayan etiketleri doğru bir şekilde koruyabildiğini gösterirken, düşük Kaynak Doğruluğu savunma yönteminin etiketleri yeterince koruyamadığı ve yanlış sınıflandırmaya sebep olduğunu gösterir.

$$\text{Kaynak Doğruluğu} = \frac{\text{Doğru Etiketlenen Hedef Etiket Sayısı}}{\text{Toplam Hedef Etiket Sayısı}} \quad (31)$$

3. BULGULAR VE TARTIŞMA

Bu bölümde, FL sistemlerinde Bizans ve Etiket Çevirme saldırılarına karşı yapılan deneylerden elde edilen görsel ve sayısal bulgular irdelenerek, literatürde alanda mevcut çalışmalarla performans karşılaştırması yapılacaktır. Önerilen yöntemlerin FL 'nin güvenliği ve dayanıklılığı üzerinde sağladığı katkılar ayrı ayrı değerlendirilecektir. Önerilen yöntemlerden elde edilen bulgular ve literatürde mevcut çalışmalarla karşılaştırması alt başlıklar altında incelenmiştir. Deneylerde biri Core i5 13400F işlemci, GeForce RTX 3060 ekran kartı ve 16GB Vengeance RGB RS RAM'e sahip, diğeri ise AMD Ryzen 9 5950X 16-Core Processor 3.40 GHz işlemci, GeForce RTX 3080 ekran kartı 64GB RAM'e sahip 2 farklı bilgisayar kullanılmıştır.

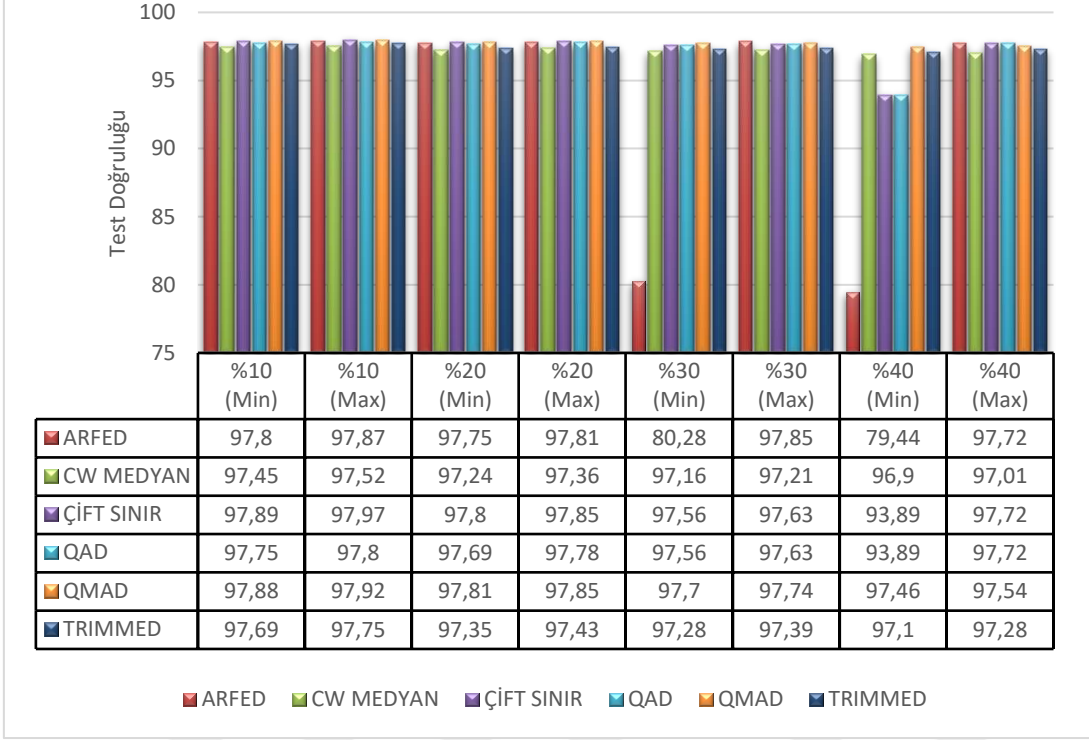
3.1. MNIST Veri Setinde Yapılan Deneysel Çalışmalar

MNIST veri setinde yapılan tüm deneylerde veri setindeki örneklerin 50.000 eğitim, 10.000 doğrulama ve 10.000 test olarak ayrılmıştır. 2 farklı mimaride yapılan 4'er tane deneyin sonuçları tablolar halinde sunularak sonuçları analiz edilmiştir. Analizler sonucunda hangi mimarinin daha yüksek performans sergilediği, MNIST veri setinde Etiket Çevirme ve Bizans saldırılarından hangisinin daha etkili olduğu ve salgınların organize çalışmasının sonuçlara etkisi tartışılarak değerlendirilen yöntemlerden hangisinin MNIST veri setinde daha başarılı savunma yaptığı analiz edilmiştir.

3.1.1. Etiket Çevirme Saldırısı

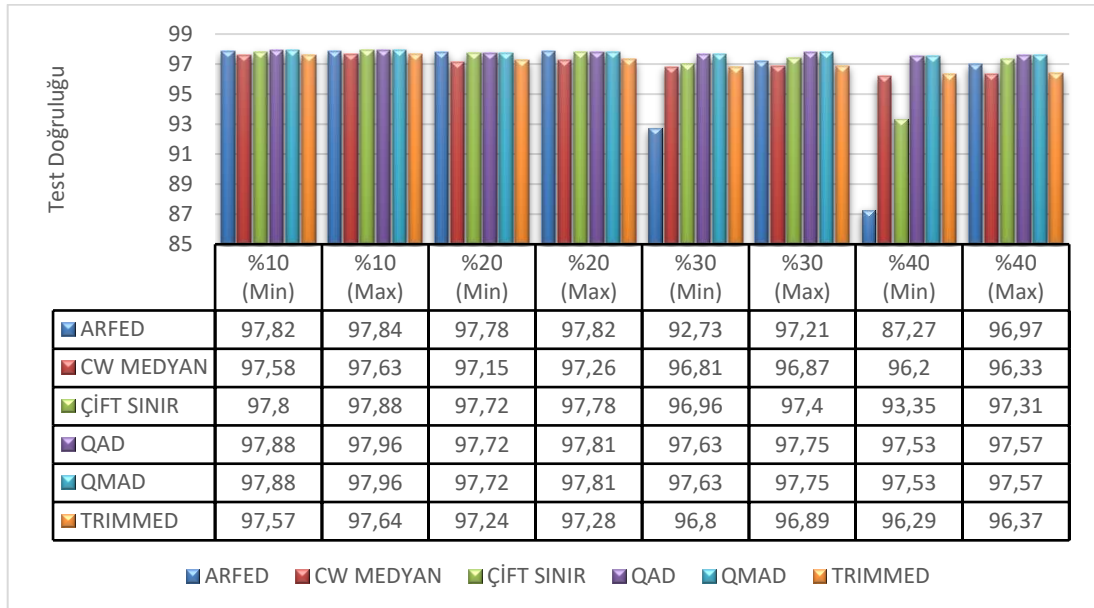
Etiket Çevirme saldırılarında saldırganların organize olması durumunda MNIST veri setinde Tablo 1'de gösterildiği üzere ikinci sınıfın verileri dokuzuncu sınıf, dokuzuncu sınıfın verileri ikinci sınıf olarak etiketlenmiştir. Saldırganların organize çalışmasının test doğruluğu sonuçları Tablo 8'de gösterilmiştir

Tablo 7 MNIST veri seti 2NN mimarisi organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



Organize olmayan saldırganlar ise rastgele olarak sınıf etiketlerini değiştirmişlerdir. Saldırganların organize olmadan çalışmasının sonuçları Tablo 8’de gösterilmiştir.

Tablo 8 MNIST veri seti 2NN mimarisi organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.87) ulaşırken, kötü niyetli kullanıcı oranının %40'a çıkmasıyla başarı oranı %74.72 seviyesine gerilemiştir. Bu, performansta yaklaşık %24'lük bir düşüşe işaret etmektedir ve oran arttıkça performansında önemli bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.84) ulaşmış, ancak %40 kötü niyetli kullanıcı oranında başarı oranı %87.27 seviyesine düşmüştür. Bu da yaklaşık %11'lik bir performans düşüşüne işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir.

CW MEDYAN yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranını (%97.52) elde etmiş olup, %40 kötü niyetli kullanıcı oranında başarı oranı %89.24 seviyesinde kalmıştır. Bu durum, performansta yaklaşık %8'lik bir düşüş olduğunu göstermektedir ve bu nedenle performansın nispeten stabil kabul edilebilir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.67) ulaşmış olup, %40 kötü niyetli kullanıcı oranında başarı oranı %86.99 seviyesine gerilemiştir, bu da yaklaşık %11'lik bir düşüşe işaret etmektedir bu düşüş ARFED yöntemi ile aynı oranda olmuştur.

ÇİFT SINIR yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.89) ulaşmış ve %40 kötü niyetli kullanıcı oranında %97.72 seviyesine düşmüştür. Bu, %0.17'lik bir performans düşüşüne işaret etmektedir ve oran arttıkça performansında çok az bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.92) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %93.55 seviyesine düşmüştür, bu da %4.37'lik bir düşüşe işaret etmektedir ARFED ve CW MEDYAN yöntemlerinin yarısı kadar bir düşüş yaşamıştır.

QAD yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.80) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %97.72 seviyesine düşmüştür. Bu, %0.08'lik bir düşüşe işaret etmektedir ve oran arttıkça performansında çok az bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.89) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %95.77 seviyesine düşmüştür, bu da %2.12'lik bir düşüş anlamına gelmektedir ve oran arttıkça performansında hafif bir azalma olduğunu göstermektedir.

QMAD yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.92) ulaşmış, %40 kötü niyetli kullanıcı oranında başarı oranı %97.72 seviyesinde

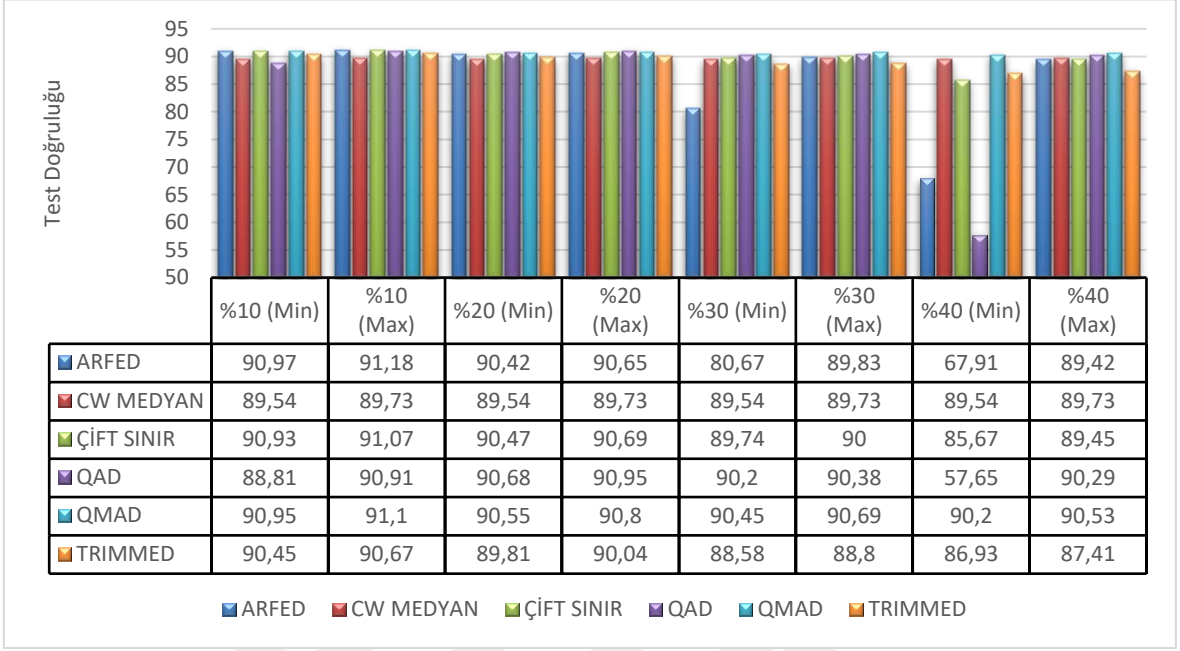
kalmıştır. Bu, %0.20'lik bir düşüşe işaret etmektedir ve bu nedenle performansın stabil kabul edilebilir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.96) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %95.77 seviyesine düşmüştür, bu da %2.19'luk bir düşüş anlamına gelmektedir ve oran arttıkça performansında hafif bir azalma olduğunu göstermektedir.

TRIMMED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.75) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %97.28 seviyesine düşmüştür, bu da %0.47'lik bir düşüş anlamına gelmektedir ve oran arttıkça performansında çok az bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%97.64) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %96.37 seviyesine inmiştir, bu da stabil sayılabilecek bir düşüş oranıdır.

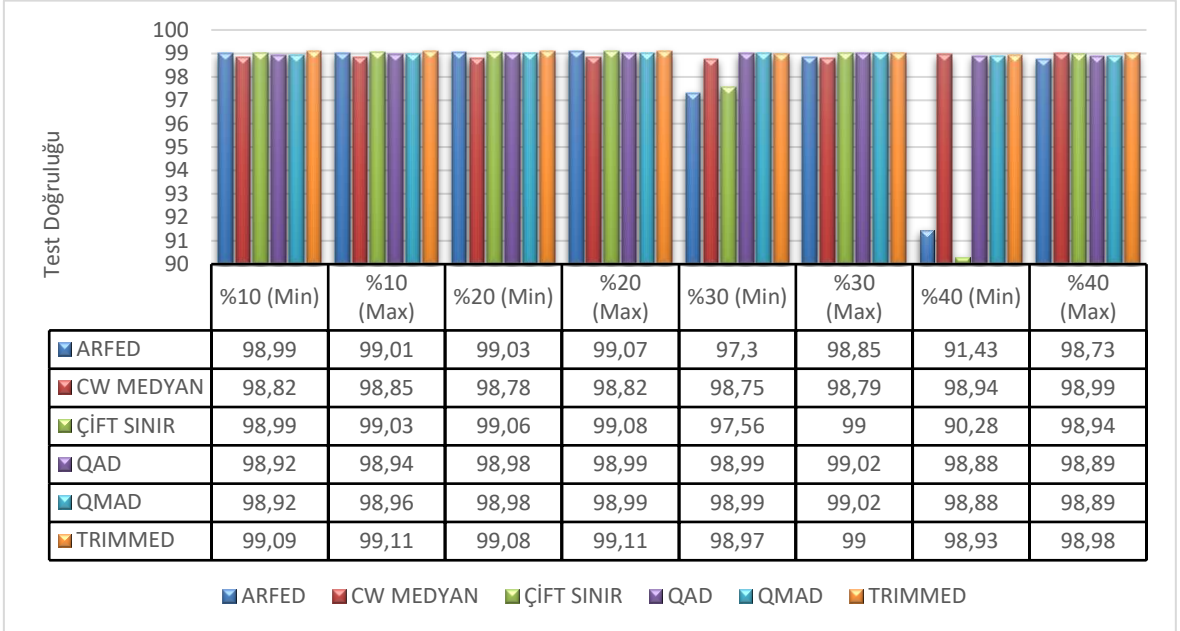
Genel olarak, organize ve organize olmayan saldırılar altında, QMAD ve TRIMMED yöntemleri kötü niyetli kullanıcı oranı arttıkça en yüksek performansı göstermektedir. Özellikle %40 kötü niyetli kullanıcı oranında bile bu yöntemler %2'den az düşüş göstermiştir. Organize olmayan saldırılar altında ise, ARFED ve ÇİFT SINIR yöntemleri %10 kötü niyetli kullanıcı oranında en yüksek performansı göstermektedir. Ancak kötü niyetli kullanıcı oranı arttıkça, özellikle ARFED yöntemi dramatik bir şekilde, yaklaşık %24'lük bir performans kaybı yaşamaktadır. QMAD yöntemi ise kötü niyetli kullanıcı oranı %40'a çıktığında bile %2'den az düşüş gösterdiği için performansı stabil kabul edilebilir.

Organize saldırılar altında, algoritmalar genellikle daha yüksek performans göstermiştir. Organize olmayan saldırılar, ARFED, CW MEDYAN ve ÇİFT SINIR gibi yöntemlerde daha fazla performans kaybına yol açmıştır. Özellikle, ARFED ve ÇİFT SINIR yöntemleri organize olmayan saldırılar altında organize saldırılara göre daha belirgin performans düşüşleri yaşamıştır. Bu durum, MNIST 2NN modelinde Etiket Çevirme saldırılarında organize olmayan saldırıların daha etkili olduğunu göstermektedir.

Tablo 9 MNIST veri seti CNN mimarisi organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



Tablo 10 MNIST veri seti CNN mimarisi organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.00) ulaşırken, kötü niyetli kullanıcı oranının %40'a çıkmasıyla başarı oranı %79.84 ile %98.92 arasında değişmiştir. Bu, performansta yaklaşık %19.16'lık bir düşüşe işaret

etmekle birlikte, aynı kötü niyetli kullanıcı oranında bile sonuçların oldukça değişken olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%98.99) ulaşmış, ancak %40 kötü niyetli kullanıcı oranında başarı oranı %84.29 ile %98.73 arasında değişmiştir. Bu da yaklaşık %14.70'lik bir performans düşüşüne işaret etmekle birlikte, sonuçların oldukça dalgalı olduğunu göstermektedir.

CW MEDYAN yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranını (%98.76) elde etmiş olup, %40 kötü niyetli kullanıcı oranında başarı oranı %87.24 ile %88.97 arasında değişmiştir. Bu durum, performansta yaklaşık %11.52'lik bir düşüş olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%98.85) ulaşmış olup, %40 kötü niyetli kullanıcı oranında başarı oranı %83.77 ile %97.91 arasında değişmiştir, bu da yaklaşık %15.08'lik bir düşüşe işaret etmekle birlikte, sonuçların oldukça değişken olduğunu göstermektedir.

ÇİFT SINIR yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%98.97) ulaşmış ve %40 kötü niyetli kullanıcı oranında %78.24 ile %98.45 arasında değişmiştir. Bu, %20.21'lik bir performans düşüşüne işaret etmekle birlikte, sonuçların oldukça dalgalı olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.07) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %93.35 ile %97.41 arasında değişmiştir, bu da %5.72'lik bir düşüşe işaret etmekle birlikte, sonuçların oldukça değişken olduğunu göstermektedir.

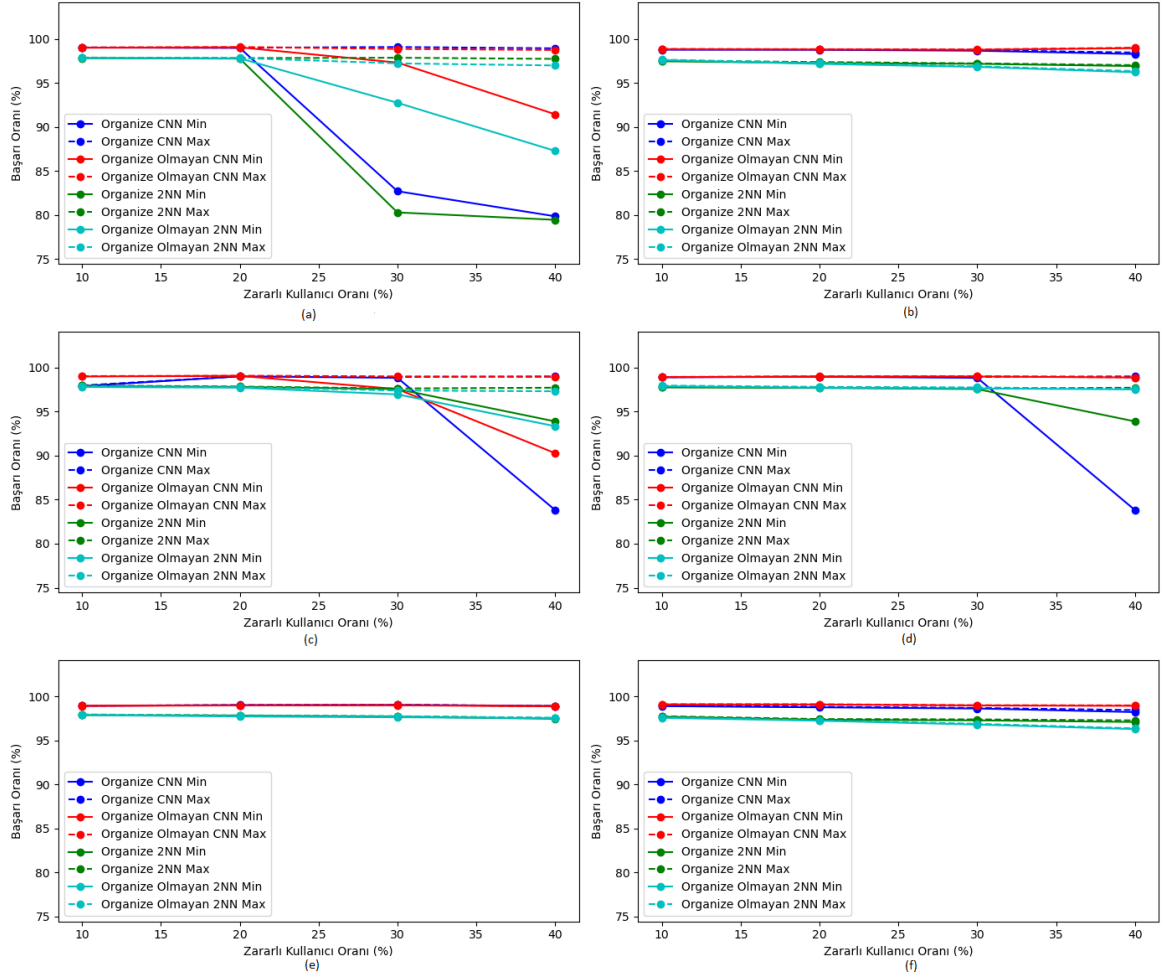
QAD yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.07) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %87.24 ile %98.45 arasında değişmiştir. Bu, %11.83'lük bir düşüşe işaret etmekle birlikte, sonuçların oldukça dalgalı olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.07) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %78.74 ile %98.72 arasında değişmiştir, bu da %20.27'lik bir düşüş anlamına gelmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir.

QMAD yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.08) ulaşmış, %40 kötü niyetli kullanıcı oranında başarı oranı %97.72 ile %98.45 arasında kalmıştır. Bu, %0.73'lük bir düşüşe işaret etmekle birlikte, sonuçların oldukça değişken olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.08) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %97.91 ile %98.88 arasında değişmiştir, bu da %0.97'lik bir düşüş

anlamına gelmektedir ve oran arttıkça performansında hafif bir azalma olduğunu göstermektedir.

TRIMMED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.09) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %88.69 ile %98.98 arasında değişmiştir. Bu, %10.29'luk bir düşüş anlamına gelmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%99.11) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %97.72 ile %98.88 arasında değişmiştir, bu da %1.16'lık bir düşüşe işaret etmekle birlikte, sonuçların oldukça değişken olduğunu göstermektedir.

Genel olarak, organize ve organize olmayan saldırılar altında, ARFED, CW MEDYAN ve ÇİFT SINIR yöntemleri kötü niyetli kullanıcı oranı arttıkça performanslarını büyük ölçüde kaybetmiştir. Özellikle %40 kötü niyetli kullanıcı oranında bile bu yöntemler %20'ye varan düşüş göstermiştir. Organize olmayan saldırılar altında ise, ARFED ve CW MEDYAN yöntemleri %10 kötü niyetli kullanıcı oranında en yüksek performansı göstermektedir. Ancak kötü niyetli kullanıcı oranı arttıkça, özellikle ARFED yöntemi %14.70'lik bir performans kaybı yaşamaktadır. QAD yöntemi ise kötü niyetli kullanıcı oranı %40'a çıktığında %20.27'lik bir düşüş göstermiştir. Bu analiz, yöntemlerin kötü niyetli kullanıcı oranlarına göre performans değişimlerini göstermektedir ve bazı yöntemler belirli kötü niyetli kullanıcı oranlarında diğerlerinden daha iyi performans sergilemektedir. Ayrıca, aynı kötü niyetli kullanıcı oranında minimum ve maksimum başarı oranları arasındaki büyük farklar, eğitim performansının standart sapmasının yüksek olduğunu ve sonuçların kararsız olduğunu Şekil 14' de gözükmemektedir.



Şekil 14 MNIST veri setinde 2NN ve CNN mimarisine ait modellerde organize olan ve organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c) organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.

• Organize saldırılar altında, algoritmalar genellikle daha yüksek performans göstermiştir. Organize olmayan saldırılar, ARFED, CW MEDYAN ve ÇİFT SINIR gibi yöntemlerde daha fazla performans kaybına yol açmıştır. Özellikle, ARFED ve ÇİFT SINIR yöntemleri organize olmayan saldırılar altında organize saldırılara göre daha belirgin performans düşüşleri yaşamıştır. Bu durum, organize olmayan saldırıların daha etkili olduğunu göstermektedir.

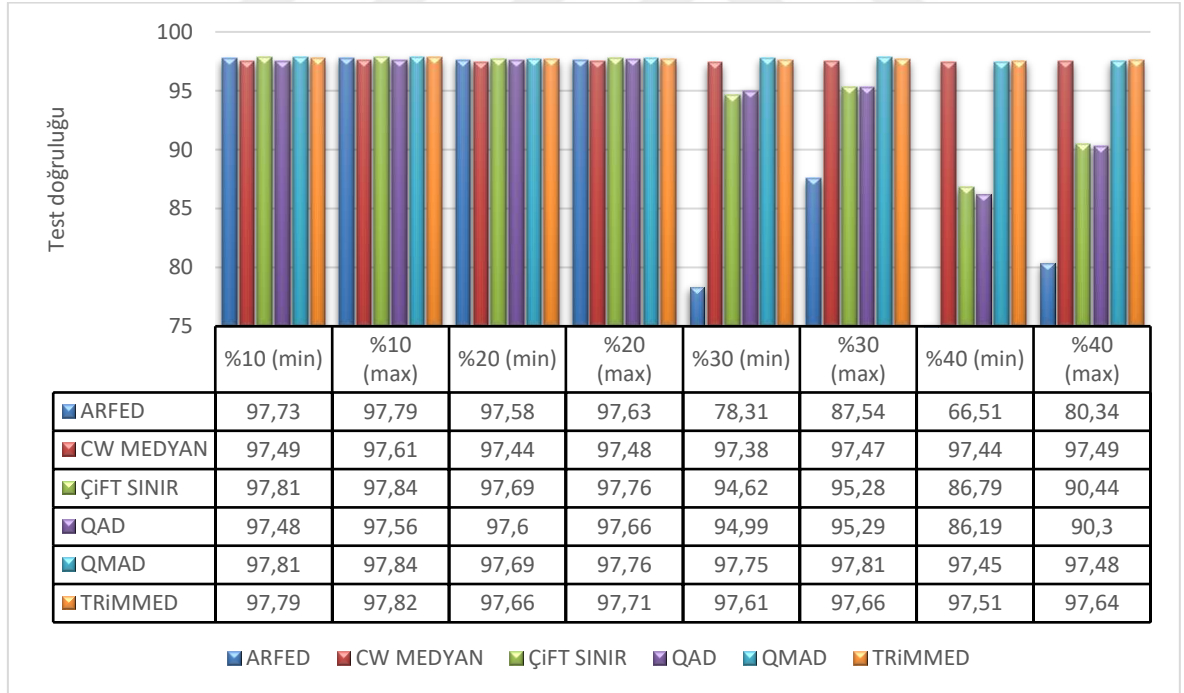
• Genel olarak, QMAD yöntemi hem organize hem de organize olmayan saldırılar altında en stabil ve yüksek performansı gösteren yöntem olarak öne çıkmaktadır.

• CNN modeli hem organize hem de organize olmayan saldırılar altında genel olarak daha yüksek performans sergilemiştir. Hem en yüksek hem de en düşük doğruluk oranlarında CNN modeli 2NN modeline göre daha başarılı olmuştur. Bu da, CNN modelinin kötü niyetli kullanıcı oranı arttıkça performansını daha iyi koruduğunu ve daha stabil olduğunu göstermektedir

3.1.2. Bizans Saldırısı

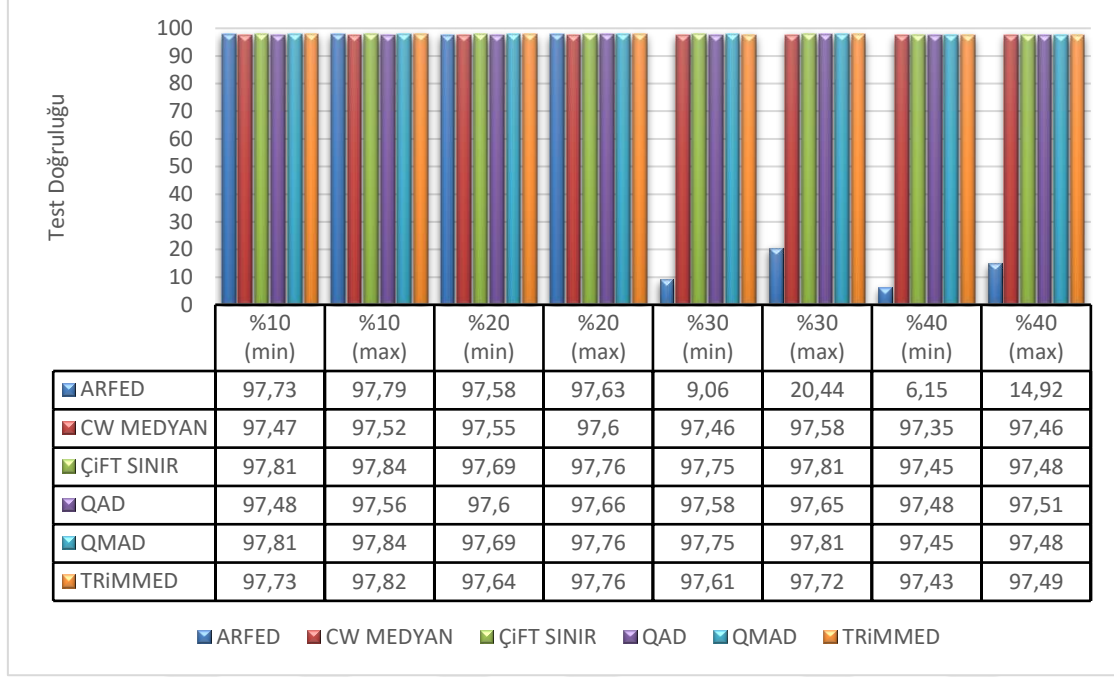
Deneylerde rastgele gradyan güncellemeleri kapsamındaki Bizans saldırıları gerçekleştirilmiştir. Yapılan Bizans saldırılarında, organize bir şekilde hareket eden kötü niyetli katılımcılar, önceden belirlenen model ağırlıklarını ana modele göndererek modelin doğruluğunu bozmayı hedeflemiştir.

Tablo 11 MNIST veri seti 2NN mimarisi organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



Organize olmayan kötü niyetli katılımcılar ise, her biri rastgele belirledikleri farklı yerel model güncellemelerini ana modele göndererek modelin doğruluğunu bozmayı hedeflemiştir.

Tablo 12 MNIST veri seti 2NN mimarisi organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi düşük kötü niyetli kullanıcı oranlarında yüksek performans sergilemektedir. Ancak %30 ve %40 kötü niyetli kullanıcı oranlarında organize saldırılar altında ciddi performans düşüşleri gözlemlenmektedir. Organize olmayan saldırılarda ise performans düşüşü daha azdır, ancak yine de belirgindir.

CW MEDYAN yöntemi tüm kötü niyetli kullanıcı oranlarında oldukça stabil ve yüksek performans göstermektedir. Hem organize hem de organize olmayan saldırılar altında güvenilir ve tutarlı sonuçlar vermektedir.

ÇİFT SINIR yöntemi düşük kötü niyetli kullanıcı oranlarında yüksek performans sergilerken, %30 ve %40 kötü niyetli kullanıcı oranlarında performansında belirgin bir düşüş yaşanmaktadır. Ancak, düşüş organize saldırılar altında bile diğer bazı yöntemler kadar ciddi değildir ve yöntem hala güvenilir sayılabilir.

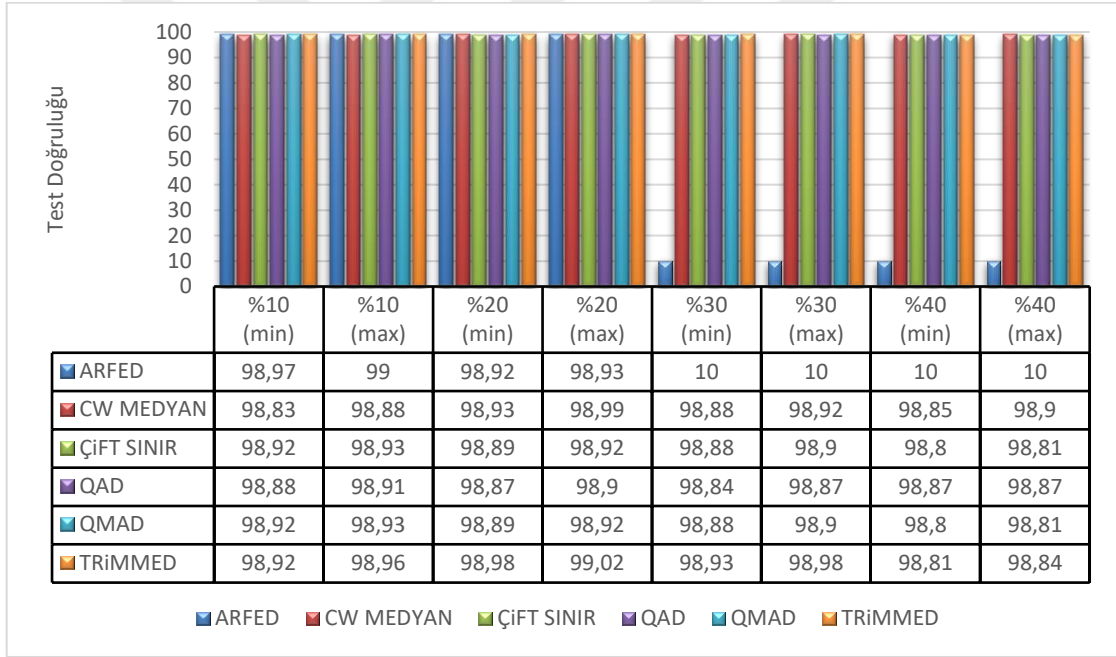
QAD yöntemi düşük kötü niyetli kullanıcı oranlarında yüksek performans sergilerken, %30 ve %40 kötü niyetli kullanıcı oranlarında performansında düşüş yaşanmaktadır. Ancak bu düşüş organize saldırılar altında bile CW MEDYAN kadar ciddi değildir ve yöntem hala güvenilir sayılabilir.

QMAD yöntemi tüm kötü niyetli kullanıcı oranlarında oldukça yüksek ve stabil performans göstermektedir. Performans düşüşü neredeyse yok denecek kadar azdır ve yöntem organize ve organize olmayan saldırılar altında oldukça güvenilirdir.

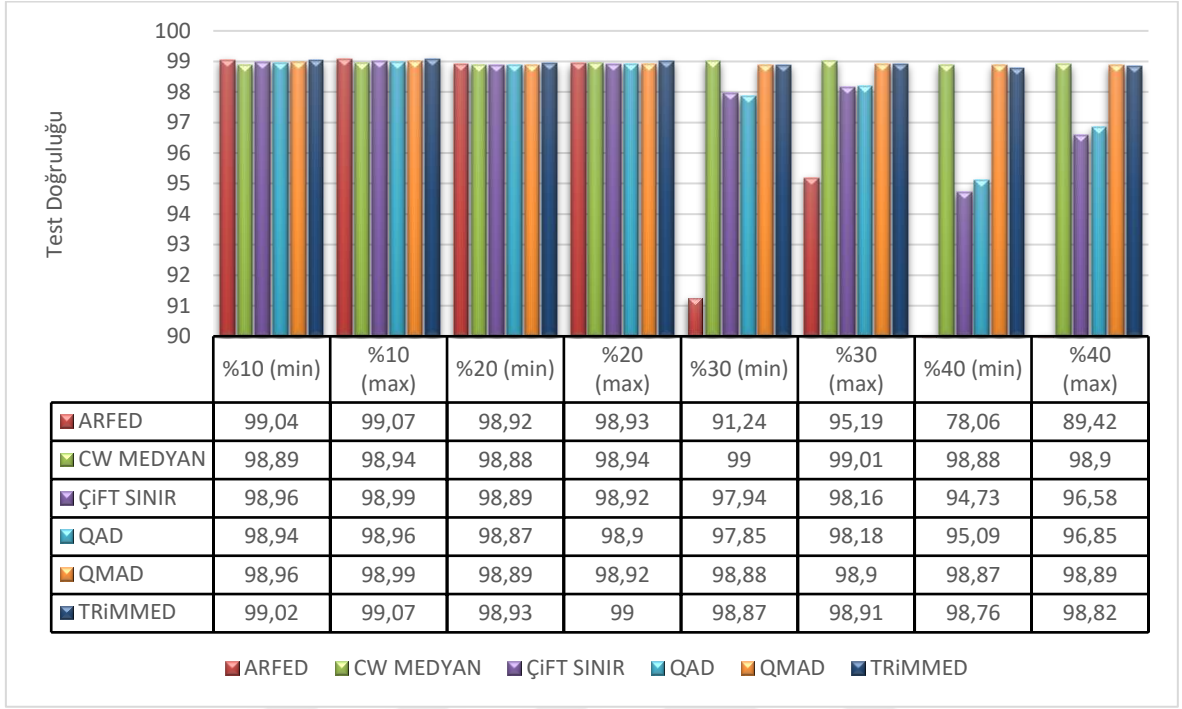
TRIMMED yöntemi tüm kötü niyetli kullanıcı oranlarında oldukça yüksek ve stabil performans göstermektedir. Performans düşüşü neredeyse yok denecek kadar azdır ve yöntem organize ve organize olmayan saldırılar altında oldukça güvenilirdir.

2NN mimarisindeki saldırgan modelleri birebir olarak CNN mimarisinde de uygulanmıştır. Yani yapılan Bizans saldırılarında, organize bir şekilde hareket eden kötü niyetli katılımcılar, önceden belirlenen model ağırlıklarını ana modele göndererek modelin doğruluğunu bozmayı hedeflemiştir.

Tablo 13 MNIST veri seti CNN mimarisi organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



Tablo 14 MNIST veri seti CNN mimarisi organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi, düşük kötü niyetli kullanıcı oranlarında yüksek performans sergilemektedir. Ancak %30 ve %40 kötü niyetli kullanıcı oranlarında, özellikle organize saldırılar altında ciddi performans düşüşleri yaşamaktadır. Organize olmayan saldırılarda performans düşüşü daha azdır, ancak yine de belirgindir. Sonuçlar, ARFED yönteminin yüksek kötü niyetli kullanıcı oranlarına karşı daha duyarlı olduğunu göstermektedir.

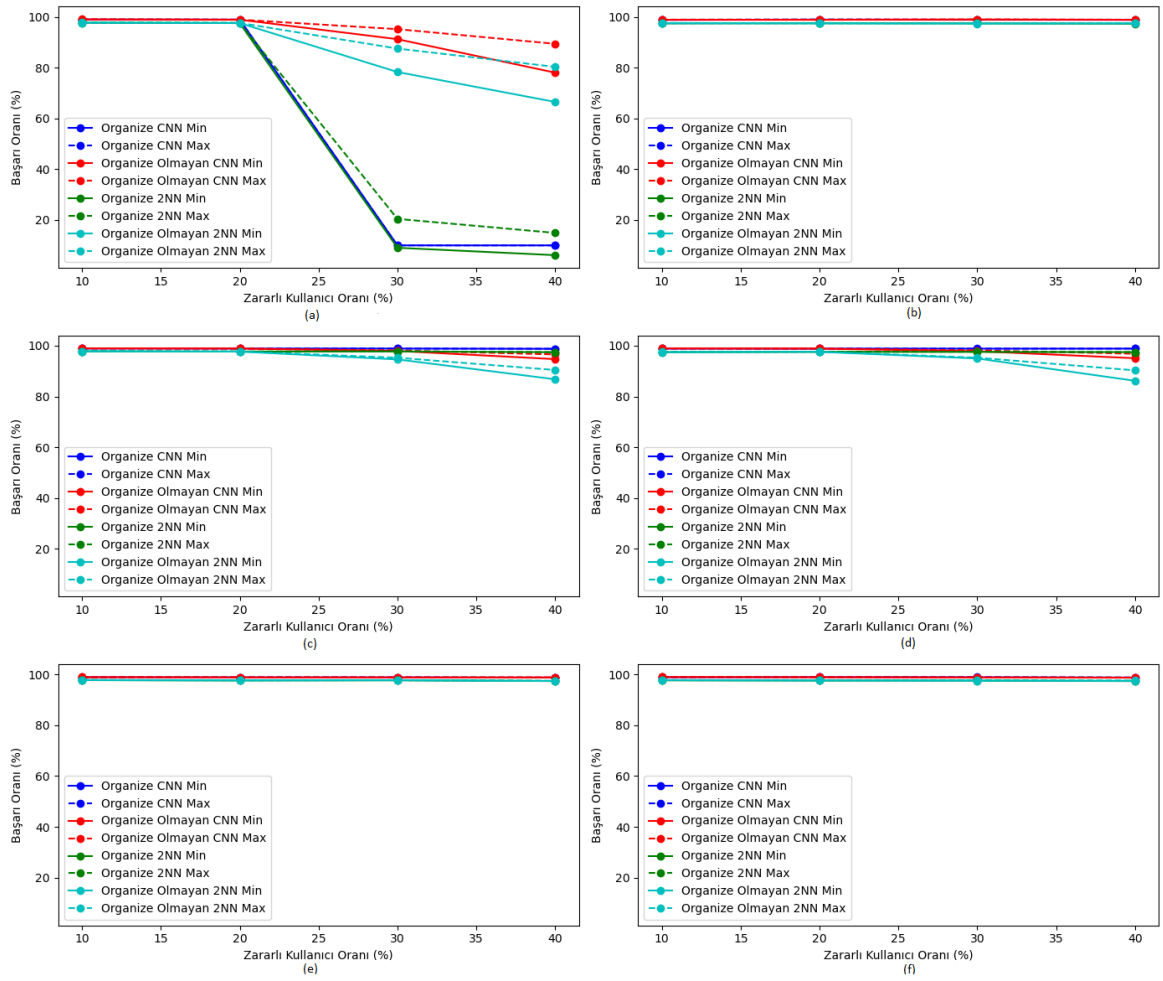
CW MEDYAN yöntemi, tüm kötü niyetli kullanıcı oranlarında oldukça stabil ve yüksek performans göstermektedir. Hem organize hem de organize olmayan saldırılar altında güvenilir sonuçlar sunmaktadır. Bu yöntem, kötü niyetli kullanıcı oranları arttıkça bile tutarlı bir performans sergilemektedir.

ÇİFT SINIR yöntemi, düşük kötü niyetli kullanıcı oranlarında yüksek performans sergilerken, %30 ve %40 oranlarında performansında belirgin bir düşüş yaşamaktadır. Organize saldırılar altında performans düşüşü minimaldir ve yöntem hala güvenilir sayılabilir.

QAD yöntemi, düşük kötü niyetli kullanıcı oranlarında yüksek performans sergilerken, %30 ve %40 oranlarında performansında düşüş yaşamaktadır. Organize olmayan saldırılarda performans düşüşü daha belirgindir, ancak organize saldırılar altında stabil performans göstermektedir.

QMAD yöntemi, tüm kötü niyetli kullanıcı oranlarında oldukça yüksek ve stabil performans göstermektedir. Hem organize hem de organize olmayan saldırılar altında güvenilir sonuçlar vermektedir. Bu yöntem, kötü niyetli kullanıcı oranları arttıkça bile tutarlı bir performans sergilemektedir ve herhangi bir belirgin düşüş yaşamamaktadır.

TRIMMED yöntemi, tüm kötü niyetli kullanıcı oranlarında yüksek performans göstermektedir. Hem organize hem de organize olmayan saldırılar altında oldukça güvenilirdir. Bu yöntem, yüksek kötü niyetli kullanıcı oranlarında bile minimal performans düşüşü yaşamaktadır. Şekil 15’de yöntemlerin başarıları tek bir grafikte değerlendirilmiştir.



Şekil 15 MNIST veri setinde 2NN ve CNN mimarisine ait modellerde organize olan ve organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c) organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.

- QMAD, yöntemi 2NN modelinde olduğu gibi en iyi sonuçları veren yöntem olmuştur
- CNN Modeli, daha yüksek performans ve dayanıklılık sunduğu için tercih edilmelidir. Özellikle CW MEDYAN, QMAD ve TRİMMED yöntemleri ile kullanıldığında, kötü niyetli kullanıcı oranları ne olursa olsun yüksek ve stabil sonuçlar elde edilebilmektedir. Bu nedenle, güvenilir ve tutarlı sonuçlar isteyen uygulamalarda CNN modeli daha uygun bir seçenek olacaktır.

- CNN Modeli, organize ve organize olmayan saldırılar altında daha yüksek performans ve dayanıklılık gösterdiğinden dolayı önerilen modeldir.

- 2NN Modeli de iyi sonuçlar verebilmesine rağmen, CNN modelinin sağladığı tutarlılık ve güvenilirlik seviyesine ulaşamamaktadır.

- CNN modelinin daha stabil ve yüksek performansı, güvenilir sonuçlar isteyen uygulamalar için ideal bir seçim yapmaktadır.

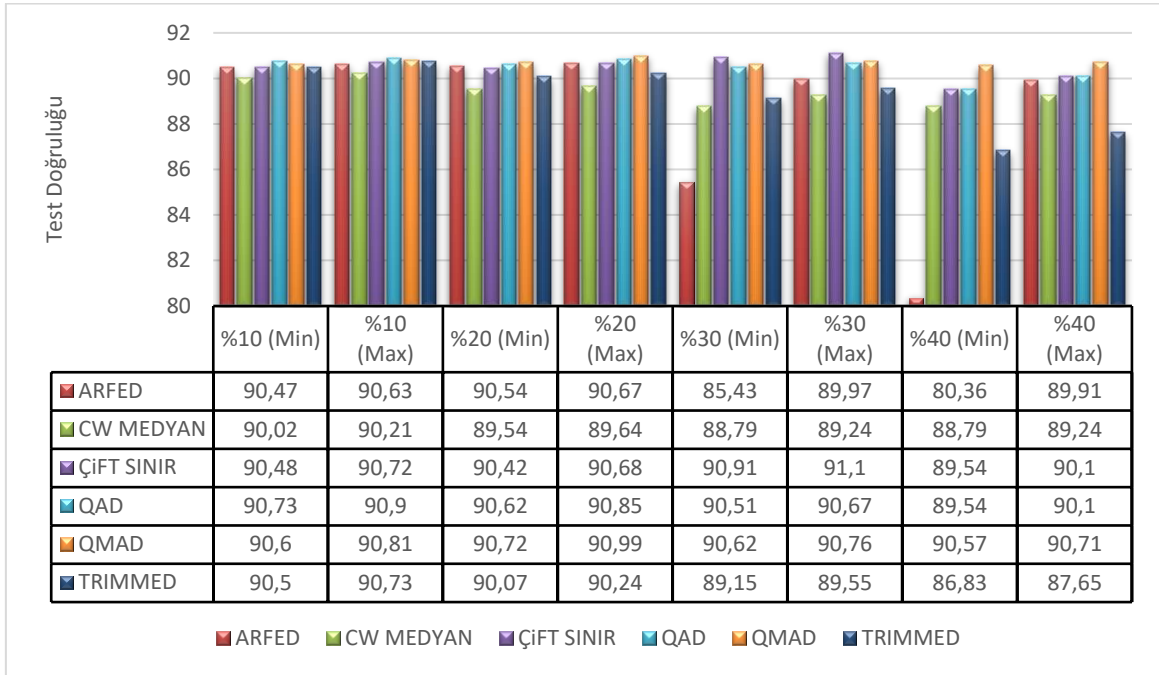
3.2. Fashion MNIST Veri Setinde Yapılan Deneysel Çalışmalar

Bu bölümde Fashion MNIST veri setinde yapılan 4 farklı deneyin sonuçları değerlendirilerek savunma yöntemlerinin farklı saldırı senaryolarında nasıl performans gösterdiğini analiz edilecektir. Analizler sonucunda saldırıların etkinlikleri ve yöntemlerin savunma başarıları değerlendirilecektir. Fashion MNIST veri setinde eğitim seti 60.000, test seti ise 10.000 görüntüden oluşmaktadır

3.2.1. Etiket Çevirme Saldırısı

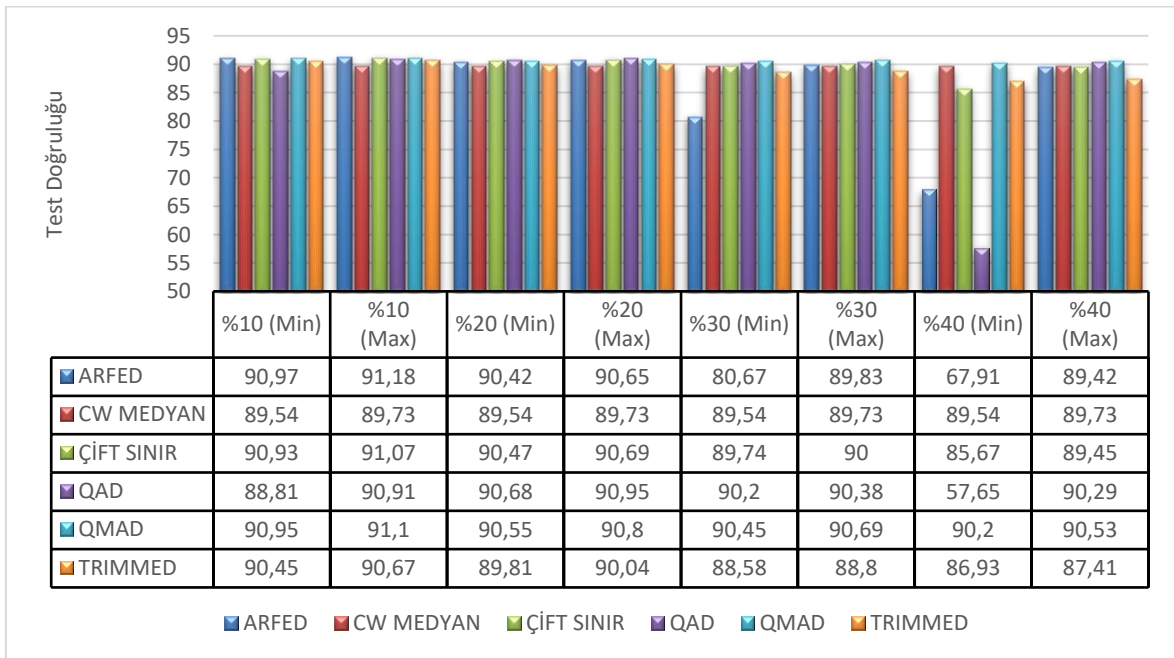
Etiket Çevirme saldırılarında saldırganların organize olması durumunda Fashion MNIST veri setinde Tablo 1’de gösterildiği üzere kazak örnekleri ceket, ceket etiketler de kazak olarak çevrilmiştir. Saldırganların organize çalışmasının test doğruluk sonuçları Tablo 16’de gösterilmiştir.

Tablo 15 Fashion MNIST veri setinde organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



Organize olmayan saldırganlar ise rastgele olarak sınıf etiketlerini değiştirmişlerdir. Saldırganların organize olmadan çalışmasının sonuçları Tablo 16’da gösterilmiştir.

Tablo 16 Fashion MNIST veri setinde organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%90.63) ulaşırken, kötü niyetli kullanıcı oranının %40'a çıkmasıyla başarı oranı %38.06 seviyesine gerilemiştir. Bu, performansta yaklaşık %58'lik bir düşüşe işaret etmektedir ve oran arttıkça performansının dramatik şekilde azaldığını göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%91.18) ulaşmış, ancak %40 kötü niyetli kullanıcı oranında başarı oranı %67.91 seviyesine düşmüştür. Bu da yaklaşık %26'lık bir performans düşüşüne işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir.

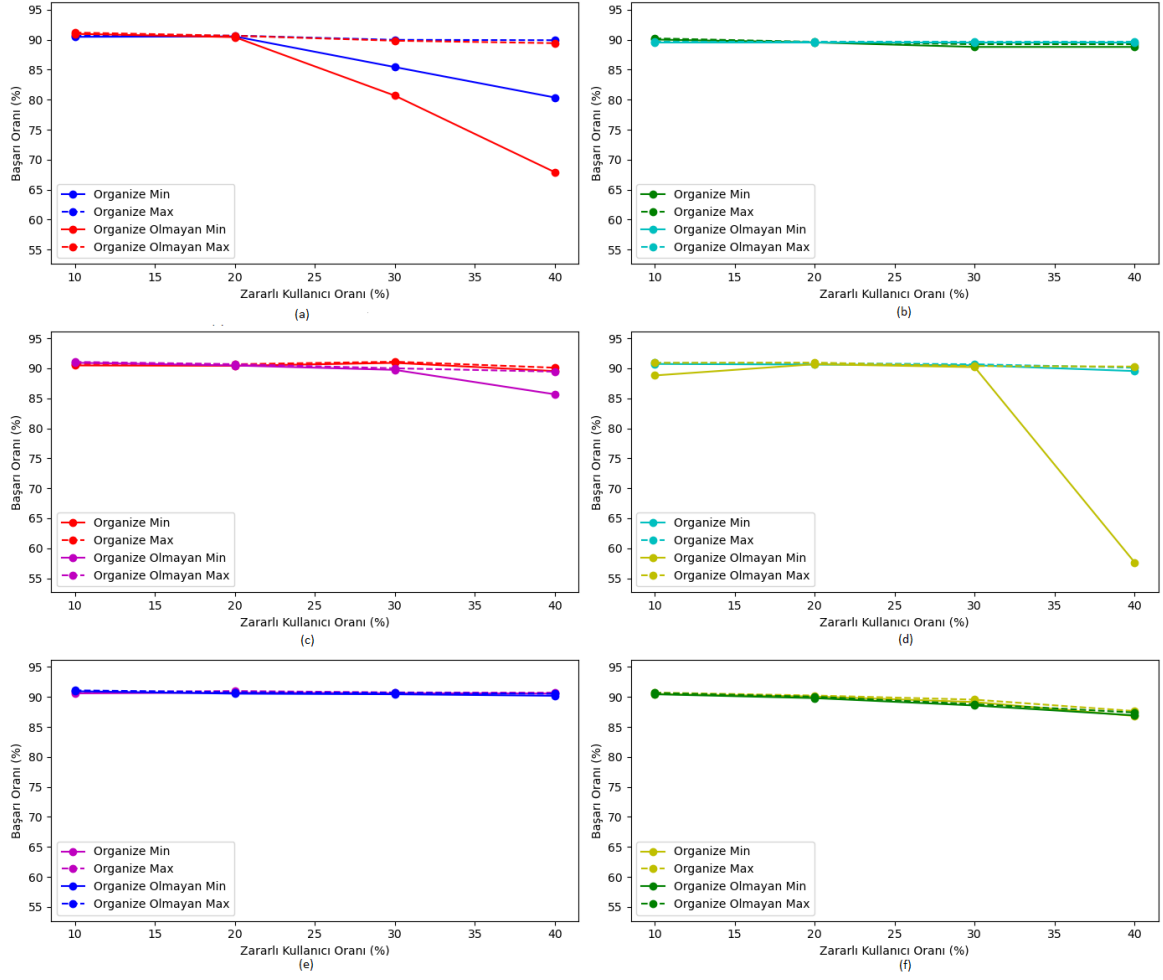
CW MEDYAN yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranını (%90.21) elde etmiş olup, %40 kötü niyetli kullanıcı oranında başarı oranı %89.24 seviyesinde kalmıştır. Bu durum, performansta sadece %1'lik bir düşüş olduğunu göstermektedir ve bu nedenle performansın stabil kabul edilebilir. Organize olmayan saldırılar altında ise, %20 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%89.73) ulaşmış olup, %40 kötü niyetli kullanıcı oranında başarı oranı %89.54 seviyesine gerilemiştir, bu da %0.2'lik bir düşüşe işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olmadığını göstermektedir.

ÇİFT SINIR yöntemi, %30 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%91.10) ulaşmış ve %40 kötü niyetli kullanıcı oranında %85.43 seviyesine düşmüştür. Bu, %6'lık bir performans düşüşüne işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%91.07) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %83.65 seviyesine düşmüştür, bu da %8'lik bir düşüşe işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir.

QAD yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%90.90) sahipken, %30 kötü niyetli kullanıcı oranında başarı oranı %90.51 ile %90.67 arasında değişmiştir. %40 kötü niyetli kullanıcı oranında ise başarı oranı %85.35 seviyesine düşmüştür, bu da %6'lık bir düşüşe işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%91.04) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %85.35 seviyesine düşmüştür, bu da %6'lık bir düşüş anlamına gelmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir.

QMAD yöntemi, %20 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%90.99) ulaşmış, %30 kötü niyetli kullanıcı oranında başarı oranı %90.62 ile %90.76 arasında değişmiştir. %40 kötü niyetli kullanıcı oranında ise başarı oranı %90.57 seviyesinde kalmıştır, bu da %0.5'lik bir düşüşe işaret etmektedir ve bu nedenle performansın stabil kabul edilebilir. Organize olmayan saldırılar altında, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%91.10) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %90.20 seviyesine düşmüştür, bu da %1'lik bir düşüş anlamına gelmektedir ve oran arttıkça performansında belirgin bir azalma olmadığını göstermektedir.

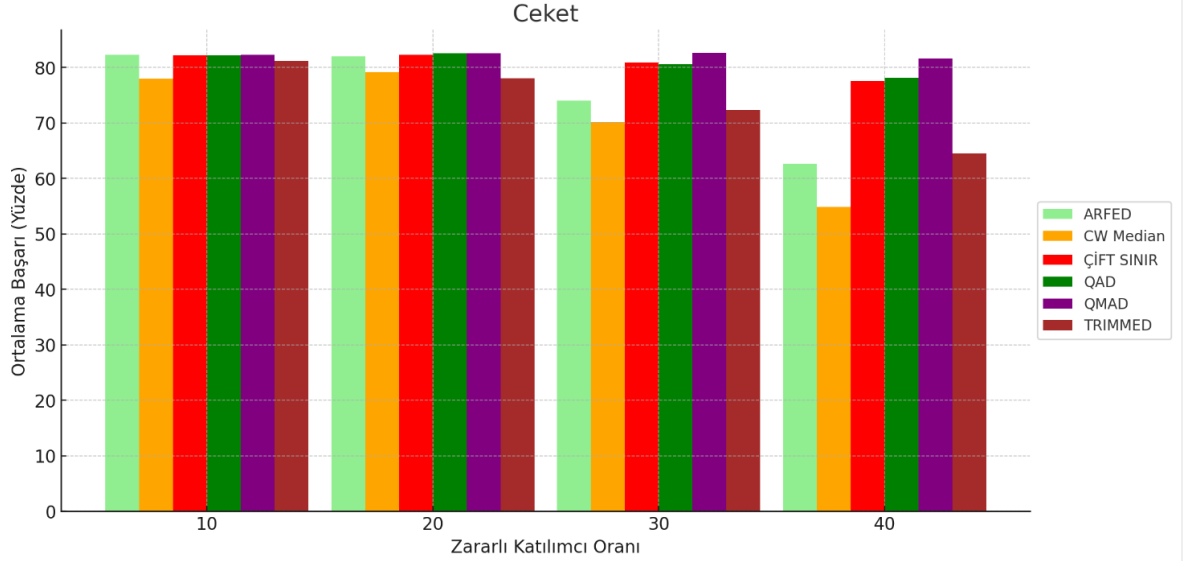
TRIMMED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%90.73) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %85.56 seviyesine düşmüştür, bu da %6'luk bir düşüş anlamına gelmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%90.67) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %86.93 seviyesine inmiştir, bu da %4'lük bir düşüşe işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir



Şekil 16 Fashion MNIST veri setinde organize olan ve organize olmayan Etiket Çevirme saldırılarında savunma yöntemlerinin test doğruluk değerleri, (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.

Genel olarak, organize ve organize olmayan saldırılar altında, ÇİFT SINIR ve QMAD yöntemleri kötü niyetli kullanıcı oranı arttıkça en yüksek performansı göstermektedir. Özellikle %30 kötü niyetli kullanıcı oranında ÇİFT SINIR yöntemi en yüksek başarı oranını sağlamaktadır. Organize olmayan saldırılar altında ise, ARFED ve ÇİFT SINIR yöntemleri %10 kötü niyetli kullanıcı oranında en yüksek performansı göstermektedir. Ancak kötü niyetli kullanıcı oranı arttıkça, özellikle ARFED yöntemi dramatik bir şekilde, yaklaşık %58'lik bir performans kaybı yaşamaktadır. CW MEDYAN ve QMAD yöntemleri ise, kötü niyetli kullanıcı oranı %40'a çıktığında bile %2'den az düşüşler gösterdiği için performansları stabil kabul edilebilir. Genel olarak, organize ve organize olmayan saldırılar

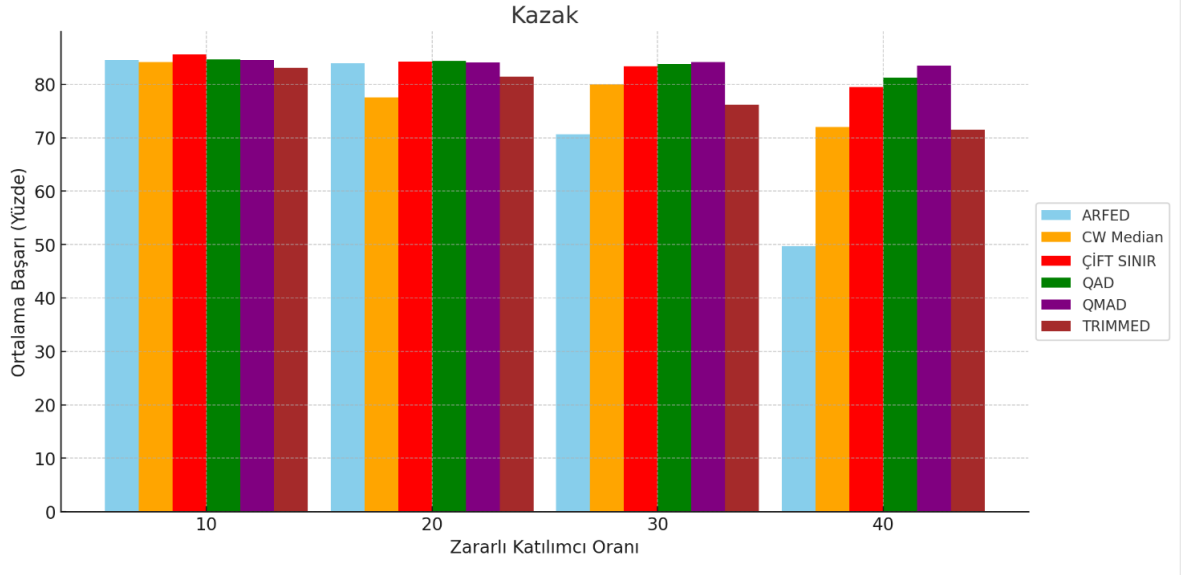
altında en yüksek performansı gösteren yöntemler Çift sınır ve QMAD'dir. Çift sınır yöntemi, %30 kötü niyetli kullanıcı oranında en yüksek başarı oranına ulaşmış-tır. QMAD yöntemi ise, kötü niyetli kullanıcı oranı %40'a çıktığında bile %2'den az düşüşler gösterdiği için performansı oldukça stabil kabul edilebilir. Algoritmalar, organize olmayan saldırılar altında organize saldırılara göre daha düşük performans göstermiştir. Bu durum, Fashion MNIST veri setinde Etiket Çevirme saldırılarında organize olmayan saldırıların daha etkili olduğunu göstermektedir. CW MEDYAN ve QMAD yöntemleri, her iki saldırı türünde de performanslarını stabil bir şekilde korumuştur. Çift sınır ve TRIMMED yöntemleri ise, kötü niyetli kullanıcı oranı arttıkça performanslarında belirgin düşüşler yaşamıştır. Yöntemlerin Test doğruluğu sonuçları ayrıntılı olarak ele alınmıştır. Saldırıların etkisi ve yöntemlerin savunma başarısının ayrıntılı incelenebilmesi için kaynak doğruluğu metrik sonuçları Şekil 17 ve Şekil 18' de gösterilerek ayrıntıları incelenecektir.



Şekil 17 Fashion MNIST veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden ceket etiketinin kaynak doğruluğu

Ceket etiketinde %10 zararlı katılımcı oranında ARFED yöntemi en yüksek başarıyı gösterirken, CW MEDIAN en düşük başarıyı elde etmiştir. %20 oranında ARFED, QAD, ÇİFT SINIR ve QMAD yöntemleri %80 başarı ile aynı seviyede performans göstermiştir. %30 ve %40 oranlarında ARFED'de belirgin bir başarı düşüşü gözlemlenmiş, CW MEDIAN yine en düşük başarıyı sergilemiştir. TRIMMED ise CW MEDIAN'dan sonra en düşük başarıyı gösteren yöntem olmuştur. Genel olarak, ÇİFT SINIR, QAD ve QMAD

yöntemleri en stabil performansı göstermiş ve zararlı katılımcı oranı artsa bile daha az düşüş yaşamıştır, bu da onları genel olarak en iyi yöntemler olarak öne çıkarmaktadır.



Şekil 18 Fashion MNIST veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden kazak etiketinin kaynak doğruluğu

Kazak etiketinde ise %10 zararlı katılımcı oranında ÇİFT SINIR yöntemi yaklaşık %82 başarı oranı ile en yüksek performansı sergilemiştir. Diğer yöntemler, ARFED, CW MEDIAN, QAD, QMAD ve TRIMMED, %80 civarında başarı göstererek birbirine oldukça yakın bir performans sergilemiştir. %20 zararlı katılımcı oranında da ÇİFT SINIR yönteminin üstünlüğü devam ederek %80 başarı oranına ulaşmıştır. Bu oranlarda ARFED, QAD ve QMAD yöntemleri de %80 civarında başarı göstermiş, ancak CW MEDIAN ve TRIMMED yöntemlerinde belirgin bir düşüş gözlemlenmiştir.

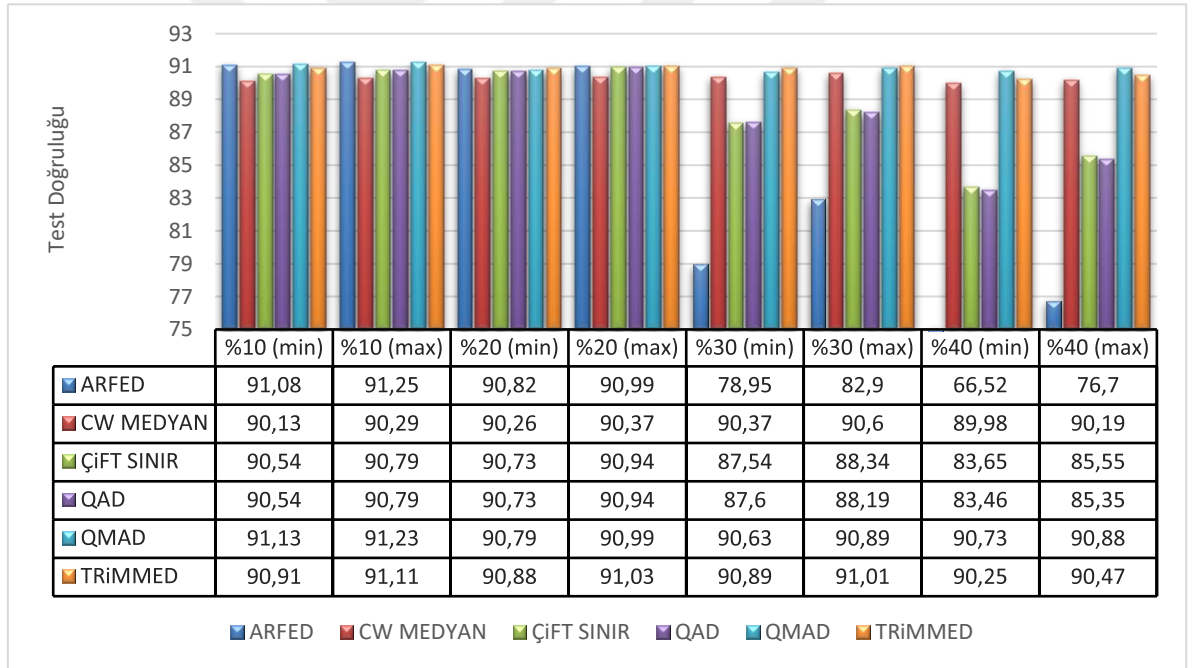
Zararlı katılımcı oranı %30'a çıktığında, ARFED yönteminin başarı oranı %70'e düşmüş ve ani bir performans kaybı yaşamıştır. Buna karşılık CW MEDIAN, %80 civarında başarı oranı yakalayarak yükselmiştir. Aynı oranda, QAD, QMAD ve ÇİFT SINIR yöntemleri %80-%83 aralığında başarı oranlarıyla en yüksek performansı göstermiştir. TRIMMED Mean yöntemi ise ARFED'den sonra en düşük başarıyı sergileyen yöntem olmuştur. %40 zararlı katılımcı oranında ise QMAD yöntemi en yüksek başarıyı sağlamış, ÇİFT SINIR ve QAD yöntemleri bir miktar düşüş yaşamalarına rağmen, QMAD'den sonra en yüksek başarı oranlarına sahip olmuştur. CW MEDIAN yönteminde düşüş gözlemlenmiş

ve TRIMMED ile benzer bir başarı seviyesine gerilemiştir. Bu oran için en düşük başarı oranı ise ARFED yöntemi tarafından gösterilmiştir.

3.2.2. Bizans Saldırısı

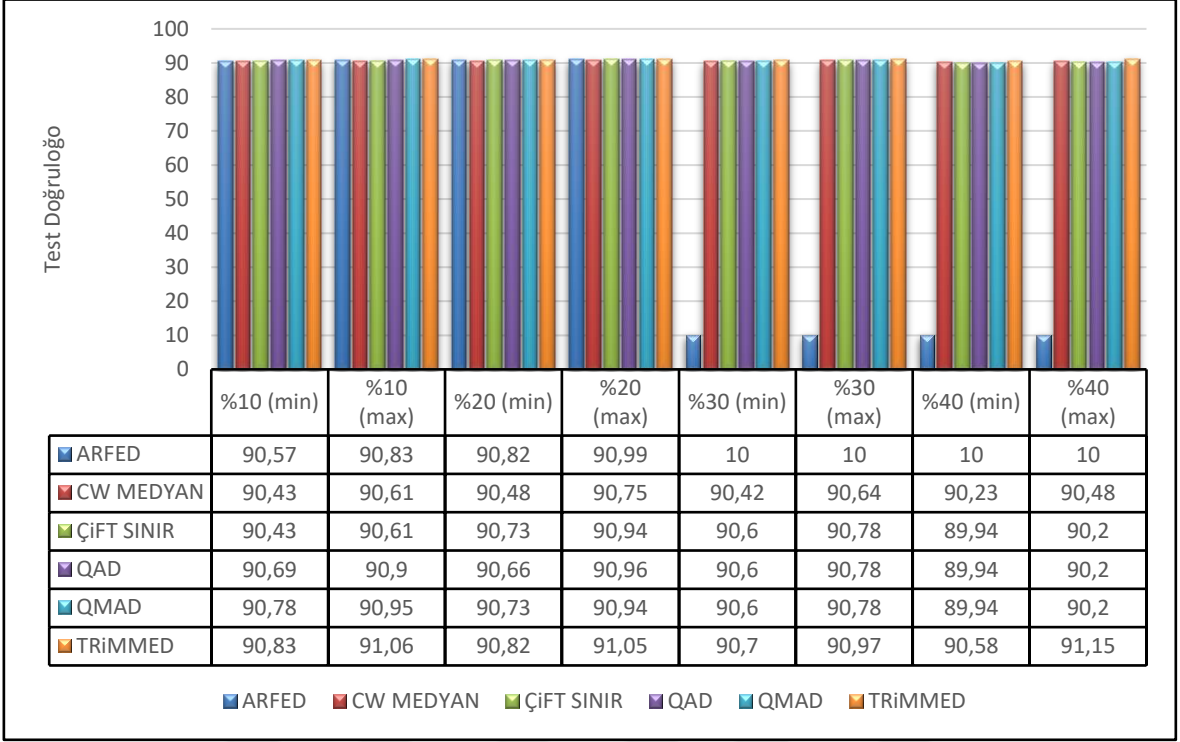
Deneylerde rastgele gradyan güncellemeleri kapsamındaki Bizans saldırıları gerçekleştirilmiştir. Yapılan Bizans saldırılarında, organize bir şekilde hareket eden kötü niyetli katılımcılar, önceden belirlenen model ağırlıklarını ana modele göndererek modelin doğruluğunu bozmayı hedeflemiştir. Veri setinin MNIST veri setinden daha zorlayıcı olması sebebi ile Bizans saldırılarının etkileri MNIST veri setine oranlara daha açık gözükmemektedir.

Tablo 17 Fashion MNIST veri setinde organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



Organize olmayan kötü niyetli katılımcılar ise, her biri rastgele belirledikleri farklı yerel model güncellemelerini ana modele göndererek modelin doğruluğunu bozmayı hedeflemiştir.

Tablo 18 Fashion MNIST veri setinde organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi, %10 ve %20 kötü niyetli kullanıcı oranlarında oldukça yüksek bir başarı göstermektedir (yaklaşık %90.57 ile %90.99 arasında). Ancak %30 ve %40 kötü niyetli kullanıcı oranlarında performansı %10'a düşerek %89'luk büyük bir düşüş yaşamıştır. Organize olmayan saldırılarda ise %10 oranında %91.08 olan başarı oranı, %30 oranında %78.95'e düşerek %13.3'lük bir düşüş, %40 oranında ise %66.52'ye düşerek %26.9'luk bir düşüş göstermiştir. Bu düşüşler, ARFED'in yüksek kötü niyetli kullanıcı oranlarına karşı savunmasız olduğunu göstermektedir.

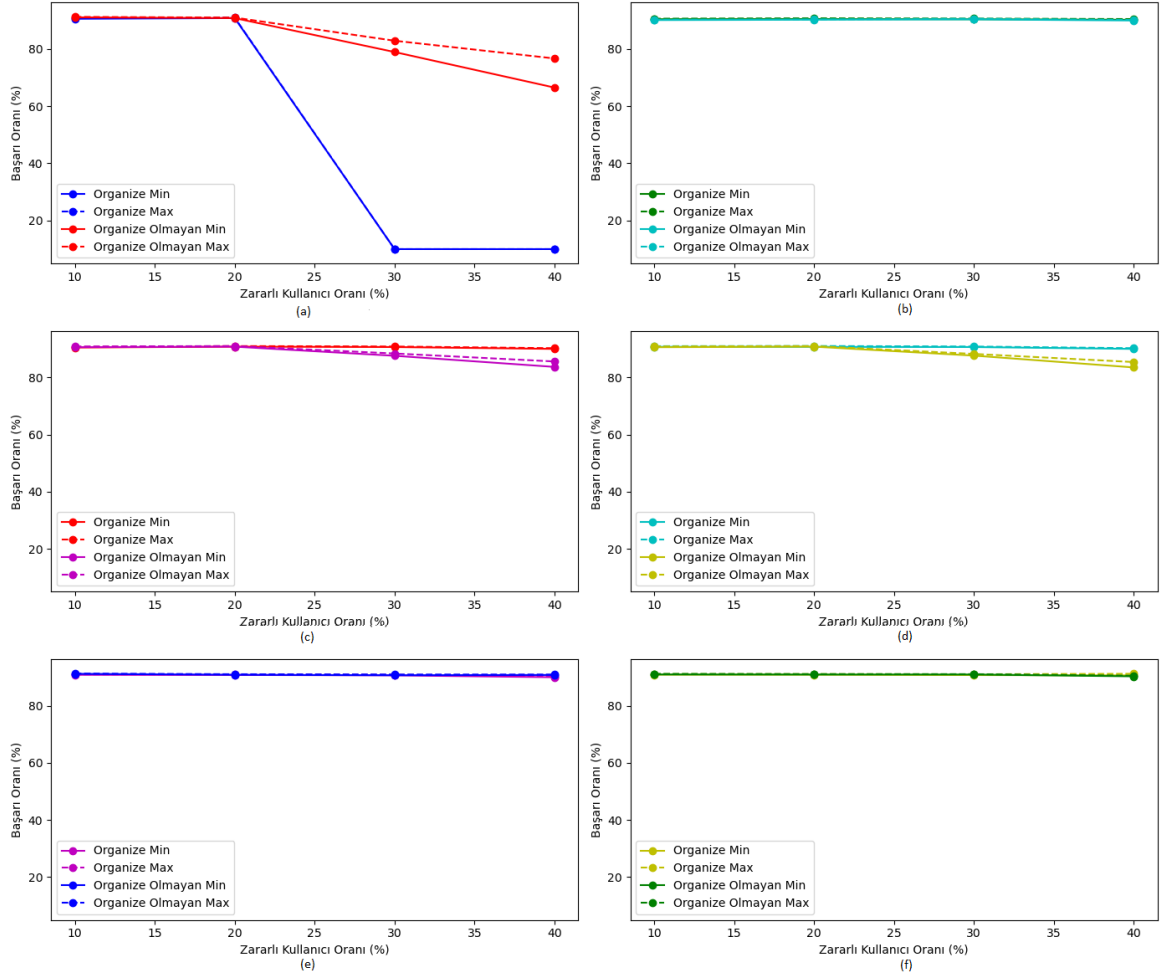
CW MEDYAN yöntemi, %10 ve %20 kötü niyetli kullanıcı oranlarında %90.43 ile %90.75 arasında oldukça tutarlı bir performans sergilemiştir. %30 ve %40 kötü niyetli kullanıcı oranlarında ise başarı oranları %90.23 ile %90.64 arasında kalmıştır. Bu, CW MEDYAN'ın yüksek kötü niyetli kullanıcı oranlarında bile tutarlılığını koruduğunu göstermektedir.

Çift sınır yöntemi, %10 ve %20 kötü niyetli kullanıcı oranlarında %90.43 ile %90.94 arasında yüksek bir başarı göstermiştir. Organize olmayan saldırılarda ise %10 oranında %90.54 olan başarı oranı, %30 oranında %87.54'e düşerek %3.3'lük bir düşüş, %40 oranında ise %83.65'e düşerek %7.6'luk bir düşüş göstermiştir.

QAD yöntemi de benzer bir performans göstermektedir. Genel olarak, tutarlı sonuçlar vermiştir. Organize olmayan saldırılarda %10 oranında %90.54 olan başarı oranı, %30 oranında %87.6'ya düşerek %3.2'lik bir düşüş, %40 oranında ise %83.46'ya düşerek %7.8'lik bir düşüş göstermiştir.

QMAD yöntemi, %10 ve %20 kötü niyetli kullanıcı oranlarında yüksek başarı göstermiştir (%90.78 ile %90.94). %30 ve %40 kötü niyetli kullanıcı oranlarında başarı oranları %90.60 ve %89.94'e düşmüş, bu düşüşler %2'den az olduğu için dikkate alınmamıştır. Bu küçük düşüşler, kötü niyetli kullanıcı sayısı artarken yararlı kullanıcı sayısının azalması nedeniyle doğal varyasyonlar olarak kabul edilmektedir. Genel olarak, QMAD yöntemi hem organize hem de organize olmayan saldırılarda yüksek başarı oranlarını korumuş ve tutarlılığını sürdürmüştür.

TRIMMED yöntemi ise Genel olarak hem organize hem de organize olmayan saldırılarda yüksek başarı oranlarını korumuş ve tutarlılığını sürdürmüştür. Bu sonuçlar, TRIMMED yönteminin kötü niyetli kullanıcı oranlarının artmasına rağmen başarılı ve tutarlı bir performans sergilediğini göstermektedir.



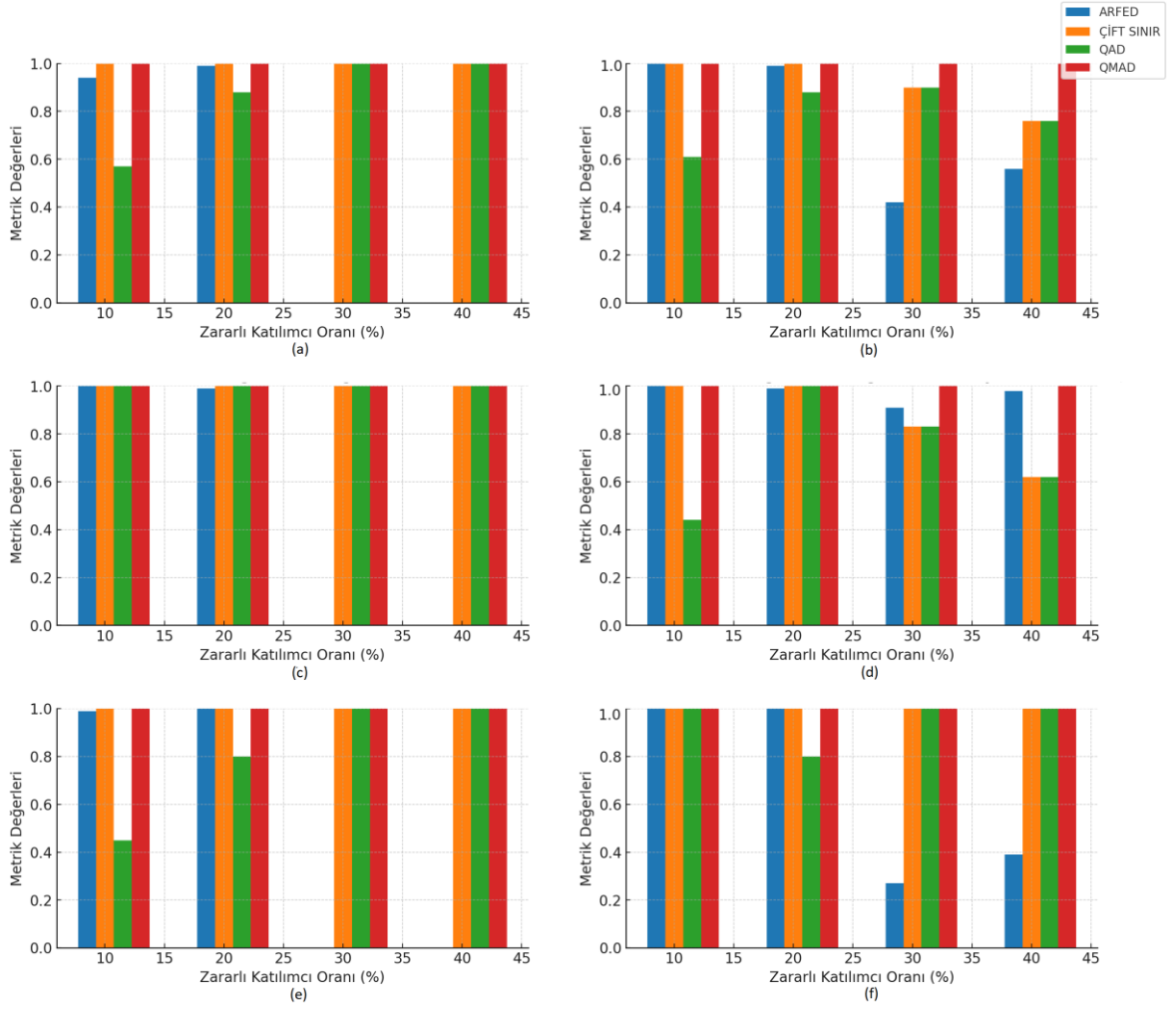
Şekil 19 Fashion MNIST veri setinde organize olan ve organize olmayan Bizans saldırılarında savunma yöntemlerinin test doğruluk değerleri, (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu, (f), organize ve organize olmayan saldırılardaki TRIMMED test doğruluğu.

Organize saldırılarda ARFED yöntemi %10 ve %20 kötü niyetli kullanıcı oranlarında yüksek başarı göstermiş olsa da %30 ve %40 kötü niyetli kullanıcı oranlarında büyük performans düşüşleri yaşamıştır. CW MEDYAN ve QMAD yöntemleri bu tür saldırılarda oldukça tutarlı performans sergilemiş ve yüksek kötü niyetli kullanıcı oranlarında bile başarı oranlarını korumuştur. Çift sınır ve QAD yöntemleri ise organize saldırılarda yüksek başarı göstermiş, ancak düşüşler yok sayılabilir düzeyde kalmıştır. TRIMMED yöntemi ise bu tür saldırılarda yüksek başarı oranlarını korumuş ve tutarlılığını sürdürmüştür.

Organize olmayan saldırılarda ARFED yöntemi %30 ve %40 kötü niyetli kullanıcı oranlarında ciddi performans düşüşleri yaşamıştır. CW MEDYAN yöntemi bu saldırılarda

da tutarlı performansını korumuş ve yüksek kötü niyetli kullanıcı oranlarında bile başarılı olmuştur. Çift sınır ve QAD yöntemleri ise organize olmayan saldırılarda belirgin performans düşüşleri yaşamışlardır. QMAD yöntemi organize olmayan saldırılarda da oldukça tutarlı bir performans sergilemiştir. TRIMMED yöntemi ise düşük kötü niyetli kullanıcı oranlarında yüksek başarı oranlarını korumuş, yüksek oranlarda ise küçük düşüşlerle tutarlılığını sürdürmüştür.

Genel olarak, en tutarlı yöntemler CW MEDYAN ve QMAD olmuştur. Her iki saldırı türünde de yüksek başarı oranlarını koruyarak en dayanıklı performansı göstermişlerdir. TRIMMED yöntemi de her iki saldırı türünde yüksek başarı oranlarını sürdürmüş ve tutarlılığını korumuştur. ARFED yöntemi ise özellikle yüksek kötü niyetli kullanıcı oranlarında büyük performans düşüşleri yaşayarak en savunmasız yöntem olarak öne çıkmıştır. Fashion MNIST veri seti için en uygun yöntemler CW MEDYAN, QMAD ve TRIMMED olarak değerlendirilebilir. Bu yöntemler, yüksek kötü niyetli kullanıcı oranlarına karşı bile tutarlı ve yüksek performans sergilemişlerdir. Test doğruluk değerlerini analiz ettiğimiz yöntemlerin F1, Doğruluk ve Kesinlik değerlerini Şekil 20’de göstererek analiz edeceğiz ve test doğruluk metriği sonuçları ile tutarlılığı değerlendirilecektir. F1 skor hesaplamasında CW MEDYAN ve TRIMMED yöntemi değerlendirilememiştir çünkü bu yöntemlerde direkt olarak bir zararlı katılımcı tespiti yapılmamaktadır.



Şekil 20 Fashion MNIST veri setindeki organize ve organize olmayan Bizans saldırılarındaki yöntemlerin F1, Doğruluk ve Keskinlik değerleri. (a), organize saldırılardaki F1 skor, (b) organize olmayan saldırılardaki F1 skor, (c), organize saldırılardaki Doğruluk, (d) organize olmayan saldırılardaki Doğruluk, (e), organize saldırılardaki Keskinlik, (f) organize olmayan saldırılardaki Keskinlik değerleri.

Şekil 20’de gösterilen F1, Doğruluk ve Keskinlik değerlerine göre ARFED yöntemi, %10 ve %20 zararlı katılımcı oranlarında hem organize hem de organize olmayan saldırılar altında yüksek doğruluk, kesinlik ve F1 skoru ile başarılı olmuştur. Ancak, %30 ve %40 zararlı katılımcı oranlarında performansı dramatik bir şekilde düşmüştür. ARFED’in özellikle yüksek zararlı katılımcı oranları altında organize ve organize olmayan saldırılara karşı yeterince dirençli olmadığı görülmektedir.

ÇİFT SINIR yöntemi, %10 ve %20 zararlı katılımcı oranlarında hem organize hem de organize olmayan saldırılar altında iyi bir performans sergilemiştir. Ancak, %30 ve %40 oranlarına gelindiğinde bu yöntemin etkisi belirgin şekilde azalmıştır. ÇİFT SINIR’ın düşük

zararlı katılımcı oranlarında etkili olduğunu, ancak daha yüksek oranlarda performans kaybı yaşadığı görülmektedir.

QAD yöntemi, %10 ve %20 zararlı katılımcı oranlarında her iki saldırı türünde de belirli bir başarı göstermiştir. Ancak, %30 ve %40 oranlarında bu yöntemin performansı hem organize hem de organize olmayan saldırılarda ciddi şekilde düşmüştür. Bu da QAD'nin yüksek zararlı katılımcı oranları altında etkili olmadığını göstermektedir.

QMAD yöntemi de benzer bir eğilim göstermiştir. %10 ve %20 zararlı katılımcı oranlarında her iki saldırı türünde de iyi performans sergilerken, %30 ve %40 oranlarında performansı büyük ölçüde azalmıştır. QMAD yönteminin de yüksek zararlı katılımcı oranları karşısında zayıf kaldığını görülmüştür.

Tüm yöntemler, %10 ve %20 zararlı katılımcı oranlarında organize ve organize olmayan saldırılara karşı başarılı olmuş, ancak %30 ve %40 oranlarında belirgin bir performans düşüşü yaşamıştır. Bu, yüksek zararlı katılımcı oranlarına karşı daha dirençli yöntemlerin geliştirilmesi gerektiğini göstermektedir. Her yöntemin performansı, saldırıların organize olup olmamasından bağımsız olarak zararlı katılımcı oranı arttıkça düşmektedir. Bu da saldırıların yoğunluğuna karşı daha dayanıklı modellerin geliştirilmesi gerekliliğini ortaya koymaktadır.

3.3. CIFAR Veri Setinde Yapılan Deneysel Çalışmalar

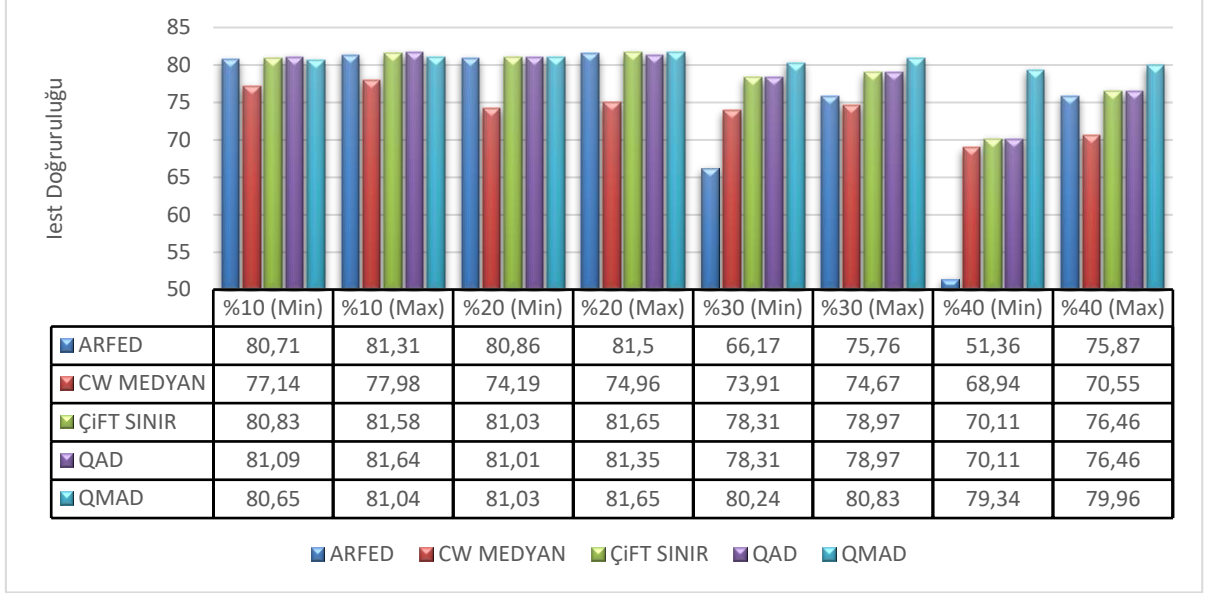
CIFAR-10 veri seti 50.000 eğitim ve 10.000 test görüntüsüne sahiptir, yapılan deneyler arasında en zorlu deney setini oluşturmaktadır. Bu nedenle, bu veri setinde kullanılan savunma yöntemlerinin başarısı büyük önem taşımaktadır. Çünkü kötü niyetli saldırganlar tespit edilemediğinde, modelin doğruluk değerlerinde ciddi düşüşler yaşanacaktır. Bu durum, savunma mekanizmalarının etkinliğini değerlendirmede CIFAR-10 veri setini kritik bir test ortamı haline getirmektedir. Savunma yöntemlerinin etkinliği 4 farklı deneyde test edilmiş ve sonuçları analiz edilmiştir.

3.3.1. Etiket Çevirme Saldırısı

Etiket Çevirme saldırılarında saldırganların organize olması durumunda CIFAR-10 veri setinde Tablo 1’de gösterildiği üzere kedi örnekleri köpek, köpek etiketler de kedi olarak

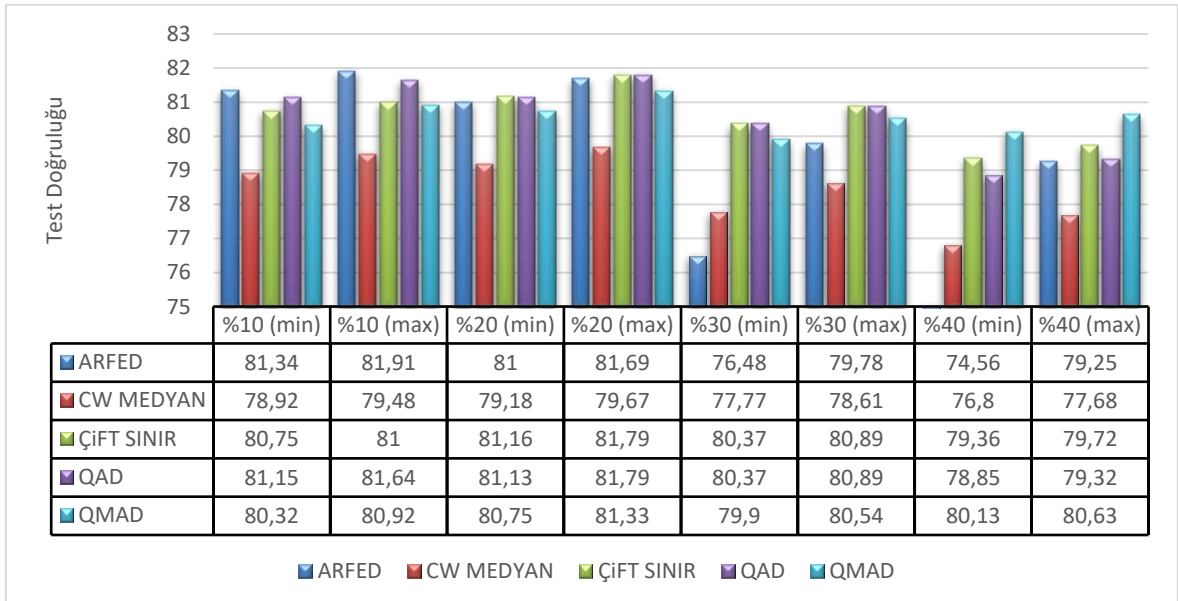
çevrilmiştir. Saldırganların organize çalışmasının test doğruluk sonuçları Tablo 20’de gösterilmiştir

Tablo 19 CIFAR veri setinde organize olmayan Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



Organize olmayan saldırganlar ise rastgele olarak sınıf etiketlerini değiştirmişlerdir. Saldırganların organize olmadan çalışmasının sonuçları Tablo 20’de gösterilmiştir.

Tablo 20 CIFAR veri setinde organize Etiket Çevirme saldırısında savunma yöntemlerinin test doğruluk değerleri



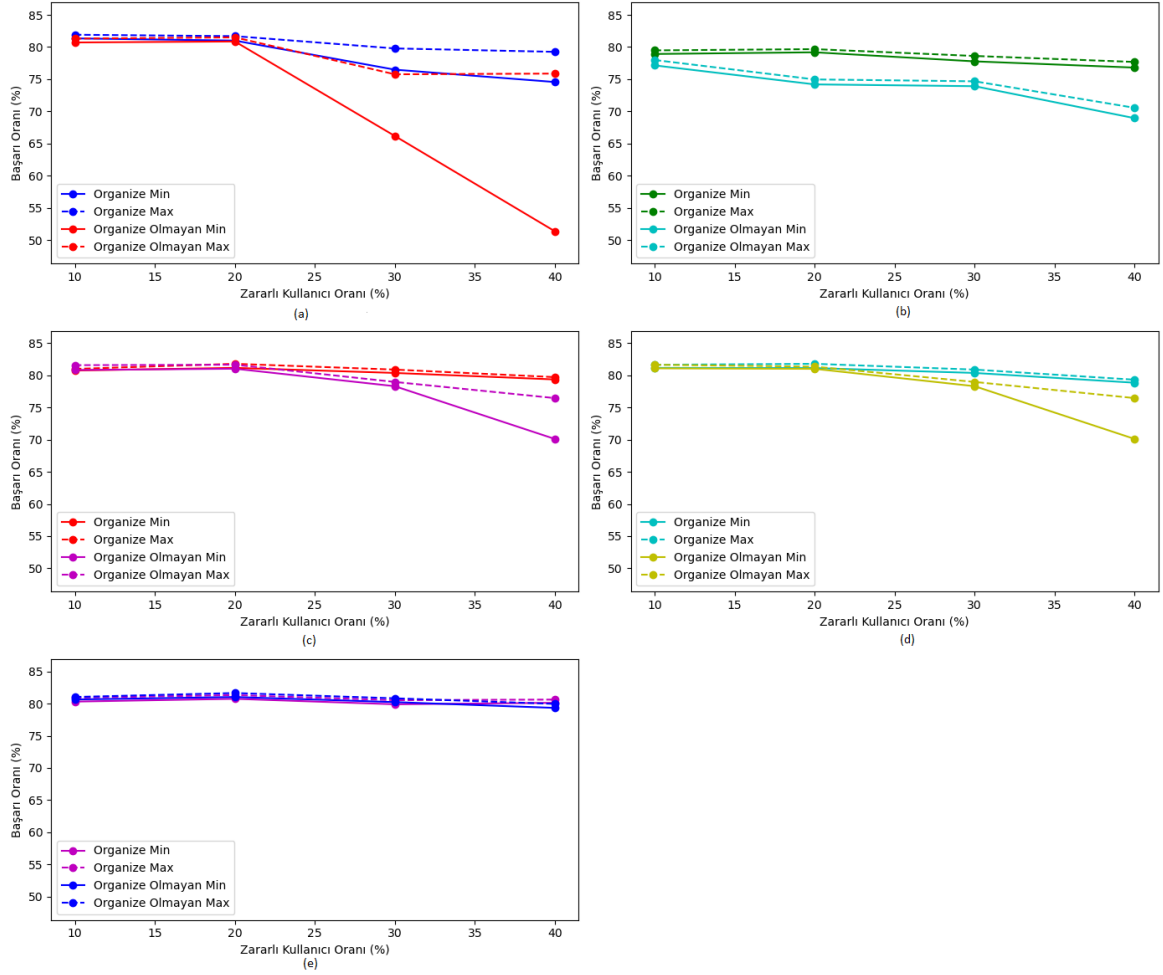
ARFED yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.91) ulaşırken, kötü niyetli kullanıcı oranının %40'a çıkmasıyla başarı oranı %74.56 seviyesine gerilemiştir. Bu, performansta yaklaşık %9'luk bir düşüşe işaret etmektedir ve oran arttıkça performansında önemli bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.31) ulaşmış, ancak %40 kötü niyetli kullanıcı oranında başarı oranı %51.36 seviyesine düşmüştür. Bu da yaklaşık %37'lik bir performans düşüşüne işaret etmektedir ve oran arttıkça performansının dramatik şekilde azaldığını göstermektedir.

CW MEDYAN yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranını (%79.48) elde etmiş olup, %40 kötü niyetli kullanıcı oranında başarı oranı %77.68 seviyesinde kalmıştır. Bu durum, performansta sadece %2'lik bir düşüş olduğunu göstermektedir ve bu nedenle performansın nispeten stabil kabul edilebilir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%77.98) ulaşmış olup, %40 kötü niyetli kullanıcı oranında başarı oranı %70.55 seviyesine gerilemiştir, bu da yaklaşık %10'luk bir düşüşe işaret etmektedir ve oran arttıkça performansında belirgin bir azalma olduğunu göstermektedir.

Çift sınır yöntemi, %30 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.79) ulaşmış ve %40 kötü niyetli kullanıcı oranında %79.72 seviyesine düşmüştür. Bu, %2'lik bir performans düşüşüne işaret etmektedir ve oran arttıkça performansında küçük bir azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.58) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %76.46 seviyesine düşmüştür, bu da %6'lık bir düşüşe işaret etmektedir ve oran arttıkça performansında kayda değer bir azalma olduğunu göstermektedir.

QAD yöntemi, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.64) sahipken, %30 kötü niyetli kullanıcı oranında başarı oranı %80.89 seviyesine düşmüştür. %40 kötü niyetli kullanıcı oranında ise başarı oranı %79.32 seviyesine düşmüştür, bu da %3'lük bir düşüşe işaret etmektedir ve oran arttıkça performansında bir miktar azalma olduğunu göstermektedir. Organize olmayan saldırılar altında ise, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.64) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %76.46 seviyesine düşmüştür, bu da %6'lık bir düşüş anlamına gelmektedir ve oran arttıkça performansında kayda değer bir azalma olduğunu göstermektedir.

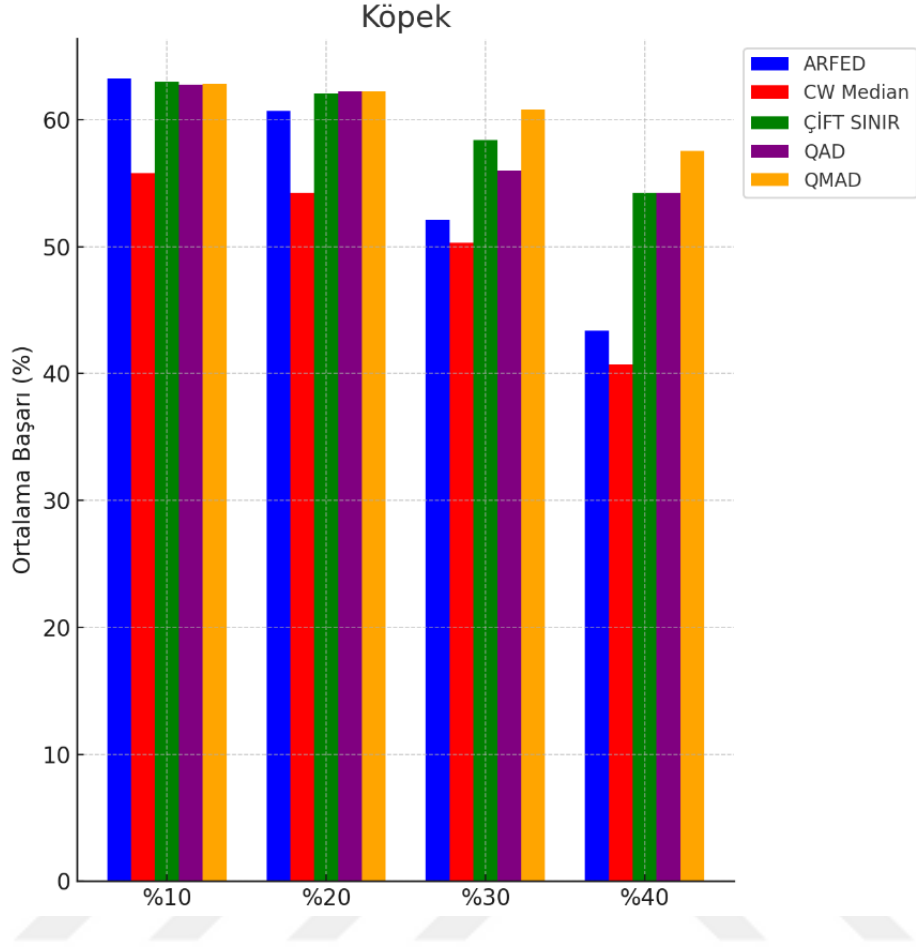
QMAD yöntemi, %20 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.33) ulaşmış, %30 kötü niyetli kullanıcı oranında başarı oranı %80.54 seviyesine düşmüştür. %40 kötü niyetli kullanıcı oranında ise başarı oranı %80.63 seviyesinde kalmıştır, bu da %0.9'luk bir düşüşe işaret etmektedir ve bu nedenle performansın stabil kabul edilebilir. Organize olmayan saldırılar altında, %10 kötü niyetli kullanıcı oranında en yüksek başarı oranına (%81.04) sahipken, %40 kötü niyetli kullanıcı oranında başarı oranı %79.96 seviyesine düşmüştür, bu da %1.3'lük bir düşüş anlamına gelmektedir ve oran arttıkça performansında hafif bir azalma olduğunu göstermektedir. Şekil 21'de her yöntemin ayrı ayrı olarak test doğrulukları gösterilmiştir.



Şekil 21 CIFAR veri setinde organize olan ve organize olmayan Etiket Çevirme saldırılarında savunma yöntemlerinin test doğruluk değerleri, (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu.

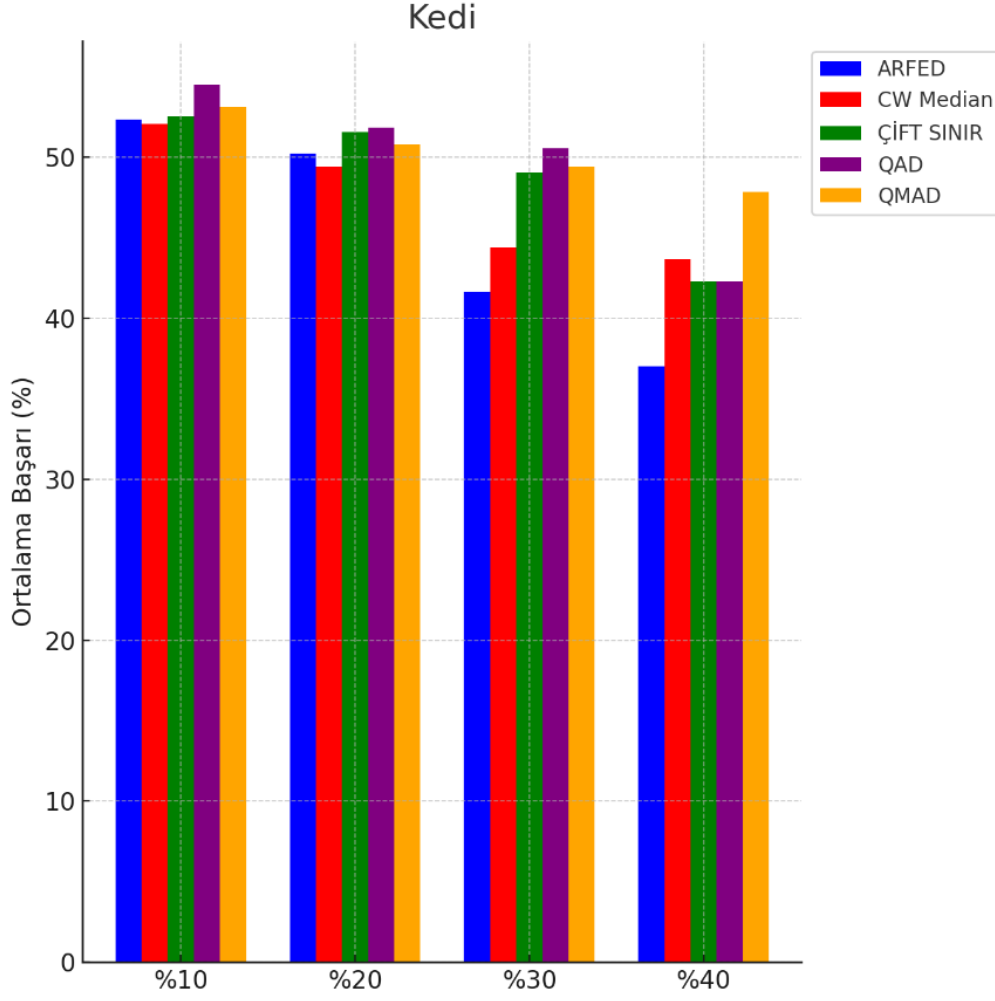
Şekil 21’da görüldüğü gibi genel olarak, organize ve organize olmayan saldırılar altında, Çift sınır ve QMAD yöntemleri kötü niyetli kullanıcı oranı arttıkça en yüksek performansı göstermektedir. Özellikle %30 kötü niyetli kullanıcı oranında Çift sınır yöntemi en yüksek başarı oranını sağlamaktadır. Organize olmayan saldırılar altında ise, ARFED ve Çift sınır yöntemleri %10 kötü niyetli kullanıcı oranında en yüksek performansı göstermektedir. Ancak kötü niyetli kullanıcı oranı arttıkça, özellikle ARFED yöntemi dramatik bir şekilde, yaklaşık %37’lik bir performans kaybı yaşamaktadır. QMAD yöntemi ise kötü niyetli kullanıcı oranı %40’a çıktığında bile %2’den az düşüş gösterdiği için performansı stabil kabul edilebilir. Bu analiz, yöntemlerin kötü niyetli kullanıcı oranlarına göre performans değişimlerini göstermektedir ve bazı yöntemler belirli kötü niyetli kullanıcı oranlarında diğerlerinden daha iyi performans sergilemektedir.

Organize saldırılar altında, algoritmalar genellikle daha yüksek performans göstermiştir. Organize olmayan saldırılar, ARFED, CW MEDYAN ve Çift sınır gibi yöntemlerde daha fazla performans kaybına yol açmıştır. Özellikle, ARFED ve ÇİFT SINIR yöntemleri organize olmayan saldırılar altında organize saldırılara göre daha belirgin performans düşüşleri yaşamıştır. Bu durum, organize olmayan saldırıların daha etkili olduğunu göstermektedir. Test doğruluğu sonuçları analiz edilmiştir. Daha ayrıntılı sonuçların incelemesi için saldırıya uğrayan kedi ve köpek etiketlerinin kaynak doğruluğu metriği Şekil 22 ve Şekil 23’de gösterilerek detaylı analiz edilecektir.



Şekil 22 CIFAR veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden köpek etiketinin kaynak doğruluğu

Şekil 22’de Köpek sınıfında, ARFED ve ÇİFT SINIR yöntemleri %10 ve %20 zararlı katılımcı oranlarında istikrarlı bir performans sergilerken, CW MEDİAN yönteminin diğerlerine göre daha düşük bir başarı gösterdiği dikkat çekiyor. %30 oranında, CW MEDİAN yöntemi ciddi bir düşüş yaşıyor, buna karşın QMAD ve ÇİFT SINIR yöntemleri bu oranda daha iyi performans gösteriyor. %40 zararlı katılımcı oranında ise QMAD yöntemi en yüksek başarıyı elde ederken, ARFED ve ÇİFT SINIR yöntemleri biraz daha düşük başarı oranlarıyla öne çıkıyor.



Şekil 23 CIFAR veri setindeki organize Etiket Çevirme saldırısına uğrayan etiketlerden kedi etiketinin kaynak doğruluğu

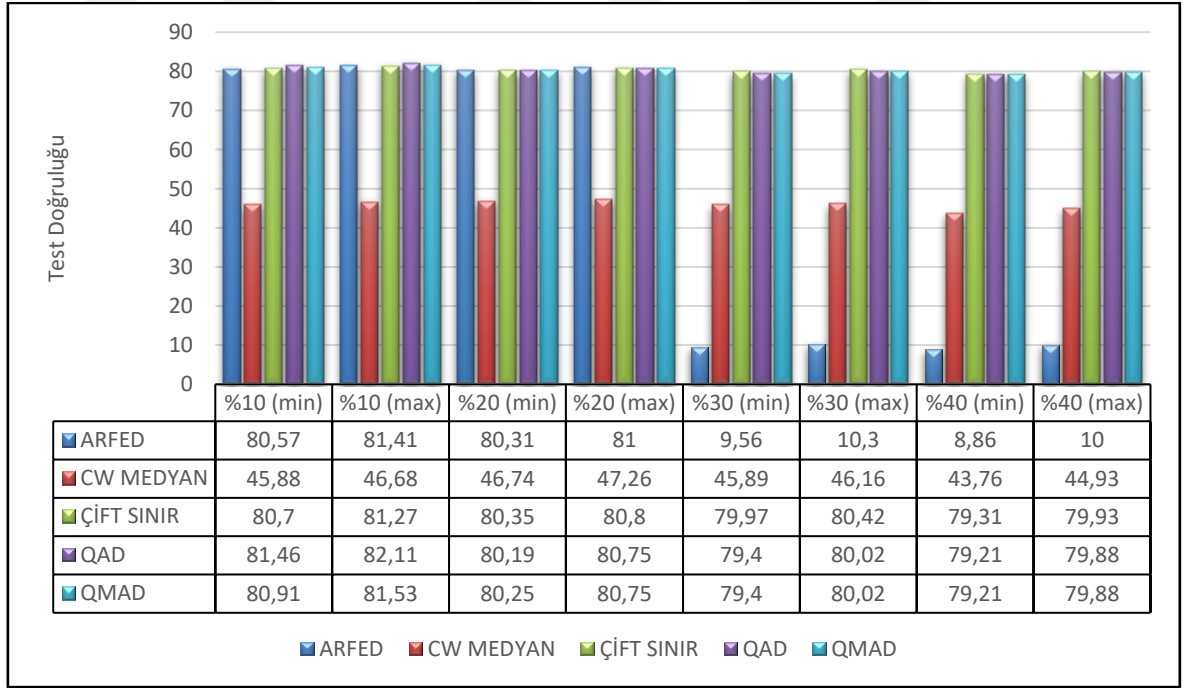
Şekil 23’de Kedi sınıfında da benzer bir eğilim görülüyor. ARFED, ÇİFT SINIR ve QMAD yöntemleri düşük zararlı katılımcı oranlarında (%10 ve %20) oldukça başarılı. Ancak %30 oranında CW MEDIAN yönteminin başarısı önemli ölçüde azalıyor ve QMAD yöntemi bu oranda da güçlü performansını sürdürüyor. %40 zararlı katılımcı oranında, yine QMAD en yüksek başarıyı elde ederken, CW MEDIAN yöntemi daha düşük bir performans gösteriyor.

Genel olarak, QMAD yöntemi, özellikle yüksek zararlı katılımcı oranlarında hem Köpek hem de Kedi sınıflarında en istikrarlı ve yüksek başarıyı sağlayan yöntem olarak öne çıkıyor. CW MEDIAN yöntemi ise özellikle %30 ve %40 oranlarında belirgin bir performans düşüşü gösteriyor, bu da onu diğer yöntemlere kıyasla daha az güvenilir kılıyor.

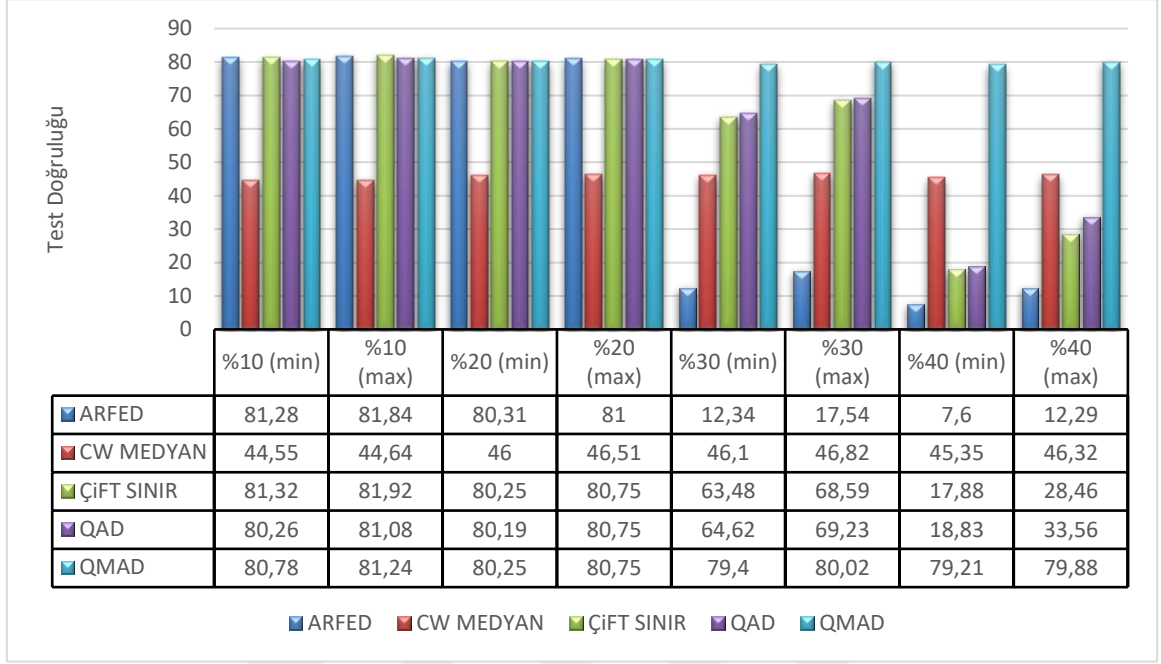
3.3.2. Bizans Saldırısı

Deneylerde diğer 2 veri setindeki saldırgan modeli kullanılmıştır, rastgele gradyan güncellemeleri kapsamındaki Bizans saldırıları gerçekleştirilmiştir. Yapılan Bizans saldırılarında, organize bir şekilde hareket eden kötü niyetli katılımcılar, önceden belirlenen model ağırlıklarını ana modele göndererek modelin doğruluğunu bozmayı hedeflemiştir. CIFAR veri setinin zorlayıcı yapısı nedeni ile kötü niyetli katılımcıların etkisinin en çok görüldüğü veri seti CIFAR olmuştur.

Tablo 21 CIFAR veri setinde organize Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



Tablo 22 CIFAR veri setinde organize olmayan Bizans saldırısında savunma yöntemlerinin test doğruluk değerleri



ARFED yöntemi, kötü niyetli kullanıcı oranının %10 ve %20 olduğu durumlarda oldukça yüksek bir başarı göstermiştir. Ancak, kötü niyetli kullanıcı oranı %30 ve %40'a çıktığında performansı dramatik bir şekilde düşmüştür. Organize saldırılarda %10 oranında %81.41 olan başarı oranı, %30 oranında %10.30'a düşmüştür ki bu yaklaşık %87'lik bir düşüştür. %40 oranında ise başarı oranı %10.00'a düşerek yaklaşık %88'lik bir düşüş göstermiştir.

Organize olmayan saldırılarda ise %10 oranında %81.84 olan başarı oranı, %30 oranında %17.54'e düşmüş ve bu %79'luk bir düşüş anlamına gelmektedir. %40 oranında ise başarı %12.29'a düşerek %85'lik bir düşüş göstermiştir. Bu, ARFED'in yüksek kötü niyetli kullanıcı oranlarında çok büyük performans kayıpları yaşadığını göstermektedir.

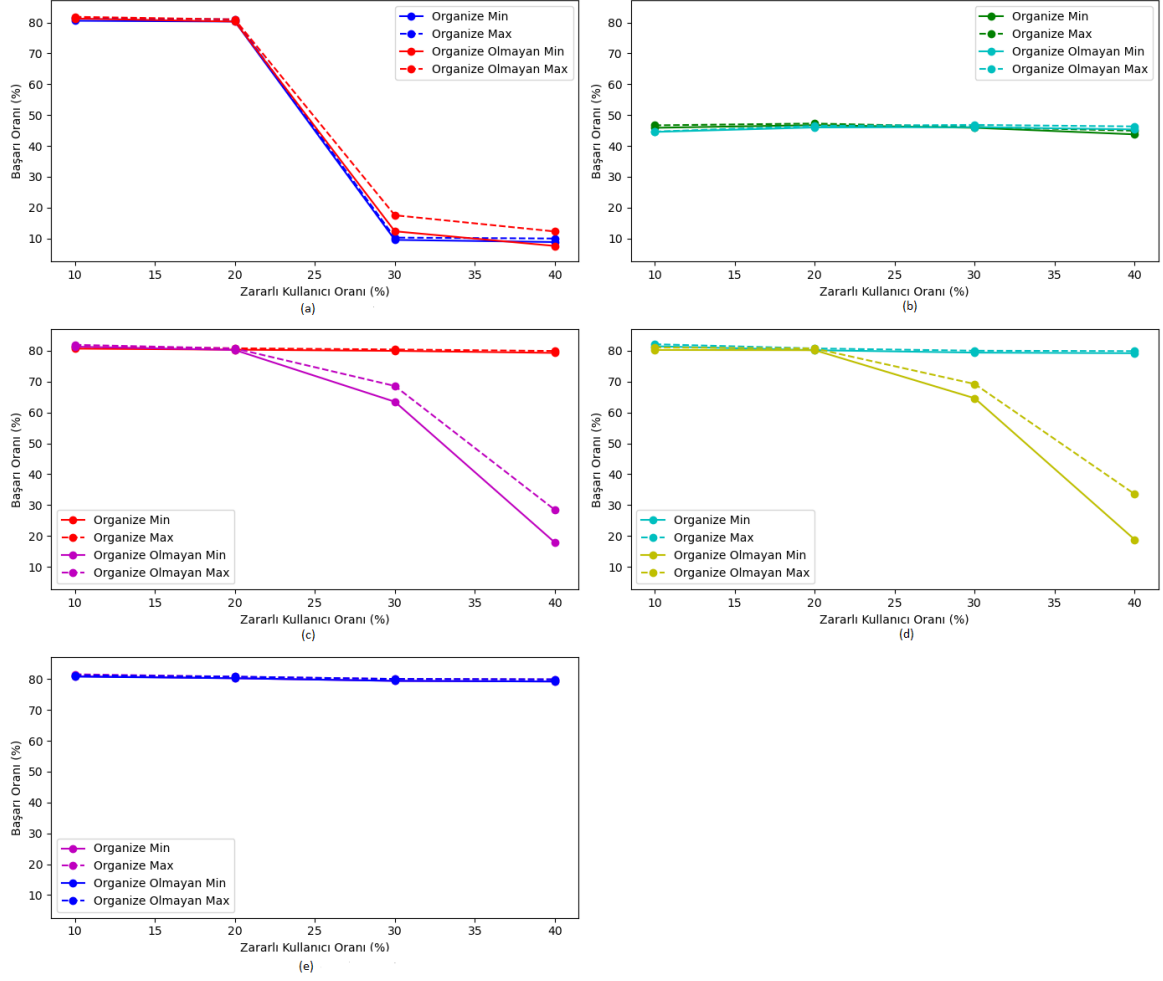
CW MEDYAN yöntemi, her iki saldırı türünde de tutarlı bir performans sergilemiştir. %10 ve %20 kötü niyetli kullanıcı oranlarında başarı oranları %44.55 ile %47.26 arasında değişmiş, %30 ve %40 oranlarında ise %43.76 ile %46.32 arasında kalmıştır. Ancak, bu tutarlılığa rağmen CW MEDYAN'ın başarı oranları diğer yöntemlere kıyasla oldukça düşüktür. Bu, CW MEDYAN'ın güvenilir ve tutarlı olmasına rağmen, genel başarı oranının düşük olduğunu ve dolayısıyla en iyi seçenek olmadığını göstermektedir.

ÇİFT SINIR yöntemi, %10 ve %20 kötü niyetli kullanıcı oranlarında yüksek başarı göstermiştir. Ancak, %30 ve %40 oranlarında organize olmayan saldırılarda performansı

ciddi şekilde düşmüştür. Örneğin, %10 oranında %81.92 olan başarı oranı, %30 oranında %68.59'a düşmüştür ki bu %16'lık bir düşüştür. %40 oranında ise başarı oranı %28.46'ya düşerek %65'lik bir düşüş göstermiştir. Bu, ÇİFT SINIR'ın organize olmayan saldırılarda %40 oranında ciddi düşüş yaşadığını göstermektedir.

QAD yöntemi, organize saldırılarda %81.46 ile %82.11 arasında en yüksek başarı oranını göstermiştir. Ancak organize olmayan saldırılarda %30 ve %40 oranlarında performansı %64.62 ile %69.23 arasında değişmiştir. Bu, %30 ve %40 kötü niyetli kullanıcı oranlarında organize saldırılarda en yüksek performansı gösteren QAD'nin, organize olmayan saldırılarda kötü niyetli katılımcı oranı %30 u aştığında %15 ile %18 arasında bir performans düşüşü yaşadığını göstermektedir.

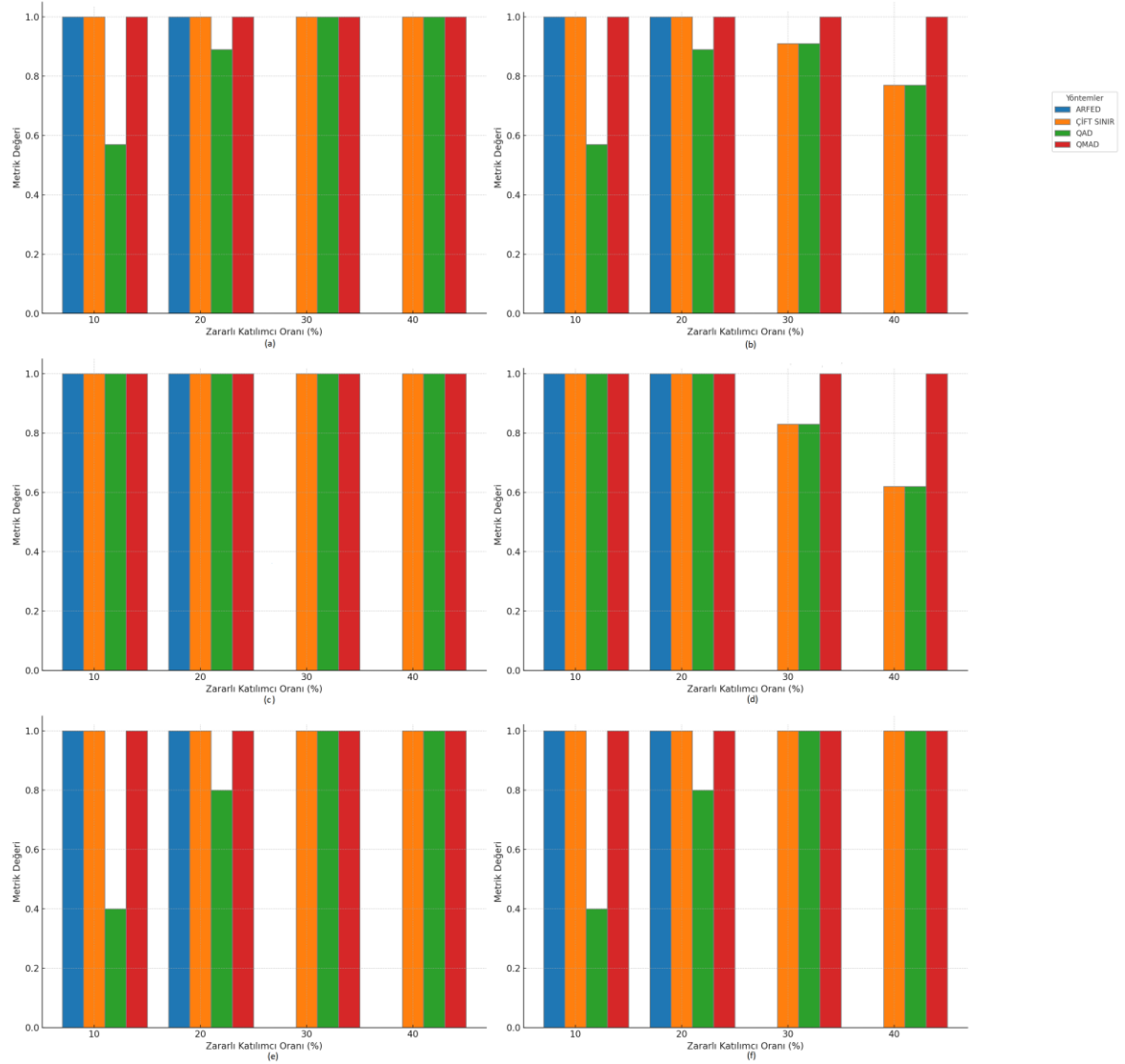
QMAD yöntemi, her iki saldırı türünde de en tutarlı performansı sergilemiştir. %10 ve %20 kötü niyetli kullanıcı oranlarında başarı oranları %80.25 ile %81.53 arasında değişmiş, %30 ve %40 oranlarında ise %79.21 ile %80.02 arasında kalmıştır. Bu, QMAD'in hem organize hem de organize olmayan saldırılarda en tutarlı yöntem olduğunu göstermektedir.



Şekil 24 CIFAR veri setinde organize olan ve organize olmayan Bizans saldırılarında savunma yöntemlerinin test doğruluk değerleri (a), organize ve organize olmayan saldırılardaki ARFED test doğruluğu, (b) organize ve organize olmayan saldırılardaki CW MEDIAN test doğruluğu, (c), organize ve organize olmayan saldırılardaki ÇİFT SINIR test doğruluğu, (d) organize ve organize olmayan saldırılardaki QAD test doğruluğu, (e), organize ve organize olmayan saldırılardaki QMAD test doğruluğu.

• Organize saldırılarda en iyi performansı QAD göstermiştir. Ancak organize olmayan saldırılarda %30 ve %40 kötü niyetli kullanıcı oranlarında performansı %15 ile %18 arasında bir düşüş yaşamıştır. QMAD, her iki saldırı türünde de en tutarlı performansı sergilemiştir. CW MEDYAN, tutarlı bir performans sergilemesine rağmen, genel başarı oranları diğer yöntemlere göre oldukça düşüktür. ARFED ise kötü niyetli kullanıcı oranı arttıkça en büyük düşüşü yaşamış ve bu nedenle yüksek kötü niyetli kullanıcı oranlarında başarısız olmuştur. ÇİFT SINIR da benzer şekilde yüksek kötü niyetli kullanıcı oranlarında performans düşüşü yaşamıştır.

• Şekil 24 Şekil 24 CIFAR veri setinde organize olan ve organize olmayan Bizans saldırılarında savunma yöntemlerinin test doğruluk değerleri’ da grafiksel olarak gösterilen sonuçlarda da açıkça görüldüğü gibi QMAD hem organize hem de organize olmayan saldırılarda en tutarlı yöntem olmuştur. CW MEDYAN, tutarlı bir performans sergilemesine rağmen genel başarı oranı düşük kalmıştır. ARFED ise yüksek kötü niyetli kullanıcı oranlarında ciddi performans düşüşleri yaşamışlardır. Yöntemlerin Test doğruluk değerleri incelenmiştir. Sonuçların F1 skoru değerleri de değerlendirilmiştir.



Şekil 25 CIFAR veri setindeki organize ve organize olmayan Bizans saldırılarındaki yöntemlerin F1, Doğruluk ve Keskinlik değerleri. (a), organize saldırılardaki F1 skor, (b) organize olmayan saldırılardaki F1 skor, (c), organize saldırılardaki Doğruluk, (d) organize olmayan saldırılardaki Doğruluk, (e), organize saldırılardaki Keskinlik, (f) organize olmayan saldırılardaki Keskinlik değerleri.

Şekil 25’de ARFED yöntemi, %10 ve %20 zararlı katılımcı oranlarında hem kesinlik hem doğruluk hem de F1 skoru açısından yüksek performans sergiliyor. Ancak %30 ve %40 oranlarında bu metriklerin tamamında ciddi bir düşüş görülüyor. Bu, ARFED’in bu oranlarda etkisiz hale geldiğini ve başarısız olduğunu gösteriyor.

ÇİFT SINIR yöntemi ise tüm oranlarda yüksek performans sergileyerek tutarlılığını koruyor. Kesinlik, doğruluk ve F1 skoru tüm zararlı katılımcı oranlarında yüksek kalıyor, bu da yöntemin hem organize hem de organize olmayan saldırılara karşı oldukça dayanıklı olduğunu gösteriyor.

QAD yöntemi, başlangıçta (%10 ve %20 oranlarında) biraz daha düşük kesinlik, doğruluk ve F1 skoru sunarken, %30 ve %40 oranlarında performansını büyük ölçüde koruyor. Bu da QAD’in daha dengeli bir yöntem olduğunu ve yüksek oranlarda bile tutarlı bir performans sergilediğini gösteriyor.

QMAD yöntemi, kesinlik, doğruluk ve F1 skoru açısından tüm zararlı katılımcı oranlarında sabit ve yüksek sonuçlar vermiş. Bu da yöntemin yüksek zararlı katılımcı oranlarına karşı bile oldukça dayanıklı olduğunu ortaya koyuyor. Yani QMAD, bu oranlarda herhangi bir düşüş yaşamadan performansını sürdürmüştür.

Bu durumda, özellikle QMAD ve ÇİFT SINIR yöntemlerinin her türlü zararlı katılımcı oranında tutarlı ve dayanıklı olduğunu, ARFED’in ise özellikle yüksek oranlarda ciddi performans kayıpları yaşadığını gözlemleyebiliriz.

4. SONUÇLAR

Bu Tez çalışması kapsamında FL’de Bizans ve Etiket Çevirme saldırıları sırasında kötü niyetli katılımcıların tespiti için ÇİFT SINIR, QAD ve QMAD olmak üzere üç farklı yöntem önerilmiştir. Önerilen yöntemler, literatürde mevcut olan ve yaygın olarak kullanılan yöntemlerle karşılaştırılarak kapsamlı bir şekilde değerlendirilmiştir. Mevcut yöntemlerin kötü niyetli katılımcı oranları düşük olduğunda başarılı sonuçlar verdiği, ancak oranlar arttığında başarı oranlarının önemli ölçüde düştüğü gözlemlenmiştir. Buna karşılık, önerilen üç yöntemin, yüksek kötü niyetli katılımcı oranlarında dahi daha üstün performans sergilediği tespit edilmiştir. Önerilen 3 yöntemde literatürdeki çoğu yöntemden farklı olarak FL’nin mantığına aykırı olan varsayımlarda bulunmamaktadır

Bu bölümde, tez çalışması kapsamında elde edilen sonuçlar özet biçimde sıralanacaktır.

Bizans saldırılarında hem organize hem de organize olmayan şekilde iki çeşit atak gerçekleştirilmiştir. Saldırıların etkileri CIFAR, Fashion MNIST ve MNIST veri setlerinde olmak üzere 3 farklı veri setinde ve MNIST veri seti için iki farklı mimaride incelenmiştir.

Bizans saldırılarında organize saldırıların test başarısını daha fazla etkilediği görülmüştür. Bizans saldırılarında QMAD yönteminin diğer yöntemlerden daha başarı elde ettiği ve yüksek oranlı kötü niyetli katılımcı oranlarına daha dayanıklı olduğu görülmüştür.

Etiket Çevirme saldırılarında da Bizans saldırılarında olduğu gibi saldırıların etkileri CIFAR, Fashion MNIST ve MNIST veri setlerinde olmak üzere 3 farklı veri setinde ve MNIST veri seti için iki farklı mimaride incelenmiştir.

Etiket Çevirme saldırılarının Bizans saldırılarına oranla test başarısını daha az etkilediği görülmüştür. Etiket Çevirme saldırılarında Çift sınır yönteminin başarısı Bizans saldırılarına oranla daha yüksek olmasına rağmen QMAD yönteminin Etiket Çevirme saldırılarında da dayanıklı olması her iki saldırıya da dayanıklı olması sebebi ile en etkili yöntem olduğu görülmüştür. Sonuç olarak, çalışmada önerilen üç yöntem, mevcut yöntemlere kıyasla FL ortamlarında kötü niyetli katılımcıların tespitinde üstün performans sergilemiştir. Özellikle QMAD yöntemi, yüksek oranlı kötü niyetli katılımcı durumlarında dahi performansını koruyarak, güvenilir bir çözüm sunmuştur. ÇİFT SINIR ve QAD yöntemleri de başarılı sonuçlar vermiş olup, FL uygulamaları için etkili çözümler olduklarını göstermişlerdir. QMAD yöntemi, kötü niyetli katılımcı oranlarına karşı en dayanıklı yöntem

olarak belirlenmiştir. Bu bağlamda, QMAD yöntemi, FL ortamlarında kötü niyetli katılımcıların etkilerini minimize etme potansiyeli ile öne çıkmaktadır.



5. ÖNERİLER VE TARTIŞMA

Bu test kapsamında üç farklı yöntem önerilmiştir. QAD yönteminde 50 ve 75 çeyrekler kullanılmıştır. Ancak ilerleyen çalışmalarda, yöntemlerin farklı çeyrek kombinasyonları (örneğin, 25 ile 50, 50 ile 25, 20 ile 75 gibi) denenerek toplamda 9 farklı varyasyonu oluşturulabilir. Varyasyonların performansları değerlendirilerek, QAD yönteminin daha da geliştirilmesi sağlanabilir.

QAD yönteminin varyasyonlarının denenmesi ile QAD yöntemi geliştirilecek ve QMAD yöntemi ile birleştirilerek çift sınır yöntemi üst bir seviyeye taşınacaktır. Yapılacak olan geliştirmeler ile hem düşük oranlı kötü niyetli katılımcılarda hem de yüksek oranlı kötü niyetli katılımcılarda en üst düzey başarı sağlayarak farklı tip saldırılarda da başarılı olmayı hedeflemekteyiz.

Aynı zamanda, bu test kapsamında homojen veriler üzerinde çalışma yapılmıştır. İlerleyen çalışmalarımızda heterojen veriler üzerinde de çalışmalar gerçekleştirilecektir. Bu, yöntemlerin daha geniş bir yelpazede etkinliğinin test edilmesine olanak tanıyacaktır.

Test kapsamında yapılan çalışmalarda, üç farklı yöntemle, literatüre QMAD yöntemi eklenmiş olup, özellikle model saldırıları ve veri zehirlenmesi saldırılarında başarılı bir yöntem kazandırıldığı düşünülmektedir. Bu başarı, ilerleyen araştırmalar ve farklı veri tipleri üzerinde yapılacak testlerle daha da pekiştirilebilir ve geliştirilebilir. Ayrıca, Çift sınır yönteminin geliştirilmesiyle farklı saldırı tiplerinde farklı yöntemlerle yapılacak karşılaştırmalar, yöntemin etkinliğini daha da artıracaktır.

6. KAYNAKLAR

- Akinshin, A. (2022). Quantile absolute deviation. *arXiv preprint arXiv:2208.13459*.
- Bagdasaryan, E. a. (2020). How to backdoor federated learning. *International conference on artificial intelligence and statistics* (s. 2938--2948). içinde PMLR.
- Blanchard, P. E. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems. Advances in Neural Information Processing Systems*.
- Bonawitz, K. a. (2019). Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 374--388.
- Cao, D. a. (2019). Understanding distributed poisoning attack in federated learning. *2019 IEEE 25th international conference on parallel and distributed systems (ICPADS)* (s. 233--239). içinde IEEE.
- Cao, X. a. (2020). Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*.
- Chen, H. a. (2023). FLRAM: Robust Aggregation Technique for Defense against Byzantine Poisoning Attacks in Federated Learning. *Electronics*.
- Chen, Y. a. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*.
- Douceur, J. R. (2002). The sybil attack. *International workshop on peer-to-peer systems* (s. 251--260). içinde Springer.
- Erdol, E. S. (2024). Low dimensional secure federated learning framework against poisoning attacks. *Elsevier*, 183--199.
- Fung, C. a. (2018). Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*.
- Gosselin, R. a. (2022). Privacy and security in federated learning: A survey. *Applied Sciences*.
- Isik-Polat, E. a. (2023). ARFED: Attack-Resistant Federated averaging based on outlier elimination. *Future Generation Computer Systems*, 626--650.
- Jebreel, N. M.-F.-J. (2022). Defending against the label-flipping attack in federated learning. *arXiv preprint arXiv:2207.01982*.
- Kairouz, P. a. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*.

- Konečný, J. M. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.
- Krizhevsky, A. a. (2009). Learning multiple layers of features from tiny images. *Toronto, ON, Canada*.
- LeCun, Y. (1998). The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.
- Li, L. a. (2019). RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. *Proceedings of the AAAI conference on artificial intelligence* (s. 1544--1551). içinde
- Li, S. a. (2020). Learning to detect malicious clients for robust federated learning. *arXiv preprint arXiv:2002.00211*.
- Lian, X. a.-J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*.
- Lin, J. a. (2019). Free-riders in federated learning: Attacks and defenses. *arXiv preprint arXiv:1911.12560*.
- Lyu, L. a. (2020). Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*.
- McMahan, B. a. (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics* (s. 1273--1282). içinde PMLR.
- Nguyen, T. D. (2022). {FLAME}: Taming backdoors in federated learning. *31st USENIX Security Symposium (USENIX Security 22)* (s. 1415--1432). içinde
- Paulik, M. S. (2021). Federated evaluation and tuning for on-device personalization: System design & applications. *arXiv*.
- Reddi, S. a. (2020). Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.
- Shafahi, A. a. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*.
- Shejwalkar, V. a. (2021). Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. *NDSS*. içinde
- Vanhaesebrouck, P. a. (2017). Decentralized collaborative learning of personalized models over networks. *Artificial Intelligence and Statistics* (s. 509--517). içinde PMLR.
- Xiao, H. a. (2017). H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017. *arXiv preprint arXiv:1708.07747*.

- Yang, Q. a. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1--19.
- Yin, D. a. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *International conference on machine learning* (s. 5650--5659). içinde Pmlr.
- Zhang, C. a. (2024). Anomaly detection and defense techniques in federated learning: a comprehensive review. *Artificial Intelligence Review*.



ÖZGEÇMİŞ

Fatma Zehra Solak, İlköğrenimini Cemile Erkunt İlköğretim Okulu'nda bitirdikten sonra lise eğitimini Atatürk Anadolu Lisesi'nde tamamladı. 2016-2020 öğretim yılında Konya Teknik Üniversitesi Bilgisayar Mühendisliği bölümünde lisans eğitimine başladı. 2020 yılında bu bölümden mezun oldu. Mezuniyet sonrası 2 sene özel sektörde Full Stack Developer olarak çalıştı. İş hayatının 2. Yılı olan 2021 yılında Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim dalında tezli yüksek lisans programına başladı. 10.04.2023 tarihinde Konya Teknik Üniversitesi Yazılım Mühendisliği Bölümü'nde Araştırma Görevlisi olarak göreve başladı. Konya Teknik Üniversitesi Yazılım Mühendisliği Bölümü'nde halen çalışmaya devam etmektedir. Yabancı dil olarak İngilizce ve Korece bilmektedir.

Hakemli konferans/sempozyumların bildiri kitaplarında yer alan yayınlar

Solak, F. Z., Üstübioğlu, B., Erdöl, E. S., Ulutaş, G., & Mühendisligi, Y. (2024, May). Federated Learning Resistant to Byzantine Attacks with Quantil-Based Statistical Propagation. In 2024 32nd Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE..