KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

SU ARITMA VE DAĞITMA SİSTEMLERİNE YÖNELİK OPTİMAL ÇEKİŞMELİ SALDIRI TASARIMI VE SAVUNMA MEKANİZMASI

YÜKSEK LİSANS TEZİ

Enis KARA

TEMMUZ 2024
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

SU ARITMA VE DAĞITMA SİSTEMLERİNE YÖNELİK OPTİMAL ÇEKİŞMELİ SALDIRI TASARIMI VE SAVUNMA MEKANİZMASI

Enis KARA

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

Tezin Enstitüye Verildiği Tarih : 31 / 07 / 2024
Tezin Savunma Tarihi : 03 / 08 / 2024

Tez Danışmanı : Doç. Dr. Mustafa Şinasi AYAS

Trabzon 2024

ÖNSÖZ

Bu yüksek lisans tezinin hazırlanmasında emeği geçen, destek ve bilgilerini esirgemeyen birçok değerli kişi bulunmaktadır. Öncelikle, tez konusunun belirlenmesi ve araştırma sürecinde beni yönlendiren, yapıcı eleştirileri ve destekleyici yaklaşımıyla her zaman bana yol gösteren danışmanım Doç. Dr. Mustafa Şinasi AYAS'a en içten teşekkürlerimi sunmak isterim. Kendisinin bilgi ve deneyimleri, bu çalışmanın şekillenmesinde büyük bir rol oynamıştır.

Ayrıca, çalışma sürecim boyunca değerli fikirleri ve önerileriyle her zaman destek olan tüm hocalarıma ve araştırma sürecimde yanımda olan arkadaşlarıma teşekkür ederim.

Süreç boyunca sabır ve anlayışla her zaman yanımda olan, bana daima destek veren aileme en içten teşekkürlerimi sunarım. Varlıkları ve verdikleri destek, çalışmamı tamamlamamda en büyük motivasyon kaynağı olmuştur.

Yüksek lisans tezimin hazırlanmasına katkılarından dolayı Türkiye Bilimsel ve Teknolojik Araştırma Kurumu'na (TÜBİTAK) teşekkür ederim. 122E383 numaralı proje kapsamında sağlanan destek, araştırmamın başarıyla tamamlanmasında önemli bir rol oynamıştır. Ayrıca, TÜBİTAK projesi süresince birlikte çalıştığım tüm ekibe de gönülden teşekkürlerimi sunarım.

Enis KARA

Trabzon 2024

TEZ ETİK BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum “Su Arıtma ve Dağıtma Sistemlerine Yönelik Optimal Çekişmeli Saldırı Tasarımı ve Savunma Mekanizması” başlıklı bu çalışmayı, danışmanım Doç. Dr. Mustafa Şinasi AYAS’ın rehberliğinde tamamladığımı beyan ederim. Bu çalışmada kullanılan verileri ve örnekleri kendim topladım, deneyleri ve analizleri ilgili laboratuvarlarda gerçekleştirdim. Başka kaynaklardan aldığım bilgileri metin içinde ve kaynakçada eksiksiz olarak gösterdim. Çalışma sürecinde bilimsel araştırma ve etik kurallara uygun davrandım. Aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim. 31/07/2024

Enis KARA

İÇİNDEKİLER

Yüksek Lisans Tezi

ÖZET

SU ARITMA VE DAĞITMA SİSTEMLERİNE YÖNELİK OPTİMAL ÇEKİŞMELİ SALDIRI TASARIMI VE SAVUNMA MEKANİZMASI

Enis KARA

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Mustafa Şinasi AYAS
2024, 101 Sayfa

Endüstriyel kontrol sistemlerinin güvenliğini artırmak amacıyla, su arıtma (SWaT) ve su dağıtma (WADI) sistemlerinde aykırılık tespiti üzerine odaklanılmıştır. Bu çalışmada, SWaT ve WADI sistemlerinde meydana gelebilecek aykırılıkları tespit etmek için CNN-LSTM modeli kullanılmaktadır. Modelin performansı kesinlik, duyarlılık, F1-skoru ve yanlış pozitif oranı gibi metriklerle değerlendirilmektedir.

Tezin ikinci aşamasında, SWaT ve WADI sistemlerine yönelik eyleyici ve sensör saldırıları gerçekleştirilmiş, bu aşamada tek nokta, süreç bazlı ve tüm süreç saldırıları ayrıntılı olarak incelenmiştir. Süreç bazlı değişken ve sabit genlikli sensör saldırılarının etkileri de detaylandırılmıştır. Kural denetçisinin performansı analiz edilerek sistem güvenliğini artırmaya yönelik önemli bulgular elde edilmiştir.

Çalışmanın özgün katkısı, optimize edilmiş çekişmeli makine öğrenmesi çerçevesi olan OptAML'nin geliştirilmesidir. Bu çerçeve, optimal pertürbasyonlar üreterek literatürde yüksek performans veren CNN-LSTM modelinin performansını düşürebileceğini ortaya koymaktadır. Çekişmeli eğitim (AT), savunma mekanizması olarak kullanılmıştır. AT, modelin performansını iyileştirmiş ve sistemin dayanıklılığını önemli ölçüde artırmıştır.

Anahtar Kelimeler: Endüstriyel kontrol sistemleri, Su arıtma sistemi, Su dağıtma sistemi, Makine öğrenimi, Derin öğrenme, CNN-LSTM, Çekişmeli makine öğrenimi, Çekişmeli saldırı, Çekişmeli eğitim

Master Thesis

SUMMARY

THE DESIGN OF OPTIMAL ADVERSARIAL ATTACKS AND DEFENSE MECHANISMS FOR WATER TREATMENT AND DISTRIBUTION SYSTEMS

Enis KARA

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Electrical-Electronics Engineering Graduate Program
Supervisor: Assoc. Prof. Mustafa Şinasi AYAS
2024, 101 Pages

The objective of this study is to enhance the security of industrial control systems by detecting anomalies in water treatment (SWaT) and water distribution (WADI) systems. To achieve this, a CNN-LSTM model is employed to identify potential anomalies in the SWaT and WADI systems. The performance of the model is evaluated using a set of metrics, including precision, recall, F1-score, and false positive rate.

In the second stage of the thesis, a series of experiments were conducted to examine the efficacy of various types of attacks on SWaT and WADI systems. These experiments included single-point, process-based, and full-process attacks, as well as the analysis of the effects of process-based variables and constant amplitude sensor attacks. Additionally, the performance of the rule checker was analyzed to identify potential avenues for enhancing system security.

This study makes a significant contribution to the field by developing OptAML, an optimized adversarial machine learning framework. This framework demonstrates that by generating optimal perturbations, it can degrade the performance of the CNN-LSTM model, which has previously demonstrated high performance in the literature. Adversarial training (AT) was used as a defense mechanism. AT improved the model's performance and significantly increased the system's robustness.

Key Words: Industrial control systems, Water treatment system, Water distribution system, Deep learning, CNN-LSTM, Adversarial machine learning, Adversarial attack, Adversarial training

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER VE KISALTMALAR DİZİNİ

AML : Çekişmeli makine öğrenmesi
APAE : Yaklaşık projeksiyon otoenkoderi
AT : Çekişmeli eğitim
AUC : Eğri altındaki alan
CDF : Kümülatif dağılım fonksiyonu
CI : Kritik altyapılar
CPS : Siber-fiziksel sistemler
CNN : Evrişimsel sinir ağı
CS : Kosinüs benzerliği
CR : Çaprazlama olasılığı
CUSUM : Kümülatif toplam
C&W : Carlini & Wagner
DADA : Derin çekişmeli alan uyarlama
DBSCAN : Yoğunluk tabanlı uzamsal kümelenme
DCTLN : Derin evrişimli transfer öğrenme ağı
DE : Diferansiyel evrim
DIF : Derin öğrenme tabanlı sızma tespiti çerçevesi
DL : Derin öğrenme
DSAE : Derin istiflenmiş autoencoder
E-GAN : Verimli üretken çekişmeli ağlar
FB : Özellik torbalama
FC : Tam bağlantılı katman
FGSM : Hızlı gradyan işaret yöntemi
FL : Dağıtık Öğrenme
FN : Yanlış negatif
FP : Yanlış pozitif
GAN : Üretken çekişmeli ağlar
GRU : Kapı kontrollü tekrarlayan birim
HCl : Hidroklorik asit
ICS : Endüstriyel kontrol sistemleri
IoT : Nesnelerin interneti

KNN : K-en yakın komşu algoritması
KS : Kolmogorov-smirnov
LSTM : Uzun-kısa süreli bellek
LSTM-ED : Uzun-kısa süreli bellek kodlayıcı-çözücü
MAD-GAN : Çoklu eşleşmeli ayrıklaştırıcı üretici ağlar
MAE : Ortalama mutlak hata
ML : Makine öğrenme
MRM : Ortalama sağlamlık ölçütü
MSSP : Çok aşamalı tek nokta
MSMP : Çok aşamalı çok nokta
MSE : Ortalama kare hata
NaCl : Sodyum klorür
NaOCl : Sodyum hipoklorit
$NaHSO_3$: Sodyum bisülfit
NCC : Normalleştirilmiş çapraz korelasyon
NLP : Doğal dil işleme
NM : NearMiss
ORP : Oksidasyon indirgeme potansiyeli
P1 : İlk aşama (Su girişi ve kontrol aşaması)
P2 : İkinci aşama (Ön arıtma aşaması)
P3 : Üçüncü aşama (İnce filtrasyon aşaması)
P4 : Dördüncü aşama (Klorsuzlaştırma aşaması)
P5 Beşinci aşama (Ters osmoz aşaması)
P6 : Altıncı aşama (Depolama ve dağıtma aşaması)
PLC : Programlanabilir mantık denetleyici
PCA : Temel bileşen analizi
PGD : Projeksiyonlu gradyan inişi
pH : Potansiyel hidrojen
ReLU : Düzleştirilmiş doğrusal birim
RNN : Tekrarlayan sinir ağı
SCADA : Veri denetleme ve toplama sistemi
SL : Denetimli Öğrenme
SSSP : Tek aşamalı tek nokta

SSMP : Tek aşamalı çok nokta
SVM : Destek vektör makineleri
SWaT : Su arıtma sistemi
TN : Doğru negatif
TO : Ters osmoz
TP : Doğru pozitif
UF : Ultra filtrasyon
UL : Denetimsiz öğrenme
YPO : Yanlış pozitif oranı
WADI : Su dağıtma sistemi

1. GENEL BİLGİLER

1.1. Giriş

Kritik altyapılar (CI) arasında yer alan su arıtma sistemleri (SWaT), su dağıtma sistemleri (WADI), akıllı şebekeler ve otonom araçlar, topluma temel hizmetlerin sağlanmasında hayati bir rol oynamaktadır [1]. Hesaplama algoritmaları ve fiziksel bileşenleri bir araya getiren siber-fiziksel sistemler (CPS) de bu tür önemli görevlerde kullanılmaktadır [2]. Nesnelerin İnterneti (IoT), günlük cihazların internete bağlanmasını sağlayan ve cihazlar arası iletişimi mümkün kılan bir teknolojidir. Bu teknoloji, cihazların ve sistemlerin birbirleriyle iletişim kurmasının yanı sıra uzaktan kontrol edilmelerini de sağlamaktadır.

Geçmişte endüstriyel kontrol sistemleri (ICS), genellikle tek bir tesiste ve küçük ölçekli kontrol işlemlerinde tercih edilmekteydi. IoT'nin ortaya çıkışıyla birlikte bu sistemler internete bağlanabilir hale gelmiş, böylece uzaktan izleme ve kontrol imkânı sunulmaya başlanmıştır. Bu teknolojinin yaygınlaşması, ülke genelinde dağıtılmış sistemlerin bilgi paylaşımını kolaylaştırmakta ve farklı konumlardaki eylemlerin kontrolünü mümkün kılmaktadır. Ancak, bu bağlantılar siber saldırı tehditlerini de beraberinde getirmektedir. Saldırganlar, iletişim kanallarına erişerek sistemin kontrolünü ele geçirip çeşitli noktalara saldırılar düzenleyebilirler. Bu saldırılar, hizmetin kullanılamaz hale gelmesinden sistem felaketlerine kadar farklı etkilere yol açabilmektedir [2], [3], [4].

Siber saldırılar, ülke ekonomisine, ulusal güvenliğe ve kamu düzenine ciddi zararlar vermektedir. Özellikle son yıllarda meydana gelen San Bruno boru patlaması, South Houston su arıtma tesisinin siber saldırıya uğraması [5], Alman çelik fabrikası siber saldırısı [5], [6] ve İran petrol terminalinin çevrimdışı olması gibi olaylar, CI'ları siber saldırılardan korumanın önemini vurgulamaktadır. Siber saldırıların tespit edilmesi konusundaki araştırmalar, dünya genelinde büyük önem kazanmaktadır [7].

SWaT ve WADI sistemleri, ICS'ler içinde büyük öneme sahiptir. Sağlıklı ve temiz içme suyu temini açısından kritik görevleri bulunan bu sistemler, su arıtma ve dağıtma süreçlerini kontrol etmek için birçok sensör ve eyleyici kullanmaktadır [8]. SWaT ve WADI sistemlerinin verimli çalışabilmesi için sensör ve eyleyicilerden elde edilen verilerin analizi büyük önem taşımaktadır. Bu sistemlerde meydana gelen aykırı durumlar, sensör veya eyleyicilerin arızalanması ya da siber saldırılar sonucunda ortaya çıkmaktadır. SWaT ve

WADI sistemlerinde oluşabilecek aykırılıklar, insan sağlığı ve diğer alanlarda ciddi zararlara yol açabilmektedir. Bu nedenle, bu tür aykırılıkları tespit edebilecek modellerin geliştirilmesi kritiktir [9]. Doğal afetler, altyapı arızaları ve siber saldırı gibi olaylar, insan sağlığının yanı sıra çeşitli sektörleri de olumsuz etkilemekte ve önemli ekonomik kayıplara yol açmaktadır. Bu nedenle, SWaT ve WADI sistemlerindeki olası aykırılıkların tespiti ve önlenmesi, büyük önem arz etmektedir.

Endüstriyel IoT alanında aykırılık tespiti amacıyla makine öğrenimi (ML) yöntemlerinin kullanımı, bu alandaki önemli araştırma konuları arasında yer almaktadır [10], [11], [12]. ML tabanlı yöntemler, verilerdeki aykırılıkları etkin bir şekilde analiz edebilme ve bu aykırılıkları tespit edebilme kapasitesine sahiptir. Günümüzde, literatürde ML tabanlı birçok aykırılık tespit yöntemi geliştirilmektedir. Bu yöntemler genellikle üç ana kategoriye ayrılmaktadır: ilişki kuralı öğrenimi tabanlı aykırılık tespiti [13], [14], kümeleme tabanlı aykırılık tespiti [15], [16] ve sınıflandırma tabanlı aykırılık tespiti [17], [18].

1.2. Literatür Araştırması

Aykırılık tespitlerinde, sistemin normal işleyişini hatasız bir şekilde kontrol edebilecek güvenilir bir modelin olması gerekmektedir. Son yıllarda, ML ve derin öğrenme (DL) modelleri, hem kestirim hem de aykırılık tespit modellerinin geliştirilmesi amacıyla başarılı bir şekilde kullanılmaktadır [15]. Goh ve arkadaşları [15], çalışmalarında SWaT sistemlerinde kullanılmak üzere tekrarlayan sinir ağı (RNN) ve kümülatif toplam (CUSUM) tabanlı bir aykırılık tespiti önermişlerdir. Bu çalışmada, kendi oluşturdukları SWaT veri setini kullanmışlardır. Erişime açık olan bu veri seti, RNN tabanlı aykırılık tespit modelinin, CUSUM yöntemi ile benzer şekilde aykırılıkları tespit edebildiğini göstermektedir. Önerilen RNN tabanlı model, düşük yanlış pozitif oranı (YPO) elde etmesine rağmen, SWaT test ortamının yalnızca bir aşamasında uygulanmış olması nedeniyle çalışmanın etkisini sınırlamaktadır.

Inoue ve arkadaşları [19], SWaT sistemlerindeki aykırılıkları tespit etmek amacıyla uzun–kısa süreli bellek (LSTM) tabanlı bir model önermişlerdir. Bu yöntemin performansını değerlendirmek için SWaT veri setini kullanmışlardır. Kesinlik, duyarlılık ve F1-skoru metrikleri hesaplanmış ve sonuçlar, tek sınıf destek vektör makinesi (SVM) yöntemiyle karşılaştırılmıştır. Deneysel sonuçlar, yüksek yanlış negatif (FN) oranının doğruluğu düşürdüğünü göstermektedir.

Kravchik ve Shabtai [20], SWaT sistemleri için tek boyutlu evrişimsel sinir ağı (CNN) kullanarak bir saldırı tespit yöntemi önermişlerdir. SWaT veri seti üzerinde yapılan hesaplamalar, F1-skoru ve doğruluk değerlerinin yüksek olduğunu, ancak duyarlılık değerinde iyileştirme yapılması gerektiğini göstermektedir. Bir başka çalışmada, SWaT sistemlerinde aykırılıkları tespit etmek amacıyla NearMiss (NM) yetersiz örnekleme tekniği ile değiştirilmiş DenseNet yaklaşımı önerilmiştir. Çeşitli değiştirilmiş DenseNet mimarileri, k-kat çapraz doğrulama tekniği kullanılarak değerlendirilmiş ve SWaT veri seti üzerinde test edilmiştir. Deneysel sonuçlar, önerilen DenseNet yapısının önceki çalışmalarla karşılaştırıldığında daha düşük YPO ve daha yüksek duyarlılık değeri sağladığını göstermektedir [21]. Ancak, kullanılan DenseNet aykırılık tespit modelinin, etiketlenmiş veriye ihtiyaç duyan bir danışmanlı ML modeli olması, veri etiketleme maliyetleri ve karmaşıklık açısından bir dezavantaj olarak değerlendirilebilir.

Birçok ML ve DL modeli, SWaT sistemlerine ek olarak WADI sistemlerinde de aykırılık tespiti için önerilmiştir. Li ve arkadaşları [22], WADI sistemlerinde aykırılık tespiti için üretken çekişmeli ağlar (GAN), verimli üretken çekişmeli ağlar (E-GAN), özellik torbalama (FB), k-en yakın komşu algoritması (KNN) ve temel bileşen analizi (PCA) gibi çeşitli çok değişkenli veri analizi için aykırılık tespit yöntemleri önermişlerdir. Çoklu eşleşmeli ayrıklaştırıcı üretici ağlar (MAD-GAN) yöntemi, WADI sistemlerinde aykırılıkları tespit edebilme yeteneği açısından diğer yöntemlere kıyasla daha üstündür. Bu üstünlük, WADI veri kümesi üzerinde elde edilen dikkate değer hassasiyet, duyarlılık ve F1-skoru değerleri ile gözlemlenmiştir. Ancak, bu metriklerin görece düşük değerleri, daha fazla iyileştirme yapılması gerektiğini göstermektedir.

Kravchik ve Shabtai [23], WADI sisteminin verilerini kullanarak, 1-D CNN otomatik kodlayıcı ve standart kodlayıcı kombinasyonu ile aykırılıkları tespit eden bir yöntem geliştirmişlerdir. Elde edilen hassasiyet, duyarlılık ve F1-skoru değerleri, bu yaklaşımın daha düşük hesaplama maliyetiyle yüksek performans gösterdiğini ortaya koymaktadır. Elnour ve arkadaşları [24], WADI sistemi için aynı metriklere odaklanarak, [22]'de sunulan alternatif yöntemlerle karşılaştırılan Derin Öğrenme Tabanlı Sızma Tespiti Çerçevesi (DIF) modelinin etkinliğini değerlendirmişlerdir. Sonuçlar, sızma tespit yeteneği ve yüksek boyutlu sistemlere uygulanabilirliği açısından DIF modelinin diğer yöntemlere göre daha üstün olduğunu göstermektedir.

Goodge ve arkadaşlarının çalışmasında [25], autoencoder'ların aykırılık tespiti sırasında çekişmeli saldırılara karşı dayanıklılığı incelenmiştir. Çekişmeli saldırılar, modelin performansını bozmak amacıyla verilerde küçük değişiklikler yapılmasını içermektedir. Çalışmada, siber-fiziksel sistem verileri kullanılarak autoencoder'ların bu saldırılara karşı savunmasız olduğu gösterilmiştir. Bu soruna çözüm olarak Yaklaşık Projeksiyon Otoenkoderi (APAE) modeli geliştirilmiştir. Bu model, hem performansı artırmak hem de çekişmeli saldırılara karşı daha dayanıklı hale gelmek için iki savunma mekanizması içermektedir. Deneyler sonucunda APAE, saldırılara karşı %9, saldırı olmadan ise %8 oranında performans artışı sağlamıştır.

Liu ve arkadaşlarının çalışmasında [26], ICS'lerde çekişmeli örnek saldırılarının etkinliği ve bu saldırılara karşı geliştirilen savunma mekanizmaları araştırılmıştır. Çalışmada, LSTM Kodlayıcı-Çözücü (LSTM-ED) algoritmasına dayalı bir saldırı ve savunma yöntemi önerilmiştir. Bu saldırı yöntemi, protokol gereksinimlerine uygun olarak sensör ve aktüatör verilerine ufak bozulmalar ekleyerek çekişmeli örnekler oluşturmayı hedeflerken, LSTM-FWED adında bir savunma modeli geliştirilmiştir. LSTM-FWED, çekişmeli örnek bilgisi olmadan savunma sağlayarak modelin dayanıklılığını artırmaktadır. Yapılan deneyler, bu savunma yöntemi sayesinde çekişmeli saldırılar altında modelin %21,83 oranında Eğri Altındaki Alan (AUC) iyileşmesi gösterdiğini ortaya koymuştur.

Abdelaty ve arkadaşlarının çalışmasında [27], ICS'ler için ML tabanlı aykırılık tespit sistemlerinin, çekişmeli ve çekişmesiz veri bozulmalarına karşı dayanıklılığı incelenmiştir. Çalışmada, ICS'lerde kullanılan ML tabanlı aykırılık tespit sistemlerinin, doğal gürültü ve cihaz yaşlanması gibi çekişmesiz bozulmalara karşı hassas olduğu vurgulanmış ve bu sorunları çözmek amacıyla DAICS adlı geniş ve derin öğrenme tabanlı bir model önerilmiştir. DAICS, değişen normal davranışlara hızlı adaptasyon sağlamakta ve operatörden gelen geri bildirimler doğrultusunda yanlış alarmları azaltmaktadır. Ayrıca, SiFA çerçevesi ile yanlış alarmların %80 oranında azaltıldığı gösterilmiştir. Çekişmeli veri bozulmalarına karşı ise GADoT adlı bir savunma mekanizması geliştirilmiştir. Bu yöntem, çekişmeli örneklerle eğitimi artırarak aykırılık tespit modellerinin DDoS saldırılarına karşı dayanıklılığını artırmıştır. GADoT, GAN kullanarak zararlı trafik örneklerini çekişmeli bir şekilde değiştirip modeli eğitmiştir.

Jia ve arkadaşlarının çalışmasında [28], CI'larda yer alan CPS'lerde kullanılan aykırılık tespit yöntemlerine karşı çekişmeli saldırıların etkileri araştırılmıştır. Çalışmada, sensör ve eyleyici verilerine ufak bozulmalar eklenerek hem aykırılık tespit sistemlerini hem

de kural denetleyicilerini yanıltmayı amaçlayan bir çekişmeli saldırı modeli önerilmiştir. Özellikle, gradyan tabanlı yöntemler kullanılarak sensör verilerinde gürültü oluşturulmuş ve ardından genetik algoritmalar kullanılarak bu bozulmalar optimize edilmiştir. Araştırma kapsamında gerçekleştirilen deneylerde, önerilen saldırıların test yataklarında kullanılan aykırılık tespit sistemlerinin doğruluğunu %50'den fazla düşürdüğü gözlemlenmiştir. Ayrıca, çekişmeli örnekler kullanılarak tespit sistemlerinin eğitilmesi yoluyla saldırıların hafifletilmesi incelenmiş, ancak düşük miktarda gürültü içeren saldırıların tespit edilmesinin zor olduğu belirlenmiştir.

Yuan ve arkadaşlarının çalışmasında [29], ICS'lerde saldırı tespiti performansını artırmak amacıyla veri dengesizliğini çözmeye yönelik bir yaklaşım geliştirilmiştir. Çalışmada, GAN ve LSTM ağları kullanılarak B-GAN adı verilen bir veri dengeleme yöntemi önerilmiştir. Bu yöntem, küçük sınıf saldırı örneklerinin üretilmesi yoluyla veri dengesizliğini gidermekte ve böylece aykırılık tespit modellerinin saldırıları daha doğru bir şekilde öğrenmesini sağlamaktadır. Yapılan deneyler, B-GAN ile dengelenmiş veri kümeleri üzerinde eğitilen modellerin performansında önemli bir iyileşme olduğunu ortaya koymuştur. Bu yöntem, özellikle SWaT ve WADI veri kümelerinde test edilerek %5'e kadar F1 skoru artışı sağlamıştır.

Zhang ve arkadaşlarının çalışmasında [30], bilinmeyen veya değişken şifreleme fidye yazılımı saldırılarını tespit etmek amacıyla çift GAN tabanlı TGAN-IDS adlı bir saldırı tespit çerçevesi geliştirilmiştir. Bu sistem, şifrelenmiş trafik içerisindeki aykırılıkları tespit etmek için DL modelleri olan DCGAN ve TGAN kullanılarak eğitilmiştir. TGAN, normal trafiği modelleyerek fide yazılımı saldırılarını yüksek doğruluk oranlarıyla tespit etmeyi amaçlamaktadır. Çalışmanın bulgularına göre, TGAN-IDS sistemi, test edilen diğer derin öğrenme ağlarına (AlexNet, ResNet, DenseNet) kıyasla daha yüksek doğruluk ve F1 puanı elde etmiştir. Özellikle KDD99, SWaT ve WADI veri kümeleri üzerinde yapılan deneyler, bu modelin yalnızca şifrelenmiş saldırıları değil, aynı zamanda bilinmeyen saldırıları da etkili bir şekilde tespit edebildiğini göstermiştir.

Haili Sun ve arkadaşlarının çalışmasında [31], CPS'lerde aykırılık tespiti amacıyla MTS-DVGAN adlı çift varyasyonel GAN modeli geliştirilmiştir. Bu model, çok değişkenli zaman serileri üzerindeki aykırılıkları tespit etmek için tasarlanmıştır. Temel yenilik olarak, aykırı örneklerin normal verilerden daha iyi ayrılmasını sağlamak amacıyla karşıt öğrenme ile modele eklenen kontrastif sınırlamalar kullanılmaktadır. MTS-DVGAN, özellikle ICS'lerde kullanılan SWaT, WADI ve NSL-KDD veri kümeleri üzerinde test edilmiş ve

mevcut yöntemlere kıyasla daha kararlı ve yüksek performans sergilemiştir. Yapılan deneyler sonucunda, modelin %2,2 oranında performans artışı sağladığı gözlemlenmiştir.

Zeng ve arkadaşlarının çalışmasında [32], ICS'lerde çekişmeli saldırılara karşı savunma geliştirmek amacıyla otomatik bir dağıtık öğrenme (FL) tabanlı saldırı ve savunma yöntemi önerilmiştir. AFL-GAS adlı bu yöntem, transfer saldırısı ilkesi üzerine kurulmuş olup çekişmeli örnekler oluştururken modelin zayıf yönlerini tam olarak ortaya çıkarmayı hedeflemektedir. Ayrıca, MoNAS-IDSAA adında, hem normal verilerdeki sınıflandırma performansını hem de çekişmeli sağlamlığı aynı anda optimize eden çok amaçlı bir sinir ağı arama tabanlı saldırı tespit modeli geliştirilmiştir. Yapılan deneyler, önerilen yöntemlerin mevcut sistemlere kıyasla daha yüksek kaçış oranları ve dayanıklılık sağladığını göstermiştir. AFL-GAS, hafif ve optimize edilmiş modeller sunarken, MoNAS-IDSAA yöntemi çekişmeli saldırılara karşı önemli bir sağlamlık artışı sağlamıştır.

Zhao ve arkadaşlarının çalışmasında [33], rulman arızası teşhisi için sinyalden-görüntüye haritalama ve DL tabanlı bir CNN modeli geliştirilmiştir. Bu modelde, bir boyutlu titreşim sinyalleri iki boyutlu gri tonlamalı görüntülere dönüştürülerek işlenmiş, böylece uzman bilgisine dayalı özellik çıkarımına olan bağımlılık azaltılmıştır. Titreşim sinyallerinden elde edilen görüntüler CNN modeli ile işlenmiş ve arıza sınıflandırması yüksek doğrulukla gerçekleştirilmiştir. Çalışmada, sinyal işleme ve veri artırımı teknikleri kullanılarak orijinal veri miktarı artırılmış ve modelin etkinliği iki farklı rulman veri seti üzerinde test edilmiştir. Sonuçlar, önerilen yöntemin geleneksel sinyal işleme yöntemlerine kıyasla daha iyi sınıflandırma performansı sunduğunu ortaya koymuştur.

Madry ve arkadaşlarının (2018) çalışmasında [34], DL modellerinin çekişmeli saldırılara karşı dayanıklılığını artırmaya yönelik bir yaklaşım geliştirilmiştir. Çalışmada, sinir ağlarının, insan tarafından fark edilemeyen küçük değişikliklerle bile kolayca yanıltılabildiği gösterilmiştir. Bu sorunu çözmek amacıyla, yazarlar sağlam optimizasyon çerçevesini kullanarak modellerin güvenliğini artırmak için yeni bir eğitim ve saldırı yöntemi geliştirmiştir. Çalışma, Yansıtılmış Gradyan İnişi (PGD) saldırısına karşı dayanıklı ağlar oluşturmayı hedeflemiş ve MNIST ile CIFAR-10 veri kümeleri üzerinde yapılan deneylerle yüksek direnç sağlanmıştır. Yazarlar, sinir ağlarının çekişmeli saldırılara karşı %89'a kadar doğruluk sağlayabildiğini belirtmiştir.

Wen ve arkadaşlarının çalışmasında [35], ICS'lerde hata teşhisi için veri tabanlı bir CNN modeli geliştirilmiştir. Bu model, özellikle LeNet-5 mimarisine dayanan ve ham titreşim sinyallerini iki boyutlu gri tonlamalı görüntülere dönüştüren yeni bir yöntem

önermektedir. Böylece geleneksel yöntemlerin aksine, uzmanlar tarafından el ile çıkarılan özelliklere ihtiyaç duymadan sinyallerin otomatik olarak işlenmesi sağlanmaktadır. Çalışma, motor rulman, kendinden emişli santrifüj pompa ve eksenel pistonlu hidrolik pompa veri kümeleri üzerinde test edilmiştir. Model, sırasıyla %99,79, %99,481 ve %100 doğruluk oranlarına ulaşarak diğer DL ve geleneksel yöntemlere kıyasla üstün performans sergilemiştir.

Guo ve arkadaşlarının çalışmasında [36], etiketlenmemiş verilerle makine arıza teşhisinde kullanılmak üzere Derin Evrişimli Transfer Öğrenme Ağı (DCTLN) adlı yeni bir yöntem geliştirilmiştir. Bu yöntem, iki modülden oluşmaktadır: Durum Tanıma Modülü ve Alan Uyarlama Modülü. Durum Tanıma Modülü, bir boyutlu bir CNN kullanarak makine sağlığıyla ilgili özellikleri otomatik olarak öğrenmekte ve bu özellikler üzerinden makinenin sağlık durumunu tanımlamaktadır. Alan Uyarlama Modülü ise, modelin farklı makinelerden elde edilen verilerle uyumlu olabilmesi için, bu özellikleri alanlar arası tutarlı hale getirmektedir. Çalışma, etiketlenmiş verilerle eğitilen modelin, etiketlenmemiş verilere sahip farklı makinelerdeki hataları yüksek doğrulukla tanımlayabildiğini göstermiştir. Üç rulman veri seti üzerinde yapılan deneyler, önerilen yöntemle hata tanıma doğruluğunun %32,1 oranında arttığını ortaya koymuştur.

Liu ve arkadaşlarının çalışmasında [37], rulman arızası teşhisinde kullanılan veri uyumsuzluğu sorununu çözmek amacıyla Derin Çekişmeli Alan Uyarlama (DADA) modeli geliştirilmiştir. Bu model, eğitim ve test verileri arasındaki dağılım farklılıklarını azaltmak için Derin İstiflenmiş Autoencoder (DSAE) ve çekişmeli uyarlama ağlarını birleştirmiştir. Model, etiketlenmemiş veriler üzerinde etkili bir şekilde arıza özelliklerini öğrenmek amacıyla denetimsiz öğrenme yöntemini kullanmakta ve Softmax sınıflayıcı ile sınıflandırma doğruluğunu artırmaktadır. Yapılan deneyler, modelin mevcut ML ve DL yöntemlerine kıyasla daha yüksek sınıflandırma doğruluğu ve genelleme kabiliyeti sağladığını göstermiştir. Bu modelin, rulman arızalarını gerçek dünya uygulamalarında daha etkin bir şekilde tespit etmek için uygun olduğu belirtilmiştir.

Szegedy ve arkadaşlarının çalışmasında [38], derin sinir ağlarının beklenmedik ve sezgilere aykırı iki önemli özelliği ortaya konulmuştur. İlk olarak, ağın yüksek seviyeli katmanlarındaki bireysel birimlerin tek başına anlamsal bilgi taşımadığı, bunun yerine tüm aktivasyon uzayının bu bilgiyi barındırdığı gösterilmiştir. İkinci olarak, sinir ağlarının giriş verilerine yapılan çok küçük, insan algısı tarafından fark edilemeyen bozucu değişikliklerle doğru sınıflandırmanın tamamen yanlış hale getirilebildiği tespit edilmiştir. Bu tür bozucu

değişikliklerle oluşturulan örneklere "çekişmeli örnekler" denmiştir. Çalışmanın en dikkat çekici bulgularından biri, çekişmeli örneklerin farklı veri kümeleri veya farklı eğitim koşulları altında eğitilmiş sinir ağları üzerinde de aynı hataları tetikleyebilmesidir. Bu durum, sinir ağlarının bu tür örneklere karşı genel bir zayıflık gösterdiğini ortaya koymaktadır.

Goodfellow ve arkadaşlarının çalışmasında [39], DL modellerinin çekişmeli örnekler karşısındaki savunmasızlığı analiz edilmiş ve bu durumun temel nedeninin modellerin doğrusal doğası olduğu ileri sürülmüştür. Çekişmeli örnekler, bir modelin doğru sınıflandırdığı girdilere çok küçük ama kasıtlı bozulmalar uygulanarak oluşturulur ve bu bozulmalar modelin yanlış sonuçlar üretmesine neden olur. Çalışmada, çekişmeli örneklerin doğrusal davranıştan kaynaklandığı açıklanmış ve bu tür örneklerin hızlı bir şekilde üretilmesine olanak sağlayan Hızlı Gradyan İşareti Yöntemi (FGSM) geliştirilmiştir. Bu yöntem, sinir ağlarının çekişmeli örnekler karşısındaki direncini artırmak için çekişmeli eğitim stratejisini de tanıtmaktadır. Deneyler, çekişmeli eğitimle eğitilen modellerin hem daha dayanıklı hem de daha düzenli hale geldiğini göstermiştir.

Moosavi-Dezfooli ve arkadaşlarının çalışmasında [40], derin sinir ağlarının çekişmeli örnekler karşısındaki hassasiyeti araştırılmış ve bu örnekleri oluşturmak için DeepFool adlı bir algoritma geliştirilmiştir. Çalışmada, mevcut DL modellerinin, küçük ve algılanamaz bozulmalar karşısında nasıl kolayca yanıltılabileceği gösterilmiştir. DeepFool algoritması, bu tür bozulmaları optimize ederek sınıflandırıcıları yanıltan minimum miktarda değişiklikle çekişmeli örnekler üretmektedir. Deneyler, DeepFool'un, diğer mevcut yöntemlere kıyasla daha küçük ve doğru çekişmeli bozulmalar üretebildiğini göstermiştir. Ayrıca, çekişmeli eğitim yoluyla derin ağların dayanıklılığının artırılabileceği de ortaya konmuştur. DeepFool, DL modellerinin güvenlik ve dayanıklılık analizleri için etkili bir araç olarak öne çıkmaktadır.

Kurakin ve arkadaşlarının çalışmasında [41], çekişmeli örneklerin büyük ölçekli ML modellerine nasıl etki ettiğini araştırmış ve ImageNet gibi büyük veri kümeleri üzerinde çekişmeli eğitim kullanarak modellerin çekişmeli saldırılara karşı dayanıklılığını artırmaya yönelik çözümler önerilmiştir. Çalışmada, tek adımlı saldırı yöntemlerine karşı yüksek dayanıklılık sağlandığı, ancak çok adımlı saldırıların transfer edilebilirliğinin daha düşük olduğu gözlemlenmiştir. Ayrıca, etiket sızdırma etkisi ile çekişmeli eğitimin bazı durumlarda çekişmeli örneklerde, temiz örneklerden daha iyi performans göstermesine neden olduğu bulunmuştur. Bu çalışma, çekişmeli eğitimin büyük ölçekli veri kümeleri ve

modeller üzerinde nasıl uygulanabileceğine dair öneriler sunarken, aynı zamanda model kapasitesinin artırılmasının çekişmeli örneklere karşı dayanıklılığı artırabileceğini de göstermektedir.

Suciu ve arkadaşlarının çalışmasında [42], ML sistemlerine yönelik kaçınma ve zehirleme saldırılarının etkileri ile bu saldırıların transfer edilebilirliği incelenmiştir. Çalışmada, farklı düzeyde bilgi ve kontrol sahibi saldırganlar için FAIL saldırgan modeli tanıtılmıştır. Bu model, saldırganın bilgi ve kontrol seviyesini dört boyutta (özellikler, algoritmalar, örnekler ve kaldıraç) tanımlayarak, ML sistemlerine karşı daha zayıf saldırganların bile başarılı olabileceği durumları araştırmayı amaçlamaktadır. Araştırma kapsamında geliştirilen StingRay adlı hedeflenmiş zehirleme saldırısı, dört farklı ML uygulamasında test edilmiş ve mevcut savunma mekanizmalarını bypass edebilmiştir. Çalışma, zehirleme ve kaçınma saldırılarının daha gerçekçi bir şekilde modellenmesi ve bu saldırılara karşı savunma mekanizmalarının geliştirilmesi için yeni yollar önermektedir.

Kuppa ve arkadaşlarının çalışmasında [43], derin aykırılık tespit sistemlerine karşı kara kutu saldırıları incelenmiştir. Çalışmada, aykırılık tespiti sırasında kullanılan DL modellerinin, küçük çekişmeli örneklerle kolayca yanıltılabileceği gösterilmiştir. Özellikle kara kutu saldırılarında, modelin iç işleyişine dair bilgi sahibi olmadan, yalnızca modelle sınırlı sayıda etkileşim kurarak çekişmeli örnekler üretmek amaçlanmıştır. Bu çalışmada, sınırlı sorgu sayısıyla çalışan bir saldırı modeli geliştirilmiştir. Araştırma, farklı derin aykırılık tespit sistemleri üzerinde test edilmiş ve çekişmeli örneklerle bu sistemlerin doğruluğunun önemli ölçüde düşürülebildiği gözlemlenmiştir. Çalışmada, DAGMM, AutoEncoder, AnoGAN ve ALAD gibi modeller üzerinde yapılan deneylerde, saldırıların etkinliği doğrulanmıştır.

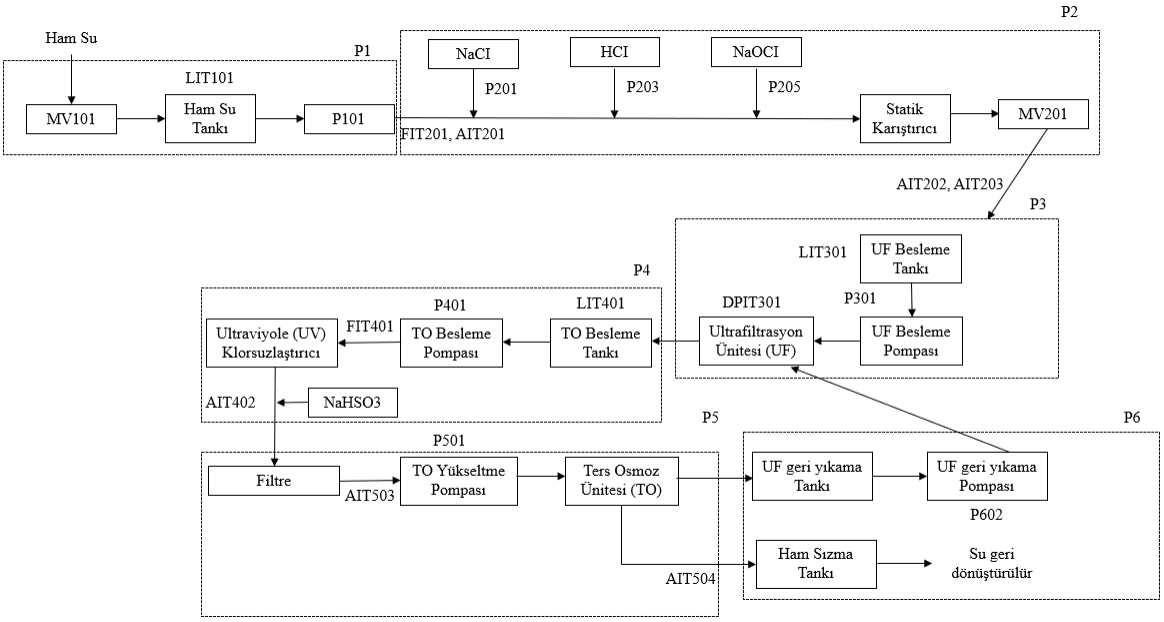
Xu ve arkadaşlarının çalışmasında [44], ML tabanlı kötü amaçlı yazılım tespit sistemlerini atlatmak amacıyla grafik yapısına dayalı bir çekişmeli örnek oluşturma çerçevesi olan MANIS geliştirilmiştir. Çalışmada, grafik tabanlı kötü amaçlı yazılım tespiti sistemlerinin, grafik verilerinin manipüle edilmesi yoluyla nasıl yanıltılabileceği incelenmiştir. MANIS, iki ana saldırı yöntemini kullanmaktadır: n-en güçlü düğüm ve gradyan işareti yöntemi. Bu yöntemler, grafik tabanlı kötü amaçlı yazılım tespit sistemlerinin yanlış sınıflandırmalar yapmasına neden olmaktadır. Deneylerde, Drebin kötü amaçlı yazılım veri kümesi kullanılarak yapılan testlerde, n-en güçlü düğüm yönteminin %72,2'lik yanlış sınıflandırma oranı sağladığı gözlemlenmiştir. Gradyan işareti yöntemi ise %33,4'lük

bir yanlış sınıflandırma oranı ile sonuçlanmıştır. Bu sonuçlar, çekişmeli örneklerin grafik yapıları üzerinde etkili olduğunu göstermektedir.

1.3. Su Arıtma Sistemi (SWaT)

SWaT sistemi, saatte 5 Amerikan galonu filtrelenmiş su üreten bir su arıtma tesisinin işletme test ortamıdır. Modern şehir su arıtma tesislerinin küçük ölçekli versiyonunu temsil eden bu ortam, Singapur Kamu Hizmetleri Kurumu ile iş birliği içinde tasarlanmıştır. Bu iş birliği sayesinde, test ortamının fiziksel süreçleri ve kontrol sistemleri gerçek dünya sistemlerine yakın olacak şekilde oluşturulmuş, elde edilen sonuçların gerçek sistemlere uygulanabilirliği sağlanmaktadır. SWaT sistemi, siber saldırılar ile sistem tepkileri üzerine araştırmalar yapmak ve yeni karşı önlem tasarımlarını test etmek amacıyla kullanılmaktadır.

Şekil 1'de gösterildiği üzere, SWaT sistemi altı aşamadan oluşmaktadır. Her bir alt süreçte sensörler ile eyleyiciler yer almaktadır. Sensörlerden elde edilen veriler, programlanabilir mantık denetleyicilere (PLC) gönderilmektedir. Bu veriler doğrultusunda eyleyicilerin hareketleri kontrol edilmektedir [45].



Şekil 1. Su arıtma sistemi süreçleri

Sistemin altı aşamadan oluşan yapısı şu şekildedir: Su girişi ve kontrol aşamasında (P1), motorlu bir vananın açılıp kapanmasıyla arıtılacak suyun akışı kontrol edilmektedir. Ardından su, ön arıtma aşamasına (P2) geçmekte ve bu süreçte suyun kalitesi

değerlendirilmektedir. Su kalitesi kabul edilebilir sınırlar içinde değilse kimyasal dozlama yapılmaktadır. İnce filtrasyon aşamasında (P3), istenmeyen maddeler ince filtrasyon membranları aracılığıyla giderilmektedir. Ultra Filtrasyon (UF) sistemi ile filtrelendikten sonra, kalan klor klorsuzlaştırma aşamasında (P4) ultraviyole ışın kullanılarak giderilmektedir. P4'ten gelen su, inorganik safsızlıkları azaltmak için ters osmoz (TO) aşamasına (P5) pompalanmaktadır. Son aşama olan depolama ve dağıtma sürecinde (P6), TO'dan gelen su depolanmakta ve WADI sisteminde dağıtıma hazır hale getirilmektedir. SWaT sisteminde ise arıtılmış su, yeniden işlenmek üzere ham tanka geri aktarılmaktadır. Ancak veri toplama amacıyla, P6'dan gelen su, çıkışa yönlendirilerek su dağıtımı taklit edilmektedir.

Her alt süreç, bir yıldız ağı aracılığıyla denetleyici ve veri toplama sistemi (SCADA) ile iletişim kuran kendi PLC'lerine sahiptir. Sensörler ve eyleyiciler, verilerini ilgili PLC'ye ileten bir halka ağı üzerinden haberleşmektedir. Bu ağlardaki iletişimler kablolu veya kablosuz sağlanmaktadır. SWaT sistemi, farklı sayıda sensör ve eyleyici içeren her bir alt süreçte çeşitli potansiyel saldırı konumları barındırmaktadır [10]. SWaT sisteminde kullanılan sensör ve eyleyicilerin isimleri ile işlevleri Tablo 1'de gösterilmektedir.

Tablo 1. Su arıtma sisteminde kullanılan sensör ve eyleyiciler

Numara	İsim	Tür	Tanım
1	FIT-101	Sensör	Akış ölçer; Ham su tankına girişi ölçer.
2	LIT-101	Sensör	Seviye ölçer; Ham su tank seviyesi.
3	MV-101	Eyleyici	Motorlu vana; Ham su tankına giden su akışını kontrol eder.
4	P-101	Eyleyici	Pompa; Ham su tankındaki suyu ikinci kademeye pompalar.
5	P-102 (yedek)	Eyleyici	Pompa; Ham su tankındaki suyu ikinci kademeye pompalar.
6	AIT-201	Sensör	İletkenlik analizörü; NaCl seviyesini ölçer.
7	AIT-202	Sensör	pH analizörü; HCI seviyesini ölçer.
8	AIT-203	Sensör	ORP analizörü; NaOCl seviyesini ölçer.
9	FIT-201	Sensör	Akış ölçer; Dozaj pompalarını kontrol eder.
10	MV-201	Eyleyici	Motorlu vana; UF besleme suyu tankına su akışını kontrol eder.
11	P-201	Eyleyici	Dozaj pompası; NaCl dozaj pompası.
12	P-202 (yedek)	Eyleyici	Dozaj pompası; NaCl dozaj pompası.
13	P-203	Eyleyici	Dozaj pompası; HCl dozaj pompası.
14	P-204 (yedek)	Eyleyici	Dozaj pompası; HCl dozaj pompası.
15	P-205	Eyleyici	Dozaj pompası; NaOCl dozaj pompası.
16	P-206 (yedek)	Eyleyici	Dozaj pompası; NaOCl dozaj pompası.

Tablo 1'in devamı

17	DPIT-301	Sensör	Diferansiyel basıncı gösteren verici; Geri yıkama işlemini kontrol eder.
18	FIT-301	Sensör	Akış ölçer; UF aşamasında su akışını ölçer.
19	LIT-301	Sensör	Seviye ölçer; UF besleme suyu tankı seviyesi.
20	MV-301	Eyleyici	Motorlu vana; UF-Geri yıkama işlemini kontrol eder.
21	MV-302	Eyleyici	Motorlu vana; UF prosesinden Klorsuzlaştırma ünitesine giden suyu kontrol eder.
22	MV-303	Eyleyici	Motorlu vana; UF-Geri yıkama tahliyesini kontrol eder.
23	MV-304	Eyleyici	Motorlu vana; UF tahliyesini kontrol eder.
24	P-301 (yedek)	Eyleyici	UF besleme pompası; Suyu UF filtrasyon sayesinde UF besleme suyu tankından TO besleme suyu tankına pompalar
25	P-302	Eyleyici	UF besleme pompası; Suyu UF filtrasyon sayesinde UF besleme suyu tankından TO besleme suyu tankına pompalar
26	AIT-401	Sensör	Suyun TO sertlik ölçümü
27	AIT-402	Sensör	ORP analizörü; NaHSO3 dozajını (P203), NaOCl dozajını(P205) kontrol eder.
28	FIT-401	Sensör	Akış ölçer; UV klor gidericiyi kontrol eder.
29	LIT-401	Sensör	Seviye ölçer; RO besleme suyu tankı seviyesi.
30	P-401 (yedek)	Eyleyici	Pompa; RO besleme tankından UV klor gidericiye su pompalar.
31	P-402	Eyleyici	Pompa; RO besleme tankından UV klor gidericiye su pompalar.
32	P-403	Eyleyici	Sodyum bi-sülfat pompası
33	P-404 (yedek)	Eyleyici	Sodyum bi-sülfat pompası
34	UV-401	Eyleyici	Klor giderici; Kloru sudan giderir.
35	AIT-501	Sensör	TO pH analizi; HCI seviyesini ölçer.
36	AIT-502	Sensör	TO besleme ORP analizörü; NaOCl seviyesini ölçer.
37	AIT-503	Sensör	TO besleme iletkenlik analizörü; NaCl seviyesini ölçer.
38	AIT-504	Sensör	TO sızma iletkenlik analizi; NaCl seviyesini ölçer.
39	FIT-501	Sensör	Akış ölçer; TO membran giriş akış ölçer
40	FIT-502	Sensör	Akış ölçer; TO sızma akış ölçer
41	FIT-503	Sensör	Akış ölçer; TO reddetme akış ölçer
42	FIT-504	Sensör	Akış ölçer; TO devridaim akış ölçer
43	P-501	Eyleyici	Pompa; Klorsuzlaştırılmış suyu TO'ya pompalar
44	P-502 (yedek)	Eyleyici	Pompa; Klorsuzlaştırılmış suyu TO'ya pompalar.
45	PIT-501	Sensör	Basınç ölçer; TO besleme basıncı.
46	PIT-502	Sensör	Basınç ölçer; TO sızma basıncı.
47	PIT-503	Sensör	Basınç ölçer; TO reddetme basıncı.
48	FIT-601	Sensör	Akış ölçer; UF Geri yıkama akış ölçer.
49	P-601	Eyleyici	Pompa; TO sızma tankından ham su tankına su pompalar (veri toplama için kullanılmaz).
50	P-602	Eyleyici	Pompa; Membranı temizlemek için UF geri yıkama tankından UF filtresine su pompalar.
51	P-603	Eyleyici	Henüz SWaT'ta uygulanmamıştır.

SWaT sistemi, bilimsel araştırmalar için araştırmacılara uygun bir çalışma ortamı sunmaktadır. SWaT sistemine yönelik her saldırının amacı ya tankın taşmasına neden olmak ya da su arıtma performansını düşürmektir. Bu doğrultuda, saldırganlar tanktaki su seviyesini, motorlu vanaların ve pompaların durumunu manipüle etmektedir. Araştırmacılar, kendi senaryolarında sistemin alt ve üst eşik değerlerini belirleme imkanına sahiplerdir. Umer ve arkadaşları [27], çalışmalarında sensörler ve eyleyiciler için Tablo 2'de belirtilen alt ve üst eşik değerlerini kullanmışlardır.

Tablo 2. Su arıtma sistemindeki sensör ve eyleyicilerin alt ve üst eşik değerleri [46]

	Parametre	Alt ve üst eşik değerleri
AIT	pH	P_h=7.05, P_l=6.95
AIT	İletkenlik	C_h=260, C_l=250 µS/cm
AIT	Oksidasyon indirgeme potansiyeli	O_h=480, O_l=440 mV
FITxxx	Akış ölçer	F_h=2, F_l=0.1 m^3/h
LIT101	Tank101 su seviyesi	HH=1000, H=800, L=500, LL=250
LIT301	Tank301 su seviyesi	HH=1200, H=1000, L=800, LL=250
LIT401	Tank401 su seviyesi	HH=1200, H=1000, L=800, LL=250
MVxxx	Durum geçiş zamanı	8s
Pxxx	Durum geçiş zamanı	2s

1.3.1. Su Arıtma Sistemine Ait Veri Seti

Goh ve arkadaşları [28], SWaT test sistemi üzerinde geliştirdikleri çeşitli senaryolar doğrultusunda saldırılar düzenleyerek SWaT veri setini oluşturmuşlardır. Erişime açık olarak sunulan bu veri seti, test sistemi üzerinde gerçekleştirilen farklı saldırılardan elde edilen verilerden oluşmaktadır. Saldırılar, dört farklı kategoriye ayrılmaktadır: Tek aşamalı tek nokta (SSSP), tek aşamalı çok nokta (SSMP), çok aşamalı tek nokta (MSSP) ve çok aşamalı çok nokta (MSMP). Bu kategorilere göre saldırı dağılımı ise şu şekildedir: SSSP kategorisinde 26 saldırı, SSMP kategorisinde 4 saldırı, MSSP kategorisinde 2 saldırı ve MSMP kategorisinde 4 saldırı gerçekleştirilmiştir. Bu saldırıların temel amacı, veri paketlerinin PLC'lere ulaşmadan önce sensör verilerini manipüle etmek amacıyla yakalanmasıdır.

SWaT veri seti, SCADA sistemi ile PLC'ler arasındaki iletişim bağlantılarındaki paketlerin ele geçirilmesiyle, toplamda 11 gün boyunca toplanan verilerden oluşmaktadır. Bu sürecin ilk 7 gününde herhangi bir saldırı düzenlenmemiş olup, SWaT test sistemi kesintisiz olarak 24 saat çalıştırılmıştır. Takip eden 4 gün boyunca ise toplam 36 farklı saldırı gerçekleştirilmiştir. Bu saldırıların hedefleri ve amaçları Tablo 3'te detaylandırılmaktadır.

SWaT veri seti, saldırılı ve normal olarak etiketlenmiş toplam 51 öznitelik içermekte olup bu öznitelikler, 25 sensörden alınan ölçümler ile 26 eyleyiciden gelen sinyalleri kapsamaktadır. İlk 7 gün boyunca toplam 496.800 örnek kaydedilirken, saldırıların uygulandığı son 4 günde ise 449.919 örnek toplanmıştır. Son 4 günde toplanan verilerin 54.584'ü saldırı, kalan 395.335'i ise normal olarak etiketlenmiştir. Sonuç olarak, SWaT veri seti, her alt sürecin özniteliklerinin kategorize edildiği toplam 946.722 örnekten oluşmaktadır.

Tablo 3. Su arıtma sistemleri için tasarlanan ve başlatılan tüm saldırılar [47]

Saldırı ID	Saldırı Açıklaması	Amaç
1	MV-101 kapalı olması gerektiği halde açık konumdadır.	Tank taşması meydana gelmiştir.
2	P-102 kapalı olması gerektiği halde açık durumdadır.	Boru patlaması.
3	LIT-101 sensöründen alınan değer, saniyede 1 mm artış göstermektedir.	Tank alt seviyede ve P-101'e zarar verir.
4	MV-504 kapalı olması gerektiği halde açık konumdadır.	Etkisi yok.
5	AIT-202 sensöründen alınan değer, nominal değerin altına düşmüştür.	AIT-504 sensöründen alınan değer artmış, ancak drenaj başlamamıştır.
6	LIT-301 sensöründen alınan değer, maksimum limitin üzerine çıkmıştır.	Tank alt seviyeye inmiş ve P-101'e zarar vermiştir.
7	DPIT-301 sensöründen alınan değer, nominal değerin üzerine çıkmıştır.	Geri yıkama süreci yeniden başlamış, bu durum Tank-401 su seviyesini düşürmüş ve Tank-301 su seviyesini artırmıştır.
8-9	FIT-401 sensöründen alınan değer, nominal değerin altına düşmüştür.	UV süreci durdurulmuş ve P-501 kapatılmıştır.
10	MV-304 açık olması gerektiği halde kapalı konumdadır.	MV-304 kapatılmıştır.
11	MV-303 kapalı konumda sıkışmıştır.	Etkisi yok.
12	LIT-301 sensöründen alınan değer, saniyede 1 mm azalmaktadır.	Tank taşması meydana gelmiştir.
13	MV-303 kapalı konumda sıkışmış durumdadır.	Aşama 3 operasyonunu durdur.
14-15	AIT-504 sensöründen alınan değer, nominal değerin üzerine çıkmıştır.	Etkisi yok.
16	MV-101 açık konumda sıkışmış durumdadır ve LIT-101 sensöründen alınan değer 0.7 m olarak belirlenmiştir.	Tank taşması meydana gelmiştir.

Tablo 3’ün devamı

17	UV-401 açık olması gerektiği halde kapalıdır, AIT-502 sensöründen alınan değer nominal değerin üzerine çıkmıştır ve P-501 açık modda sıkışmıştır.	FIT-502’deki çıktı azaltılmıştır.
18	DPIT-301 sensöründen alınan değer, nominal değerin üzerine çıkmıştır, MV-302 açık konumda sıkışmıştır ve P-602 kapalı modda sıkışmıştır.	Sistem donması meydana gelmiştir.
19	P-203 ve P-205 açık olması gerektiği halde kapalı durumdadır.	Etkisi yok.
20	LIT-401 sensöründen alınan değer, nominal değerin üzerine çıkmıştır ve P-205 açık modda sıkışmıştır.	Tank taşması meydana gelmiştir.
21	P-101 kapalı olması gerektiği halde açık modda sıkışmıştır ve P-205 açık modda sıkışmıştır.	Tank-101'de alt seviye sorunu ve Tank-301'de taşma meydana gelmiştir.
22	P-302 açık modda sıkışmıştır ve LIT-401 sensöründen alınan değer 0.8 m olarak belirlenmiştir.	Tank taşması meydana gelmiştir.
23	P-302 açık olması gerektiği halde kapalı durumdadır.	Tank-401'e giriş durmuştur.
24	P-201, P-203 ve P-205 kapalı olması gerektiği halde açık durumdadır.	Etkisi yok.
25	P-101 ve MV-101 kapalı olması gerektiği halde açık modda sıkışmıştır ve LIT-101 sensöründen alınan değer 0.7 m olarak belirlenmiştir.	Tank-101'de alt seviye sorunu ve Tank-301'de taşma meydana gelmiştir.
26	LIT-401 sensöründen alınan değer, minimum seviyenin altına düşmüştür.	Tank taşması meydana gelmiştir.
27	LIT-301 sensöründen alınan değer, maksimum seviyenin üzerine çıkmıştır.	Tank alt seviyeye inmiş ve P-302'ye zarar vermiştir.
28	LIT-101 sensöründen alınan değer, maksimum seviyenin üzerine çıkmıştır.	Tank alt seviyeye inmiş ve P-101'e zarar vermiştir.
29	P-101 açık olması gerektiği halde kapalı durumdadır.	Yedek pompa P-102 açılmıştır.
30	P-101 açık olması gerektiği halde kapalı durumdadır ve P-102 kapalı modda sıkışmıştır.	Çıkış akışı durmuştur.
31	LIT-101 sensöründen alınan değer, minimum seviyenin altına düşmüştür.	Tank taşması meydana gelmiştir.
32	P-501 açık olması gerektiği halde kapalı durumdadır ve FIT-502 sensöründen alınan değer nominal değerin üzerine çıkmıştır.	P-501 kapatılmamış, FIT-502 azalmış ve P-501'in hızı artmıştır.
33	AIT-402 ve AIT-502 sensörlerinden alınan değerler 260 olarak belirlenmiştir.	Etkisi yok.
34	FIT-401 sensöründen alınan değer 0.5, AIT-501 sensöründen alınan değer ise 140 mV olarak belirlenmiştir.	Etkisi yok.

Tablo 3'ün devamı

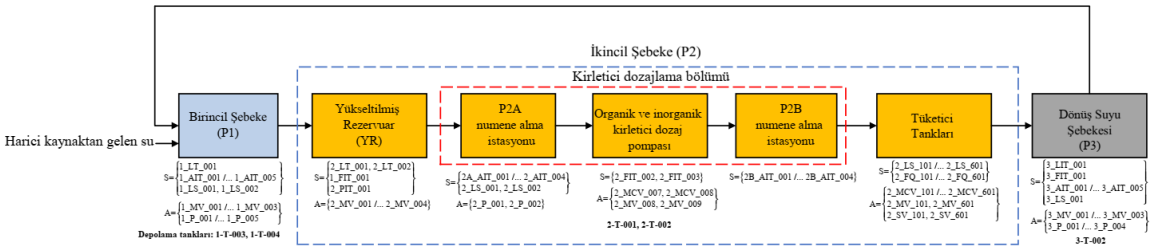
35	FIT-401 sensöründen alınan değer sıfır olarak belirlenmiştir.	P-402 kapanmamıştır.
36	LIT-301 sensöründen alınan değer, saniyede 0.55 birim azalmaktadır.	Tank su seviyesi azalmaktadır.

1.4. Su Dağıtma Sistemi (WADI)

WADI sistemi, fiziksel yapısı ve güvenlik riskleri açısından SWaT sistemine göre farklılık göstermektedir. SWaT sistemi genellikle güvenli ve sınırlı alanlarda işletilirken, WADI sistemi geniş bir alana yayılan boru hatlarından oluşan kapsamlı bir ağ yapısına sahiptir. Bu geniş altyapı, WADI sistemini fiziksel saldırılara karşı daha savunmasız hale getirmektedir. SWaT sisteminin bir uzantısı olarak geliştirilen WADI sistemi [30], şehre dakikada 37,8 litre filtrelenmiş su sağlayan bir test platformudur. WADI sisteminde üç farklı kontrol aşaması bulunmaktadır: birincil şebeke (P1), ikincil şebeke (P2) ve dönüş suyu şebekesi (P3). Her bir aşama, bir PLC tarafından yönetilmektedir.

İlk aşama olan P1'de, SWaT sisteminden veya harici bir kaynaktan gelen su, 2,5 litrelik iki ham su tankına aktarılmakta ve P2 aşamasındaki yükseltilmiş rezervuar tankını doldurmak üzere depolanmaktadır. Depolama işlemi öncesinde, suyun içilebilir hale gelmesi amacıyla kimyasal arıtma yapılmaktadır. Ana şebekedeki seviye sensörü (1_LT_001), ham su tanklarının seviyelerini izlemektedir.

P2 aşaması, iki yükseltilmiş rezervuar tankı, tüketici tankları ve bir kirlilik örnekleme istasyonundan oluşmaktadır. Ham su tanklarından alınan su, birincil şebekede yer alan ham su pompası (1_P_003) aracılığıyla yükseltilmiş rezervuar tanklarına iletilmektedir. Bu tanklardaki su seviyeleri, 2_LT_001 ve 2_LT_002 olarak adlandırılan iki seviye sensörü ile ölçülmektedir. Yükseltilmiş rezervuarlardan gelen su, önceden belirlenmiş bir talep modeline göre tüketici tanklarına yönlendirilmektedir. Suyun tüketici tanklarına akmadan önce kalitesini kontrol etmek amacıyla P2A ve P2B olarak adlandırılan iki su kalitesi izleme istasyonu bulunmaktadır. Su kalitesi uygun olmadığında su boşaltılmakta; uygun olduğunda ise tüketici tanklarına yönlendirilmektedir. Her bir tüketici tankında bir seviye anahtarı bulunmakta olup, tank dolduğunda bu anahtar tetiklenmekte ve su, dönüş suyu şebekesine boşaltılmaktadır.



Şekil 2. Su dağıtma sistemi süreçleri

Son aşama olan P3, dönüş suyu aşamasını temsil etmektedir. Bu aşamada, tüketiciler tarafından kullanılan su, dönüş suyu deposunda toplanmakta ve geri dönüştürülmek üzere birincil şebekeye pompalanmaktadır. Pompalama işlemi öncesinde, suyun kalitesini değerlendirmek için dönüş suyu şebekesine su kalitesi analizörleri yerleştirilmiştir. SWaT sistemine benzer şekilde, WADI sistemi de kimyasal dozaj sistemleri, hidrofor pompaları, vanalar, enstrümantasyon ve analizörler gibi benzer donanımlarla yapılandırılmıştır. Bu kapsamlı kurulum, ağ tabanlı saldırıların simülasyonunu gerçekleştirme imkânı sunmaktadır. WADI sisteminin süreçleri Şekil 2'de gösterilmekte, WADI sisteminde kullanılan sensörler ve eyleyicilerin tanımları ise Tablo 4'te yer almaktadır.

Tablo 4. Su dağıtma sistemlerine ait sensör ve eyleyiciler

Numara	İsim	Tür	Tanım
1	x_LT_xxx	Sensör	Seviye ölçer
2	x_FIT_xxx	Sensör	Akış ölçer
3	x_AIT_001	Sensör	İletkenlik analizörü
4	x_AIT_002	Sensör	Bulanıklık analizörü
5	x_AIT_003	Sensör	pH analizörü
6	x_AIT_004	Sensör	ORP analizörü
7	x_AIT_005	Sensör	Toplam kalan klor
8	x_LS_xxx	Sensör	Seviye anahtarı
9	x_PIT_xxx	Sensör	Basınç ölçer
10	x_MV_xxx	Eyleyici	Motorlu vana
11	x_P_xxx	Eyleyici	Ham su pompası
12	x_MCV_xxx	Eyleyici	Oransal kontrol vana
13	x_SV_xxx	Eyleyici	Solenoid vana
14	x_FQ_xxx	Eyleyici	Debi toplayıcısı

1.4.1. Su Dağıtma Sistemine Ait Veri Seti

WADI veri seti, Singapur Kamu Hizmetleri Komisyonu tarafından yönetilen gerçek bir şehir su dağıtma şebekesinden toplanan verilerden oluşmaktadır. Bu veri seti, 16 günlük bir süre boyunca 123 sensörden elde edilen verileri kapsamaktadır. İlk 14 gün, sistemin standart operasyonel davranışını ve normal işleyişini yansıtan bir senaryoyu temsil etmekte olup, bu süreçte toplam 1.048.571 veri örneği kaydedilmiştir. Kalan gün ise 172.801 örnek içeren bir saldırı senaryosuna ayrılmıştır. Bu iki günlük saldırı senaryosu sırasında toplanan 172.801 örneğin yalnızca 9.977'si "saldırı" olarak etiketlenirken, geri kalan 162.941 örnek "normal" olarak etiketlenmiştir. Bu saldırılar, sensör verilerinin manipülasyonu ve kontrol komutlarının değiştirilmesi gibi çeşitli teknikleri içermektedir. WADI sistemi için tasarlanan ve uygulanan tüm saldırılar Tablo 5'te ayrıntılı olarak sunulmaktadır.

Tablo 5. Su dağıtma sistemleri için tasarlanan ve gerçekleştirilen tüm saldırılar [24]

Saldırı ID	Saldırı Açıklaması	Amaç
1	1_MV_001 kapalı olması gerekirken açıktır	Tank taşması
2	1_FIT_001 okuma değeriyle oynanmıştır.	Su kalitesini değiştirmek
3-4	1_AIT_001 okuma değeriyle oynanmıştır.	Tank taşması
5	2_MCV_101, 2_MCV_201, 2_MCV_301, 2_MCV_401, 2_MCV_501 ve 2_MCV_601 açık olması gerekirken kapalıdır.	Su dağıtma sürecine müdahale etmek.
6	2_MCV_101, 2_MCV_201 kapalı olması gerekirken açıktır.	Su dağıtma sürecine müdahale etmek.
7	1_AIT_002 okuma ile oynanmıştır ve 2_MV_003 kapalı olması gerekirken açıktır.	Su kalitesini değiştirmek.
8,11 ve 12	2_MCV_007 kapalı olması gerekirken açıktır.	Su dağıtma sürecine müdahale etmek.
9	1_P_006 kapalı olması gerekirken açıktır.	Boru patlaması
10	1_MV_001'e ve ham su pompasına zarar verme	Tank taşması
13	Basınç pompasının ayar noktasını azaltma	Su dağıtma sürecine müdahale etmek.
14	Kimyasal dozajlama pompalarını durdurma.	Su kalitesini değiştirmek
15	AIT_001 okuma değeriyle oynanmıştır.	Tank taşması

1.5. Endüstriyel Kontrol Sistemlerinin Güvenliğinde Aykırılık Tespit Yaklaşımları

ICS güvenliğinde, sistemin normal işleyişinden sapmaları tespit etmek amacıyla çeşitli aykırılık tespit yöntemleri kullanılmaktadır. Bu yöntemler, genel olarak beş ana kategoriye ayrılmaktadır: İstatistiksel yöntemler, ML yöntemleri, kural tabanlı yöntemler, hibrit

yaklaşımlar ve imza tabanlı yöntemler. Proje kapsamında, ML yöntemleri ve kural tabanlı yöntemler kullanıldığı için bu iki yaklaşım detaylı olarak ele alınmaktadır. ML yöntemleri kapsamında denetimli öğrenme (SL), denetimsiz öğrenme (UL) ve DL modelleri incelenmektedir. Kural tabanlı yöntemler bağlamında, iş mantığı kuralları ve durum tabanlı kurallar değerlendirilmektedir.

1.5.1. Denetimli Öğrenme

SL, etiketli veriler kullanarak normal ve aykırı durumları sınıflandırmak amacıyla sınıflandırıcılar oluşturmaktadır. Bu yöntem, eğitim sürecinde etiketli verilerden öğrendiği bilgiyi yeni verilere uygulayarak aykırılıkları tespit etmektedir. Literatürde, SVM ve Naive Bayes gibi algoritmaların, yüksek doğruluk oranları ve belirli aykırılık türlerine karşı güçlü performans sergilediği belirtilmektedir [48]. SVM, özellikle doğrusal olmayan veri kümelerinin sınıflandırılmasında etkili olmakta ve genellikle Kernel yöntemlerini kullanarak veriyi yüksek boyutlu bir uzaya haritalamaktadır. Bu da karmaşık aykırılıkların tespit edilmesini mümkün kılmaktadır. Naive Bayes algoritması ise özellikle büyük veri kümelerinde hızlı ve etkili çalışmasıyla bilinmekte ancak, bağımsızlık varsayımı nedeniyle bazı durumlarda performans kayıpları yaşayabilmektedir [49].

SL yöntemlerinin dezavantajları arasında, etiketli veri gereksinimi ile büyük veri kümeleri üzerinde çalışırken zaman ve kaynak yoğunluğu yer almaktadır. Örneğin, büyük veri kümeleri ile çalışırken eğitim süresi ve hesaplama maliyetleri önemli ölçüde artmaktadır. Ayrıca, etiketli verilerin elde edilmesi zaman alıcı ve maliyetli bir süreç olmaktadır [50].

1.5.2. Denetimsiz Öğrenme

UL, etiketli verilerin olmadığı durumlarda sistemin normal davranışını öğrenmek ve aykırılıkları tespit etmek amacıyla kullanılan bir yaklaşım olarak bilinmektedir. Bu yöntemde, veri noktaları arasındaki benzerliklere dayanarak gruplar oluşturulmakta ve bu gruplardan sapmalar aykırılık olarak değerlendirilmektedir. K-ortalama ve yoğunluk tabanlı mekânsal kümeleme (DBSCAN) gibi algoritmalar, etiketli veriye ihtiyaç duymaması ve farklı türdeki aykırılıkları tespit edebilmeleriyle önemli avantajlar sunmaktadır.

K-ortalama algoritması, veri noktalarını önceden belirlenmiş sayıda kümeye ayırarak her küme için bir merkez belirlemekte ve bu merkezler, kümedeki veri noktalarının

ortalamalarına göre sürekli olarak güncellenmektedir. Ancak, K-ortalama algoritmasının bazı dezavantajları bulunmaktadır. Özellikle önceden belirlenmesi gereken küme sayısına bağımlılığı ve küresel olmayan küme yapılarında düşük performans sergilemesi bu dezavantajlardan bazılarıdır [51]. DBSCAN algoritması ise, veri noktalarını yoğunluklarına göre kümelemekte ve düşük yoğunluklu bölgeleri aykırılık olarak değerlendirmektedir. Bu algoritmanın en önemli avantajı, küme sayısının önceden belirlenmesine gerek kalmadan çalışabilmesidir. Ayrıca, gürültü verilerini dikkate alması, performansını artıran bir diğer önemli özellik olarak öne çıkmaktadır [52].

UL yöntemlerinin avantajları arasında, etiketli veriye ihtiyaç duymamaları ve farklı türde aykırılıkları tespit edebilme kabiliyetleri bulunmaktadır. Ancak, aykırılıkların doğru şekilde tanımlanması zor olabildiğinden, bu yöntemler bazen düşük hassasiyetle sonuçlanabilmektedir. Ayrıca, bu yöntemler genellikle yüksek boyutlu veri setlerinde performans kayıpları yaşayabilmektedir [53].

1.5.3. Derin Öğrenme Modelleri

Karmaşık veri yapılarında aykırılık tespiti için DL modelleri kullanılmaktadır. CNN, RNN ve Autoencoder gibi DL yöntemleri, büyük veri setlerinde yüksek doğruluk oranlarıyla aykırılıkları tespit edebilmektedir. Özellikle CNN modeli, görüntü ve zaman serisi verilerinde başarılı sonuçlar sunmaktadır. Bu modeller, veri içindeki yerel bağlantıları ve uzamsal hiyerarşileri öğrenme kapasiteleri sayesinde karmaşık veri yapılarında aykırılık tespiti için önemli avantajlar sağlamaktadır [54].

RNN modeli, özellikle zaman serisi verilerinde kullanılan DL modelleri olup ardışık veri noktaları arasındaki bağımlılıkları öğrenmektedir. LSTM ve kapı kontrollü tekrarlayan birim (GRU) gibi RNN türevi modeller, uzun vadeli bağımlılıkları öğrenme kapasiteleriyle öne çıkmakta ve bu nedenle zaman serisi verilerinde yaygın olarak tercih edilmektedir [55].

Autoencoder, veri sıkıştırma ve gürültü temizleme gibi görevlerde kullanılan DL modelleridir. Bu modeller, veriyi düşük boyutlu bir temsille yeniden inşa etmekte ve yeniden inşa edilen veri ile orijinal veri arasındaki farkları analiz ederek aykırılıkları tespit etmektedir. DL modellerinin avantajları arasında, büyük veri setlerinde yüksek doğruluk ve karmaşık veri yapılarının derinlemesine analizi yer almaktadır. Ancak, bu yöntemlerin yüksek hesaplama maliyeti ve büyük miktarda eğitim verisi gereksinimi gibi dezavantajları

bulunmaktadır. Ayrıca, modelin karmaşıklığı arttıkça aşırı uyum ve genelleme problemleri de artış göstermektedir [54].

1.5.4. İş Mantığı Kuralları

İş mantığı kuralları, ICS'lerde belirli olayların gerçekleşme sıralarına veya mantığına dayalı olarak aykırılıkları tespit etmek için kullanılmaktadır. Bu kurallar, sistemin işleyişine dair spesifik süreçleri ve adımları tanımlamakta ve adımların belirli bir sırayla gerçekleşmesi gerektiğini belirtmektedir. Örneğin, SWaT sisteminde suyun belirli bir aşamadan geçtikten sonra bir sonraki aşamaya ilerlemesi gerektiği iş mantığı kuralları ile tanımlanmaktadır. Bu sırada bir sapma meydana geldiğinde, sistem bu sapmayı aykırılık olarak tespit etmektedir [56].

İş mantığı kuralları, sistemin spesifik süreçlerinin kontrolünde etkili bir yöntem olarak değerlendirilmektedir. Genellikle işletme prosedürleri ve süreçlerinin bir yansıması olan bu kurallar, sistemdeki olayların belirli bir mantık çerçevesinde gerçekleşmesini sağlamaktadır. Mevcut operasyonel prosedürler üzerine inşa edildikleri için hızlı tespit ve düşük maliyet avantajları sunmaktadırlar [57]. Ancak, kapsamlı kural setlerine ihtiyaç duyulması ve sistemdeki değişikliklere hızlı uyum sağlama gerekliliği gibi zorluklar da bulunmaktadır. Ayrıca, kuralların manuel olarak güncellenmesi ve bakımı gerekmektedir.

1.5.5. Durum Tabanlı Kurallar

Durum tabanlı kurallar, sistemin belirli durumlarına yönelik olarak tanımlanmış kurallar aracılığıyla aykırılıkları tespit etmektedir. Bu kurallar, sistemin fiziksel ve operasyonel kısıtlamalarını esas alarak belirli koşullar altında etkin hale gelmektedir. Örneğin, bir tankın su seviyesinin belirli bir aralıkta tutulması gerekliliği, aykırılık tespiti amacıyla kullanılan bir kuraldır. Bu tür kurallar, sistemin mevcut durumuna göre aktifleşmekte ve o durumun sınırlarını belirlemektedir [58].

Durum tabanlı kurallar, sistemin belirli fiziksel parametrelerini ve operasyonel sınırlamalarını dikkate almaktadır. Bu kurallar, belirli bir sensör değeri, basınç seviyesi veya sıcaklık gibi fiziksel parametrelere odaklanmakta ve bu parametrelerin belirli aralıklar içinde olup olmadığını kontrol etmektedir. Bu sayede, sistemin normal çalışma koşullarının dışında bir durum tespit edildiğinde hızlı bir şekilde aykırılık alarmı verebilmektedir [59]. Ancak,

bu kuralların değişen sistem koşullarına uyum sağlaması zor olabilmekte ve bu nedenle kuralların sürekli olarak güncellenmesi gerekebilmektedir.

1.6. Performans Metrikleri

Karışıklık matrisi, bir modelin performansını değerlendirmede önemli bir analiz aracı olarak kullanılmaktadır. Bu matris, modelin doğru ve yanlış tahminlerini gerçek gözlem değerleriyle karşılaştırarak performansın ayrıntılı bir şekilde incelenmesine olanak tanımaktadır. Böylece, modelin güçlü ve zayıf yönleri belirlenebilmektedir. Karışıklık matrisi, doğru pozitif (TP), yanlış pozitif (FP), doğru negatif (TN) ve FN gibi tahminleri ortaya koyarak modelin genel doğruluğu ve hata türleri hakkında daha derinlemesine bir anlayış sağlamaktadır. Bu durum, modelin iyileştirilmesine ve karar verme süreçlerinin daha etkili bir şekilde yönetilmesine katkı sunmaktadır.

Her bir sınıf için doğru ve yanlış tahminlerin detaylı bir şekilde sunulması, modelin sınıflandırma yeteneğinin kapsamlı bir şekilde değerlendirilmesine olanak tanımaktadır. Özellikle karmaşık veri setleri üzerindeki başarıyı değerlendirmek için karışıklık matrisi kritik bir gösterge olarak değerlendirilmektedir.

TP değeri hem gerçekte pozitif olan hem de model tarafından pozitif olarak tahmin edilen değerleri ifade etmektedir. Bu durum, modelin aykırılıkları başarıyla tespit ettiğini göstermektedir. Benzer şekilde, TN değeri ise gerçekte negatif olan ve model tarafından da negatif olarak tahmin edilen değerleri ifade etmektedir. Bu, modelin normal durumları doğru bir şekilde tanımladığını belirtmektedir. FP değeri, gerçekte negatif olan ancak model tarafından pozitif olarak tahmin edilen değerleri ifade etmektedir. Bu, modelin yanlış alarm durumlarına neden olduğunu göstermektedir. FN değeri ise gerçekte pozitif olan ancak model tarafından negatif olarak tahmin edilen değerleri ifade etmektedir. Bu durum, modelin bazı aykırılıkları gözden kaçırdığını ortaya koymaktadır.

DL modellerinin performansını değerlendirmek için çeşitli metrikler kullanılmaktadır. Bu metrikler, modelin doğruluğunu, hassasiyetini, duyarlılığını ve genel başarısını ölçmeye olanak tanımaktadır. Yaygın olarak kullanılan performans metrikleri aşağıda detaylı bir şekilde ele alınmaktadır.

Doğruluk: Genel model başarısını değerlendirmek için yaygın olarak kullanılan bir performans metriği doğruluk olup, modelin doğru tahmin yapma yeteneğini yansıtmaktadır. Ancak, pozitif ve negatif örneklerin sayısının dengesiz olduğu veri setlerinde doğruluk

metriği yanıltıcı olabilmektedir. Örneğin, veri setinde çoğunlukla negatif örnekler bulunuyorsa, modelin tüm tahminlerini negatif yapması yüksek bir doğruluk oranı gösterebilmektedir. Bu, modelin pozitif örnekleri doğru bir şekilde tespit edemediği anlamına gelmektedir. Matematiksel olarak doğruluk, doğru tahminlerin toplam tahminlere bölünmesiyle hesaplanmakta ve şu formülle ifade edilmektedir.

$$Doğruluk = \frac{Doğru\ Pozitifler + Doğru\ Negatifler}{Toplam\ Örnek\ Sayısı} \quad (1)$$

Hassasiyet: Hassasiyet, modelin pozitif tahminlerinin ne kadarının gerçekten pozitif olduğunu belirleyen bir performans metriğidir. Bu metrik, özellikle dengesiz veri setlerinde ve FP sonuçlarının maliyetinin yüksek olduğu senaryolarda büyük önem taşımaktadır. Hassasiyet, modelin TP tahminlerine odaklanarak FP değerlerini en aza indirme kapasitesini göstermektedir. Örneğin, SWaT sisteminde aykırılık tespiti için kullanılan bir modelde, yüksek FP oranı gereksiz alarmlara ve artan operasyonel maliyetlere neden olabilmektedir. Bu bağlamda, hassasiyet metriği, bir sistemin TP değerlerini doğru şekilde tanımlama başarısını değerlendirmek için kritik bir ölçüt olarak kabul edilmektedir.

$$Hassasiyet = \frac{Doğru\ Pozitifler}{Doğru\ Pozitifler + Yanlış\ Pozitifler} \quad (2)$$

Duyarlılık: Duyarlılık, bir modelin TP örnekleri tespit etme oranını ölçen önemli bir performans göstergesidir. Bu metrik, FN sonuçların maliyetinin yüksek olduğu durumlarda özellikle önem taşımaktadır. Duyarlılık, modelin pozitif sınıfa ait tüm örnekleri doğru bir şekilde bulma kapasitesini ifade etmektedir. Örneğin, SWaT sisteminde patojenlerin tespiti için kullanılan bir modelin düşük duyarlılık oranı, tehlikeli patojenlerin gözden kaçmasına ve potansiyel sağlık risklerine yol açabilmektedir. Bu nedenle duyarlılık metriği, bir sistemin doğru pozitif örnekleri yakalama başarısını değerlendirmek için belirleyici bir kriterdir.

$$Duyarlılık = \frac{Doğru\ Pozitifler}{Doğru\ Pozitifler + Yanlış\ Negatifler} \quad (3)$$

F1 Skoru: Modelin performansını değerlendirmek için kullanılan F1 skoru, hassasiyet ve duyarlılık metriklerinin harmonik ortalamasını alan bir metriktir. Özellikle dengesiz veri

setleriyle çalışırken veya FP ve FN sonuçlarının maliyetinin yüksek olduğu durumlarda F1 skoru büyük önem taşımaktadır. Bu metrik, TP değerlerin yanı sıra FP ve FN sonuçlarını da dikkate alarak dengeli bir performans değerlendirmesi sunmaktadır. Örneğin, SWaT sisteminde tehlikeli kimyasalların tespiti için kullanılan bir modelde yüksek bir F1 skoru, modelin hem tehlikeli kimyasalları doğru tespit etme hem de yanlış alarmları azaltma konusundaki etkinliğini göstermektedir. Bu nedenle F1 skoru, bir modelin genel doğruluk ve güvenilirlik düzeyini değerlendirmek için kritik bir metriktir.

$$F1\ Skoru = 2\ x\ \frac{Hassasiyet\ x\ Duyarlılık}{Hassasiyet + Duyarlılık} \tag{4}$$

Kesinlik: Modelin, negatif sınıfa ait örnekleri doğru bir şekilde tespit etme oranını ölçen bir performans göstergesi olmaktadır. Özellikle FP sonuçlarının maliyetinin yüksek olduğu durumlarda büyük önem taşımaktadır. Kesinlik, modelin FP değerlerini en aza indirme kapasitesini ifade etmektedir. Örneğin, SWaT sisteminde zararlı maddelerin yanlışlıkla varmış gibi algılanması, operasyonel maliyetleri artırabilmektedir. Bu tür durumlarda, yüksek bir kesinlik oranı, modelin TN değerlerini doğru bir şekilde belirleme yeteneğini göstermektedir. Bu nedenle kesinlik metriği, modelin yanlış alarm oranını düşük tutma başarısını değerlendirmek için başlıca performans göstergelerinden biri olmaktadır.

$$Kesinlik = \frac{Doğru\ Negatifler}{Doğru\ Negatifler + Yanlış\ Pozitifler} \tag{5}$$

Ortalama Mutlak Hata (MAE): MAE, bir modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasını hesaplayan bir performans metriği olmaktadır. MAE, modelin tahminlerinin ne kadar hatalı olduğunu gösteren basit fakat etkili bir ölçüm sunmaktadır. Özellikle sürekli verilerle çalışan regresyon modellerinin performansını değerlendirmek için yaygın olarak kullanılmaktadır. MAE, hataların büyüklüğüne duyarlıdır, ancak hataların yönünü dikkate almamaktadır; bu nedenle, pozitif ve negatif hatalar aynı derecede değerlendirilmektedir.

Örneğin, SWaT sisteminde su kalitesini belirlemeye yönelik kullanılan bir modelde düşük bir MAE değeri, modelin su kalitesini doğru bir şekilde tahmin ettiğini ve tahminlerin gerçek değerlere yakın olduğunu göstermektedir. Bu bağlamda, MAE, bir modelin ortalama

tahmin hatasını değerlendirmek için güvenilir bir ölçüt olup, aynı zamanda model performansının iyileştirilmesi gereken alanların belirlenmesine de katkı sağlamaktadır.

Ortalama Kare Hatası (MSE): MSE, modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkların karelerinin ortalamasını hesaplayan bir performans metriği olmaktadır. MSE, regresyon analizinde ve ML modellerinde, modelin tahmin hatalarının büyüklüğünü değerlendirmek amacıyla yaygın olarak kullanılmaktadır. Hataların karelerinin alınması, MSE metriğinin büyük hatalara daha fazla ağırlık vermesine neden olmaktadır. Bu özellik, modelin büyük sapmalara karşı duyarlılığını artırmakta ve hataların ortalamadan ne kadar uzaklaştığını belirlemeye yardımcı olmaktadır.

Örneğin, SWaT sisteminde su kalitesini tahmin eden bir modelde düşük bir MSE değeri, tahminlerin gerçek değerlere yakın olduğunu ve hataların genellikle küçük olduğunu ifade etmektedir. Buna karşın, yüksek bir MSE değeri, modelin tahminlerinde büyük hataların bulunduğunu belirtmektedir. Bu durum, model performansını olumsuz etkilemektedir. Bu bağlamda, MSE, bir modelin doğruluğunu ve güvenilirliğini değerlendirmek için önemli bir ölçüt olmaktadır.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

Yanlış Pozitif Oranı: YPO, bir modelin negatif sınıfa ait örnekleri yanlışlıkla pozitif olarak sınıflandırdığı durumların oranını ölçen bir performans metriğidir. YPO, özellikle FP sonuçların maliyetinin yüksek olduğu durumlarda model performansını değerlendirmek için kritik bir göstergedir. Düşük bir YPO değeri, modelin yanlış pozitif hata yapma olasılığının düşük olduğunu ve negatif örnekleri doğru bir şekilde tanımladığını göstermektedir. Buna karşılık, yüksek bir YPO değeri, modelin birçok negatif örneği yanlışlıkla pozitif olarak sınıflandırdığını ve bu nedenle gereksiz uyarılara veya yanlış alarmlara yol açtığını ifade etmektedir. Örneğin, SWaT sisteminde kullanılan bir aykırılık tespit modelinde, yüksek bir YPO değeri, modelin hatalı alarmlar üretebileceğini göstermektedir. Bu durum, sistemin gereksiz müdahalelere yol açmasına ve operasyonel maliyetlerin artmasına neden olmaktadır.

$$Yanlış\ Pozitif\ Oranı = \frac{Yanlış\ Pozitifler}{Yanlış\ Pozitifler + Doğru\ Negatifler} \quad (8)$$

Kosinüs Benzerliği (CS): Ölçüm değeri (x) ile çekişmeli örnek (x_{adv}) arasındaki CS, iki vektör arasındaki açının kosinüsünü hesaplayan bir metrik olmaktadır. Bu yöntem, sinyallerin yönelimleri arasındaki benzerliği değerlendirmek için kullanılmaktadır. CS değeri 1'e yaklaştıkça, iki vektörün büyük ölçüde benzer olduğu anlaşılmaktadır. Buna karşılık, CS değeri 0'a yaklaştığında, vektörler arasındaki benzerliğin düşük olduğu ve farklı yönelimlere sahip oldukları sonucuna varılmaktadır. CS, sinyallerin büyüklüklerinden bağımsız olarak yalnızca yönelimlerini dikkate almakta; bu nedenle, büyüklük değişimlerinden etkilenmeyen bir benzerlik ölçüsü olarak tercih edilmektedir.

CS, özellikle sinyal işleme ve ML gibi alanlarda, sinyallerin veya veri noktalarının benzerliğini değerlendirmek için yaygın olarak kullanılmaktadır. Bu metrik, veri noktalarının aynı yönde ne kadar hizalandığını belirleyerek, farklı büyüklüklere sahip veri noktalarının benzerliğini etkili bir şekilde ölçmektedir.

$$CS = \frac{x.x_{adv}}{\|x\|.\|x_{adv}\|} \quad (9)$$

Normalleştirilmiş Çapraz Korelasyon (NCC): Ölçüm değeri (x) ile çekişmeli örnek (x_{adv}) arasındaki NCC, iki sinyal arasındaki benzerliği ölçen bir metrik olmaktadır. NCC, sinyallerin ortalama ve standart sapmalarını kullanarak korelasyon değerini hesaplamakta ve sonuçların -1 ile 1 arasında olmasını sağlamaktadır. Bu bağlamda, 1 değeri, yüksek pozitif korelasyonu ifade etmekte ve iki sinyalin aynı doğrultuda değiştiğini göstermektedir. Buna karşılık, -1 değeri, yüksek negatif korelasyonu belirtmekte ve iki sinyalin ters doğrultuda değiştiğini ifade etmektedir.

Bu metrik, sinyallerin büyüklük ve birim farklılıklarından bağımsız olarak yalnızca benzerliklerini dikkate aldığı için tercih edilmektedir. Bu nedenle, NCC, özellikle sinyal işleme ve zaman serisi analizlerinde, iki veri seti arasındaki ilişkiyi analiz etmek için yaygın olarak kullanılmaktadır. NCC, veri setleri arasındaki benzerlikleri ve ilişkileri değerlendirme konusunda güçlü bir araç olmaktadır.

$$NCC = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\left(x_{adv,i} - \overline{x_{adv}}\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}\left(x_{adv,i} - \overline{x_{adv}}\right)^2}} \quad (10)$$

Öklid Normu (L_2 Normu): Pertürbasyon vektörünün L_2 normu, yani Öklid normu, vektörün büyüklüğünü ölçen bir metriktir. L_2 normu, vektörün elemanlarının karelerinin toplamının karekökü alınarak hesaplanmakta ve bu hesaplama, pertürbasyonun büyüklüğünü belirlemek için kullanılmaktadır. Bu norm, özellikle vektörlerin genel büyüklüklerini ve uzunluklarını değerlendirmede yaygın olarak tercih edilmektedir. Pertürbasyon vektörlerinin büyüklüğünü bu şekilde hesaplamak, sinyal işleme ve ML gibi alanlarda, modelin veya sistemin dış etkiler karşısındaki duyarlılığını analiz etmek açısından önemli olmaktadır. L_2 normu, modelin dayanıklılığını ve değişikliklere tepkisini anlamak için temel bir ölçüt olarak kullanılmakta; bu sayede bir modelin küçük değişiklikler karşısında nasıl tepki vereceği daha iyi anlaşılabilmektedir.

$$\|\Delta_\alpha\|_2 = \sqrt{\sum_{i=1}^{n}\left(\Delta_{\alpha,i}\right)^2} \quad (11)$$

Ortalama Sağlamlık Ölçütü (MRM): Test kümesi D üzerinde kullanılan MRM, çekişmeli pertürbasyonların etkisine karşı bir sistemin dayanıklılığını değerlendiren bir metriktir. Bu ölçüt, bir sistemin dış etkilere ne kadar dirençli olduğunu belirlemeye yardımcı olmaktadır. Bu bağlamda, ρ değeri, test kümesindeki tüm pertürbasyonların ortalama büyüklüğünü hesaplamakta ve bu pertürbasyonların sistem üzerindeki genel etkisini ölçmektedir. MRM, özellikle modelin güvenilirliğini ve çeşitli koşullar altında performansını değerlendirmek amacıyla kullanılmaktadır. Pertürbasyonların büyüklüğünü ve etkisini bu şekilde analiz etmek, bir sistemin veya modelin çekişmeli saldırılara karşı ne kadar dayanıklı olduğunu anlamak için kritik öneme sahip olmaktadır.

$$\rho_{adv} = \frac{1}{|D|}\sum_{x \in D}\left(\frac{\|\Delta_\alpha\|_2}{\|x\|_2}\right) \quad (12)$$

1.7. Tezin Kapsamı

Bu yüksek lisans tezi, su sistemlerinin güvenliği kapsamında yaygınca kullanılan derin öğrenme modellerinin çekişmeli saldırılar karşısındaki davranışlarını incelemek ve bu saldırılara karşı savuma mekanizmasının geliştirilmesine odaklanmaktadır. Tez kapsamındaki çalışmalar literatürde yaygınca kullanılan SWaT ve WADI sistemlerinden elde edilmiş halka açık veri seti üzerinde gerçekleştirilmiştir. Tez çalışmasında, bu sistemlerde meydana gelebilecek aykırılıkları tespit etmek için CNN-LSTM modeli tabanlı bir yaklaşım kullanılmıştır. Modelin performansı; kesinlik, duyarlılık, F1-skoru ve YPO gibi metrikler ile kapsamlı bir şekilde değerlendirilmiştir. İlk aşamada, CNN-LSTM modeli ayrıntılı olarak ele alınmakta ve SWaT ile WADI sistemlerinde aykırılık tespiti için optimize edilmiştir. Modelin performansı, veri ön işleme, model eğitimi ve hiper parametre optimizasyonu gibi adımlarla iyileştirilmiştir.

İkinci aşamada, eyleyici ve sensör saldırıları gerçekleştirilmiş ve bu saldırıların CNN-LSTM modeli üzerindeki etkileri analiz edilmiştir. Bu kapsamda, tek nokta, süreç bazlı ve tüm süreç saldırıları incelenmiştir. Ayrıca, sabit ve değişken genlikli sensör saldırıları karşısında CNN-LSTM modelinin performansı değerlendirilmiştir.

Üçüncü aşamada, kural denetçisinin saldırılara karşı etkinliği değerlendirilmiş ve sistem güvenliğine katkıları incelenmiştir. Kural denetçili ve denetçisiz yapıların performans farkları karşılaştırılış, bu sayede sistemin zayıf noktaları ve bu noktaların iyileştirilmesine yönelik kritik bulgular elde edilmiştir. Tezin özgün katkısı, "OptAML" olarak adlandırılan, optimize edilmiş bir çekişmeli makine öğrenimi (AML) çerçevesinin geliştirilmedir. OptAML, modelin hassas bir şekilde manipüle edilmesine yönelik optimal pertürbasyonlar üretmektedir. Bu çerçeve hem sensör hem de eyleyici saldırıları yoluyla sistemi manipüle ederek modelin performansını düşürmeyi amaçlamaktadır. Son olarak, çekişmeli eğitim (AT) yöntemi kullanılarak modelin performansının artırılabileceği ve sistemin saldırılara karşı daha dirençli hâle getirilebileceği gösterilmiştir.

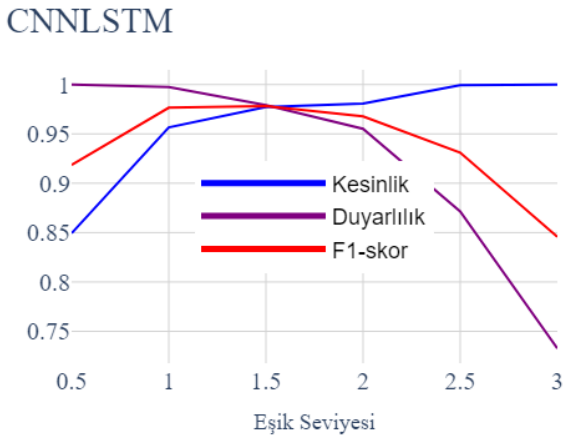
2. YAPILAN ÇALIŞMALAR

2.1. Eşik Değerinin Aykırılık Tespitine Etkisi

2.1.1. Su Arıtma Sistemlerinde Eşik Değerinin Aykırılık Tespitine Etkisi

SWaT sisteminde aykırılıkların tespiti amacıyla kullanılan CNN-LSTM modelinin kesinlik, duyarlılık ve F1-skoru gibi performans metrikleri, kullanılan eşik değer vektörüne bağlı olarak değişiklik göstermektedir. Eşik değer vektörünün (τ) etkisini incelemek için gerçekleştirilen deneylerde, bu vektör z-skor vektörü ile ilişkilendirilmekte ve bir ağırlık katsayısı (W) ile çarpılarak oluşturulmaktadır. Bu hesaplama μ_z×W şeklinde ifade edilmektedir; burada μ_z, z-skorunun ortalamasını ve W ise ağırlık katsayısını temsil etmektedir.

CNN-LSTM kullanılarak gerçekleştirilen aykırılık tespiti deneylerinde, farklı W değerlerine (0.5, 1, 1.5, 2, 2.5 ve 3) göre elde edilen kesinlik, duyarlılık ve F1-skor metrikleri Şekil 3'te gösterilmektedir. Bu metriklere ek olarak, YPO ile birlikte TN, TP, FN ve FP değerleri de Tablo 6'da detaylı bir şekilde sunulmaktadır.



Şekil 3. Su arıtma sisteminde kullanılan CNN-LSTM modelinin farklı eşik seviyelerine göre performans metrikleri

Eşik değeri, modelin bir durumu aykırı olarak sınıflandırmasını belirleyen kritik bir parametre olmaktadır. Eşik seviyesindeki değişiklikler, modelin performans metriklerini önemli ölçüde etkileyebilmektedir. Düşük eşik seviyelerinde daha fazla aykırılık tespit edilirken, yüksek eşik seviyelerinde tespit oranı azalmaktadır. Modelin, yüksek kesinlik ve

duyarlılık sunarak dengeli bir performans sergilemesini sağlamak amacıyla, eşik seviyesi W=1.5 olarak belirlenmiştir.

Şekil 3 ve Tablo 6 incelendiğinde, CNN-LSTM modelinin kesinlik, duyarlılık ve F1-skoru değerlerinin, düşük eşik değerlerinde (0.5, 1 ve 1.5) yaklaşık %85'in üzerinde bir performans sergilediği görülmektedir. Eşik değeri arttıkça, kesinlik değerinin yükseldiği ancak duyarlılık ve F1-skoru metriklerinin azaldığı gözlemlenmektedir. Bu durum, modelin aykırılık tespitinde daha az hatalı pozitif üretmekte başarılı olduğunu, ancak potansiyel aykırılıkların bir kısmını kaçırabileceğini göstermektedir.

Tablo 6. Su arıtma sisteminde CNN-LSTM modelinin farklı eşik seviyelerindeki aykırılık tespit performansının karşılaştırılması

Model	Eşik Seviyesi	TN	FP	TP	FN	Kesinlik	Duyarlılık	F1-skor	YPO
CNN-LSTM	$\mu_z x0.5$	85701	9299	52446	0	0.84940	1	0.91857	0.09788
	$\mu_z x1$	92622	2378	52309	137	0.95652	0.99739	0.97652	0.02503
	$\mu_z x1.5$	93803	1197	51349	1097	0.97722	0.97908	0.97815	0.01260
	$\mu_z x2$	94013	987	50096	2350	0.98068	0.95519	0.96777	0.01039
	$\mu_z x2.5$	94969	31	45704	6742	0.99932	0.87145	0.93102	0.00033
	$\mu_z x3$	95000	0	38422	14024	1	0.73260	0.84567	0

Tablo 7, literatürde yer alan ve SWaT sistemi üzerinde kullanılan çeşitli modellerin aykırılık tespit performanslarını karşılaştırmaktadır. Bu tabloda, kesinlik, duyarlılık ve F1-skoru değerleri incelenerek, farklı modellerin aykırılık tespitinde ne kadar etkili olduğu değerlendirilmektedir. Yapılan bu karşılaştırma, hangi modelin SWaT sistemlerinde aykırılıkları daha doğru ve güvenilir bir şekilde tespit edebildiğine dair önemli bilgiler sunmaktadır. Böylece, SWaT sisteminin güvenliği için en uygun modelin seçimi konusunda rehberlik sağlamaktadır.

Tablo 7. Su arıtma sisteminde kullanılan aykırılık tespit modellerinin performans karşılaştırması

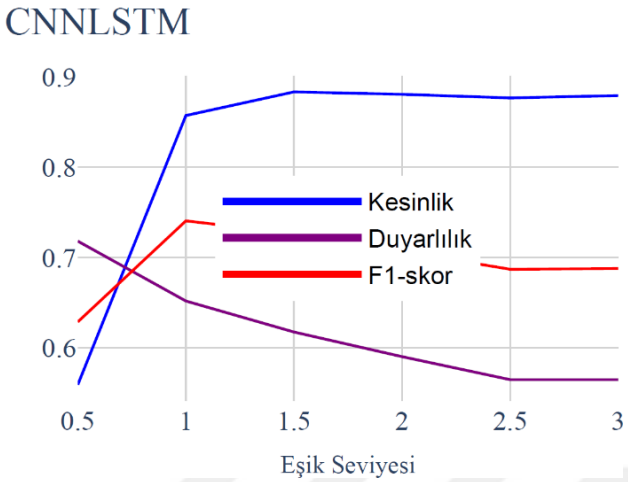
	Model	Kesinlik	Duyarlılık	F1-skor
Literatürde yapılan çalışmalar	SVM [19]	0.925	0.699	0.796
	DNN [19]	0.983	0.678	0.803
	MLP [60]	0.967	0.696	0.812
	MAD-GAN [22]	0.986	0.637	0.770
	STAE-AD [61]	0.960	0.815	0.880
	CNN+GRU [4]	0.997	0.833	0.919
	GDN [62]	0.993	0.681	0.810
	UAEF [63]	0.911	0.860	0.885
	NN-OC [3]	0.940	0.820	0.870
	LW-LSTM-VAE-M [23]	0.899	0.796	0.845
	InterFusion [64]	0.911	0.903	0.907
	USAD[65]	0.997	0.688	0.814
	AFNF [66]	0.964	0.854	0.906
	PGRGAT [67]	0.983	0.787	0.874
	OmniAnamoly [68]	0.814	0.843	0.828
	CNN[19]	1.000	0.853	0.920
	TABOR[69]	0.862	0.788	0.823
Proje kapsamında önerilen model	CNN-LSTM [8]	0.994	0.973	0.983

2.1.2. Su Dağıtma Sistemlerinde Eşik Değerinin Aykırılık Tespitine Etkisi

WADI sisteminde aykırılık tespiti amacıyla kullanılan CNN-LSTM modelinin kesinlik, duyarlılık ve F1-skoru metrikleri açısından performansı, seçilen eşik değer vektörüne bağlı olarak değişiklik göstermektedir. Eşik değer vektörünün etkisini değerlendirmek amacıyla, SWaT veri setinde uygulanan deneylerin benzerleri WADI veri setinde de gerçekleştirilmiş ve elde edilen sonuçlar Şekil 4 ve Tablo 9'da sunulmaktadır. Tablo 9, CNN-LSTM modelinin WADI sistemindeki aykırılık tespit performansını ayrıntılı bir şekilde göstermektedir.

Eşik değeri μ_z×1.5 olarak seçildiğinde, CNN-LSTM modelinin kesinlik, duyarlılık ve F1-skoru değerlerinin sırasıyla yaklaşık %88, %65 ve %73 olduğu gözlemlenmektedir. Eşik değeri arttıkça kesinlik değerinin yükseldiği ancak duyarlılık ve F1-skoru metriklerinin belirli bir eşik değerinden sonra azaldığı dikkat çekmektedir. Bu sonuçlar, SWaT veri setinde elde edilenlerle karşılaştırıldığında, CNN-LSTM modelinin WADI veri setindeki performansının görece daha düşük olduğunu göstermektedir. Modelin, yüksek kesinlik ve

duyarlılık arasında dengeli bir performans sergilemesi için eşik seviyesi 1.5 olarak belirlenmiştir.



Şekil 4. Su dağıtma sisteminde kullanılan CNN-LSTM modelinin farklı eşik seviyelerine göre performans metrikleri

Tablo 8, literatürde WADI sisteminde kullanılan çeşitli modellerin performans metriklerini sunmaktadır. Tablo 8 incelendiğinde, CNN-LSTM modelinin genel olarak diğer modellerden daha iyi performans gösterdiği görülmektedir. Bu durum, CNN-LSTM modelinin AML saldırılarına karşı davranışlarının incelenmesi için uygun bir temel sağladığını ortaya koymaktadır.

Tablo 8. Su dağıtma sisteminde kullanılan aykırılık tespit modellerinin performans karşılaştırması

	Model	Kesinlik	Duyarlılık	F1-skor
Literatürde yapılan çalışmalar	UAE [23]	0.916	0.640	0.754
	LW-LSTM-VAE-M [70]	0.727	0.313	0.437
	GRN [71]	0.358	0.740	0.483
	PCA [22]	0.395	0.056	0.100
	MAD-GAN [22]	0.414	0.339	0.370
	KNN [22]	0.078	0.077	0.080
	FB [22]	0.086	0.086	0.090
	GDN	0.975	0.402	0.570
	EGAN [22]	0.113	0.378	0.170
	PCA-Reconstruction[24]	0.763	0.571	0.653
	Windowed-PCA[24]	0.807	0.593	0.683
	1D CNN[24]	0.697	0.731	0.714
	AE[24]	0.834	0.681	0.750

Tablo 8'in devamı

Proje kapsamında önerilen model	CNN-LSTM [72]	0.883	0.617	0.727

Tablo 9'da CNN-LSTM modelinin WADI sistemindeki performansı farklı eşik değerleri için ayrıntılı olarak sunulmaktadır. Bu tablo, kesinlik, duyarlılık ve F1-skoru metriklerinin farklı eşik değerlerine göre nasıl değiştiğini göstermektedir. Eşik değeri $\mu_z \times 1.5$ olarak seçildiğinde, modelin kesinlik değeri artarken, duyarlılık ve F1-skoru metriklerinin daha düşük seviyelere indiği tekrar gözlemlenmiştir. Bu eğilim, modelin yüksek kesinlik ve düşük YPO elde etme pahasına bazı potansiyel aykırılıkları kaçırabileceğini göstermektedir.

Tablo 9. Su dağıtma sisteminde CNN-LSTM modelinin farklı eşik seviyelerindeki aykırılık tespit performansının karşılaştırılması

Model	Eşik Seviyesi	TN	FP	TP	FN	Kesinlik	Duyarlılık	F1-skor	YPO
CNN-LSTM	$\mu_z x0.5$	36609	5318	6739	2649	0.55893	0.71783	0.62849	0.12684
	$\mu_z x1$	40905	1022	6117	3271	0.85684	0.65158	0.74024	0.02438
	$\mu_z x1.5$	41159	768	5795	3593	0.88298	0.61728	0.72660	0.01832
	$\mu_z x2$	41174	753	5539	3849	0.88032	0.59001	0.70651	0.01796
	$\mu_z x2.5$	41179	748	5300	4088	0.87632	0.56455	0.68671	0.01784
	$\mu_z x3$	41197	730	5300	4088	0.87894	0.56455	0.68751	0.01741

2.2. Fiziksel Kısıtlar ve Değişmezler

2.2.1. Su Arıtma Sistemlerine Ait Fiziksel Kısıtlar ve Değişmezler

ICS'lerin güvenliğini ve işleyişini sağlamak için fiziksel kısıtlar ve değişmezler kritik öneme sahiptir. Fiziksel kısıtlar ve değişmezler, bir sistemin işleyişi boyunca sabit kalan ve her durumda doğru olan özellikleri veya koşulları tanımlamaktadır. Başka bir deyişle, bu kısıtlar ve değişmezler, bir sürecin belirli bir durumunda mutlaka doğru olması gereken fiziksel durumları ifade etmektedir. Fiziksel kısıtlar ve değişmezler göz ardı edildiğinde, sistemde oluşan aykırılıklar hata kontrol mekanizmaları tarafından tespit edilmektedir. Bu kısıtlardan ve değişmezlerden oluşturulan kural setleri, sensör ve eyleyici verilerine dayanarak sistemin normal ve aykırı durumlarını belirlemekte ve bu sayede aykırılıklara karşı dayanıklılığı artırmaktadır. Bu özellikleri sayesinde, fiziksel kısıtlar ve değişmezler,

aykırılıkların tespiti için etkin bir savunma mekanizması sunmakta ve sistem güvenliğinin önceden sağlanmasına olanak tanımaktadır.

Literatürde sunulan fiziksel kısıtlar ve değişmezler temel alınarak, SWaT sisteminde %100 doğruluk oranıyla çalışan 69 kural geliştirilmiştir. Bu kuralların etkinliği, SWaT sistemi üzerinde yapılan testlerle kanıtlanmıştır. Tablo 10, SWaT sistemine ait fiziksel kısıtların ve değişmezlerin tamamını detaylı olarak sunmaktadır. Tabloda, eşik değeri delta 0.5 olarak belirlenmiş ve bu eşik değerine göre sistemin normal ve aykırı durumları değerlendirilmektedir.

Tablo 10. Su arıtma sistemlerine ait fiziksel kısıtlar ve değişmezler

1	P602=2 => P302=1, MV304=1, MV302=1, MV301=2, FIT301<Delta, MV101=2, FIT101>Delta
2	MV301=2 => P302=1, MV304=1, MV302=1, FIT301<Delta, MV101=2, FIT101>Delta
3	FIT601>Delta => P302=1, MV304=1, MV303=2, MV302=1, FIT301<Delta, MV101=2, FIT101>Delta
4	MV304=2 => P602=1, FIT601<Delta, MV301=1
5	FIT301<Delta => MV302=1
6	MV201=1 => P205=1, P203=1, P101=1
7	FIT201<Delta => P205=1, P203=1
8	MV101=1 => P602=1, FIT601<Delta, MV303=1, MV301=1
9	MV101=2 => FIT101>Delta
10	P101=2 => MV201=2
11	MV302=2 => P602=1, FIT601<Delta, MV303=1, MV301=1, FIT301>Delta
12	P302=2 => FIT601<Delta, P602=1, MV301=1
13	MV303=1 => FIT601<Delta
14	MV301=1 => P602=1
15	MV301=1, MV302=2 => MV303=1, FIT301>Delta
16	MV302=2, MV303=1 => FIT301>Delta
17	MV303=1, FIT301>Delta => MV301=1
18	P602=2, FIT101>Delta => MV101=2, MV301=2, MV302=1, MV304=1, P302=1, FIT301<Delta
19	P602=2, MV101=2 => FIT301<Delta, MV301=2, MV302=1, MV304=1, P302=1
20	P602=2, FIT301<Delta => MV301=2, MV302=1, MV304=1, P302=1
21	P602=2, MV301=2 => MV302=1, MV304=1, P302=1
22	P602=2, MV304=1 => MV302=1
23	P602=2, P302=1 => MV302=1, MV304=1
24	MV301=2, FIT101>Delta => MV101=2, FIT301<Delta, MV302=1, MV304=1, P302=1
25	MV301=2, MV101=2 => FIT301<Delta, MV302=1, MV304=1, P302=1
26	MV301=2, FIT301<Delta => MV302=1, MV304=1, P302=1
27	MV301=2, MV302=1 => MV304=1, P302=1
28	MV301=2, MV304=1 => P302=1

Tablo 10'un devamı

29	FIT601>Delta, FIT101>Delta => MV101=2, FIT301<Delta, MV302=1, MV303=2, MV304=1, P302=1
30	FIT601>Delta, MV101=2 => FIT301<Delta, MV302=1, MV303=2, MV304=1, P302=1
31	FIT601>Delta, FIT301<Delta => MV302=1, MV303=2, MV304=1, P302=1
32	FIT601>Delta, MV302=1 => MV303=2, MV304=1, P302=1
33	FIT601>Delta, MV304=1 => MV303=2
34	FIT601>Delta, P302=1 => MV303=2, MV304=1
35	MV303=2, FIT301<Delta => MV101=2, MV302=1
36	MV303=2, MV302=1 => MV101=2
37	MV304=2, MV301=1 => FIT601<Delta, P602=1
38	MV304=2, FIT601<Delta => P602=1
39	MV201=1, P101=1 => P203=1, P205=1
40	MV201=1, P203=1 => P205=1
41	FIT201<Delta, P203=1 => P205=1
43	FIT101<Delta, MV301=1 => MV303=1, FIT601<Delta, P602=1
44	FIT101<Delta, MV303=1 => FIT601<Delta, P602=1
45	FIT101<Delta, FIT601<Delta => P602=1
46	MV101=1, MV301=1 => MV303=1, FIT601<Delta, P602=1
47	MV101=1, MV303=1 => FIT601<Delta, P602=1
48	MV101=1, FIT601<Delta => P602=1
49	MV302=2, FIT301>Delta => MV301=1, MV303=1
50	MV302=2, MV301=1 => MV303=1, FIT601<Delta, P602=1
51	MV302=2, FIT301>Delta => FIT601<Delta, P602=1
52	MV302=2, MV303=1 => FIT601<Delta, P602=1
53	MV302=2, FIT601<Delta => P602=1
54	P302=2, MV301=1 => FIT601<Delta, P602=1
55	P302=2, FIT601<Delta => P602=1
56	FIT301>Delta, MV303=1 => FIT601<Delta, P602=1
57	FIT301>Delta, FIT601<Delta => P602=1
58	LIT301 ≤800 => P101 or P102= 2
59	LIT301 ≥1000 => P101 or P102= 1
60	FIT201 ≤0.1 => P201, P202, P204, P206= 1
61	AIT201 > 260 => P201 or P202= 1
62	AIT503 ≥ 260 => P201 or P202= 1
63	AIT203 > 500 => P205 or P206= 1
64	LIT301 ≤ 250 => P301 or P302= 1
65	LIT401 ≥ 1000 => P301 or P302 =1
66	P401 or P402= 2 => FIT401 > delta
67	AIT402 ≤ 440 => P403 or P404= 1
68	UV401 =2 => P501 =2
69	P301= 2 => FIT301 > delta

2.2.2. Su Dağıtma Sistemlerine Ait Fiziksel Kısıtlar ve Değişmezler

Fiziksel kısıtlar ve değişmezler, bir sistemin belirli durumlarda koruması gereken temel koşulları ifade etmektedir. Sistemlerin fiziksel ve kimyasal durumlarının bu değişmezlere göre izlenmesi, beklenmeyen aykırılıkların tespiti için önemli bir temel teşkil etmektedir. Örneğin, bir su tankının seviyesi her zaman minimum ve maksimum ayar noktaları arasında kalmalıdır. Bu değişmez, sistemin normal işleyişi süresince korunması gereken bir koşul olmaktadır. Ayrıca, tankın giriş ve çıkış akışları da bu koşula bağlı olarak düzenlenmelidir. Örneğin, tank yüksek ayar noktasına ulaştığında giriş vanasının açık olması, aykırı bir durum olarak değerlendirilmektedir. Bu tür senaryolarda, değişmezler, sistemde meydana gelebilecek aykırılıkların tespit edilmesinde kritik bir rol oynamaktadır.

Tablo 11'de, WADI sistemi için belirlenen 9 fiziksel kısıt ve değişmez listelenmiştir. Bu kurallar, WADI üzerindeki hem normal hem de saldırı durumlarına ait verilerle test edilmiş ve %100'e yakın bir doğruluk oranı sergilemiştir. Elde edilen sonuçlar, geliştirilen kısıtların ve değişmezlerin sistemin güvenliğini sağlamada etkili olduğunu ortaya koymaktadır.

Tablo 11. Su dağıtma sistemleri için oluşturulan değişmez kısıtların normal ve saldırılı verideki doğruluk oranları

Kısıt Numarası	Değişmez Kısıtlar	Normal verideki doğruluk oranı (%)	Saldırılı verideki doğruluk oranı (%)
1	2_P_003: AÇIK ⇒ 2_MV_006: AÇIK	99.46	99.84
2	1_FIT_001(<0.5) v 1_AIT_005(>3) ⇒ (1_P_001 ∧ 1_P_003): KAPALI	100	100
3	2_MV_003: AÇIK v 2_MV_001: AÇIK ⇒ 1_P_005: AÇIK	99.84	100
4	1_LIT_001(<60) ⇒ 1_MV_004: KAPALI	99.79	99.85

Tablo 11'in devamı

5	1_LIT_001(<40) v {SU KALİTESİ ALARM DURUMU} v 2_LIT_001> 80 ⇒ 2_MV_001: KAPALI	100	100
6	3_FIT_001(<0.5) ⇒ 3_P_001: KAPALI v 3_P_003: KAPALI	100	100
7	2_MCV_101: KAPALI ∧ 2_MCV_201: KAPALI ∧ 2_MCV_301: KAPALI ∧ 2_MCV_401: KAPALI ∧ 2_MCV_501: KAPALI ∧ 2_MCV_601: KAPALI ⇒ 2_MV_006: KAPALI	99.89	100
8	2_LIT_002(>40) ⇒2_MV_004: AÇIK	100	100
9	2_MV_005: OTOMATİK OLARAK AÇIK	100	100

2.3. Veri Ön İşleme

ML, algoritmalar ve istatistiksel modeller kullanarak veri analizi yapmayı ve verilen girdi verilerine dayanarak sonuçları tahmin edebilecek modeller geliştirmeyi amaçlamaktadır. ML tabanlı modellerin performansı ve doğruluğu, kullanılan verilere büyük ölçüde bağımlıdır. Bu modellerin başarısı, verilerin kalitesi ve temsil kabiliyeti ile doğrudan ilişkilidir. Eğer ham veriler karmaşık, gereksiz ya da gürültü ile kirlenmişse, modelin uygulama amacına yönelik anlamlı bilgi öğrenme ve çıkarma yeteneği sınırlı kalabilmektedir [24].

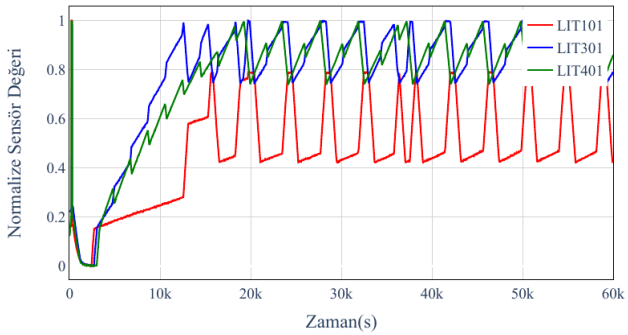
Veri ön işleme, veri bilimi ve ML çalışmalarında başarının sağlanabilmesi için kritik bir adım olarak kabul edilmektedir. Veri ön işlemenin temel adımları arasında veri temizleme, veri dönüşümü, veri entegrasyonu, veri azaltma ve veri bölme işlemleri yer almaktadır. Bu adımlar, ham verileri eksiksiz, tutarlı ve analiz edilebilir hâle getirmekte ve verilerin kalitesini artırarak daha doğru ve güvenilir analiz sonuçları elde edilmesine katkıda bulunmaktadır. Eksik ve hatalı verilerin temizlenmesi, verilerin ölçeklendirilmesi ve dönüştürülmesi, farklı veri kaynaklarının birleştirilmesi ve veri setlerinin uygun şekilde bölünmesi, analiz sürecinin etkinliğini ve doğruluğunu artırmaktadır [22].

Veri temizleme aşaması, eksik verilerin giderilmesi ve gürültünün azaltılması yoluyla analizlerin doğruluğunu artırmayı hedeflemektedir. Veri dönüşümü, verilerin standart bir formatta olmasını sağlayarak kıyaslama kabiliyetini ve analiz edilebilirliğini güçlendirmektedir. Veri entegrasyonu, farklı kaynaklardan gelen verilerin birleştirilmesi ve uyumlu hâle getirilmesiyle veri setlerinin bütünlüğünü sağlamaktadır. Veri azaltma ise gereksiz özelliklerin çıkarılması ve anlamlı olanların seçilmesi ile veri setinin boyutunu küçültmektedir. Bu aşama, analiz sürecini hızlandırırken model performansını da iyileştirmektedir. Son olarak, veri bölme işlemi, modelin performansını değerlendirmek ve doğrulamak amacıyla veri setinin eğitim ve test setlerine ayrılmasına olanak tanımaktadır.

Bu adımların doğru bir şekilde uygulanması, veri analizi ve modelleme süreçlerinin başarıyla gerçekleştirilmesini sağlamaktadır. Veri ön işleme sürecinin titizlikle yapılması, verilerin kalitesini artırarak daha güvenilir ve doğru sonuçlar elde edilmesine katkıda bulunmaktadır. Bu nedenle, veri ön işleme süreci, veri bilimi projelerinin vazgeçilmez bir parçası olarak kabul edilmektedir. Tüm süreçte yapılan her bir adımın dikkatlice ve özenle gerçekleştirilmesi büyük önem taşımaktadır [22].

2.3.1. Su Arıtma Sistemlerinde Veri Ön İşleme Süreci

SWaT sisteminde normal koşullar altında toplanan ve 496.800 örnek içeren eğitim verilerinin, LIT101, LIT301 ve LIT401 sensör verileri için 60.000 saniye ile sınırlandırılmış zaman-değer çizimleri Şekil 5'te sunulmaktadır.



Şekil 5. Su arıtma sisteminde başlangıç anında LITx seviye sensörü tarafından ölçülen veriler [8]

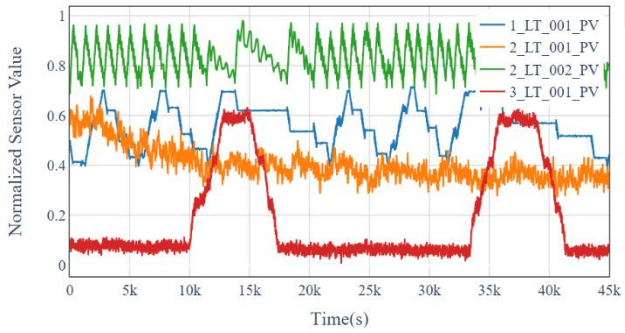
Grafikten gözlemlendiği üzere, sistem nominal çalışma noktasına yaklaşık 21.000 saniyede ulaştığı için eğitim verilerinin ilk 21.000 saniyelik kaydı çıkarılmış bulunmaktadır. Ardından, sürekli ve kategorik veriler, ortalama ve standart sapma değerleri kullanılarak (0,

1) aralığında normalize edilmiş ve bu normalizasyon işlemi, test verilerinin ölçeklendirilmesi için de referans olarak kaydedilmiştir. Ek olarak, UV401, P402 ve P501 kategorik değere sahip eyleyici verileri, normal koşullarda toplanan örneklerde 0 ve saldırı sırasında toplanan örneklerde 1 değerine sahip oldukları için eğitim sürecine dahil edilmemiştir. Normal veriler, %80'i eğitim ve %20'si test olmak üzere iki gruba ayrılmıştır. 36 farklı saldırı sonucu kaydedilen veriler ise yalnızca test aşamasında kullanılmaktadır. Bu çalışmada, SIU çalışmasında izlenen adımların dikkatlice uygulanmasına özen gösterilmiştir [8].

2.3.2. Su Dağıtma Sistemlerinde Veri Ön İşleme Süreci

WADI veri kümesinin ön işleme süreci, dört temel adımdan oluşmaktadır:

Eksik Verilerin Yönetimi: WADI veri kümesinde toplamda 127 özellik bulunmaktadır ve bu özelliklerden 9 tanesinde eksik veri satırları tespit edilmiştir. Bu 9 özelliğin 4'ü (2S001AL, 2LS002AL, 2P001STATUS, 2P002STATUS) tamamen eksik verilere sahiptir.



Şekil 6. Su dağıtma sisteminde başlangıç anında x_LT_xxx seviye sensörü tarafından ölçülen veriler [72]

Geriye kalan 5 özellikte ise az sayıda eksik veri satırı bulunmaktadır. Bu 5 özellikten 4'ünde (1AIT002PV, 1AIT004PV, 2BAIT004PV, 3AIT002PV), eksik verilerin bulunduğu noktaların öncesi ve sonrası lineer bir eğilim göstermektedir. Ancak, 1 özellik (3AIT004PV) lineer bir eğilim sergilememektedir. Lineer eğilim gösteren özellikler için eksik veriler, önceki ve sonraki değerlerin ortalaması alınarak tamamlanmıştır. Lineer eğilim göstermeyen özellikler için ise eksik veriler, aşamalı yükseltme regresyon modeli yöntemi ile tahmin edilmiştir.

Veri Görselleştirme: Sensör ve eyleyici verilerinin kaydedilmesi, sistemin çalıştırılması ile başlamaktadır. Bu durum, sistemdeki boruların ve tankların dolum aşaması

olan geçici durumu ifade etmektedir. Sistem kararlı duruma ulaştığında, sensör ve eyleyici verileri geçici durumdan farklılık göstermektedir. Geçici duruma ait veriler, çalışmanın başarı oranını olumsuz etkileyebileceği için veri kümesinin görselleştirilmesi ve sistemin başlangıç verilerinin dikkatle incelenmesi önemli olmaktadır. Şekil 6'da LT sensörlerinden elde edilen veriler görselleştirilmiştir. Şekil 4'teki veriler, yukarıda belirtilen geçici durumun veri kümesinde bulunmadığını göstermektedir. Bu nedenle, çalışmada veri kümesinin tamamı kullanılmıştır.

Sağlam Z-Skoru Yöntemi ile Aykırı Değer Tespiti: Z-skoru ve sağlam z-skoru yöntemleri, aykırı değer tespitinde yaygın olarak kullanılmaktadır. Z-skoru yöntemi, bir verinin ortalamadan sapmasını standart sapma cinsinden ölçer ve denklem (13)'te verilmektedir.

$$z = \frac{x - \mu(x)}{\sigma(x)} \tag{13}$$

Burada, x bir özelliğin değerini, $\mu(x)$ o özelliğin ortalama değerini ve $\sigma(x)$ ise özelliğin standart sapmasını temsil etmektedir. Z-skoru ile sağlam z-skoru yöntemleri arasındaki fark, medyan ile ortalama değer arasındaki farklılıktan kaynaklanmaktadır. Sağlam z-skoru yöntemi, denklem (14)'te verilmiştir.

$$rz = \frac{x - med(x)}{mad\ (x)} \tag{14}$$

Denklemde belirtilen $med(x)$ özelliğin medyanını, $mad(x)$ ise özelliğin medyan mutlak sapmasını temsil etmektedir. Aykırı değerler, bir veri kümesinin ortalama değerini önemli ölçüde etkileyebilmektedir ancak medyan bu değerlerden büyük ölçüde etkilenmediği için ortalamaya göre daha dayanıklı kabul edilmektedir [73]. Z-skoru yöntemi, aykırı değerlerden yüksek oranda etkilendiği için bu değerleri her zaman tespit edemeyebilmektedir [74]. Bu nedenle, sağlam z-skoru yöntemi, aykırı değerleri başarıyla tespit etmek için kullanılmaktadır.

Bu çalışmada, aykırı değerler her bir özellik için ayrı ayrı ele alınarak sağlam z-skoru yöntemi ile tespit edilmiştir. Veri kümesinin büyüklüğüne göre, yüksek aykırı değer oranına sahip özellikler veri kümesinden çıkarılmıştır. Özellikle %4 ve üzeri aykırı değere sahip

özellikler (2PIC003PV, 2PIT003PV, 2P003SPEED, 2HC301PV, 2FQ301PV, 2HC601PV, 2FQ601PV, 2F1C501PV, 2FQ501PV, 2FIC201PV, 2FQ201PV) analizden çıkarılarak, temiz bir veri kümesi elde edilmiştir.

Kolmogorov-Smirnov Testi ile Eğitim ve Test Veri Kümesinin Uyum İyiliği: Kolmogorov-Smirnov (KS), bir ampirik veri kümesinin teorik bir dağılımla uyumunu karşılaştırmak için kullanılan istatistiksel bir testtir [75]. Bu test ile iki farklı veri kümesi arasındaki uyum karşılaştırılmaktadır. KS testi, iki veri kümesinin kümülatif dağılım fonksiyonlarını (CDF) karşılaştırarak çalışmaktadır. Karşılaştırma sonucunda, dağılım fonksiyonları arasındaki maksimum fark belirlenmekte ve bu fark, KS test istatistiğini oluşturmaktadır. KS testinin denklemi, denklem (15)'te verilmektedir.

$$KS = max|F_1(x) - F_2(x)| \tag{15}$$

Denklemlerde belirtilen F(x) ve G(x), iki veri kümesinin CDF'lerini temsil etmektedir. CDF'nin hesaplanması aşağıda sunulmaktadır.

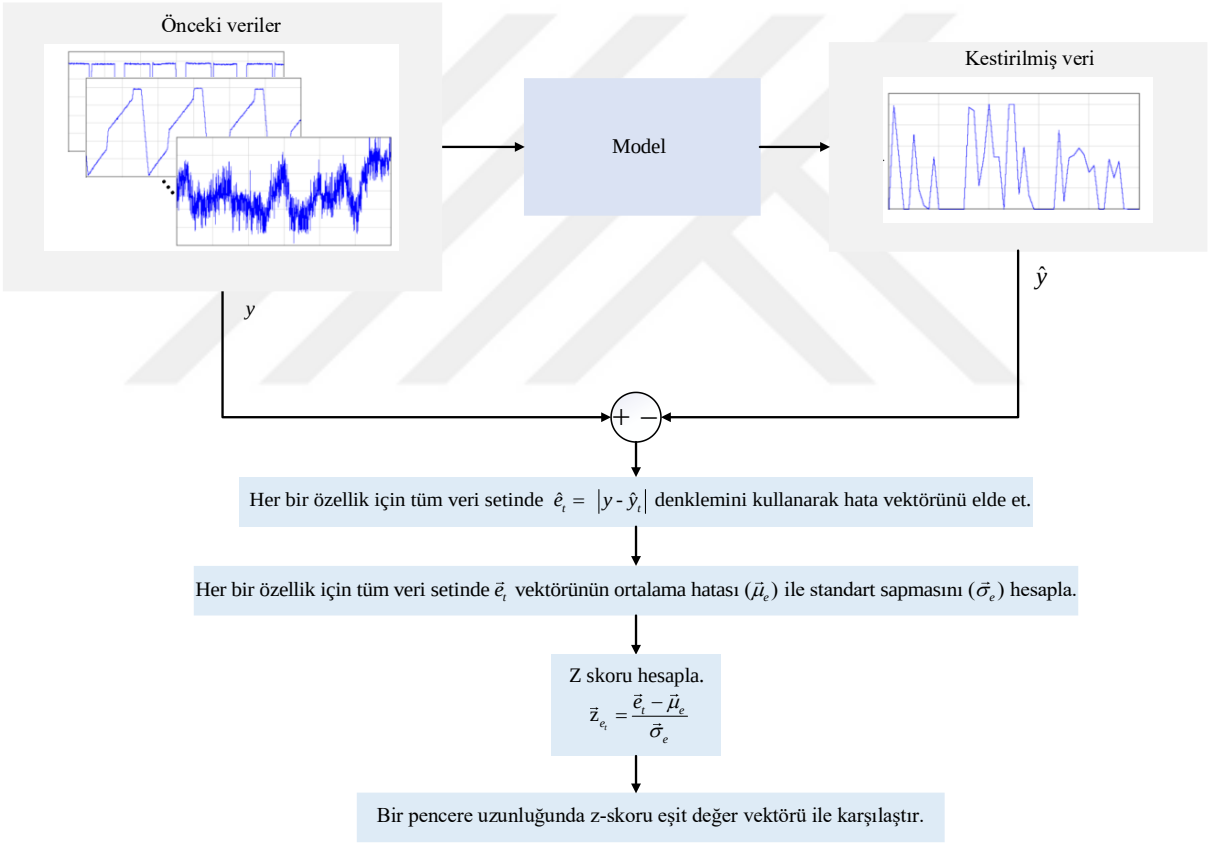
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i \leq x} \tag{16}$$

$$\delta_{x_i \leq x} = \begin{cases} 1 & x_i \leq x \\ 0 & aksi\ taktirde \end{cases} \tag{17}$$

Bu çalışmada, veri kümesi %80'i eğitim ve %20'si test olacak şekilde ikiye ayrılmaktadır. KS testi uygulanarak, eğitim ve test kümeleri karşılaştırılmakta ve iyi bir uyum göstermeyen özellikler veri kümesinden çıkarılmaktadır. Test sonuçlarına göre, 7 özellik (1AIT003PV, 3AIT005PV, 1AIT004PV, 1AIT001PV, 3AIT002PV, 2AAIT003PV, 2AAIT001PV) veri kümesinden çıkarılmaktadır. Bu çalışmada, ASYU çalışmasında uygulanan yöntemlere benzer veri işleme süreçleri gerçekleştirilmektedir [22].

2.4. Aykırılık Tespit Yaklaşımı

Aykırılık tespit yaklaşımı, temel olarak, sistemin önceki zamandaki sensör ve eyleyici verilerini kullanarak bir sonraki zaman adımında bu verilerin beklenen değerlerini kestirmeye dayanmaktadır. Derin ağ modeliyle kestirilen sensör ve eyleyici verileri, ilgili zamanda alınan gerçek ölçümlerle karşılaştırılarak bir hata vektörü oluşturulmaktadır. Bu hata vektörüne dayanarak, bir z-skor vektörü hesaplanmaktadır. Daha sonra, bu z-skor vektörü, deneysel olarak belirlenmiş eşik seviye vektörü ile karşılaştırılarak olası aykırılıklar tespit edilmektedir. SWaT ve WADI sistemlerinde aykırılık tespiti için kullanılan şema, Şekil 7'de gösterilmektedir.



Şekil 7. Aykırılık tespit yaklaşımı

Modelin çıkış aşamasında, geçmiş sensör ve eyleyici verileri kullanılarak gelecekteki değerler tahmin edilmektedir. Her bir özellik için zaman t'de tahmin edilen değer ($\hat{y}_t$) ile gerçek değer (y_t) arasındaki mutlak fark, $e_t = |y_t - \hat{y}_t|$ formülü ile hesaplanmaktadır. Bu farklardan oluşan hata vektörü ($\vec{e_t}$) oluşturulmaktadır. Modelin çıkış aşamasında, geçmiş sensör ve eyleyici verileri kullanılarak gelecekteki değerler tahmin edilmektedir. Her bir

özellik için zaman *t*'de tahmin edilen değer ($\hat{y}_t$) ile gerçek değer (y_t) arasındaki mutlak fark, $e_t = |y_t - \hat{y}_t|$ formülü ile hesaplanmaktadır. Bu farklardan oluşan hata vektörü ($\overrightarrow{e_t}$) oluşturulmaktadır.

Her bir özellik için, tüm veriler üzerinde tahmin hatasının ortalama değeri (μ_e) ve standart sapması (σ_e) hesaplanmaktadır. Bu hesaplamalar sonucunda, test aşamasında kullanılmak üzere $\overrightarrow{\mu_e}$ ve $\overrightarrow{\sigma_e}$ vektörleri elde edilmektedir. Her bir özelliği kapsayan z-skor vektörü, tahmin hatasının olasılığını temsil etmek üzere hesaplanmaktadır. Bu hesaplama, denklem 18'de verilmektedir.

$$\overrightarrow{z_{e_t}} = \frac{\overrightarrow{e_t} - \overrightarrow{\mu_e}}{\overrightarrow{\sigma_e}} \tag{18}$$

Saldırı tespiti için z-skor vektörü, önceden belirlenen bir eşik değer vektörü (τ) ile karşılaştırılmaktadır. Yanlış alarmların önüne geçmek için bu karşılaştırma, 10 uzunluklu bir pencere süresince yapılmaktadır. Bu pencere boyunca, z-skor değerlerinin eşik değerinden büyük olması gerekmektedir. τ vektörünün değeri, deneysel olarak incelenmekte ve aykırılık tespiti üzerindeki etkisi gösterilmektedir. Z-skor vektörü, tahmin hatalarının normalleştirilerek standart sapma biriminde ifade edilmesini sağlamakta; böylece, farklı ölçeklerdeki verilerde bile tutarlı bir aykırılık tespiti yapılabilmektedir. Ayrıca, pencere uzunluğu boyunca yapılan karşılaştırmalar, anlık sapmalardan kaynaklanan yanlış alarmların önüne geçerek daha güvenilir bir aykırılık tespiti sağlamaktadır.

Bu yaklaşımın avantajları arasında yüksek doğruluk ve geniş veri setlerinde etkili performans göstermesi yer almaktadır. Ancak, bu yöntemin yüksek hesaplama maliyeti ve büyük miktarda eğitim verisi gereksinimi gibi bazı dezavantajları da bulunmaktadır. Bu nedenle, yöntemin etkili bir şekilde uygulanabilmesi için yeterli hesaplama kaynaklarına ve kaliteli veri setlerine ihtiyaç duyulmaktadır.

2.5. CNN-LSTM Modeli

Aykırılık tespitinde kullanılmak üzere CNN-LSTM modeli tercih edilmektedir. CNN-LSTM modeli, zaman serisi verilerinde aykırılık tespiti için iki farklı DL mimarisinin avantajlarını birleştiren güçlü bir modeldir. Bu model, CNN ve LSTM katmanlarını kullanarak verilerin hem kısa hem de uzun vadeli bağımlılıklarını öğrenebilmektedir.

Modelin ilk aşaması, verilerin işlenmesi ve uygun formata dönüştürülmesidir. Zaman serisi verileri, genellikle sabit uzunluktaki zaman dilimlerine bölünmekte, ardından normalize edilmektedir. Bu süreç, verilerin ölçeklenmesini ve modellenmesini kolaylaştırarak CNN ve LSTM katmanlarının birlikte çalışmasına olanak tanımaktadır.

CNN katmanları, zaman serisi verilerindeki kısa vadeli bağımlılıkları ve örüntüleri çıkarmayı amaçlamaktadır. Giriş verisi üzerinde kayan filtreler kullanarak özellik haritaları oluşturmakta ve bu haritalar, verinin yerel örüntülerini ve kısa vadeli bağımlılıklarını yakalamaktadır. Konvolüsyon işlemi, filtrelerin veri üzerinde kaydırılması ve özellik haritalarının oluşturulması sürecini kapsamaktadır. CNN katmanlarında genellikle düzleştirilmiş doğrusal birim (ReLU) aktivasyon fonksiyonu kullanılmaktadır. ReLU, negatif değerleri sıfıra çekerken pozitif değerleri olduğu gibi bırakmakta ve böylece doğrusal olmayan ilişkilerin öğrenilmesini sağlamaktadır. CNN katmanlarının bir diğer önemli bileşeni olan havuzlama katmanı, özellik haritalarını küçülterek hesaplama maliyetini azaltmakta ve modelin genelleme yeteneğini artırmaktadır. Havuzlama işlemi, verinin boyutunu azaltırken en önemli özellikleri korumaktadır.

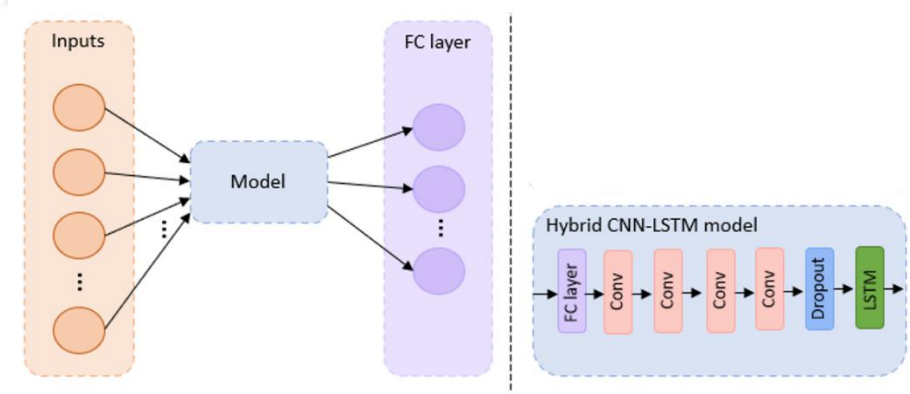
LSTM katmanları, zaman serisi verilerindeki uzun vadeli bağımlılıkları ve ardışık bilgileri modellemek için kullanılmaktadır. LSTM, geleneksel RNN modellerinin uzun vadeli bağımlılık sorunlarını aşmak amacıyla geliştirilmiş bir yapıdır. LSTM katmanlarının temel bileşenleri arasında hücre durumu, giriş kapısı, unutma kapısı ve çıkış kapısı bulunmaktadır. Hücre durumu, zaman serisi boyunca bilgiyi saklayan ve taşıyan bir bileşen olmaktadır. Giriş kapısı, yeni bilgilerin hücre durumuna ne kadar katkıda bulunacağını belirlerken; unutma kapısı, hücre durumundan hangi bilgilerin silineceğine karar vermektedir. Çıkış kapısı ise hücre durumundaki bilgilerden hangilerinin bir sonraki adımda çıkış olarak kullanılacağını belirlemektedir. Bu kapılar, verinin uzun vadeli bağımlılıklarını modellemek için kritik öneme sahip olup, LSTM katmanının güçlü hafıza kapasitesini sağlayan temel unsurlarını oluşturmaktadır.

LSTM katmanlarından çıkan veriler, tam bağlı katmanlara gönderilmektedir. Bu katmanlar, verilerin daha yüksek seviyede özellik çıkarımını yapmakta ve nihai çıktıyı üretmektedir. Genellikle, modelin son katmanı bir sınıflandırıcı veya regresyon katmanı olarak yapılandırılmaktadır. Model, normal verilerle eğitilerek, normal örüntüleri öğrenmeyi ve aykırılıkları tespit edebilecek bir yapıya sahip olmayı amaçlamaktadır. Eğitim sürecinde, kayıp fonksiyonu minimize edilerek modelin parametreleri optimize edilmektedir. Yaygın olarak kullanılan kayıp fonksiyonları arasında MSE ve çapraz entropi bulunmaktadır. Bu

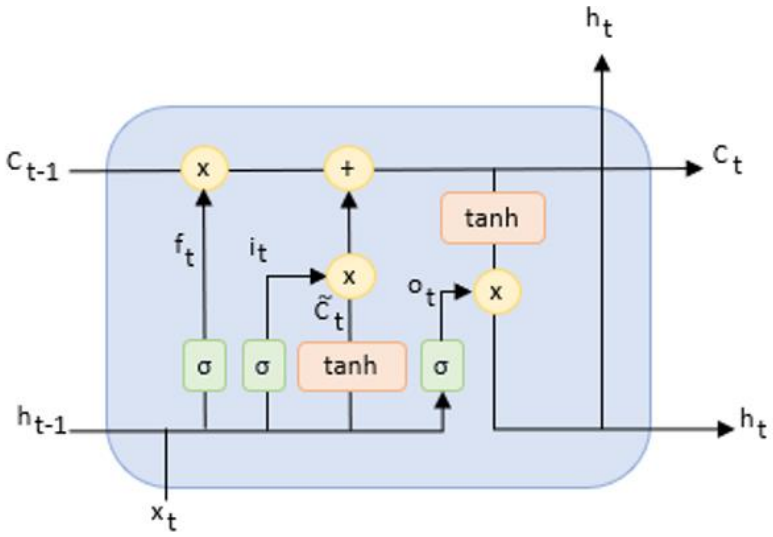
fonksiyonlar, modelin tahmin performansını iyileştirmek için hata oranını en aza indirmeye çalışmaktadır.

Modelin performansı, test verileri üzerinde değerlendirilmektedir. Aykırı veriler, modelin öğrenmiş olduğu normal örüntülerden sapmalar gösterdiğinde, model yüksek hata değerleri üretmektedir. Modelin performansı, kesinlik, duyarlılık, F1 skoru ve YPO gibi metrikler kullanılarak ölçülmektedir. CNN modelinin kısa vadeli bağımlılıkları ve örüntüleri çıkarma yeteneği ile LSTM modelinin uzun vadeli bağımlılıkları modelleme yeteneğinin birleşimi, elde edilen CNN-LSTM modelinin yüksek doğruluk ve dayanıklılık sağlamasına olanak tanımaktadır.

Sonuç olarak, CNN-LSTM modeli, zaman serisi verilerindeki aykırılıkların tespiti için güçlü bir model olmaktadır. Eğitim ve değerlendirme süreçlerinde uygun teknikler kullanıldığında, CNN-LSTM modeli aykırılık tespitinde etkili sonuçlar vermektedir. CNN-LSTM modeli, veri bilimi ve ML alanlarında aykırılık tespiti için yaygın olarak kullanılmakta ve çeşitli endüstriyel uygulamalarda başarıyla uygulanmaktadır. Şekil 8, aykırılık tespiti için kullanılan CNN-LSTM ağ mimarisini; Şekil 9 ise bu mimaride kullanılan LSTM hücresinin yapısını göstermektedir.



Şekil 8. Tahmin için kullanılan CNN-LSTM modelinin mimarisi



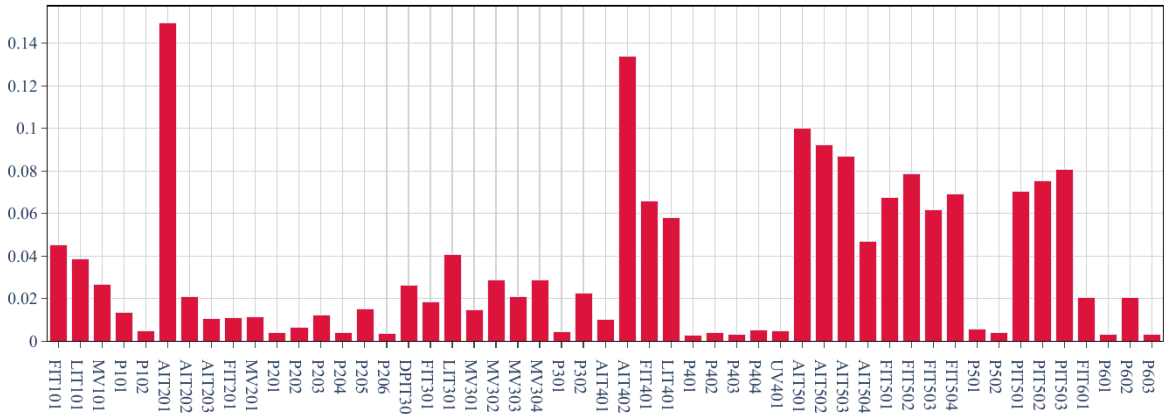
Şekil 9. LSTM hücresinin yapısı

Derin sinir ağı modelleri için SWaT sisteminde ağın girişi, 51 özellikten oluşan ve N uzunluğunda bir dizi ile beslenmektedir. Çıkışta ise 51 özellikten oluşan tek bir değer tahmin edilmektedir. WADI veri setinde ise ağın girişi, 105 özellikten oluşan ve N uzunluğunda bir dizi ile beslenmekte; çıkışta ise 105 özellikten oluşan tek bir değer tahmin edilmektedir. Deneyler sırasında, her iki sistem için de N değeri 30 olarak seçilmektedir. İlk iterasyonda, 1 ile 30. saniyeler arasındaki veriler ağın girişine verilmekte, bir sonraki iterasyonda ise bu örüntü, 2 ile 31. saniyeler şeklinde kaydırılarak devam etmektedir. Modelin çıkışında, çıkış özellik haritasının boyutunu istenen özellik sayısına indirmek için tam bağlantılı bir katman (FC) kullanılmaktadır.

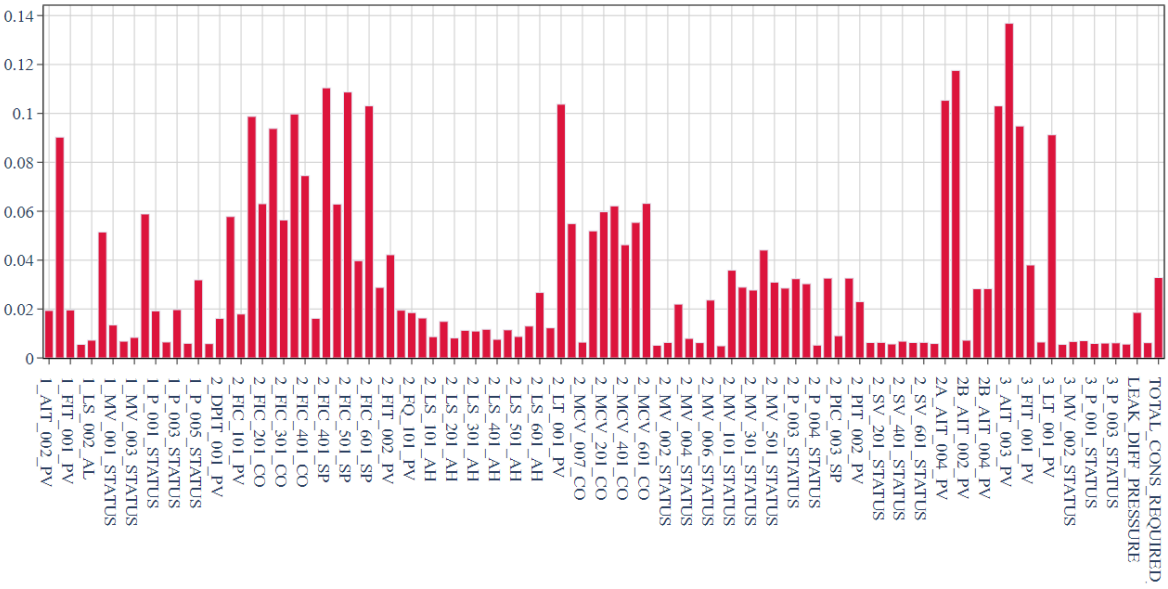
Giriş dizisi, özellikler arasındaki bağımlılıkları öğrenmek amacıyla, özellik sayısını üçe katlayan bir FC katmanından geçirilmektedir. FC katmanından sonra, uzamsal özellikleri çıkarmak için 1D konvolüsyonel katmanlar kullanılmaktadır. Bu katmanların giriş ve çıkış özellik haritalarının aynı boyutta olması için parametre değerleri; filtre boyutu 3, atlama boyutu 1 ve doldurma boyutu 1 olarak ayarlanmaktadır. Aşırı uyma problemini önlemek için konvolüsyonel katmanların ardından ağa bir dropout katmanı eklenmekte ve bir elemanın sıfır olma olasılığı $p = 0.2$ olarak belirlenmektedir. Dropout katmanının çıkışında elde edilen veriler, 256 nörona sahip 3 katmanlı LSTM modeline giriş olarak verilmekte ve zamansal özellikler bu katmanlar aracılığıyla çıkarılmaktadır.

2.5.1. CNN-LSTM Modelinin Eğitim Süreci

CNN-LSTM modelinin SWaT ve WADI sistemlerindeki sensör ve eyleyici verilerini tahmin etme performansı bu bölümde sunulmaktadır. Bu çalışmada, ağ modeli normal etiketli verilerle eğitilmekte ve modelin performansını değerlendirmek amacıyla test için ayrılan veriler kullanılmaktadır. SWaT sistemindeki tüm sensör ve eyleyici verileri için elde edilen MAE değerlerinin ortalaması 0.035 olarak hesaplanmıştır. MAE değeri, 0 ile sonsuz arasında bir değer almakta olup, MAE değerinin düşük olması tahminin daha başarılı olduğunu göstermektedir. Bu sonuç, önerilen CNN-LSTM modelinin sensör ve eyleyici verileri üzerinde yüksek bir tahmin performansı sergilediğini ortaya koymaktadır.



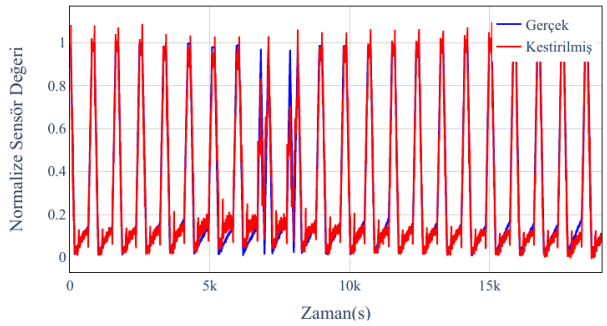
Şekil 10. Su arıtma sistemlerine ait sensör ve eyleyici verilerinin MAE değerleri [8]



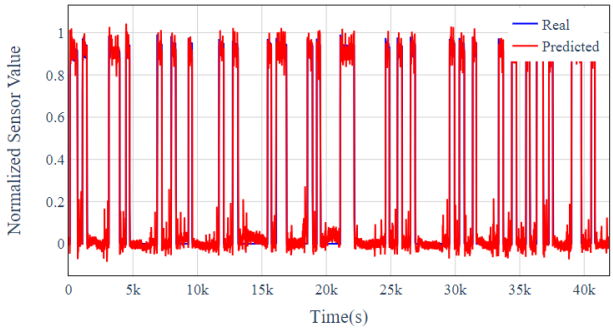
Şekil 11. Su dağıtma sistemlerinde sensör ve eyleyici verilerinin MAE değerleri [72]

Şekil 10 ve 11, SWaT ve WADI sistemlerindeki sensör ve eyleyici verilerinin MAE değerlerini detaylandırmaktadır. Bu şekiller, farklı sensör ve eyleyiciler için modelin tahmin ettiği değerler ile gerçek değerler arasındaki mutlak farkları göstermektedir. Düşük MAE değerleri, modelin ilgili bileşenler üzerindeki tahminlerinin doğruluğunun yüksek olduğunu ve bu bileşenlerde daha az hata ürettiğini ifade etmektedir. Şekillerin analizi, hangi sensör ve eyleyicilerin model tarafından daha iyi tahmin edildiğini ve hangilerinin iyileştirilmesi gerektiğini belirlemeye yardımcı olmaktadır. Özellikle yüksek MAE değerlerine sahip sensör ve eyleyiciler, modelin bu alanlardaki performansının artırılması gerektiğini göstermektedir.

Grafikler incelendiğinde, SWaT sisteminde en iyi tahminin P601 eyleyicisi için yapıldığı, en kötü tahminin ise AIT201 sensöründe elde edildiği görülmektedir. Örnek bir tahmin sonucu olarak, LIT101 ve 1FIT001PV sensörlerine ait gerçek ve tahmin edilen verileri gösteren grafikler, Şekil 12 ve Şekil 13'te sunulmaktadır. Şekil 12'den görüleceği üzere, 0.039 MAE değerine sahip LIT101 sensör verisi, gerçek değerine oldukça yakın bir şekilde tahmin edilmektedir.



Şekil 12. Su arıtma sistemlerinde LIT101 sensörüne ait gerçek ve tahmin edilen veriler [8]



Şekil 13. Su dağıtma sistemlerinde 1FIT001PV sensörüne ait gerçek ve tahmin edilen veriler [72]

2.6. Eyleyici Saldırıları ve Sonuçları

Eyleyici saldırıları, ICS'lerin kritik bileşenleri olan eyleyicilere yönelik gerçekleştirilen ve bu bileşenlerin normal işleyişini bozarak modelin genel performansına ciddi zararlar verebilen saldırılar olmaktadır. SWaT ve WADI sistemlerinde, farklı kapsam ve etki düzeylerine sahip eyleyici saldırı senaryoları uygulanmakta ve bu saldırılar, modelin güvenliğini test etmek amacıyla gerçekleştirilmektedir. Bu çalışmada, uygulanan üç farklı eyleyici saldırı türü detaylı bir şekilde incelenmektedir: tek nokta eyleyici saldırısı, süreç bazlı eyleyici saldırısı ve tüm süreç eyleyici saldırısı.

ICS'nin güvenliği ve operasyonel sürekliliği açısından büyük önem taşıyan eyleyici saldırıları, SWaT ve WADI sistemlerinde gerçekleştirilen deneylerle analiz edilmektedir. Tek nokta, süreç bazlı ve tüm süreç eyleyici saldırıları, bu tür saldırıların etkilerini anlamak ve bu etkileri en aza indirmek için kritik veriler sağlamaktadır. Bu tür saldırılar, sistemin zayıf noktalarını belirlemek ve daha güvenilir modeller geliştirmek amacıyla önemli bilgiler sunmaktadır. Ayrıca, bu saldırılara karşı CNN-LSTM modelinin performansını değerlendirmek, modelin güvenilirliğini ve etkinliğini ölçmek açısından oldukça önemli olmaktadır.

2.6.1. Tek Nokta Eyleyici Saldırısı

Tek nokta eyleyici saldırısı, SWaT ve WADI sistemlerinde belirli bir eyleyiciye odaklanarak gerçekleştirilen ve bu eyleyicinin işlevini manipüle eden bir saldırı türü olmaktadır. Bu tür saldırılar, tek bir eyleyicinin normal işleyişini bozarak genel sistem performansını ciddi şekilde olumsuz etkileyebilmektedir. Bu bağlamda, tek nokta eyleyici saldırıları genellikle "tersleme saldırısı" olarak adlandırılmakta ve eyleyicinin mevcut durumunu tersine çevirerek sistemin normal işleyişine müdahale etmektedir. Örneğin, SWaT sistemlerinde normalde açık olan bir valfin kapatılması veya kapalı olan bir valfin açılması gibi eylemler, sistemin güvenliğini riske atabilmektedir.

Tek nokta eyleyici saldırıları, belirli bir eyleyiciye odaklandıkları ve hedeflenmiş bir değişiklik içerdiği için aykırılık tespit modellerinin bu saldırıları fark etmesini zorlaştırmaktadır. Özellikle saldırı küçük bir değişiklikle gerçekleştiğinde, standart eşik tabanlı veya istatistiksel aykırılık tespit modelleri tarafından gözden kaçabilmektedir. Bu durum, saldırıların tespit edilmesini güçleştirerek potansiyel hasar riskini artırmaktadır.

Bu tür saldırılara karşı savunma mekanizmaları genellikle aykırılık tespiti ve çeşitli izleme tekniklerine dayanmaktadır. DL tabanlı modeller ve kural tabanlı sistemler, bu saldırıların belirlenmesinde kullanılmakta ve birlikte kullanıldıklarında daha etkili sonuçlar sağlamaktadır.

Örneğin, kritik bir kimyasal dozaj pompası (P601) veya su akışını kontrol eden bir valf gibi eyleyicilere yönelik bir saldırı senaryosu tasarlanabilmektedir. Bu senaryoda, saldırgan, eyleyicilerin sürekli açık kalmasını sağlayarak suya aşırı kimyasal karışmasına neden olabilmekte veya suyun belirli bir noktada aşırı ya da yetersiz akmasına yol açabilmektedir. Bu durum, suyun kimyasal dengesinin bozulmasına, su kalitesinin düşmesine ve sistem performansının olumsuz etkilenmesine yol açmaktadır. Yanıltıcı bilgiler alan sistem operatörleri, hatalı kararlar vererek operasyonel verimliliğin düşmesine ve maliyetlerin yükselmesine sebep olabilmektedir.

a) Su Arıtma Sistemi - SWaT

Tablo 12, CNN-LSTM modelinin çeşitli saldırı noktaları altındaki performans değerlendirmelerini içermektedir. Genel olarak, modelin çoğu saldırı noktası için yüksek kesinlik (0.97 ve üzeri) ve düşük YPO (0.03 ve altında) değerleri sergilediği gözlemlenmektedir. Bu durum, CNN-LSTM modelinin genellikle doğru tahminler yaptığını ve düşük YPO değerlerine sahip olduğunu göstermektedir.

Bununla birlikte, bazı saldırı noktalarında modelin zayıf performans gösterdiği belirlenmektedir. Özellikle P-101, P-203, P-205 ve P-302 eyleyicileri bu bağlamda dikkat çekmektedir. Örneğin, P-101 için kesinlik değeri 0.67901 iken, YPO değeri 0.25552 olarak hesaplanmaktadır. Benzer şekilde, P-302 için kesinlik değeri 0.45096, YPO değeri ise 0.65806'dır. Bu veriler, modelin bu noktalar için oldukça düşük performans sergilediğini ve yüksek YPO değerlerine sahip olduğunu ortaya koymaktadır. Bu saldırı noktaları, modelin en zayıf olduğu ve saldırılara karşı en savunmasız kaldığı yerler olarak değerlendirilmektedir.

Öte yandan, MV-101, P-102, MV-201, P-202 ve P-204 gibi birçok noktada model, yüksek kesinlik (yaklaşık 0.97 ve üzeri) ve düşük YPO değerleri (yaklaşık 0.02) ile iyi performans göstermektedir. Bu durum, modelin bu saldırı noktalarına karşı oldukça dayanıklı olduğunu ve doğru sınıflandırma yaptığını göstermektedir. Modelin belirli saldırı noktalarındaki performansı, sistem güvenliği açısından kritik bir öneme sahip olmaktadır. Örneğin, P-302 gibi yüksek YPO değerine sahip noktalar, modelin saldırılara karşı

etkinliğini azaltabilmekte ve bu noktaların güvenlik önlemlerinin güçlendirilmesi gerekebilmektedir.

Tablo 12, CNN-LSTM modelinin tek nokta eyleyici saldırıları altında, çeşitli saldırı noktalarındaki performansını kapsamlı bir şekilde özetlemektedir. Genel olarak, modelin çoğu noktada iyi performans sergilediği, ancak P-101, P-203, P-205 ve P-302 gibi belirli noktalarda zayıf kaldığı görülmektedir. Sonuçlar, modelin zayıf kaldığı saldırı noktalarının belirlenmesi ve bu noktalara yönelik güvenlik önlemlerinin geliştirilmesi gerektiğini ortaya koymaktadır.

Tablo 12. Su arıtma sisteminde tek nokta eyleyici saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	CNN-LSTM			
	TN	FP	Kesinlik	YPO
MV-101	93826	1174	0.97765	0.01236
P-101	70726	24274	0.67901	0.25552
P-102	93798	1202	0.97713	0.01265
MV-201	93781	1219	0.97681	0.01283
P-201	92984	2016	0.96222	0.02122
P-202	93789	1211	0.97905	0.01275
P-203	83637	11363	0.81881	0.11961
P-204	93806	1194	0.97728	0.01257
P-205	80243	14757	0.77676	0.15534
P-206	93793	1207	0.97703	0.01271
MV-301	93809	1191	0.97733	0.01254
MV-302	93798	1202	0.97711	0.01265
MV-303	93784	1216	0.97687	0.01280
MV-304	93712	1288	0.97553	0.01356
P-301	93813	1187	0.97741	0.01249
P-302	32484	62516	0.45096	0.65806
P-401	92283	2717	0.94975	0.02860
P-402	93708	1292	0.97546	0.01360
P-403	93809	1191	0.97733	0.01254
P-404	93561	1439	0.97274	0.01515
UV-401	93719	1281	0.97566	0.01348
P-501	93729	1271	0.97585	0.01338
P-502	93747	1253	0.97618	0.01319
P-601	93755	1245	0.97633	0.01311
P-602	93701	1299	0.97533	0.01367

b) Su Dağıtma Sistemi - WADI

Tablo 13, WADI sisteminde tek nokta eyleyici saldırıları altında CNN-LSTM modelinin performansını detaylı bir şekilde sunmaktadır. Bu tablo, saldırı noktaları için TN, FP, kesinlik ve YPO değerlerini içermektedir.

Tablo incelendiğinde, tüm saldırı noktaları için TN değerinin 40905 ve FP değerinin 1022 olduğu görülmektedir. Bu durum, modelin saldırı noktalarında tutarlı bir performans sergilediğini göstermektedir. Kesinlik değerinin tüm saldırı noktaları için 0.857 olması, modelin pozitif olarak sınıflandırdığı örneklerin yaklaşık %85'inin gerçekten pozitif olduğunu ifade etmektedir. Bu yüksek kesinlik oranı, modelin doğru tahmin yapma yeteneğinin güçlü olduğunu göstermektedir.

YPO değerinin 0.02 olarak belirlenmesi, modelin negatif örnekleri pozitif olarak sınıflandırma oranının düşük olduğunu göstermektedir. Düşük YPO değeri, modelin yanlış alarm üretme oranının düşük olduğunu ve güvenilir bir aykırılık tespit performansı sunduğunu ortaya koymaktadır.

1_MV_001_STATUS, 1_MV_002_STATUS ve 1_MV_003_STATUS gibi saldırı noktalarında aynı performans değerlerinin elde edilmesi, modelin bu noktalarda istikrarlı ve güvenilir bir performans sergilediğini göstermektedir. Benzer şekilde, 2_FQ_101_PV, 2_FQ_401_PV ve 2_MCV_007_CO gibi diğer saldırı noktalarında da benzer performans sonuçlarının gözlemlenmesi, modelin yalnızca belirli noktalarda değil, sistem genelinde de tutarlı ve yüksek performans gösterdiğini kanıtlamaktadır.

Sonuç olarak, CNN-LSTM modelinin WADI veri setindeki tek nokta eyleyici saldırıları karşısında sergilediği performans, modelin aykırılık tespitinde güçlü ve tutarlı olduğunu ortaya koymaktadır. Tüm saldırı noktalarında elde edilen yüksek kesinlik ve düşük YPO değerleri, modelin yanlış alarmları en aza indirerek doğru sınıflandırma yapma kabiliyetini desteklemektedir.

Tablo 13. Su dağıtma sisteminde tek nokta eyleyici saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	CNN-LSTM			
	TN	FP	Kesinlik	YPO
1_MV_001_STATUS	40905	1022	0.85684	0.02438
1_MV_002_STATUS	40905	1022	0.85684	0.02438
1_MV_003_STATUS	40905	1022	0.85684	0.02438
1_MV_004_STATUS	40905	1022	0.85684	0.02438
1_P_001_STATUS	40905	1022	0.85684	0.02438
1_P_002_STATUS	40905	1022	0.85684	0.02438

Tablo 13'ün devamı

1_P_003_STATUS	40905	1022	0.85684	0.02438
1_P_004_STATUS	40905	1022	0.85684	0.02438
1_P_005_STATUS	40905	1022	0.85684	0.02438
1_P_006_STATUS	40905	1022	0.85684	0.02438
2_FQ_101_PV	40905	1022	0.85684	0.02438
2_FQ_401_PV	40905	1022	0.85684	0.02438
2_MCV_007_CO	40905	1022	0.85684	0.02438
2_MCV_101_CO	40905	1022	0.85684	0.02438
2_MCV_201_CO	40905	1022	0.85684	0.02438
2_MCV_301_CO	40905	1022	0.85684	0.02438
2_MCV_401_CO	40905	1022	0.85684	0.02438
2_MCV_501_CO	40905	1022	0.85684	0.02438
2_MCV_601_CO	40905	1022	0.85684	0.02438
2_MV_001_STATUS	40905	1022	0.85684	0.02438
2_MV_002_STATUS	40905	1022	0.85684	0.02438
2_MV_003_STATUS	40905	1022	0.85684	0.02438
2_MV_004_STATUS	40905	1022	0.85684	0.02438
2_MV_005_STATUS	40905	1022	0.85684	0.02438
2_MV_006_STATUS	40905	1022	0.85684	0.02438
2_MV_009_STATUS	40905	1022	0.85684	0.02438
2_MV_101_STATUS	40905	1022	0.85684	0.02438
2_MV_201_STATUS	40905	1022	0.85684	0.02438
2_MV_301_STATUS	40905	1022	0.85684	0.02438
2_MV_401_STATUS	40905	1022	0.85684	0.02438
2_MV_501_STATUS	40905	1022	0.85684	0.02438
2_MV_601_STATUS	40905	1022	0.85684	0.02438
2_P_003_STATUS	40905	1022	0.85684	0.02438
2_P_004_SPEED	40905	1022	0.85684	0.02438
2_P_004_STATUS	40905	1022	0.85684	0.02438
2_SV_101_STATUS	40905	1022	0.85684	0.02438
2_SV_201_STATUS	40905	1022	0.85684	0.02438
2_SV_301_STATUS	40905	1022	0.85684	0.02438
2_SV_401_STATUS	40905	1022	0.85684	0.02438
2_SV_501_STATUS	40905	1022	0.85684	0.02438
2_SV_601_STATUS	40905	1022	0.85684	0.02438
3_MV_001_STATUS	40905	1022	0.85684	0.02438
3_MV_002_STATUS	40905	1022	0.85684	0.02438
3_MV_003_STATUS	40905	1022	0.85684	0.02438
3_P_001_STATUS	40905	1022	0.85684	0.02438
3_P_002_STATUS	40905	1022	0.85684	0.02438
3_P_003_STATUS	40905	1022	0.85684	0.02438
3_P_004_STATUS	40905	1022	0.85684	0.02438

2.6.2. Süreç Bazlı Eyleyici Saldırısı

Süreç bazlı eyleyici saldırıları, SWaT ve WADI sistemlerinde belirli bir sürece dahil olan tüm eyleyicileri hedef alan, koordineli ve senkronize saldırılardır. Bu tür saldırılar, bir sürece bağlı olan eyleyicilerin işlevlerini manipüle ederek sistemin normal işleyişini kesintiye uğratmayı ve genel performansı olumsuz yönde etkilemeyi amaçlamaktadır. Süreç bazlı saldırılar, çoğunlukla daha karmaşık bir yapıya sahip olup, tespiti tek bir noktaya odaklanan saldırılara kıyasla daha zor olmaktadır. Bunun temel nedeni, birden fazla eyleyicinin koordineli bir şekilde manipüle edilmesi durumunun, aykırılık tespit modelleri için zorlu bir senaryo oluşturmasıdır.

Bu tür saldırılarda, saldırganlar öncelikle hedef alınacak süreci belirlemekte ve bu süreçte yer alan tüm eyleyicileri tespit etmektedir. Eyleyicilerin durumları, tersleme saldırısı olarak bilinen yöntemle değiştirilmekte; örneğin, normalde açık olan bir eyleyici kapalı duruma getirilmekte veya kapalı olan bir eyleyici açılmaktadır. Bu yöntemle, sürecin tüm eyleyicileri eş zamanlı olarak yanlış komutlar almakta ve koordineli bir manipülasyon gerçekleştirilmektedir. Saldırının etkileri ise sistemin genel işleyişi üzerinde gözlemlenmekte ve analiz edilmektedir. Bu tür saldırılara karşı, sistem operatörlerinin süreç izleme ve erken uyarı sistemleri kullanarak hızlı bir şekilde yanıt verebilmesi kritik öneme sahiptir.

Örneğin, SWaT sistemindeki P2 sürecine yönelik bir saldırı senaryosu incelendiğinde, saldırganın bu süreçte yer alan tüm eyleyicilerin durumlarını tersleyerek sistem dengesini bozduğu görülebilmektedir. Böyle bir durumda, kimyasal dozlama pompalarının işleyişi etkilenebilmekte, bu da suyun kimyasal dengesinin bozulmasına ve nihayetinde su kalitesinin ciddi şekilde düşmesine neden olabilmektedir.

a) Su Arıtma Sistemi - SWaT

Tablo 14, CNN-LSTM modelinin SWaT sistemindeki süreç bazlı eyleyici saldırıları altında sergilediği performansı ayrıntılı bir şekilde sunmaktadır. Tabloya göre, Süreç-1 ve Süreç-2'de modelin performans seviyeleri benzerlik göstermektedir. Süreç-1 için kesinlik değeri 0.679ve YPO değeri 0.255 olarak hesaplanırken, Süreç-2'de kesinlik değeri 0.680 ve YPO değeri 0.254 olarak belirlenmektedir. Bu sonuçlar, her iki sürecin de modelin düşük doğrulukla çalıştığı ve yüksek YPO değerine sahip olduğu noktalar olduğunu göstermektedir. Bu durum, modelin bu süreçlerdeki saldırıları tespit etmekte zorlandığını ve yanlış pozitif alarm oranının yüksek olduğunu ortaya koymaktadır.

Süreç-3, CNN-LSTM modelinin en düşük performans sergilediği süreç olarak öne çıkmaktadır. Bu süreçte kesinlik değeri 0.429 ve YPO değeri 0.719 olarak belirlenmiştir. Bu veriler, modelin Süreç-3'te saldırıları tespit etme konusunda ciddi zorluklar yaşadığını ve YPO değerinin oldukça yüksek olduğunu göstermektedir. Süreç-3'teki bu düşük performans, modelin bu sürece özgü güvenlik zayıflıkları olabileceğine işaret etmektedir.

Öte yandan, Süreç-4 ve Süreç-5, modelin yüksek performans gösterdiği süreçler olarak dikkat çekmektedir. Süreç-4'te kesinlik değeri 0.844 ve YPO değeri 0.099 olarak kaydedilirken, Süreç-5'te kesinlik değeri 0.974 ve YPO değeri 0.014 olarak hesaplanmıştır. Bu sonuçlar, derin ağ modelinin bu süreçlerde saldırıları tespit etmede oldukça başarılı olduğunu ve düşük YPO değerine sahip olduğunu göstermektedir. Özellikle Süreç-5, modelin en yüksek performans sergilediği süreç olarak öne çıkmaktadır.

Son olarak, Süreç-6'da kesinlik değeri 0.947 ve YPO değeri 0.029 olarak belirlenmiştir. Bu değerler, modelin bu süreçte de yüksek doğrulukla çalıştığını ve düşük YPO değerine sahip olduğunu ortaya koymaktadır. Süreç-6'da modelin performansı, Süreç-4 ve Süreç-5 kadar yüksek olmasa da oldukça tatmin edici bir düzeydedir.

Genel olarak, Tablo 14'te sunulan veriler, CNN-LSTM modelinin SWaT veri setindeki performansının süreçler arasında önemli farklılıklar gösterdiğini ortaya koymaktadır. Özellikle Süreç-3'teki düşük performans ve Süreç-5'teki yüksek performans dikkat çekicidir.

Tablo 14. Su arıtma sisteminde süreç bazlı eyleyici saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	CNN-LSTM			
	TN	FP	Kesinlik	YPO
Süreç-1	70742	24258	0.67913	0.25535
Süreç-2	70849	24151	0.68012	0.25422
Süreç-3	26723	68277	0.42924	0.71871
Süreç-4	85539	9461	0.84441	0.09959
Süreç-5	93631	1369	0.97403	0.01441
Süreç-6	92171	2829	0.94778	0.02978

b) Su Dağıtma Sistemi – WADI

Tablo 15, WADI veri setindeki süreç bazlı eyleyici saldırıları altında CNN-LSTM modelinin performansını detaylı bir şekilde sunmaktadır. Tablonun incelenmesi, CNN-

LSTM modelinin süreç bazlı eyleyici saldırılarını tespit etme yeteneğinin oldukça düşük olduğunu göstermektedir. Süreç-1 için TN değeri 259, FP değeri ise 41668 olarak hesaplanmıştır. Bu süreçte, kesinlik değeri 0.183 ve YPO değeri 0.993 olarak belirlenmektedir. Düşük kesinlik değeri, modelin TP tespit yapma kapasitesinin zayıf olduğunu ortaya koymaktadır. Yüksek YPO değeri ise modelin çok sayıda yanlış alarm ürettiğini ve güvenilir bir performans sunmadığını göstermektedir.

Süreç-2'de modelin performansı daha da düşüktür. Bu süreçte, TN değeri sadece 32 iken, FP değeri 41895 olarak hesaplanmıştır. Kesinlik değeri 0.183 ve YPO değeri 0.999 olarak belirlenmiştir. Bu durum, modelin neredeyse hiç TN tespit edemediğini ve negatif örneklerin neredeyse tamamını FP olarak sınıflandırdığını göstermektedir.

Süreç-3 ise en düşük performansı sergilemektedir. Bu süreçte, TN değeri 10, FP değeri ise 41917 olarak belirlenmiştir. Kesinlik değeri 0.182, YPO değeri ise 0.999 olarak hesaplanmıştır. Bu sonuçlar, modelin bu süreçte de neredeyse tamamen başarısız olduğunu ve doğru tespit yapma yeteneğinin son derece zayıf olduğunu göstermektedir.

Genel olarak, Tablo 15'te sunulan veriler, CNN-LSTM modelinin WADI sistemindeki süreç bazlı eyleyici saldırılarını tespit etmede yetersiz kaldığını ortaya koymaktadır. Model, yüksek YPO ve düşük kesinlik değerleri ile dikkat çekmekte olup, TP tespit yapma yeteneği oldukça sınırlı kalmaktadır. Bu düşük performans, modelin süreç bazlı eyleyici saldırıları tespitinde yetersiz kaldığını ve özellikle yüksek YPO değeri nedeniyle güvenilirlikten uzak olduğunu göstermektedir.

Tablo 15. Su dağıtma sisteminde süreç bazlı eyleyici saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	CNN-LSTM			
	TN	FP	Kesinlik	YPO
Süreç-1	259	41668	0.18388	0.99382
Süreç-2	32	41895	0.18306	0.99924
Süreç-3	10	41917	0.18298	0.99976

2.6.3. Tüm Süreç Eyleyici Saldırısı

Tüm süreç eyleyici saldırıları, SWaT ve WADI sistemlerindeki tüm eyleyicilere yönelik gerçekleştirilen kapsamlı ve koordineli saldırılar olarak tanımlanmaktadır. Bu tür

saldırılar, bir sürecin tamamında yer alan eyleyicilerin eşzamanlı olarak manipüle edilmesi yoluyla, sistemin genel işleyişini durdurmayı veya ciddi şekilde bozmayı amaçlamaktadır. Tüm süreç eyleyici saldırıları, genellikle en yıkıcı ve tespit edilmesi en zor saldırı türleri arasında yer almakta olup, tüm eyleyicilerin koordineli bir şekilde manipüle edilmesini gerektirmektedir.

Bu saldırılarda, saldırganlar tüm eyleyicilerin durumlarını tersine çevirerek sistemin normal işleyişini bozmayı hedeflemektedir. İlk olarak, saldırıya dahil edilecek tüm eyleyiciler belirlenmekte ve ardından bu eyleyicilerin durumları tersleme saldırısı kullanılarak manipüle edilmektedir. Yani, açık konumdaki eyleyiciler kapalı duruma getirilirken, kapalı konumdakiler açılmaktadır. Bu şekilde, tüm eyleyicilerin aynı anda yanlış komutlar alması sağlanarak koordineli bir manipülasyon gerçekleştirilmektedir. Son aşamada, saldırının etkileri izlenmekte ve bu eylemlerin sürecin genel işleyişine nasıl zarar verdiği değerlendirilmektedir.

Tüm süreç eyleyici saldırıları, operasyonel verimliliği azaltırken sistemde büyük çapta aksamalara neden olabilmektedir. SWaT sistemindeki bu aksamalar, suyun tüketiciye ulaşmadan önce yeniden arıtılmasını veya tamamen bertaraf edilmesini gerektirebilmektedir. Bu durum, maliyetleri artırmakta ve operasyonel kesintilere yol açmaktadır.

a) Su Arıtma Sistemi - SWaT

Tablo 16, SWaT sisteminde tüm eyleyici saldırıları altında CNN-LSTM modelinin performansını göstermektedir. Tablo incelendiğinde, TN değeri yalnızca 10 olarak belirlenmiştir. Bu durum, modelin negatif olan örneklerin çok az bir kısmını doğru bir şekilde negatif olarak sınıflandırabildiğini göstermektedir. TN değerinin bu kadar düşük olması, modelin güvenilirliğinin de düşük olduğunu ortaya koymaktadır.

FP değeri 94990 olarak hesaplanmıştır. Bu yüksek FP değeri, modelin negatif örneklerin çoğunu yanlış bir şekilde pozitif olarak sınıflandırdığını ve güvenilirliğinin ciddi şekilde sorgulanmasına yol açtığını göstermektedir. Kesinlik değeri ise 0.351 olarak hesaplanmıştır. Bu, modelin pozitif olarak sınıflandırdığı örneklerin yalnızca yaklaşık %35'inin gerçekten pozitif olduğunu göstermektedir. Düşük kesinlik değeri, modelin aykırılık tespitinde sıkça yanlış alarm verdiğini ve güvenilir olmadığını göstermektedir.

Ayrıca YPO değeri 0.999 olarak bulunmuştur. Bu yüksek oran, modelin negatif örneklerin neredeyse tamamını pozitif olarak sınıflandırdığını ve gerçek saldırıları tespit

etmekte başarısız olduğunu göstermektedir. Yüksek YPO değeri, modelin güvenilir bir aykırılık tespit mekanizması sunamadığını açıkça ortaya koymaktadır.

Genel olarak, Tablo 16'daki veriler, CNN-LSTM modelinin tüm eyleyici saldırıları altında performansının oldukça düşük olduğunu göstermektedir. Model, negatif örnekleri büyük oranda FP olarak sınıflandırmakta ve bu da modelin doğruluğunu ve güvenilirliğini ciddi şekilde azaltmaktadır. Bu sonuçlar, modelin özellikle bu tür kapsamlı saldırılar karşısında yetersiz kaldığını ve iyileştirilmesi gerektiğini göstermektedir.

Tablo 16. Su arıtma sisteminde tüm süreç eyleyici saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	CNN-LSTM			
	TN	FP	Kesinlik	YPO
Tüm Eyleyiciler	10	94990	0.35089	0.99989

b) Su Dağıtma Sistemi - WADI

Tablo 17, WADI veri setindeki tüm eyleyici saldırıları altında CNN-LSTM modelinin performansını detaylı olarak sunmaktadır. İncelemeler, CNN-LSTM modelinin tüm eyleyici saldırılarını tespit etme performansının oldukça düşük olduğunu göstermektedir. TN değeri, 0 olarak belirlenmiştir. Bu durum, modelin hiçbir TN tespitte bulunamadığını ortaya koymaktadır. Bu sonuç, modelin saldırıları doğru bir şekilde tespit edemediğini ve saldırılara karşı tamamen etkisiz olduğunu göstermektedir.

FP değeri 41927 olarak hesaplanmıştır. Bu yüksek FP değeri, modelin negatif örnekleri büyük oranda pozitif olarak sınıflandırdığını ve çok sayıda yanlış alarm ürettiğini göstermektedir. Kesinlik değeri ise 0.183 olarak bulunmuştur. Bu düşük kesinlik değeri, modelin pozitif olarak sınıflandırdığı örneklerin yalnızca %18'inin gerçekten pozitif olduğunu göstermektedir. Ayrıca, modelin doğru tespit yapma yeteneğinin çok zayıf olduğunu da vurgulamaktadır.

YPO değeri ise 1 olarak hesaplanmıştır. YPO değerinin 1 olması, modelin negatif olması gereken tüm örnekleri pozitif olarak sınıflandırdığı ve bu nedenle %100 YPO oranına sahip olduğu anlamına gelmektedir. Bu sonuçlar, modelin negatif örnekleri sınıflandırmada tamamen başarısız olduğunu açıkça göstermektedir.

Genel olarak, Tablo 17'deki veriler, CNN-LSTM modelinin WADI veri setindeki tüm eyleyici saldırılarını tespit etmede tamamen başarısız olduğunu ortaya koymaktadır. Yüksek YPO ve düşük TP değerleri, CNN-LSTM modelinin WADI sistemindeki tüm eyleyici saldırılarını doğru bir şekilde tespit etmede başarısız olduğunu ve bu nedenle güvenilir bir aykırılık tespit mekanizması sunamadığını göstermektedir. Bu durum, modelin özellikle geniş kapsamlı saldırılar karşısında negatif örnekleri doğru sınıflandırma yeteneğinin iyileştirilmesi gerektiğine işaret etmektedir.

Tablo 17. Su dağıtma sisteminde tüm süreç eyleyici saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	CNN-LSTM			
	TN	FP	Kesinlik	YPO
Tüm Eyleyiciler	0	41927	0.18295	1

2.7. Sensör Saldırıları ve Sonuçları

Sensör saldırıları, SWaT ve WADI sistemlerinde sensörlerin gerçek değerlerini manipüle ederek sistemin işleyişini bozan siber saldırılar olarak tanımlanmaktadır. Sensörler, sistemin durumunu izleyip elde ettikleri bilgileri kontrol ünitesine ileterek sistemin doğru ve güvenli bir şekilde çalışmasına katkı sağlamaktadır. Sensörlerin manipüle edilmesi, modelin yanlış bilgi almasına ve bu yanıltıcı bilgilere dayalı hatalı kararlar vermesine yol açabilmektedir. Bu durum, sistemin genel performansını ve güvenliğini ciddi şekilde tehlikeye atabilmektedir.

Sensör saldırıları, çeşitli yöntemlerle gerçekleştirilebilmektedir. Süreç bazlı değişken genlikli sensör saldırılarında, sensör verilerine belirli bir yüzdelik sapma eklenerek manipülasyon yapılmakta ve bu durum, sistemin dinamik tepkisini bozmaktadır. Sabit genlikli sensör saldırılarında ise, sensör verileri gerçek değerlerinden bağımsız olarak sabit bir değerde tutulmakta ve bu şekilde sistem yanıltılmaktadır. Her iki saldırı türü de sistemin güvenliği ve performansı üzerinde olumsuz etkilere neden olmaktadır.

2.7.1. Süreç Bazlı Değişken Genlikli Sensör Saldırısı

Süreç bazlı değişken genlikli sensör saldırıları, SWaT ve WADI sistemlerinde, belirli bir sürece ait sensörlerin verilerine yüzdelik sapmalar ekleyerek gerçekleştirilen siber saldırılar olarak tanımlanmaktadır. Bu saldırılar, sistemin dinamik tepkisini bozmayı ve doğru kararlar almasını engellemeyi hedeflemektedir. Saldırganlar, sistemin normal çalışma aralığından sapmasına yol açacak şekilde sensör verilerini manipüle ederek güvenlik riskleri oluşturmakta ve performans düşüşlerine neden olmaktadır. Böylece, sistemin istikrarlı ve güvenli çalışması tehlikeye girmektedir.

Bu saldırı türü genellikle hedef sensörlerin belirlenmesi ve bu sensörlerin verilerine değişken büyüklükte sapmalar eklenmesiyle gerçekleştirilmektedir. Saldırganlar, doğru verilere belirli yüzdelik sapmalar ekleyen bir manipülasyon stratejisi uygulamaktadır. Bu değişken sapmalar, sistemin dinamik tepkisinin tutarsız olmasına ve kararsız davranışlar sergilemesine yol açmaktadır. Örneğin, SWaT sisteminde su seviyesini izleyen bir sensörün verisine yüzdelik sapmalar eklemek, sistemin su pompalarını gereksiz yere çalıştırmasına veya durdurmasına sebep olabilmektedir. Bu durum enerji kaybına ve olası bir sistem arızasına yol açabilmektedir.

a) Su Arıtma Sistemi - SWaT

Tablo 18, SWaT sisteminde gerçekleştirilen süreç bazlı değişken genlikli sensör saldırılarına karşı CNN-LSTM modelinin performansını ayrıntılı olarak sunmaktadır. Tablo incelendiğinde, genlik değeri %25 olduğunda modelin performansının süreçler arasında benzer olduğu görülmektedir. Örneğin, Süreç-1'de kesinlik değeri 0.369 ve YPO değeri 0.979 olarak kaydedilmiştir. Diğer süreçlerde ve tüm süreçler genelinde de bu değerlere çok yakın sonuçlar elde edilmiştir. Bu durum, modelin %25 genlik değeri altında düşük kesinlik ve yüksek YPO değerleriyle düşük performans sergilediğini göstermektedir.

Genlik değeri %50'ye çıkarıldığında da benzer bir durum gözlemlenmiştir. Süreç-1 için kesinlik değeri 0.370 ve YPO değeri 0.978 olarak hesaplanmıştır. Diğer süreçlerdeki değerler de bu sonuçlara oldukça yakındır ve tüm süreçler için kesinlik değeri 0.370 ve YPO değeri 0.978 olarak hesaplanmıştır. Bu sonuçlar, modelin %50 genlik seviyesinde de düşük kesinlik ve yüksek YPO değerlerine sahip olduğunu göstermektedir.

Genlik değeri %75 olduğunda da modelin performansında belirgin bir değişiklik olmadığı anlaşılmaktadır. Süreç-1'de kesinlik değeri 0.370 ve YPO değeri 0.977 olarak kaydedilmiştir. Diğer süreçler için hesaplanan değerler de yine bu değerlere çok yakındır ve

kesinlik değeri 0.370, YPO değeri ise 0.977 olarak hesaplanmıştır. Bu, modelin %75 genlik seviyesinde de benzer performans sergilediğini ortaya koymaktadır.

Genlik değeri %100 çıkarıldığında modelin performansı genel olarak tutarlılığını korumaktadır. Süreç-1'de kesinlik değeri 0.370 ve YPO değeri 0.978 olarak tespit edilmiştir. Diğer süreçlerde de benzer değerler elde edilmiş olup, tüm süreçler için kesinlik değeri 0.371 ve YPO değeri 0.977 olarak belirlenmiştir. Bu bulgular, en yüksek genlik seviyesinde bile modelin düşük kesinlik ve yüksek YPO değerlerine sahip olduğunu göstermektedir.

Genel olarak, Tablo 18'de sunulan veriler, CNN-LSTM modelinin genlik değeri arttıkça performansının tutarlı kaldığını ve her genlik seviyesinde düşük kesinlik ve yüksek YPO değerleri sergilediğini ortaya koymaktadır. Modelin farklı genlik seviyelerinde gösterdiği benzer performans, genlik değişimlerine karşı duyarlılığının düşük olduğunu ve genel olarak zayıf bir sonuç verdiğini göstermektedir.

Tablo 18. Su arıtma sisteminde süreç bazlı değişken genlikli sensör saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	Genlik	CNN-LSTM			
	%	TN	FP	Kesinlik	YPO
Süreç-1	25	2026	92974	0.36944	0.97867
Süreç-2		2226	92774	0.37079	0.97657
Süreç-3		2247	92753	0.37094	0.97635
Süreç-4		2191	92809	0.37056	0.97694
Süreç-5		2231	92769	0.37083	0.97652
Süreç-6		2155	92845	0.37031	0.97732
Tüm süreçler		2133	92867	0.37016	0.97755
Süreç-1	50	2102	92898	0.36995	0.97787
Süreç-2		2155	92845	0.37031	0.97732
Süreç-3		2224	92776	0.37078	0.97659
Süreç-4		2230	92770	0.37082	0.97653
Süreç-5		2139	92861	0.37020	0.97748
Süreç-6		2079	92921	0.36980	0.97812

Tablo 18'in devamı

Tüm süreçler		2119	92881	0.37007	0.97769
Süreç-1	75	2151	92849	0.37028	0.97736
Süreç-2		2238	92762	0.37087	0.97644
Süreç-3		2088	92912	0.36986	0.97802
Süreç-4		2219	92781	0.37075	0.97644
Süreç-5		2116	92884	0.37005	0.97773
Süreç-6		2142	92858	0.37022	0.97745
Tüm süreçler		2152	92848	0.37029	0.97735
Süreç-1	100	2105	92895	0.36997	0.97784
Süreç-2		2143	92857	0.37023	0.97744
Süreç-3		2250	92750	0.37096	0.97632
Süreç-4		2159	92841	0.37034	0.97727
Süreç-5		2237	92763	0.37087	0.97645
Süreç-6		2155	92845	0.37031	0.97732
Tüm süreçler		2192	92808	0.37056	0.97693

b) Su Dağıtma Sistemi – WADI

Tablo 19, WADI veri seti üzerindeki süreç bazlı değişken genlikli sensör saldırılarına karşı CNN-LSTM modelinin performansını detaylı olarak sunmaktadır. Tablonun incelenmesi sonucunda, genlik değeri %25 alındığında modelin performansının süreçler arasında benzerlik gösterdiği gözlemlenmiştir. Örneğin, Süreç-1 için kesinlik değeri 0.227 ve YPO değeri 0.763 olarak kaydedilmiştir. Diğer süreçlerde de benzer değerler elde edilmiştir. Süreç-2 ve Süreç-3'te de kesinlik ve YPO değerleri Süreç-1'e oldukça yakındır. Tüm süreçler için kesinlik değeri 0.227 ve YPO değeri 0.762 olarak belirlenmiştir. Bu durum, modelin düşük genlik seviyelerinde genel olarak düşük performans sergilediğini göstermektedir.

Genlik değeri %50'ye çıkarıldığında, modelin performansında büyük bir değişiklik olmadığı görülmektedir. Süreç-1 için kesinlik değeri 0.224 ve YPO değeri 0.778 olarak hesaplanmıştır. Diğer süreçlerde de benzer sonuçlar elde edilmiştir; Süreç-2 ve Süreç-3'teki

kesinlik ve YPO değerleri, Süreç-1'e yakın değerlere sahiptir. Tüm süreçler için genel kesinlik değeri 0.223 ve YPO değeri 0.781 olarak hesaplanmıştır. Bu sonuçlar, modelin %50 genlik seviyesinde de düşük kesinlik ve yüksek YPO değerlerine sahip olduğunu göstermektedir.

Genlik değeri %75'e yükseldiğinde modelin performansında anlamlı bir değişim gerçekleşmemektedir. Süreç-1'de kesinlik değeri 0.220 ve YPO değeri 0.794 olarak kaydedilmiştir. Diğer süreçlerde de benzer değerler elde edilmiştir ve genel olarak tüm süreçler için kesinlik değeri 0.222, YPO değeri ise 0.783 olarak hesaplanmıştır. Bu, modelin %75 genlik seviyesinde de iyi bir performans sergileyemediğini ortaya koymaktadır.

Genlik değeri %100 olarak alındığında modelin performansı genellikle benzer kalmaktadır. Süreç-1'de kesinlik değeri 0.220 ve YPO değeri 0.793 olarak belirlenmiştir. Diğer süreçler de benzer değerlere sahiptir ve genel performans tüm süreçler için kesinlik değeri 0.224 ve YPO değeri 0.777 olarak belirlenmiştir. Bu bulgular, modelin en yüksek genlik seviyesinde bile düşük kesinlik ve yüksek YPO değerlerine sahip olduğunu göstermektedir.

Genel olarak, Tablo 19'da yer alan veriler, CNN-LSTM modelinin genlik değeri arttıkça performansının tutarlı olduğunu ortaya koymaktadır. CNN-LSTM modeli, tüm genlik seviyelerinde benzer şekilde düşük kesinlik ve yüksek YPO değerlerine sahiptir. Bu değerler incelendiğinde, modelin süreç bazlı değişken genlikli sensör saldırıları altında iyi bir performans sergileyemediği görülmektedir.

Tablo 19. Su dağıtma sisteminde süreç bazlı değişken genlikli sensör saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	Genlik	CNN-LSTM			
	%	TN	FP	Kesinlik	YPO
Süreç-1	25	9949	31978	0.22695	0.76271
Süreç-2		10198	31729	0.22832	0.75677
Süreç-3		10596	31331	0.23056	0.74728
Tüm süreçler		9964	31963	0.22703	0.76235
Süreç-1	50	9321	32606	0.22356	0.77769
Süreç-2		10381	31546	0.22934	0.75240

Tablo 19'un devamı

Süreç-3		10461	31466	0.22979	0.75049
Tüm süreçler		9187	32740	0.22284	0.78088
Süreç-1	75	8656	33271	0.22007	0.79355
Süreç-2		10150	31777	0.22806	0.75791
Süreç-3		9924	32003	0.22681	0.76330
Tüm süreçler		9099	32828	0.22238	0.78298
Süreç-1	100	8664	33263	0.22011	0.79336
Süreç-2		10149	31778	0.22805	0.75794
Süreç-3		9659	32268	0.22537	0.76962
Tüm süreçler		9356	32571	0.22374	0.77685

2.7.2. Sabit Genlikli Sensör Saldırısı

Süreç bazlı sabit genlikli sensör saldırıları, SWaT ve WADI sistemlerinde, belirli bir sürece ait sensörlerin verilerinin gerçek değerlerinden bağımsız olarak sabit bir değerde tutulmasıyla gerçekleştirilen siber saldırılardır. Bu tür saldırılar, sistemin normal işleyişini bozmayı ve kontrol mekanizmalarının doğru kararlar vermesini engellemeyi amaçlamaktadır. Saldırganlar, sensörlerin okuduğu verileri sabitleyerek sistemin koşullara uygun tepki vermesini önlemekte, böylece güvenlik açıkları ortaya çıkmaktadır. Bu durum, sistem performansında düşüşlere neden olmakta ve güvenli, istikrarlı bir işleyişi tehlikeye sokmaktadır.

Bu tür saldırılar, genellikle hedef sensörlerin seçilmesi ve bu sensörlerden gelen verilerin sabit bir değerde kalacak şekilde manipüle edilmesiyle gerçekleştirilmektedir. Saldırganlar, sensörlerin gerçek zamanlı ölçümlerini değişmeyen bir değer olarak kontrol ünitesine iletecek şekilde manipüle etmektedir. Sabit verilerle yanıltılan sistem, değişen koşulları algılayamamakta; bu da hatalı kontrol kararlarına yol açmaktadır. Örneğin, SWaT sisteminde su seviyesini kontrol eden bir sensörün sürekli sabit bir değeri göstermesi, su pompalarının yanlış zamanda çalışmasına veya durmasına neden olabilmektedir. Bu durum, enerji israfına ve potansiyel sistem arızalarına yol açabilmektedir.

a) Su Arıtma Sistemi – SWaT

Tablo 20, SWaT veri setindeki süreç bazlı sabit genlikli sensör saldırılarına karşı CNN-LSTM modelinin performansını kapsamlı bir şekilde sunmaktadır. Tablonun incelenmesi sonucunda, genlik değeri 0.25 olduğunda modelin tek tek süreçlerde performansının yüksek olduğu gözlemlenmektedir. Örneğin, Süreç-2 için kesinlik değeri 0.977 ve YPO değeri 0.012 olarak kaydedilmiştir. Diğer süreçlerde de benzer değerler elde edilmiştir. Ancak, tüm süreçler için kesinlik değeri 0.414 ve YPO değeri 0.765 olarak belirlenmiş olup, modelin genel performansının düşük olduğunu göstermektedir. Bu durum, düşük genlik seviyelerinde bile modelin tüm süreçlerde tutarlı bir performans sergileyemediğini ve yüksek YPO değerleri ürettiğini ortaya koymaktadır.

Genlik değeri 0.50'ye çıkarıldığında, modelin tek tek süreçlerde performansı genel olarak yüksek kalmaktadır. Süreç-2 için kesinlik değeri 0.977 ve YPO değeri 0.013 olarak hesaplanmıştır. Diğer süreçlerde de benzer sonuçlar elde edilmiştir. Ancak, tüm süreçler için kesinlik değeri 0.364 ve YPO değeri 0.946 olarak hesaplanmıştır. Bu durum, modelin %50 genlik seviyesinde tek tek süreçlerde yüksek doğruluk ve düşük YPO değerleri sergilese de genel performansının düşük olduğunu göstermektedir.

Genlik değeri 0.75 olduğunda, modelin tek tek süreçlerdeki performansının tüm süreçlerde yüksek kaldığı görülmektedir. Süreç-1 için kesinlik değeri 0.977 ve YPO değeri 0.013 olarak belirlenmiştir. Diğer süreçlerde de benzer değerler elde edilmiştir. Ancak, tüm süreçler için kesinlik değeri 0.351 ve YPO değeri 0.999 olarak hesaplanmıştır. Bu, modelin %75 genlik seviyesinde tüm süreçlerde çok fazla yanlış pozitif alarmlar verdiğini ve güvenilirliğinin azaldığını göstermektedir.

Genlik değeri 1.00'e ulaştığında modelin performansı en düşük seviyelere inmektedir. Süreç-1 için kesinlik değeri 0.977 ve YPO değeri 0.013 olarak belirlenmiştir. Diğer süreçlerde de benzer sonuçlar elde edilmesine rağmen, Süreç-2 ve Süreç-5'te belirgin performans düşüşleri gözlemlenmiştir. Tüm süreçler için ortalama kesinlik değeri 0.351 ve YPO değeri 1.000 olarak hesaplanmıştır. Bu bulgular, modelin en yüksek genlik seviyesinde bile belirli süreçlerde iyi performans gösterse de, genel olarak tüm süreçlerde zayıf performans sergilediğini ortaya koymaktadır.

Genel olarak, Tablo 20'deki veriler, CNN-LSTM modelinin tüm süreçler genelinde genlik değeri arttıkça performansının düştüğünü göstermektedir. Tek tek süreçlerde model yüksek doğruluk ve düşük YPO değerlerine sahip olsa da, tüm süreçler genelinde saldırılar karşısında zayıf kalmaktadır.

Tablo 20. Su arıtma sisteminde süreç bazlı sabit genlikli sensör saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	Genlik	CNN-LSTM			
		TN	FP	Kesinlik	YPO
Süreç-1	0.25	93467	1533	0.97101	0.01614
Süreç-2		93813	1187	0.97741	0.01249
Süreç-3		93708	1292	0.97546	0.01360
Süreç-4		91857	3143	0.94232	0.03308
Süreç-5		93568	1432	0.97287	0.01507
Süreç-6		93790	1210	0.97698	0.01274
Tüm süreçler		22356	72644	0.41413	0.76467
Süreç-1	0.50	93664	1336	0.97464	0.01406
Süreç-2		93812	1188	0.97739	0.01251
Süreç-3		93736	1264	0.97598	0.01331
Süreç-4		93620	1380	0.97383	0.01453
Süreç-5		93814	1186	0.97742	0.01248
Süreç-6		93779	1221	0.97677	0.01285
Tüm süreçler		5103	89897	0.36355	0.94628
Süreç-1	0.75	93770	1230	0.97661	0.01295
Süreç-2		93234	1766	0.96675	0.01859
Süreç-3		93718	1282	0.97564	0.01349
Süreç-4		93719	1281	0.97566	0.01348
Süreç-5		93505	1495	0.97171	0.01574
Süreç-6		93772	1228	0.97664	0.01293
Tüm süreçler		28	94972	0.35093	0.99971

Tablo 20’nin devamı

Süreç-1	1.00	93809	1191	0.97733	0.01254
Süreç-2		88958	6042	0.89473	0.06360
Süreç-3		93681	1319	0.97496	0.01388
Süreç-4		93515	1485	0.97189	0.01563
Süreç-5		77075	17925	0.74125	0.18868
Süreç-6		93739	1261	0.97603	0.01327
Tüm süreçler		10	94990	0.35089	0.99989

b) Su Dağıtma Sistemi – WADI

Tablo 21, WADI veri setindeki süreç bazlı sabit genlikli sensör saldırılarına karşı CNN-LSTM modelinin performansını ayrıntılı bir şekilde sunmaktadır. Genlik değeri 0.25 olduğunda, modelin performansının çeşitli süreçler arasında değişkenlik gösterdiği gözlemlenmektedir. Örneğin, Süreç-1 için kesinlik değeri 0.352 ve YPO değeri 0.357 iken, Süreç-2 için kesinlik 0.284 ve YPO 0.489 olarak kaydedilmiştir. Süreç-3 için ise kesinlik 0.350 ve YPO 0.360 olarak belirlenmiştir. Tüm süreçler için kesinlik değeri 0.283 ve YPO değeri 0.491 olarak hesaplanmıştır. Bu sonuçlar, modelin düşük genlik seviyelerinde tüm saldırılar karşısında genel olarak düşük performans sergilediğini göstermektedir. Model, 0.25 genlik değerinde iyi bir performans gösterememekte olup, düşük kesinlik ve yüksek YPO değerleri de bunu desteklemektedir.

Genlik değeri 0.50'ye çıkarıldığında, modelin performansında belirgin bir değişiklik görülmemektedir. Süreç-1 için kesinlik değeri 0.348 ve YPO değeri 0.364 olarak elde edilmiştir. Süreç-2 ve Süreç-3’teki değerler de Süreç-1’e yakındır. Tüm süreçler için kesinlik değeri 0.333 ve YPO değeri 0.389 olarak hesaplanmıştır. Bu genlik seviyesinde, modelin genel performansı, 0.25 genlik değerine kıyasla belirgin bir farklılık göstermemektedir.

Genlik değeri 0.75 olduğunda, modelin performansında önceki genlik değerlerine göre belirgin bir farklılık gözlenmemektedir. Süreç-1 için kesinlik değeri 0.345 ve YPO değeri 0.369, Süreç-3 için ise kesinlik değeri 0.359 ve YPO değeri 0.347 olarak belirlenmiştir. Ancak, Süreç-2'de kesinlik değeri 0.271 ve YPO değeri 0.523 olarak hesaplanmış olup, bu süreçteki performans düşüklüğü dikkat çekmektedir. Tüm süreçler için kesinlik değeri 0.187 ve YPO değeri 0.844 olarak belirlenmiş olup, bu genlik seviyesinde modelin genel performansının ciddi şekilde düştüğü görülmektedir.

Genlik değeri 1.00'e ulaştığında, modelin performansı genel olarak en düşük seviyelere inmektedir. Süreç-1 için kesinlik değeri 0.339 ve YPO değeri 0.379, Süreç-3 için ise kesinlik değeri 0.349 ve YPO değeri 0.362 olarak belirlenmiştir. Ancak, Süreç-2'de kesinlik değeri 0.166 ve YPO değeri 0.978 olarak kaydedilmiştir; bu da modelin performansının bu süreçte ciddi şekilde düştüğünü göstermektedir. Tüm süreçler için ortalama kesinlik değeri 0.163 ve YPO değeri 0.999 olarak hesaplanmıştır. Bu sonuçlar, modelin genlik seviyesi 1.00'de genel olarak düşük performans sergilediğini ve güvenilir olmadığını ortaya koymaktadır.

Genel olarak, Tablo 21'deki veriler, CNN-LSTM modelinin sabit genlikli sensör saldırıları altında performansının süreçler ve genlik seviyeleri arasında değişkenlik gösterdiğini ortaya koymaktadır. Model genellikle düşük kesinlik ve yüksek YPO değerlerine sahip olup, özellikle Süreç-2'nin birçok seviyede düşük performans sergilediği görülmektedir. Bu sonuçlar, modelin sabit genlikli sensör saldırısı altında genel olarak iyi bir performans sergileyemediğini göstermektedir.

Tablo 21. Su dağıtma sisteminde süreç bazlı sabit genlikli sensör saldırısı altında CNN-LSTM modelinin performansı

Saldırı Noktası	Genlik	CNN-LSTM			
		TN	FP	Kesinlik	YPO
Süreç-1	0.25	26957	14970	0.35223	0.35705
Süreç-2		21417	20510	0.28412	0.48918
Süreç-3		26843	15084	0.35047	0.35977
Tüm süreçler		21350	20577	0.28346	0.49078
Süreç-1	0.50	26651	15276	0.34763	0.36435
Süreç-2		24302	17625	0.31593	0.42037
Süreç-3		27183	14744	0.35571	0.35166
Tüm süreçler		25623	16304	0.33301	0.38887
Süreç-1	0.75	26462	15465	0.34484	0.36886
Süreç-2		20015	21912	0.27086	0.52262
Süreç-3		27366	14561	0.35857	0.34729

Tablo 21'in devamı

Tüm süreçler		6553	35374	0.18707	0.84370
Süreç-1	1.00	26049	15878	0.33888	0.37871
Süreç-2		904	41023	0.16557	0.97844
Süreç-3		26767	15160	0.34936	0.36158
Tüm süreçler		60	41867	0.16278	0.99857

2.7.3. Kural Denetçisi Performansının İncelenmesi

a) Su Arıtma Sistemi - SWaT

Tablo 22, SWaT sistemindeki eyleyicilerin kural denetçisi performanslarını detaylı bir şekilde sunmaktadır. Bu tablo, saldırı noktalarının tespit edilebilme oranlarını, başarısız ve geçerli kuralların sayısını, başarısız kuralların numaralarını ve tespit sayılarını içermektedir. SWaT sistemindeki eyleyicilere, açık olan durumların kapalı ya da kapalı olan durumların açık hale getirildiği tersleme saldırıları gerçekleştirilmiştir. Bu saldırılar sonucunda, hedeflenen noktaların kural denetçisinde nasıl bir performans sergilediği incelenmektedir.

Tablo incelendiğinde, bazı eyleyicilerin kural denetçisi tarafından tespit edilemediği gözlemlenmektedir. Örneğin, MV-101, P-101, P-203, P-205, MV-301 ve MV-303 eyleyicilerinde tespit edilebilme oranı 0 olarak belirlenmiştir. Bu, kural denetçisinin bu eyleyicileri tamamen gözden kaçırdığını göstermektedir. Bu eyleyiciler için başarısız kuralların sayısı 1 ile 8 arasında değişmekte olup, ilgili kuralların numaraları tabloda detaylı şekilde gösterilmektedir.

P-102, MV-201, P-201, P-202, P-204, P-206, MV-302, MV-304, P-302, P-401, P-402, P-403, P-404, UV-401, P-501, P-502, P-601 ve P-603 eyleyicilerinde ise tespit edilebilme oranı 1 olarak belirlenmiştir. Bu eyleyicilerde başarısız kural bulunmamakta, geçerli kuralların sayısı 69 olarak belirtilmektedir. Bu durum, bu eyleyicilerin kural denetçisi tarafından tamamen tespit edildiğini ve kurallara uygun çalıştığını göstermektedir. Tespit sayısı 95010 olarak belirlenmiş olup, tespit edilen toplam pozitif sayısını ifade etmektedir. Ayrıca, rastgele bir saldırı yapmanın, yanlış alarm üretme oranını arttırmadan kural denetçisi tarafından yakalanma ihtimalinin yüksek olduğunu göstermektedir. Bu nedenle, tasarlanacak

saldırılarda su sistemlerinin fiziksel değişmezlerinin dikkate alınması gerektiği vurgulanmaktadır.

P-602 eyleyicisi ise özel bir dikkat gerektirmektedir. Bu eyleyici için tespit edilebilme oranı 0 olarak belirlenmiş, 21 başarısız kural kaydedilmiştir. Başarısız kuralların sayısının fazla olması, bu eyleyicinin denetim sürecinde önemli zorluklar yaşadığını göstermektedir. Geçerli kuralların sayısı 48 olup, diğer eyleyicilere kıyasla oldukça düşüktür.

Genel olarak, Tablo 22'deki veriler, bazı eyleyicilerin kural denetçisi tarafından etkin bir şekilde tespit edilemediğini ortaya koymaktadır. Özellikle MV-101, P-101, P-203, P-205, MV-301 ve MV-303 eyleyicilerinde tespit edilebilme oranlarının düşük olması, bu eyleyicilerin kural denetçisi tarafından gözden kaçırıldığını ve güvenilir olmadığını göstermektedir. Buna karşılık, P-102, MV-201, P-201, P-202 gibi eyleyicilerde tespit edilebilme oranlarının yüksek olması, bu eyleyicilerin kural denetçisi tarafından etkin bir şekilde tespit edildiğini ve güvenilir olduğunu göstermektedir. Bu veriler, kural denetçisi performansının artırılması için kullanılmaktadır.

Tablo 22. Su arıtma sisteminde kural denetçisi performansları

Saldırı Noktası	Tespit Edilebilme Oranı	Başarısız Kurallar	Geçerli Kurallar	Başarısız Kuralların Numaraları	Tespit Sayısı
MV-101	0	1	68	8	0
P-101	0	2	67	5,7	0
P-102	1	0	69	-	95010
MV-201	1	0	69	-	95010
P-201	1	0	69	-	95010
P-202	1	0	69	-	95010
P-203	0	3	66	5,6,38	0
P-204	1	0	69	-	95010
P-205	0	5	64	5,6,38,39,40	0
P-206	1	0	69	-	95010
MV-301	0	8	61	1,7,10,11,16,27,41,48	0
MV-302	1	0	69	-	95010
MV-303	0	8	61	7,10,14,41,42,45,48,49	0
MV-304	1	0	69	-	95010

Tablo 22’nin devamı

P-301	0	1	68	63	0
P-302	1	0	69	-	95010
P-401	1	0	69	-	95010
P-402	1	0	69	-	95010
P-403	1	0	69	-	95010
P-404	1	0	69	-	95010
UV-401	1	0	69	-	95010
P-501	1	0	69	-	95010
P-502	1	0	69	-	95010
P-601	1	0	69	-	95010
P-602	0	21	48	0,7,10,11,13,21,41,42,43,44,43, 6,47,49,50,51,52,53,54,55,56	0
P-603	1	0	69	-	95010

b) Su Dağıtma Sistemi - WADI

Tablo 23, WADI sistemindeki eyleyicilerin kural denetçisi performanslarını ayrıntılı bir biçimde sunmaktadır. Bu tablo, saldırı noktalarının tespit edilme oranlarını, geçerli ve başarısız kuralların miktarını, başarısız kuralların numaralarını, tespit edilen olay sayılarını içermektedir. WADI sistemindeki eyleyicilere, SWaT sistemindeki eyleyicilere de uygulandığı gibi kapalı olan durumların açıldığı veya açık olan durumların kapandığı tersleme saldırıları uygulanmıştır. Bu saldırılar sonucunda, hedeflenen noktaların kural denetçisi performansları analiz edilmektedir.

Genel olarak, Tablo 23'teki veriler, çoğu eyleyicinin kural denetçisi tarafından etkin bir şekilde tespit edilemediğini göstermektedir. Eyleyicilerin büyük bir kısmının tespit edilememesi, kural denetçisi tarafından gözden kaçırıldıklarına ve güvenilir olmadıklarına işaret etmektedir. Buna karşılık, 1_P_005_STATUS eyleyicisinin tespit edilebilme oranının yüksek olması, kural denetçisi tarafından etkin bir şekilde tespit edildiğini ve güvenilir olduğunu ortaya koymaktadır.

Tablo 23. Su dağıtma sisteminde kural denetçisi performansları

Saldırı Noktası	Tespit Edilebilme Oranı	Başarısız Kurallar	Geçerli Kurallar	Başarısız Kuralların Numaraları	Tespit Sayısı
1_MV_001_STATUS	0	1	12	4	0
1_MV_002_STATUS	0	1	12	4	0
1_MV_003_STATUS	0	1	12	4	0
1_MV_004_STATUS	0	2	11	4,5	0
1_P_001_STATUS	0	2	11	1,4	0
1_P_002_STATUS	0	1	12	4	0
1_P_003_STATUS	0	2	11	1,4	0
1_P_004_STATUS	0	1	12	4	0
1_P_005_STATUS	1	0	13	-	41937
1_P_006_STATUS	0	1	12	4	0
2_FQ_101_PV	0	1	12	4	0
2_FQ_401_PV	0	1	12	4	0
2_MCV_007_CO	0	1	12	4	0
2_MCV_101_CO	0	1	12	4	0
2_MCV_201_CO	0	1	12	4	0
2_MCV_301_CO	0	1	12	4	0
2_MCV_401_CO	0	2	11	4,7	0
2_MCV_501_CO	0	1	12	4	0
2_MCV_601_CO	0	1	12	4	0
2_MV_001_STATUS	0	1	12	4	0
2_MV_002_STATUS	0	1	12	4	0
2_MV_003_STATUS	0	1	12	4	0
2_MV_004_STATUS	0	1	12	4	0
2_MV_005_STATUS	0	1	12	4	0
2_MV_006_STATUS	0	1	12	4	0
2_MV_009_STATUS	0	1	12	4	0
2_MV_101_STATUS	0	1	12	4	0
2_MV_201_STATUS	0	2	11	3,4	0
2_MV_301_STATUS	0	1	12	4	0
2_MV_401_STATUS	0	1	12	4	0
2_MV_501_STATUS	0	1	12	4	0
2_MV_601_STATUS	0	1	12	4	0
2_P_003_STATUS	0	1	12	4	0
2_P_004_SPEED	0	1	12	4	0
2_P_004_STATUS	0	1	12	4	0
2_SV_101_STATUS	0	1	12	4	0
2_SV_201_STATUS	0	1	12	4	0
2_SV_301_STATUS	0	1	12	4	0
2_SV_401_STATUS	0	1	12	4	0
2_SV_501_STATUS	0	1	12	4	0
2_SV_601_STATUS	0	1	12	4	0
3_MV_001_STATUS	0	1	12	4	0
3_MV_002_STATUS	0	1	12	4	0

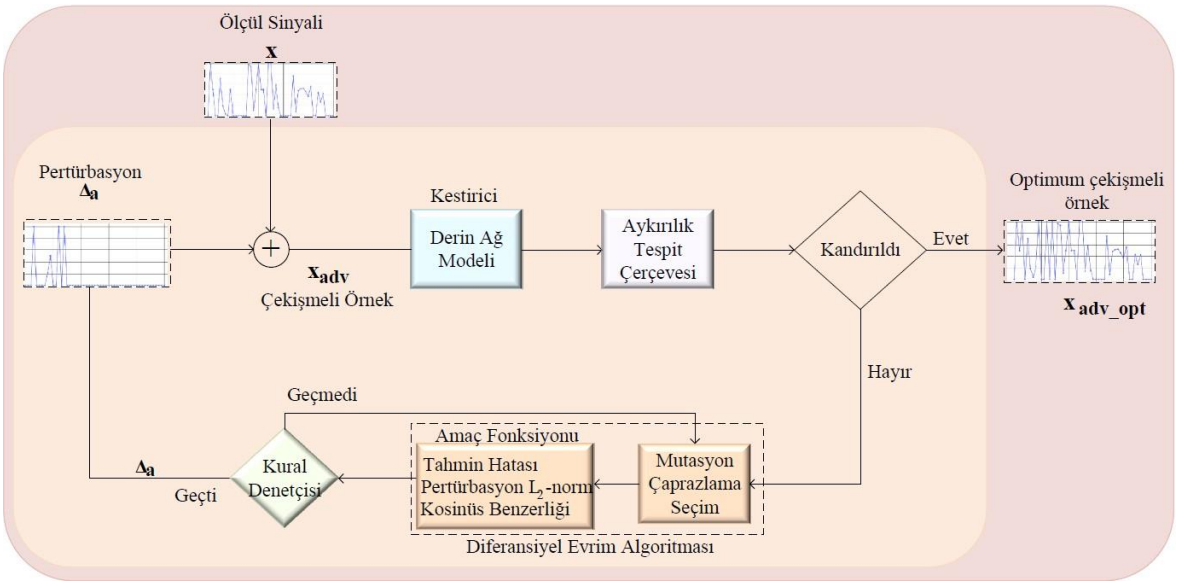
Tablo 23'ün devamı

3_MV_003_STATUS	0	1	12	4	0
3_P_001_STATUS	0	1	12	4	0
3_P_002_STATUS	0	1	12	4	0
3_P_003_STATUS	0	1	12	4	0
3_P_004_STATUS	0	1	12	4	0

2.8. Optimal Pertürbasyonun Üretimi ve Etkisi

OptAML olarak adlandırılan optimize edilmiş AML çerçevesi, SWaT ve WADI sistemlerinde hassas manipülasyonlar yapabilmek için özel olarak tasarlanmıştır. OptAML, ağ modelinde yanıltıcı davranışlara neden olan pertürbasyonlar üreterek, CPS'lerde yanlış alarmlara yol açmayı hedeflemektedir. Bu çerçeve, hem sensörlere ve eyleyicilere sınırlı erişimi olan saldırganların hem de sistemin doğal değişmezlerine dayanan kural denetçisinin stratejilerini dikkatlice ele almaktadır.

Optimal çekişmeli örneklerin üretim sürecinin blok diyagramı Şekil 14'te verilmiştir. OptAML yapısını daha iyi anlamak için optimal çekişmeli örnek üretim algoritması kullanılmaktadır. Bu algoritma, en etkili çekişmeli örneklerin nasıl oluşturulacağını adım adım özetlemektedir.



Şekil 14. Optimal çekişmeli örnek üretim sürecinin blok diyagramı

Optimal Çekişmeli Örnek Üretim Algoritması, t zamanında alınan ölçüm sinyali (x_t) ve t-1 zamanındaki pertürbasyon sinyalini ($\Delta_{a,t-1}$) girdi olarak kullanmaktadır. İlk olarak,

önceki pertürbasyon sinyalinin ($\Delta_{a,t-1}$) derin ağ modelini yanıltmada etkili olup olmadığı değerlendirilmektedir. Eğer sinyal etkiliyse, oluşturulan çekişmeli örnek ($x_{adv,t}$), doğrudan optimal çekişmeli örnek ($x_{opt\_adv,t}$) olarak kullanılmaktadır. Aksi taktirde, algoritma bir sonraki adıma geçmektedir.

İkinci adımda, algoritma, başarılı bir çekişmeli örnek bulana veya maksimum iterasyon sayısına ulaşana kadar diferansiyel evrim (DE) algoritmasını kullanarak yeni bir pertürbasyon sinyali ($\Delta_{a,t}$) üretmektedir. Eğer çekişmeli bir örnek bulunursa, kestirim hatası, pertürbasyonun L_2 normu, CS ve aykırılık tespit çerçevesinin aldatılma durumu dikkate alınarak amaç fonksiyonun değeri hesaplanır. Çerçeve aldatılmışsa, mevcut çekişmeli örnek en uygun çekişmeli örnek olarak kabul edilmektedir. Aksi taktirde, algoritma üçüncü adıma geçmektedir.

Üçüncü adımda, amaç fonksiyonu değeri sıfır ise, bu durum başarılı bir çekişmeli örnek bulunamadığını göstermektedir ve algoritma, önceki optimal çekişmeli örneği ($x_{opt\_adv,t-1}$) kullanarak yeniden saldırı gerçekleştirmektedir. Sonuç olarak, algoritma en uygun çekişmeli örneği ($x_{opt\_adv,t}$) çıktı olarak sunmaktadır.

Optimal çekişmeli örnek üretim algoritması, belirli hiper parametrelere sahip DE algoritmasını kullanmaktadır. Keşif ve çeşitlilik sağlamak için popülasyon boyutu 20 olarak belirlenmiştir. Hesaplama kaynakları ve yakınsama hızı arasında bir denge sağlamak adına maksimum iterasyon sayısı 20 olarak ayarlanmıştır. Mutasyon sırasında keşif ve sömürü arasındaki dengeyi korumak için ölçeklendirme faktörü F=0.5 olarak seçilmiştir. Yakınsama istikrarını sağlamak ve yeterli keşfi teşvik etmek için çaprazlama olasılığı CR=0.9 olarak sabitlenmiştir.

DE algoritması için tanımlanan amaç fonksiyonu Denklem (18) ile gösterilmektedir. Denklemde, $DeceptionFlag(x_{adv,t}) = 1$ (aykırılık tespit çerçevesinin başarıyla aldatıldığını belirtir) olduğunda, kestirim hatası ve pertürbasyon normunun yüksek değerleri ve CS'nin düşük değerleri maliyet fonksiyonunda bir cezaya yol açmaktadır. Aksi durumda, $DeceptionFlag(x_{adv,t}) = 0$ ise (çerçevenin aldatılamadığını gösterir) ise, amaç fonksiyonu 0 olarak değerlendirilmektedir ve bu başarısız bir saldırıyı ifade etmektedir. Bu formülasyon, amaç fonksiyonunu minimize ederek optimizasyon sürecinin saldırının etkinliğini en üst düzeye çıkarmasını sağlamaktadır.

$$Cost(\Delta_{a,t}) = -DeceptionFlag(x_{adv,t}).1/(\alpha.PredictionError(x_{adv,t}) \dots + \beta.|\Delta_{a,t}|_2 + \gamma.CosineSimilarity(x_t, x_{adv,t})) \quad (18)$$

Burada, α, β ve γ ağırlık katsayılarını, $PredictionError(x_{adv,t})$ derin ağ modelinin kestirim hatasını, $|\Delta_{a,t}|_2$ pertürbasyon sinyalinin L_2 normunu, $CosineSimilarity(x_t, x_{adv,t})$ ise orijinal ölçüm sinyali (x_t) ile çekişmeli örnek ($x_{adv,t}$) arasındaki CS'yi ifade etmektedir. $DeceptionFlag(x_{adv,t})$ ise aykırılık tespit çerçevesinin başarılı bir şekilde aldatıldığını (1) veya aldatılamadığını (0) gösteren bir bayrak işaretini temsil etmektedir.

2.8.1. Optimal Pertürbasyon Sonuçları

a) Su Arıtma Sistemi - SWaT

Tablo 24, OptAML algoritmasının SWaT sisteminde farklı saldırı senaryoları altındaki performansını ayrıntılı bir şekilde sunmaktadır. Bu bağlamda, OptAML algoritmasının kural denetçisi olan ve olmayan iki farklı yapı altındaki performansı incelenmiş ve CNN-LSTM modelinin saldırılara karşı dayanıklılığı kapsamlı bir şekilde değerlendirilmiştir. SWaT'ın tüm süreçlerine, çeşitli saldırı noktalarında, tek tek ve eş zamanlı olarak saldırılar gerçekleştirilmiştir. OptAML yapısı ile optimal pertürbasyon üretimi hem kural denetçisi ile hem de kural denetçisi olmadan uygulanmaktadır. Kural denetçisi olmayan yapı, Şekil 14'te gösterilen blok diyagramdan kural denetçisinin çıkarılmasıyla oluşturulmaktadır. Kural denetçisi içeren yapı, sistemin güvenliğini artırmayı hedeflemektedir. Kural denetçisi olmayan yapı daha esnek, ancak potansiyel olarak daha savunmasız bir performans sergilemektedir.

Tablolar, farklı süreçler ve saldırı noktaları için CNN-LSTM modelinin kesinlik, duyarlılık, F1-skoru ve YPO gibi kritik performans metriklerini içermektedir. Bu metrikler, OptAML algoritması kullanılarak SWaT sisteminde gerçekleştirilen saldırıların etkisini değerlendirmek açısından büyük önem taşımaktadır. Tablolarda, A ve S harfleri sırasıyla eyleyicilere ve sensörlere yönelik saldırıları temsil etmektedir. %10 ve %25 olarak belirtilen pertürbasyon seviyeleri ise, sensörlere eklenen maksimum genlik değerlerini ifade etmektedir ve bu seviyeler, modeli yanıltarak performansını etkileyen önemli faktörlerdir.

Tablo 24’te 35 farklı saldırı senaryosu için elde edilen veriler incelendiğinde aşağıdaki sonuçlar çıkarılmıştır:

Amacımız, OptAML saldırıları uygulanmadan önce literatürde yüksek performans metriklerine sahip olan CNN-LSTM’nin, OptAML saldırılarının gerçekleştirilmesi sonucunda performansının düşebileceğini göstermektir. Saldırısız durumda, modelin performans metrikleri oldukça yüksektir. Kesinlik, duyarlılık ve f1-skoru değerleri sırasıyla 0.981, 0.955 ve 0.968 olarak belirlenmiştir; YPO değeri ise 0.010’dur. Bu değerler, modelin OptAML algoritması tarafından üretilen optimum saldırılar olmadan yüksek doğruluk ve düşük YPO değerleri ile çalıştığını göstermektedir. Ancak, saldırılar uygulandığında bu performans metriklerinde belirgin bir düşüş gözlemlenmektedir.

Eyleyici saldırıları (A) incelendiğinde, kural denetçisi olmayan yapı altında kesinlik değerlerinin genellikle düşük olduğu (örneğin, Süreç 1'de 0.331), duyarlılık değerlerinin ise orta seviyede kaldığı (0.695) görülmektedir. Bu durum, modelin saldırıları tespit etmekte zorlandığını göstermektedir. Kural denetçisi olan yapı altında ise bu metriklerde belirgin bir iyileşme gözlenmektedir (Süreç 1'de kesinlik değeri 0.419, duyarlılık değeri 0.849). Bu da kural denetçili yapının saldırı tespitinde daha etkili olduğunu ortaya koymaktadır.

Sensör saldırıları (S) değerlendirildiğinde, %10 pertürbasyon ile gerçekleştirilen saldırılarda modelin performansı Süreç 6'ya yönelik saldırılarda oldukça yüksektir (kural denetçili kesinlik değeri 0.932, duyarlılık değeri 0.989). Ancak, Süreç 5'e yönelik saldırılarda kesinlik değeri 0.396'ya düşmekte ve YPO değeri %60'ın üzerine çıkmaktadır. Bu sonuçlar, modelin düşük seviyeli sensör saldırılarına karşı kısmen dirençli olduğunu, ancak Süreç 5'in özel dikkat gerektirdiğini göstermektedir. Ayrıca, kural denetçili yapı altındaki değerlerin, kural denetçisiz yapıya göre daha iyi olduğu görülmektedir. Bu da kural denetçili yapının amacına uygun bir savunma mekanizması işlevi gördüğünü ortaya koymaktadır.

%25 pertürbasyon ile yapılan sensör saldırılarında, metriklerin %10 pertürbasyona göre daha da düştüğü gözlemlenmektedir. Örneğin, Süreç 5'te kural denetçisiz yapı altında kesinlik değeri 0.293'e düşerken, YPO değeri %83'ün üzerine çıkmıştır. Eş zamanlı gerçekleştirilen eyleyici ve sensör saldırılarının (A+S) sonuçlarına bakıldığında, modelin performansında daha belirgin düşüşler ve artan yanlış alarmlar görülmektedir. Beklendiği gibi, kural denetçili yapı altında bu metrikler, kural denetçisiz yapıya kıyasla daha düşüktür.

Sensörlere yönelik saldırılar, özellikle düşük pertürbasyon seviyelerinde, eyleyicilere yönelik saldırılardan daha az etkili olmaktadır. Bu, sensörlerin model performansında daha

kritik bir rol oynadığını göstermektedir. Özellikle %10 pertürbasyon ile yapılan sensör saldırıları, modelin genel performansını önemli ölçüde düşürmemektedir. Buna karşılık, eyleyicilere yönelik saldırılar ve eyleyici-sensör birleşik saldırıları, modelin performansını daha ciddi şekilde etkileyebilmektedir.

Özetle, literatürde saldırısız durumda yüksek performans gösteren CNN-LSTM modelinin, OptAML saldırıları uygulandığında performansının düştüğü görülmektedir. OptAML saldırıları, güvenilir performans sergileyen DL modellerinin etkinliğini düşürmektedir. Ancak, algoritmada yer alan kural denetçisi yapısının bir savunma mekanizması olarak işlev görerek modelin performansını kısmen iyileştirdiği Tablo 24'ten açıkça anlaşılmaktadır. Tablo 24 incelendiğinde, OptAML algoritmasının ve kural denetçisinin işlevlerini başarılı bir şekilde yerine getirdiği gözlemlenmektedir.

Tablo 24. Su arıtma sistemlerinde OptAML yapısı kullanılarak üretilen çekişmeli örneklerin kural denetçisiz ve kural denetçili performansı

Kural Denetçisiz / Kural Denetçili					
Hedef Süreç	Saldırı Noktaları	Kesinlik	Duyarlılık	F1-skor	YPO
Süreç-1	A	0.331/0.419	0.695/0.849	0.449/0.561	0.774/0.650
	S(%10)	0.806/0.907	0.899/0.918	0.850/0.912	0.119/0.052
	S(%25)	0.497/0.715	0.746/0.904	0.596/0.799	0.417/0.198
	A+S(%10)	0.317/0.400	0.684/0.803	0.433/0.534	0.814/0.666
	A+S(%25)	0.293/0.378	0.647/0.770	0.403/0.507	0.863/0.700
Süreç-2	A	0.220/0.312	0.505/0.689	0.307/0.429	0.987/0.839
	S(%10)	0.561/0.740	0.834/0.881	0,671/0.804	0,360/0.171
	S(%25)	0.423/0.618	0.701/0.842	0.528/0.713	0.527/0.287
	A+S(%10)	0.200/0.298	0.453/0.658	0.278/0.410	0.998/0.856
	A+S(%25)	0.170/0.266	0.370/0.581	0.233/0.365	0.999/0.884
Süreç-3	A	0.235/0.321	0.544/0.709	0.328/0.442	0.981/0.827
	S(%10)	0.604/0.769	0.863/0.904	0,710/0.831	0.313/0.150
	S(%25)	0.455/0.653	0.723/0.871	0.559/0.747	0.478/0.255
	A+S(%10)	0.209/0.308	0.475/0.675	0.290/0.423	0.996/0.838
	A+S(%25)	0.179/0.277	0.393/0.600	0.246/0.379	0.999/0.866

Tablo 24'ün devamı

Süreç-4	A	0.246/0.414	0.572/0.762	0.344/0.537	0.970/0.595
	S(%10)	0.574/0.747	0,836/0.890	0,681/0.812	0.342/0.167
	S(%25)	0.426/0.629	0.704/0.849	0.531/0.723	0.525/0.277
	A+S(%10)	0.205/0.304	0.465/0.668	0.284/0.418	0.996/0.843
	A+S(%25)	0.175/0.271	0.383/0.589	0.240/0.371	0.999/0.875
Süreç-5	A	0.395/0.507	0.728/0.886	0.512/0.645	0.616/0.475
	S(%10)	0.392/0.510	0,739/0.818	0,512/0.628	0,633/0.433
	S(%25)	0.293/0.409	0.625/0.750	0.399/0.529	0.833/0.599
	A+S(%10)	0.290/0.360	0.640/0.763	0.399/0.489	0.866/0.750
	A+S(%25)	0.215/0.329	0.463/0.709	0.294/0.449	0.933/0.798
Süreç-6	A	0.375/0.466	0.805/0.939	0.512/0.623	0.741/0.594
	S(%10)	0.923/0.977	0.856/0.932	0.888/0.954	0.040/0.012
	S(%25)	0.605/0.760	0.782/0.926	0.682/0.834	0.282/0.162
	A+S(%10)	0.377/0.444	0.809/0.912	0.514/0.597	0.738/0.630
	A+S(%25)	0.354/0.429	0.779/0.874	0.487/0.575	0.786/0.643
Tüm Süreçler	A	0.204/0.259	0.464/0.560	0.284/0.355	1.000/0.882
	S(%10)	0.284/0.397	0.623/0.708	0.390/0.508	0.869/0.595
	S(%25)	0.231/0.303	0.537/0.642	0.323/0.411	0.986/0.815
	A+S(%10)	0.175/0.229	0.383/0.494	0.240/0.313	1.000/0.920
	A+S(%25)	0.145/0.194	0.308/0.417	0.197/0.265	1.000/0.956
Saldırısız	-	0.981	0.955	0.968	0.010

Tablo 25'te sunulan verilere göre, 35 farklı saldırı senaryosu kapsamında üretilen çekişmeli örneklerin pertürbasyonları, CS, NCC, Öklid normu ve MRM değerleri ile bir çekişmeli örneğin üretilmesi için gereken ortalama süre açısından değerlendirilmektedir. Tablodaki sonuçlar incelendiğinde, kural denetçili yapıların, kural denetçisiz yapılara kıyasla genellikle daha yüksek CS ve NCC değerlerine sahip olduğu görülmektedir. Bu durum, kural denetçili yapıların çekişmeli örneklerin gerçek ölçümlere olan benzerliğini artırdığını göstermektedir.

Özellikle kural denetçili yapılarda pertürbasyonların Öklid normu değerlerinin genellikle daha düşük olduğu gözlemlenmektedir. Bu durum, kural denetçisiz yapıların daha

az agresif ancak etkili pertürbasyonlar ürettiğini işaret etmektedir. Bu bağlamda, kural denetçili yapının pertürbasyonların kapsamını daraltarak daha kontrollü bir saldırı stratejisi oluşturduğu sonucuna varılmaktadır. Ayrıca, bir çekişmeli örneğin üretilmesi için geçen ortalama süre de kural denetçili yapıda genellikle daha uzun olmaktadır. Bu da kural denetim mekanizmasının ek hesaplama ve kontrol süreçleri gerektirdiğini göstermektedir.

Çekişmeli örneklerin, gerçek verilere yakınlığını değerlendirmede kullanılan CS ve NCC değerlerinin genellikle 0,9'un üzerinde olduğu tespit edilmiştir. Bu yüksek değerler, üretilen çekişmeli örneklerin orijinal verilere oldukça benzer olduğunu ve bu sayede modelin yanıltılabilirliğini önemli ölçüde artırdığını ortaya koymaktadır. Bilindiği üzere, CS ve NCC değerleri en fazla 1 olabilmekte olup, bu değere yaklaşıldıkça benzerlik oranı artmaktadır. Bu nedenle 0,9'un üzerindeki CS ve NCC değerleri oldukça iyi bir performansı temsil etmektedir. Bununla birlikte, tüm süreçlerin hem sensörlerine hem de eyleyicilerine yönelik eşzamanlı saldırılar tasarlandığında, benzerlik değerlerinin sırasıyla 0,72 ve 0,71'e kadar düştüğü gözlemlenmiştir. Bu durum, eşzamanlı saldırıların, çekişmeli örneklerin gerçek verilere olan benzerliğini azaltma potansiyeline sahip olduğunu göstermektedir.

Tablo 25. Su arıtma sistemlerinde OptAML yapısı kullanılarak üretilen çekişmeli örneklerle ilgili metrikler

Kural Denetçisiz / Kural Denetçili						
Hedef Süreç	Saldırı Noktaları	CS	NCC_0	$\|\|\Delta_a\|\|_2$	ρ_{adv}	t(ms)
Süreç-1	A	0.951/0.971	0.948/0.966	1.500/0.500	0.402/0.134	42/87
	S(%10)	0.996/0.998	0.995/0.998	0.140/0.099	0.038/0.027	40/76
	S(%25)	0.990/0.994	0.987/0.993	0.352/0.254	0.094/0.068	44/83
	A+S(%10)	0.947/0.969	0.941/0.964	1.506/1.122	0.404/0.301	71/184
	A+S(%25)	0.941/0.965	0.937/0.959	1.541/1.147	0.413/0.307	82/201
Süreç-2	A	0.882/0.951	0.878/0.944	2.291/1.803	0.614/0.483	95/205
	S(%10)	0.992/0.996	0.992/0.995	0.191/0.150	0.051/0.040	69/151
	S(%25)	0.981/0.987	0.974/0.983	0.498/0.382	0.133/0.102	76/166
	A+S(%10)	0.875/0.947	0.868/0.941	2.353/1.506	0.631/0.404	138/357
	A+S(%25)	0.864/0.938	0.858/0.934	2.394/1.548	0.642/0.415	151/391

Tablo 25'in devamı

Süreç-3	A	0.892/0.941	0.889/0.937	2.121/1.500	0.568/0.402	82/184
	S(%10)	0.994/0.997	0.993/0.997	0.168/0.141	0.045/0.038	53/133
	S(%25)	0.986/0.991	0.981/0.988	0.409/0.325	0.110/0.087	59/149
	A+S(%10)	0.896/0.938	0.890/0.934	2.297/1.735	0.616/0.465	122/301
	A+S(%25)	0.878/0.932	0.869/0.926	2.328/1.762	0.624/0.472	129/335
Süreç-4	A	0.902/0.978	0.895/0.973	2.236/1.414	0.599/0.379	75/169
	S(%10)	0.993/0.995	0.993/0.995	0.199/0.164	0.053/0.044	66/147
	S(%25)	0.981/0.987	0.975/0.820	0.479/0.376	0.128/0.101	76/161
	A+S(%10)	0.895/0.926	0.890/0.921	2.243/1.738	0.601/0.466	127/325
	A+S(%25)	0.883/0.928	0.870/0.922	2.287/1.627	0.613/0.436	133/348
Süreç-5	A	0.961/0.980	0.952/0.973	1.414/1.000	0.379/0.268	33/80
	S(%10)	0.978/0.990	0.973/0.987	0.321/0.223	0.086/0.060	129/283
	S(%25)	0.951/0.975	0.948/0.970	0.787/0.557	0.211/0.149	145/328
	A+S(%10)	0.931/0.971	0.926/0.965	1.445/1.022	0.387/0.274	168/461
	A+S(%25)	0.912/0.956	0.908/0.947	1.619/1.145	0.434/0.307	189/507
Süreç-6	A	0.961/0.980	0.952/0.973	1.732/1.000	0.464/0.268	41/89
	S(%10)	0.998/0.999	0.998/0.999	0.100/0.048	0.027/0.013	34/71
	S(%25)	0.995/0.998	0.995/0.998	0.250/0.173	0.067/0.046	37/79
	A+S(%10)	0.959/0.979	0.950/0.974	1.418/1.001	0.380/0.268	62/153
	A+S(%25)	0.956/0.978	0.947/0.973	1.436/1.020	0.385/0.273	67/163
Tüm Süreçler	A	0.814/0.922	0.806/0.917	3.605/2.872	0.966/0.770	156/485
	S(%10)	0.955/0.982	0.947/0.976	0.470/0.279	0.126/0.075	171/538
	S(%25)	0.912/0.945	0.908/0.940	1.060/0.826	0.284/0.221	193/584
	A+S(%10)	0.769/0.904	0.761/0.898	3.071/2.020	0.823/0.541	465/1126
	A+S(%25)	0.726/0.867	0.715/0.862	3.151/2.164	0.845/0.580	511/1284

b) Su Dağıtma Sistemi – WADI

Saldırı olmadığı durumlarda, WADI veri setinde kullanılan CNN-LSTM modelinin performans metrikleri, SWaT veri setindeki sonuçlarla benzer şekilde oldukça yüksektir. Modelin kesinlik, duyarlılık ve F1-skoru değerleri sırasıyla 0,883, 0,617 ve 0,727 olarak tespit edilmiştir. Bununla birlikte, YPO 0,018 gibi düşük bir değere sahiptir. Bu sonuçlar, saldırıların bulunmadığı koşullarda modelin oldukça yüksek bir performans sergilediğini ve

düşük bir YPO değerine sahip olduğunu göstermektedir. Ancak, saldırılar gerçekleştirildiğinde, bu performans metriklerinde belirgin bir düşüş gözlemlenmektedir. Özellikle, eyleyici ve sensör saldırılarının eşzamanlı olarak uygulandığı durumlarda model performansı ciddi şekilde azalmaktadır.

Eyleyici saldırıları (A) incelendiğinde, kural denetçisinin bulunmadığı yapı altında kesinlik ve duyarlılık değerlerinin oldukça düşük olduğu görülmektedir. Örneğin, Süreç 1'de kesinlik değeri 0,115 ve duyarlılık değeri 0,406 olarak hesaplanmıştır. Bu sonuçlar, modelin eyleyici saldırılarını tespit etmede zorlandığını ve çok sayıda FN ürettiğini göstermektedir. Öte yandan, kural denetçili yapı altında bu metriklerde bir miktar iyileşme gözlemlenmektedir. Süreç 1'de kesinlik değeri 0,178'e ve duyarlılık değeri 0,489'a yükselmiştir. Bu artışlar, kural denetçisinin bir savunma mekanizması olarak işlev gördüğünü ve modelin performansını bir dereceye kadar iyileştirdiğini göstermektedir. Ancak, bu iyileşmelerin, modelin saldırılara karşı etkin bir savunma sağladığını kanıtlamak için yeterli olmadığı açıktır.

Sensör saldırıları (S) değerlendirildiğinde, %10 pertürbasyon ile yapılan saldırılarda modelin performansı, kural denetçisinin bulunmadığı yapı altında nispeten daha yüksek olarak gözlemlenmiştir. Örneğin, Süreç 1'de kesinlik değeri 0,723 ve duyarlılık değeri 0,605 olarak belirlenmiştir. Kural denetçili yapı altında ise bu metriklerde hafif bir iyileşme görülmektedir. Kesinlik değeri 0,735'e, duyarlılık değeri ise 0,664'e yükselmektedir. Ancak, %25 pertürbasyon ile gerçekleştirilen saldırılarda performans metriklerinde belirgin bir düşüş gözlemlenmektedir. Örneğin, Süreç 1'de kesinlik değeri 0,437'ye ve duyarlılık değeri 0,558'e gerilemiştir. Bu bulgular, daha yüksek pertürbasyon seviyelerinin model üzerinde daha fazla olumsuz etkiye sahip olduğunu ortaya koymaktadır.

Eyleyici ve sensör saldırılarının eşzamanlı olarak uygulandığı durumlarda (A+S), model performansında daha belirgin düşüşler ve yanlış alarm oranlarında artışlar kaydedilmiştir. Kural denetçisinin bulunmadığı yapı altında Süreç 1'de kesinlik değeri 0,107 ve duyarlılık değeri 0,338 olarak hesaplanırken, kural denetçili yapı altında bu değerler sırasıyla 0,148 ve 0,454'e yükselmektedir. Bu değerler, yine de modelin saldırılara karşı yeterince dirençli olmadığını göstermektedir. Birleşik saldırıların etkisi, genellikle ayrı ayrı gerçekleştirilen eyleyici veya sensör saldırılarının etkisinden daha büyük olmaktadır.

Genel olarak, WADI ortamında yapılan değerlendirmeler, eyleyici ve sensör saldırılarının model performansını ciddi şekilde etkilediğini göstermektedir. Sensörlere yönelik düşük pertürbasyon seviyeleri nispeten daha az etkili olurken, eyleyicilere yönelik

saldırılar ve yüksek seviyeli pertürbasyonlar model performansını daha ciddi şekilde düşürmektedir. Kural denetçili yapı, genel olarak saldırı tespitinde daha etkili olsa da modelin tam anlamıyla korunmasını sağlamamaktadır. OptAML saldırıları, WADI sisteminde de hedeflerine ulaşarak, saldırısız durumda iyi performans gösteren CNN-LSTM modelinin performans metriklerinde düşüşlere yol açmaktadır. Kural denetçisi ise savunma mekanizması olarak işlev görmekte ve performans metriklerini kısmen iyileştirmektedir. Sonuç olarak, OptAML algoritması ve kural denetçisinin amaçlarını destekleyen değerler elde edilmiştir.

Tablo 26. Su dağıtma sistemlerinde OptAML yapısı kullanılarak üretilen çekişmeli örneklerin kural denetçisiz ve kural denetçili performansı

Kural Denetçisiz / Kural Denetçili					
Hedef Süreç	Saldırı Noktaları	Kesinlik	Duyarlılık	F1-skor	YPO
Süreç-1	A	0.115/0.178	0.406/0.489	0.179/0.261	0.700/0.505
	S(%10)	0.723/0.735	0.605/0.606	0.659/0.664	0.052/0.049
	S(%25)	0.437/0.590	0.558/0.583	0.490/0.586	0.161/0.091
	A+S(%10)	0.107/0.148	0.388/0.454	0.168/0.223	0.722/0.586
	A+S(%25)	0.083/0.122	0.325/0.402	0.132/0.188	0.806/0.646
Süreç-2	A	0.029/0.030	0.134/0.139	0.048/0.049	1.000/1.000
	S(%10)	0.081/0.091	0.349/0.369	0.132/0.146	0.884/0.827
	S(%25)	0.049/0.068	0.228/0.303	0.081/0.111	0.987/0.930
	A+S(%10)	0.024/0.028	0.110/0.127	0.039/0.045	1.000/1.000
	A+S(%25)	0.023/0.026	0.104/0.117	0.037/0.042	1.000/1.000
Süreç-3	A	0.120/0.182	0.411/0.495	0.185/0.266	0.677/0.500
	S(%10)	0.773/0.831	0.606/0.607	0.679/0.702	0.040/0.028
	S(%25)	0.449/0.719	0.565/0.595	0.500/0.651	0.155/0.052
	A+S(%10)	0.108/0.153	0.389/0.467	0.169/0.230	0.717/0.580
	A+S(%25)	0.084/0.124	0.329/0.406	0.134/0.190	0.804/0.643

Tablo 26’nın devamı

Tüm Süreçler	A	0.017/0.026	0.079/0.120	0.029/0.043	1.000/1.000
	S(%10)	0.078/0.087	0.342/0.362	0.128/0.141	0.900/0.847
	S(%25)	0.045/0.060	0.208/0.270	0.074/0.099	0.996/0.941
	A+S(%10)	0.017/0.024	0.076/0.111	0.027/0.040	1.000/1.000
	A+S(%25)	0.015/0.020	0.068/0.090	0.025/0.032	1.000/1.000
Saldırısız	-	0.883	0.617	0.727	0.018

Tablo 27, WADI veri seti için OptAML algoritması kullanılarak üretilen çeşitli pertürbasyonların kural denetçili ve kural denetçisiz yapıdaki performansını kapsamlı bir şekilde sunmaktadır. Bu analiz, kural denetçili ve kural denetçisiz yapıların farklı saldırı türleri altında performanslarını karşılaştırmakta ve kural denetçili yapının güvenlik açısından sağladığı avantajları açıkça ortaya koymaktadır. Kural denetçisinin etkisi incelendiğinde, kural denetçili yapının, kural denetçisiz yapıya kıyasla genellikle daha yüksek CS ve NCC değerlerine sahip olduğu görülmektedir. Bu yüksek CS ve NCC değerleri, üretilen çekişmeli örneklerin orijinal örneklere oldukça benzer olduğunu ve dolayısıyla modelin yanıltılabilirliğini artırabileceğini göstermektedir.

Kural denetçili yapı altında üretilen pertürbasyonların Öklid normunun genellikle daha düşük olduğu gözlemlenmiştir. Bu durum, kural denetçili yapının daha az agresif ancak etkili bir savunma mekanizması sunduğunu ortaya koymaktadır. Örneğin, Süreç 1'de sensör saldırılarında (%10 pertürbasyon) kural denetçili yapıda genlik değeri 0,384 olarak hesaplanırken, kural denetçisiz yapıda bu değer 0,579 olarak belirlenmiştir. Bu fark, kural denetçili yapının saldırıların etkisini daha etkin bir şekilde sınırlayarak modelin güvenliğini artırdığını göstermektedir.

MRM değerleri açısından da benzer bir eğilim gözlemlenmiştir. Kural denetçili yapıda bu değerler, kural denetçisiz yapıya göre genellikle daha düşük çıkmaktadır. Bu da kural denetçili yapının saldırıların etkisini azaltmada başarılı olduğunu göstermektedir. Örneğin, Süreç 2'de eyleyici ve sensör birleşik saldırılarında (A+S(%25)), MRM değeri kural denetçili yapı altında 0,560 olarak gözlemlenirken, kural denetçisiz yapıda bu değer 0,742 olarak hesaplanmıştır. Bu sonuçlar, kural denetçili yapının, saldırıların etkisini minimize etmek için daha etkin bir kontrol sağladığını göstermektedir.

Çekişmeli örnek üretimi için geçen süreler değerlendirildiğinde, kural denetçili yapının çekişmeli örnek üretim sürelerini artırdığı görülmektedir. Bu artış, kural denetçili

yapının daha karmaşık ve dirençli bir savunma mekanizması sunmasından kaynaklanmaktadır. Örneğin, Süreç 3'te eyleyici saldırıları için kural denetçisiz yapı altında üretim süresi 182 ms iken, kural denetçili yapı altında bu süre 415 ms'ye yükselmektedir. Bu durum, kural denetçili yapının, saldırıların etkisini kontrol etmek için daha fazla hesaplama gücü ve zaman gerektirdiğini ancak güvenlik açısından daha güçlü bir savunma sunduğunu göstermektedir.

Tüm süreçlerde hem eyleyicilere hem de sensörlere yönelik eşzamanlı saldırılar gerçekleştirildiğinde, benzerlik değerlerinin (CS ve NCC) kural denetçisiz yapıya kıyasla oldukça düştüğü ve genellikle 0,8'in altında kaldığı gözlemlenmiştir. Bu düşüş, saldırıların karmaşıklığı ve etkisinin artması ile ilişkilidir. Aynı zamanda, çekişmeli örnek üretim sürelerinin bu tür karmaşık saldırılar için oldukça uzun olduğu anlaşılmaktadır. Bu bulgular, kural denetçili yapının zaman açısından daha fazla hesaplama gücü gerektirdiğini, ancak saldırılara karşı daha güvenli bir yapı sağladığını ortaya koymaktadır.

Tablo 27. Su dağıtma sistemlerinde OptAML yapısı kullanılarak üretilen çekişmeli örneklerle ilgili metrikler

Kural Denetçisiz / Kural Denetçili						
Hedef Süreç	Saldırı Noktaları	CS	NCC_0	$\|\|\Delta_a\|\|_2$	ρ_{adv}	t(ms)
Süreç-1	A	0.957/0.967	0.949/0.961	1.936/1.802	0.435/0.405	117/161
	S(%10)	0.994/0.995	0.993/0.995	0.237/0.215	0.053/0.048	84/117
	S(%25)	0.987/0.994	0.983/0.993	0.579/0.384	0.130/0.086	91/124
	A+S(%10)	0.952/0.962	0.949/0.954	1.818/1.748	0.409/0.393	173/231
	A+S(%25)	0.944/0.961	0.940/0.953	2.021/1.843	0.454/0.414	205/264
Süreç-2	A	0.900/0.952	0.984/0.949	2.958/2.121	0.665/0.477	183/245
	S(%10)	0.966/0.970	0.960/0.965	0.596/0.554	0.134/0.125	241/334
	S(%25)	0.920/0.935	0.914/0.930	1.458/1.323	0.328/0.298	253/349
	A+S(%10)	0.866/0.923	0.861/0.918	3.017/2.192	0.678/0.493	682/872
	A+S(%25)	0.820/0.887	0.812/0.879	3.298/2.491	0.742/0.560	778/1059

Tablo 27'nin devamı

Süreç-3	A	0.962/0.976	0.952/0.971	1.871/1.500	0.421/0.337	121/167
	S(%10)	0.996/0.997	0.995/0.997	0.194/0.169	0.044/0.038	81/120
	S(%25)	0.991/0.995	0.988/0.995	0.473/0.333	0.106/0.075	89/129
	A+S(%10)	0.958/0.973	0.950/0.969	1.881/1.509	0.423/0.339	168/220
	A+S(%25)	0.953/0.971	0.949/0.967	1.929/0.537	0.434/0.346	197/247
Tüm Süreçler	A	0.867/0.938	0.862/0.934	3.464/2.398	0.779/0.539	345/453
	S(%10)	0.954/0.961	0.947/0.952	0.695/0.632	0.156/0.142	359/492
	S(%25)	0.911/0.930	0.992/0.926	1.561/1.392	0.351/0.313	372/511
	A+S(%10)	0.820/0.899	0.812/0.893	3.533/2.479	0.794/0.557	823/1085
	A+S(%25)	0.778/0.868	0.769/0.863	3.791/2.773	0.852/0.624	981/1386

2.8.2. Çekişmeli Makine Öğrenmesi

AML, ML modellerinin güvenliğini ve dayanıklılığını artırmak amacıyla geliştirilen bir alan olarak hızla önem kazanmaktadır. Bu yaklaşım, özellikle DL modellerinin güvenlik açıklarını ve zayıflıklarını ortaya çıkarmak için kullanılan teknikleri içermektedir. Çekişmeli örnekler, ML modellerinin yanıltılması amacıyla tasarlanmış, insan gözünün neredeyse fark edemeyeceği küçük pertürbasyonlarla değiştirilmiş veri örnekleridir. Szegedy ve arkadaşlarının [38] çalışması, küçük pertürbasyonların bile DL modellerinde büyük hata oranlarına neden olabileceğini göstererek, bu tür örneklerin modellerin güvenilirliğini ciddi şekilde azaltabileceğini ortaya koymaktadır.

AML'de kullanılan çeşitli teknikler, çekişmeli örnekler oluşturmayı ve bu örneklerle modelleri test etmeyi amaçlamaktadır. En yaygın kullanılan yöntemlerden biri olan FGSM, Goodfellow ve arkadaşları [39] tarafından geliştirilmiştir. FGSM, modelin kayıp fonksiyonunun gradyanını kullanarak çekişmeli örnekler üretir ve böylece modelin zayıf yönlerini ortaya çıkarmaktadır. Diğer önemli yöntemler arasında PGD ve Carlini & Wagner (C&W) saldırıları yer almaktadır. Bu yöntemler, çekişmeli örnekler üretirken farklı optimizasyon stratejilerini kullanarak modelin güvenlik açıklarını test etmektedir. Bu saldırı teknikleri, modelin güvenilirliğini sınamak ve zayıf noktalarını belirlemek için güçlü araçlar sunmaktadır.

AML, çeşitli alanlarda önemli bir araştırma konusu haline gelmiştir. Görüntü işleme alanında AML, görüntü sınıflandırma ve nesne tanıma sistemlerinin güvenilirliğini artırmak

için kullanılmaktadır. Kurakin ve arkadaşları [76] , DL modellerine karşı fiziksel dünyada yapılan saldırılarının modelin performansını nasıl etkilediğini göstermişlerdir. Bu çalışmada, fiziksel nesneler üzerinde yapılan çekişmeli saldırıların, modellerin yanıltılmasında ne kadar etkili olabileceği vurgulanmaktadır. Benzer şekilde, Athalye ve arkadaşları [77], 3D baskı teknolojisi kullanarak fiziksel nesneler üzerinde çekişmeli örnekler oluşturmuş ve modellerin güvenlik açıklarını test etmişlerdir. Bu tür çalışmalar, AML'nin gerçek dünya uygulamalarını ve karşılaşılan zorlukları gözler önüne sermektedir.

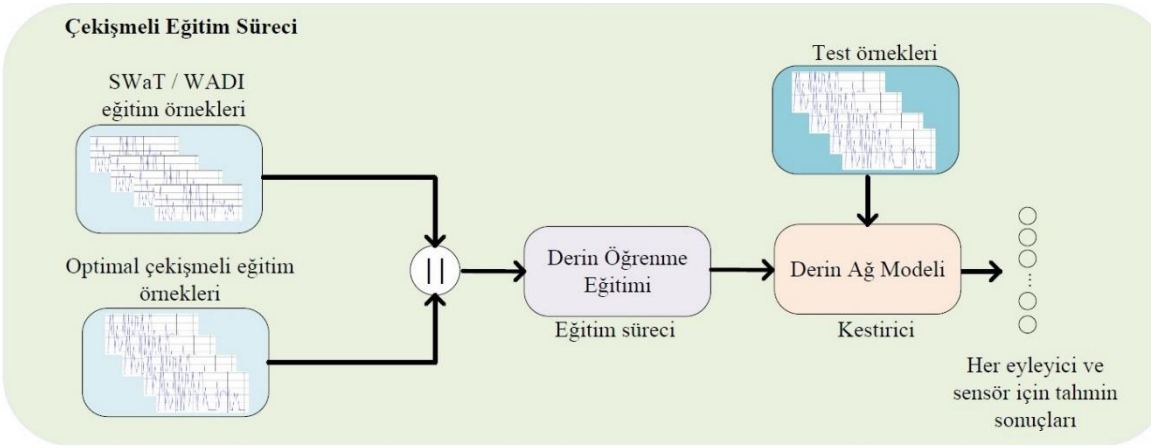
Medikal ses işaretleri üzerinde yapılan çalışmalar, AML'nin bu alandaki potansiyelini ortaya koymaktadır. Hannun ve arkadaşları [56], konuşma tanıma sistemlerine yönelik çekişmeli saldırıların modelin doğruluğunu büyük ölçüde düşürebileceğini göstermişlerdir. Bu araştırma, çekişmeli örneklerin medikal uygulamalarda oluşturabileceği tehditleri ve bu tür saldırılara karşı alınabilecek önlemleri tartışmaktadır. Benzer şekilde, Li ve arkadaşları [78], medikal ses işaretlerinde çekişmeli örneklerin nasıl kullanılabileceğini ve bu tür saldırılara karşı sistemlerin nasıl korunabileceğini incelemişlerdir. Bu çalışmalar, medikal alanda AML'nin potansiyel uygulamalarını ve bu alanda karşılaşılan zorlukları vurgulamaktadır.

AML çalışmalarında, sadece saldırı teknikleri değil, aynı zamanda bu saldırılara karşı geliştirilen savunma mekanizmaları da kritik önem taşımaktadır. Madry ve arkadaşları [41] tarafından önerilen çekişmeli eğitim (AT), modellerin çekişmeli örneklere karşı dayanıklılığını artırmak için kullanılan etkili bir yöntemdir. Bu yöntem, modelin eğitim sürecinde çekişmeli örneklerin de dahil edilmesini sağlamakta ve böylece modelin bu tür saldırılara karşı daha dirençli hale gelmesi hedeflemektedir. AT, modelin çeşitli saldırı türlerine karşı dayanıklılığını artırmak için güçlü bir savunma stratejisi olarak öne çıkmaktadır.

Sonuç olarak, AML, DL modellerinin güvenliğini ve dayanıklılığını artırmak için kritik bir alan olarak ortaya çıkmıştır. Görüntü işleme ve medikal ses işaretleri gibi çeşitli uygulama alanlarında yapılan çalışmalar, bu alandaki potansiyeli ve zorlukları gözler önüne sermektedir. Çekişmeli örneklerin üretilmesi ve bu tür saldırılara karşı geliştirilen savunma mekanizmaları, gelecekte daha güvenli ve sağlam ML sistemlerinin geliştirilmesine katkı sağlayacaktır.

2.9. Çekişmeli Eğitim

AT, ML ve DL modellerinin çekişmeli örneklere karşı dayanıklılığını artırmayı amaçlayan bir tekniktir. Bu teknik, modelin eğitim sürecine çekişmeli örneklerin dahil edilmesiyle, modelin bu tür saldırılara karşı dirençli hale gelmesini sağlamaktadır. Çekişmeli örnekler, orijinal verilerin insan gözünün fark edemeyeceği derecede küçük pertürbasyonlarla değiştirilmesiyle oluşturulmaktadır. Ancak, bu ufak değişiklikler bile modelin karar mekanizmalarını ciddi şekilde etkileyebilmekte ve hata oranlarını artırabilmektedir. Şekil 15'te çekişmeli eğitim süreci detaylı gösterilmektedir.



Şekil 15. Savunma mekanizması için kullanılan çekişmeli eğitim sürecinin genel yapısı

Goodfellow ve arkadaşları [39] tarafından geliştirilen FGSM, AT'de kullanılan ilk yöntemlerden biridir. FGSM, modelin kayıp fonksiyonunun gradyanını kullanarak hızlı bir şekilde çekişmeli örnekler üretir ve böylece modelin zayıflıklarını belirgin bir hale getirmektedir. Eğitim süreci boyunca model, hem normal veriler hem de FGSM tarafından oluşturulan çekişmeli örneklerle eğitilerek, bu tür saldırıları tanıma ve onlara karşı direnç gösterme kapasitesi kazanmaktadır.

AT'nin daha gelişmiş bir versiyonu olan PGD, Madry ve arkadaşları [34] tarafından geliştirilmiştir. PGD tabanlı AT, modelin çeşitli çekişmeli saldırılara karşı dayanıklılığını büyük ölçüde artırmaktadır. FGSM modelinin aksine PGD, iteratif bir süreç kullanarak her adımda pertürbasyonları küçük miktarlarda güncellemektedir. Bu iteratif yaklaşım, daha karmaşık ve güçlü çekişmeli örnekler oluşturarak modelin savunma kapasitesini daha da artırmaktadır.

Tramer ve arkadaşları [79] ise, AT'nin model genelleme kapasitesini artırdığını ve modelin yalnızca çekişmeli örneklere değil, aynı zamanda normal veri dağılımı üzerinde de

daha iyi performans sergilediğini göstermişlerdir. Bu çalışma, AT'nin modelin genel dayanıklılığını ve esnekliğini artırarak, farklı veri dağılımlarına karşı daha dirençli hale gelmesini sağladığını ortaya koymaktadır. AT ile eğitilen modellerin, veri dağılımındaki değişimlere ve beklenmedik varyasyonlara daha iyi uyum sağlayabildiği gözlemlenmiştir.

AT, farklı veri türleri ve uygulama alanlarında da etkili bir yöntemdir. Örneğin, Xie ve ekibi [34] , AT'nin doğal dil işleme (NLP) modellerinde kullanılabileceğini göstermiştir. Bu çalışmada, NLP modellerine uygulanan AT'nin, yanıltıcı metin saldırılarına karşı dayanıklılığı artırdığı gösterilmiştir. Bu durum, AT'nin yalnızca görüntü işleme değil, aynı zamanda metin işleme gibi farklı veri türlerinde de etkinliğini ortaya koymaktadır.

Özetle, AT, ML ve DL modellerinin güvenlik ve dayanıklılığını artırmada güçlü bir yöntemdir. Bu strateji, yalnızca çekişmeli saldırılara karşı değil, aynı zamanda genel performanslarını da iyileştirerek daha güvenilir ve sağlam sistemlerin geliştirilmesine katkıda bulunmaktadır. AT'nin temel amacı, eğitim sürecine çekişmeli örnekleri dahil ederek, modellerin bu tür örneklere karşı daha dayanıklı hale gelmesini sağlamaktır. Bu strateji, çekişmeli örneklerin oluşturulmasını ve bu örneklerin eğitim veri kümesine eklenmesini içermektedir. Model hem temiz hem de çekişmeli örnekler üzerinde eğitilerek potansiyel saldırılara karşı uyum sağlama ve onlarla başa çıkma yeteneğini geliştirmektedir. Böylece, model hem temiz verilerde hem de bozulmuş örneklerde yüksek doğruluğu koruyarak daha dayanıklı hale gelmektedir.

AT, bir min-maks optimizasyon problemi olarak tanımlanmaktadır. Denklem 19'da gösterildiği üzere, AT, iç maksimizasyon ve dış minimizasyondan oluşur. İç maksimizasyon, çekişmeli örneklerin üretilmesini sağlarken, dış minimizasyon, bu çekişmeli örnekler üzerinde dayanıklı bir model eğitmeye odaklanmaktadır.

$$min_{\theta} E_{x,y\sim D}\left[max_{\delta:\|\delta\|_p \leq \epsilon} J(\theta, x+\delta, y)\right] \tag{19}$$

Burada D veri dağılımını, θ modeli ve onun parametrelerini, $J(..,..,.)$, $x+\delta$ bozulmuş girdiyi ve y gerçek çıktısı için hesaplanan kayıp fonksiyonunu, $max_{\delta:\|\delta\|_p \leq \epsilon}$ pertürbasyonlar üzerindeki maksimumlaştırmayı ifade etmektedir.

2.9.1. Çekişmeli Eğitim Sonuçları

a) Su Arıtma Sistemi - SWaT

AT için CNN-LSTM modeli, SWaT ve WADI sistemlerinin eğitim örnekleri kullanılarak oluşturulan ve optimal saldırılarla üretilmiş çekişmeli örnekleri içeren bir veri setiyle yeniden eğitilmektedir. Bu doğrultuda, eğitim veri seti, saldırı uygulanmamış ve çekişmeli örneklerin eşit dağılımından oluşturulmuş ve birleştirilmiştir. Çekişmeli örnekler setine, Tablo 24'te belirtilen A, S(%10), S(%25), A+S(%10) ve A+S(%25) saldırı noktalarının her biri için oluşturulan çekişmeli örnekler %20 oranında eklenmiştir.

AT sürecinde, eğitilmiş derin ağ modeli bir kestirici olarak kullanılarak her bir test örneği için tahmin sonuçları elde edilmiştir. Bu bağlamda, Tablo 24'te yer alan saldırılar altındaki CNN-LSTM modelinin aykırılık tespit performansı Tablo 28'de detaylandırılmaktadır. Bu tablo, SWaT sistemi için CNN-LSTM modelinin 35 farklı saldırı senaryosunda, AT öncesi ve sonrası performansını özetlemektedir.

AT öncesi, CNN-LSTM modelinin genel olarak daha düşük bir performans sergilediği görülmektedir. Kesinlik, duyarlılık ve F1 skorlarının daha düşük, YPO değerinin ise daha yüksek olduğu gözlemlenmektedir. AT sonrasında ise modelin kesinlik, duyarlılık ve F1 skor değerlerinde artış gözlemlenirken, YPO değerinde ise düşüş gözlemlenmektedir. Süreç bazında yapılan değerlendirmeler, her bir süreç için bu eğilimin geçerli olduğunu göstermektedir. AT sonrasında, kesinlik değerlerinin çoğunlukla 0.5'in üzerinde, YPO değerlerinin ise 0.5'in altında kaldığı tespit edilmiştir. Sonuç olarak, saldırı içeren veri setleriyle eğitilen derin ağ modeli, AT sonrasında dayanıklılığını ve genel performansını önemli ölçüde geliştirmektedir. CNN-LSTM modelinin AT sonrasında elde ettiği performans metriklerine bakıldığında, derin ağ modelinin performansında belirgin bir iyileşme olduğu görülmektedir. Tablo 28'de sunulan veriler de bu durumu doğrulamaktadır. Bu sonuçlar, AT işleminin hedefine uygun bir şekilde gerçekleştirildiğini göstermektedir.

Tablo 28. Su arıtma sistemlerinde CNN-LSTM modelinin çekişmeli eğitimli ve çekişmeli eğitimsiz performansı

Çekişmeli Eğitim VAR/YOK					
Hedef Süreç	Saldırı Noktaları	Kesinlik	Duyarlılık	F1-skor	YPO
Süreç-1	A	0.631/0.419	0.906/0.849	0.744/0.561	0.292/0.650
	S(%10)	0.938/0.907	0.937/0.918	0.937/0.912	0.034/0.052
	S(%25)	0.784/0.715	0.939/0.904	0.855/0.799	0.143/0.198
	A+S(%10)	0.595/0.400	0.873/0.803	0.708/0.534	0.328/0.666
	A+S(%25)	0.574/0.378	0.855/0.770	0.687/0.507	0.350/0.700
Süreç-2	A	0.510/0.312	0.824/0.689	0.630/0.429	0.437/0.839
	S(%10)	0.868/0.740	0.927/0.881	0.896/0.804	0.078/0.171
	S(%25)	0.723/0.618	0.886/0.842	0.796/0.713	0.187/0.287
	A+S(%10)	0.486/0.298	0.816/0.658	0.609/0.410	0.476/0.856
	A+S(%25)	0.472/0.266	0.801/0.581	0.594/0.365	0.495/0.884
Süreç-3	A	0.523/0.321	0.839/0.709	0.644/0.442	0.422/0.827
	S(%10)	0.901/0.769	0.933/0.904	0.917/0.831	0.057/0.150
	S(%25)	0.761/0.653	0.912/0.871	0.830/0.747	0.158/0.255
	A+S(%10)	0.506/0.308	0.824/0.675	0.627/0.423	0.444/0.838
	A+S(%25)	0.494/0.277	0.811/0.600	0.614/0.379	0.459/0.866
Süreç-4	A	0.683/0.414	0.926/0.762	0.786/0.537	0.237/0.595
	S(%10)	0.878/0.747	0.922/0.890	0.899/0.812	0.071/0.167
	S(%25)	0.734/0.629	0.894/0.849	0.806/0.723	0.179/0.277
	A+S(%10)	0.502/0.304	0.823/0.668	0.623/0.418	0.451/0.843
	A+S(%25)	0.491/0.271	0.808/0.589	0.610/0.371	0.463/0.875
Süreç-5	A	0.797/0.507	0.936/0.886	0.861/0.645	0.132/0.475
	S(%10)	0.717/0.510	0.860/0.818	0.782/0.628	0.188/0.433
	S(%25)	0.558/0.409	0.815/0.750	0.662/0.529	0.357/0.599
	A+S(%10)	0.549/0.360	0.842/0.763	0.665/0.489	0.382/0.750
	A+S(%25)	0.525/0.329	0.834/0.709	0.645/0.449	0.416/0.798

Tablo 28’in devamı

Süreç-6	A	0.703/0.466	0.956/0.939	0.810/0.623	0.223/0.594
	S(%10)	0.975/0.977	0.944/0.932	0.960/0.954	0.013/0.012
	S(%25)	0.856/0.760	0.935/0.926	0.894/0.834	0.087/0.162
	A+S(%10)	0.634/0.444	0.929/0.912	0.754/0.597	0.296/0.630
	A+S(%25)	0.607/0.429	0.915/0.874	0.730/0.575	0.328/0.643
Tüm Süreçler	A	0.501/0.259	0.822/0.560	0.623/0.355	0.451/0.882
	S(%10)	0.597/0.397	0.823/0.708	0.692/0.508	0.307/0.595
	S(%25)	0.505/0.303	0.781/0.642	0.613/0.411	0.422/0.815
	A+S(%10)	0.470/0.229	0.797/0.494	0.591/0.313	0.496/0.920
	A+S(%25)	0.451/0.194	0.784/0.417	0.572/0.265	0.527/0.956
Saldırısız	-	0.981	0.955	0.968	0.010

b) Su Dağıtma Sistemi – WADI

WADI veri seti üzerinde CNN-LSTM modeli, AT uygulanmadan önce ve sonrasında, çeşitli saldırı senaryoları altında değerlendirilmektedir. AT yapılmadan önce, derin ağ modelinin kesinlik, duyarlılık ve F1 skoru gibi performans ölçütleri genellikle düşükken, YPO değeri yüksektir. Örneğin, Süreç-1'de eyleyici saldırısı (A) için AT olmadan önce kesinlik değeri 0.178, duyarlılık değeri 0.489, F1 skoru değeri 0.261 ve YPO değeri 0.505 olarak belirlenmiştir. Ancak, AT'nin uygulanmasıyla performans metriklerinde kayda değer bir iyileşme gözlemlenmektedir.

AT uygulandıktan sonra, kesinlik, duyarlılık ve F1 skoru değerlerinde artış görülürken, YPO değerinde bir azalma meydana gelmektedir. Örneğin, Süreç-1'de eyleyici saldırısı (A) altında AT sonrası kesinlik değeri 0.715'e, duyarlılık değeri 0.555'e ve F1 skoru değeri 0.625'e yükselirken, YPO değeri 0.050'ye düşmüştür. Bu sonuçlar, AT’nin CNN-LSTM modelini saldırılara karşı daha dayanıklı ve doğru bir yapıya kavuşturduğunu göstermektedir.

Her bir süreçte, AT öncesi ve sonrasındaki performans eğilimleri incelendiğinde, tüm süreçlerde metriklerde iyileşme olduğu görülmüştür. Özellikle, Süreç-3'te %25 pertürbasyonlu sensör saldırıları (S(%25)) altında AT sonrası kesinlik değeri 0.817 ve YPO değeri 0.052 olarak hesaplanmıştır. Bu sonuçlar, modelin her bir süreçte saldırı senaryoları karşısında daha yüksek bir performans sergilediğini göstermektedir.

Ayrıca, eyleyici ve sensör birleşik saldırıları (A+S) altında da AT'nin etkileri olumlu olmuştur. Özellikle tüm süreçlerin hedef alındığı senaryolarda, AT sonrasında kesinlik ve duyarlılık değerleri artmış, YPO ise azalmıştır. Örneğin, A+S(%25) saldırısı altında tüm süreçler için AT öncesi ve sonrası YPO değerleri sırasıyla 1.000 ve 0.448 olarak hesaplanmıştır.

Özetle, WADI veri seti üzerinde AT uygulanan CNN-LSTM modelinin performansında dikkate değer bir artış kaydedilmektedir. AT sonrasında derin ağ modelinin kesinlik değerleri genellikle 0.5'in üzerinde kalırken, YPO değerleri çoğunlukla 0.5'in altına düşmüştür. Bu bulgular, CNN-LSTM modelinin saldırı içeren veri setleriyle eğitilmesinin ardından dayanıklılığını ve genel başarımını artırdığını göstermektedir. AT uygulaması, modeli daha sağlam hale getirmiş ve saldırıları daha etkili bir şekilde tespit etme kapasitesini güçlendirmektedir.

Tablo 29. Su dağıtma sistemlerinde CNN-LSTM modelinin çekişmeli eğitimli ve çekişmeli eğitimsiz performansı

Çekişmeli Eğitim VAR/YOK					
Hedef Süreç	Saldırı Noktaları	Kesinlik	Duyarlılık	F1-skor	YPO
Süreç-1	A	0.715/0.178	0.555/0.489	0.625/0.261	0.050/0.505
	S(%10)	0.822/0.735	0.612/0.606	0.702/0.664	0.030/0.049
	S(%25)	0.789/0.590	0.598/0.583	0.680/0.586	0.036/0.091
	A+S(%10)	0.601/0.148	0.543/0.454	0.571/0.223	0.081/0.586
	A+S(%25)	0.552/0.122	0.533/0.402	0.542/0.188	0.097/0.646
Süreç-2	A	0.249/0.030	0.521/0.139	0.337/0.049	0.353/1.000
	S(%10)	0.553/0.091	0.578/0.369	0.565/0.146	0.104/0.827
	S(%25)	0.518/0.068	0.548/0.303	0.532/0.111	0.114/0.930
	A+S(%10)	0.220/0.028	0.505/0.127	0.307/0.045	0.400/1.000
	A+S(%25)	0.203/0.026	0.493/0.117	0.288/0.042	0.433/1.000
Süreç-3	A	0.750/0.182	0.569/0.495	0.647/0.266	0.042/0.500
	S(%10)	0.834/0.831	0.612/0.607	0.706/0.702	0.027/0.028
	S(%25)	0.817/0.719	0.603/0.595	0.694/0.651	0.030/0.052
	A+S(%10)	0.635/0.153	0.550/0.467	0.589/0.230	0.071/0.580
	A+S(%25)	0.562/0.124	0.537/0.406	0.549/0.190	0.094/0.643

Tablo 29’un devamı

Tüm Süreçler	A	0.243/0.026	0.516/0.120	0.331/0.043	0.360/1.000
	S(%10)	0.529/0.087	0.575/0.362	0.551/0.141	0.115/0.847
	S(%25)	0.502/0.060	0.541/0.270	0.521/0.099	0.120/0.941
	A+S(%10)	0.216/0.024	0.499/0.111	0.301/0.040	0.406/1.000
	A+S(%25)	0.194/0.020	0.482/0.090	0.277/0.032	0.448/1.000
Saldırısız	-	0.883	0.617	0.727	0.018

3. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında, su sistemlerinin güvenliğini artırmak amacıyla aykırılık tespit yapılarında kullanılan derin ağ modellerinin çekişmeli saldırılara karşı dayanıklılığı irdelenmiş ve savunma mekanizması geliştirilmiştir. Bu kapsamda öncelikle CNN-LSTM modeli kullanılarak SWaT ve WADI sistemlerindeki olası aykırılıklar tespit edilmiş ve modelin performansı kesinlik, duyarlılık, F1-skoru ve YPO gibi kritik metriklerle değerlendirilmiştir. Tek nokta, süreç bazlı ve tüm süreç eyleyici saldırıları gibi çeşitli saldırı senaryoları karşısında modelin genellikle düşük performans sergilediği ve bu sistemlerin saldırılara karşı savunmasız olduğu görülmüştür.

Tezde, OptAML yapısı önerilerek optimal çekişmeli örnekler üretilmiştir. İlgili çekişmeli örnekler, sistemin normal işleyişini yanıltacak şekilde üretilmiştir. Bu örnekler kullanılarak modelin performansı hem kural denetçili hem de kural denetçisiz yapılar altında incelenmiştir. Elde edilen sonuçlar, kural denetçili yapının saldırı tespitinde daha etkili olduğunu ancak daha katı bir güvenlik politikası uyguladığını; kural denetçisiz yapının ise daha esnek fakat daha savunmasız olduğunu göstermiştir.

Bu bağlamda, AT ile modellenen sistemler, saldırılara karşı daha dirençli hale getirilmiş ve güvenlik açısından daha yüksek kesinlik, duyarlılık ve F1-skor değerlerine ulaşılmıştır. Bu bulgular, su sistemlerinin güvenliğini artırmak için AT'nin önemini vurgulamaktadır.

Sonuç olarak, bu çalışma ile, su sistemlerinin güvenliğini artırmak için yeni stratejiler sunmakta ve özellikle AML ve AT yaklaşımları yardımıyla su sistemlerinin saldırılara karşı daha dayanıklı hale getirilebileceğini gösterilmiştir. Çalışma boyunca elde edilen bulgular ve geliştirilen yöntemler, ICS güvenliği alanında yapılacak gelecekteki araştırmalar için önemli bir temel oluşturmaktadır.

4. KAYNAKÇA

[1] Raman, G., Somu, N., & Mathur, A. P. (2020). A multilayer perceptron model for anomaly detection in water treatment plants. *International Journal of Critical Infrastructure Protection,* 31, 100393.

[2] Bozdal, M., Ileri, K., & Ozkahraman, A. (2024). Comparative analysis of dimensionality reduction techniques for cybersecurity in the SWaT dataset. *Journal of Supercomputing,* 80(1), 1059–1079.

[3] Aboah Boateng, E., Bruce, J. W., & Talbert, D. A. (2022). Anomaly detection for a water treatment system based on one-class neural network. *IEEE Access,* 10, 115179–115191.

[4] Xie, X., Wang, B., Wan, T., & Tang, W. (2020). Multivariate abnormal detection for industrial control systems using 1D CNN and GRU. *IEEE Access*, 8, 88348–88359.

[5] Shabani, H., Goudarzi, N., & Shabani, M. (2018). Failure analysis of a natural gas pipeline. *Engineering Failure Analysis*, 84, 167–184.

[6] Cárdenas, A. A., Amin, S., Lin, Z. S., Huang, Y. L., Huang, C. Y., & Sastry, S. (2011, March). Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the 6th ACM symposium on information, computer and communications security* (pp. 355-366).

[7] Raman, M. G., Dong, W., & Mathur, A. (2020). Deep autoencoders as anomaly detectors: Method and case study in a distributed water treatment plant. *Computers & Security*, *99*, 102055.

[8] Özgenç, B., Ayas, S., Doğan, R. Ö., Çavdar, B., Şahin, A. K., & Ayas, M. Ş. (2023, July). Anomaly Detection in Predicted Water Treatment Data Using Hybrid CNN-LSTM Network Model. In *2023 31st Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[9] Mathur, A. P., & Tippenhauer, N. O. (2016, April). SWaT: A water treatment testbed for research and training on ICS security. In *2016 international workshop on cyber-physical systems for smart water networks (CySWater)* (pp. 31-36). IEEE.

[10] Zhang, J., Gardner, R., & Vukotic, I. (2019). Anomaly detection in wide area network meshes using two machine learning algorithms. *Future Generation Computer Systems, 93*, 418-426.

[11] Imamverdiyev, Y., & Sukhostat, L. (2016, October). Anomaly detection in network traffic using extreme learning machine. In *2016 IEEE 10th international conference on application of information and communication technologies (AICT)* (pp. 1-4). IEEE.

[12] Buczak, A. L., Berman, D. S., Yen, S. W., Watkins, L. A., Duong, L. T., & Chavis, J. S. (2017, April). Using sequential pattern mining for common event format (CEF) cyber data. In *Proceedings of the 12th annual conference on cyber and information security research* (pp. 1-4).

[13] Liu, S., Chen, X., Peng, X., & Xiao, R. (2019, December). Network log anomaly detection based on gru and svdd. In *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)* (pp. 1244-1249). IEEE.

[14] Lin, W. C., Ke, S. W., & Tsai, C. F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-based systems*, *78*, 13-21.

[15] Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A survey of deep learning-based network anomaly detection. *Cluster Computing*, *22*, 949-961.

[16] Alalshekmubarak, A., & Smith, L. S. (2013, March). A novel approach combining recurrent neural network and support vector machines for time series classification. In *2013 9th International Conference on Innovations in Information Technology (IIT)* (pp. 42-47). IEEE.

[17] Zou, M., Wang, C., Li, F., & Song, W. (2018). Network phenotyping for network traffic classification and anomaly detection. *arXiv preprint arXiv:1803.01528*.

[18] Meshram, A., & Haas, C. (2017). Anomaly detection in industrial networks using machine learning: a roadmap. In *Machine learning for cyber physical systems: selected papers from the international conference ML4CPS 2016* (pp. 65-72). Springer Berlin Heidelberg.

[19] Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017, November). Anomaly detection for a water treatment system using unsupervised machine learning. In *2017 IEEE international conference on data mining workshops (ICDMW)* (pp. 1058-1065). IEEE.

[20] Kravchik, M., & Shabtai, A. (2018, January). Detecting cyber attacks in industrial control systems using convolutional neural networks. In *Proceedings of the 2018 workshop on cyber-physical systems security and privacy* (pp. 72-83).

[21] Ayas, S., & Ayas, M. S. (2022). A modified densenet approach with nearmiss for anomaly detection in industrial control systems. *Multimedia Tools and Applications*, 1-14.

[22] Li, D., Chen, D., Jin, B., Shi, L., Goh, J., & Ng, S. K. (2019, September). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks* (pp. 703-716). Cham: Springer International Publishing.

[23] Kravchik, M., & Shabtai, A. (2021). Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. *IEEE transactions on dependable and secure computing*, *19*(4), 2179-2197.

[24] Elnour, M., Meskin, N., Khan, K., & Jain, R. (2020). A dual-isolation-forests-based attack detection framework for industrial control systems. *IEEE Access*, *8*, 36639-36651.

[25] Goodge, A., Hooi, B., Ng, S. K., & Ng, W. S. (2021, January). Robustness of autoencoders for anomaly detection under adversarial impact. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 1244-1250).

[26] Liu, Y., Xu, L., Yang, S., Zhao, D., & Li, X. (2024). Adversarial sample attacks and defenses based on LSTM-ED in industrial control systems. *Computers & Security*, *140*, 103750.

[27] Abdelaty, M. F. Y. (2022). Robust Anomaly Detection in Critical Infrastructure.

[28] Jia, Y., Wang, J., Poskitt, C. M., Chattopadhyay, S., Sun, J., & Chen, Y. (2021). Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. *International Journal of Critical Infrastructure Protection*, *34*, 100452.

[29] Yuan, L., Yu, S., Yang, Z., Duan, M., & Li, K. (2023). A data balancing approach based on generative adversarial network. *Future Generation Computer Systems*, *141*, 768-776.

[30] Zhang, X., Wang, J., & Zhu, S. (2021). Dual generative adversarial networks based unknown encryption ransomware attack detection. *IEEE Access*, *10*, 900-913.

[31] Sun, H., Huang, Y., Han, L., Fu, C., Liu, H., & Long, X. (2024). MTS-DVGAN: Anomaly detection in cyber-physical systems using a dual variational generative adversarial network. *Computers & Security*, *139*, 103570.

[32] Zeng, G. Q., Shao, J. M., Lu, K. D., Geng, G. G., & Weng, J. (2024). Automated federated learning-based adversarial attack and defence in industrial control systems. *IET Cyber-Systems and Robotics*, *6*(2), e12117.

[33] Zhao, J., Yang, S., Li, Q., Liu, Y., Gu, X., & Liu, W. (2021). A new bearing fault diagnosis method based on signal-to-image mapping and convolutional neural network. *Measurement*, *176*, 109088.

[34] Mądry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *stat*, *1050*(9).

[35] Wen, L., Li, X., Gao, L., & Zhang, Y. (2017). A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics*, *65*(7), 5990-5998.

[36] Guo, L., Lei, Y., Xing, S., Yan, T., & Li, N. (2018). Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics, 66*(9), 7316-7325.

[37] Liu, Z. H., Lu, B. L., Wei, H. L., Chen, L., Li, X. H., & Rätsch, M. (2019). Deep adversarial domain adaptation model for bearing fault diagnosis. *IEEE transactions on systems, man, and cybernetics: Systems, 51*(7), 4217-4226.

[38] Szegedy, C. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[39] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[40] Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574-2582).

[41] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

[42] Suciu, O., Marginean, R., Kaya, Y., Daume III, H., & Dumitras, T. (2018). When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)* (pp. 1299-1316).

[43] Kuppa, A., Grzonkowski, S., Asghar, M. R., & Le-Khac, N. A. (2019, August). Black box attacks on deep anomaly detectors. In *Proceedings of the 14th international conference on availability, reliability and security* (pp. 1-10).

[44] Xu, P., Kolosnjaji, B., Eckert, C., & Zarras, A. (2020, March). Manis: Evading malware detection system on graph structure. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (pp. 1688-1695).

[45] Adepu, S., & Mathur, A. (2016, May). Distributed detection of single-stage multipoint cyber attacks in a water treatment plant. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security* (pp. 449-460).

[46] Umer, M. A., Mathur, A., Junejo, K. N., & Adepu, S. (2020). Generating invariants using design and data-centric approaches for distributed attack detection. *International Journal of Critical Infrastructure Protection, 28*, 100341.

[47] Elnour, M., Meskin, N., Khan, K., & Jain, R. (2020). A dual-isolation-forests-based attack detection framework for industrial control systems. *IEEE Access, 8*, 36639-36651.

[48] Sun, H., Huang, Y., Han, L., Fu, C., Liu, H., & Long, X. (2024). MTS-DVGAN: Anomaly detection in cyber-physical systems using a dual variational generative adversarial network. *Computers & Security, 139*, 103570.

[49] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

[50] Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*, 85-126.

[51] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

[52] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1-21.

[53] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58.

[54] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[55] Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.

[56] Hannun, A. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[57] Igure, V. M., Laughter, S. A., & Williams, R. D. (2006). Security issues in SCADA networks. *computers & security*, *25*(7), 498-506.

[58] Kreibich, C., & Crowcroft, J. (2004). Honeycomb: creating intrusion detection signatures using honeypots. *ACM SIGCOMM computer communication review*, *34*(1), 51-56.

[59] Venkatasubramanian, S., Raja, S., Sumanth, V., Dwivedi, J. N., Sathiaparkavi, J., Modak, S., & Kejela, M. L. (2022). Fault diagnosis using data fusion with ensemble deep learning technique in IIoT. *Mathematical Problems in Engineering*, *2022*(1), 1682874.

[60] Shalyga, D., Filonov, P., & Lavrentyev, A. (2018). Anomaly detection for water treatment system based on neural network with automatic architecture optimization. *arXiv preprint arXiv:1807.07282*.

[61] Macas, M., & Wu, C. (2019, December). An unsupervised framework for anomaly detection in a water treatment system. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 1298-1305). IEEE.

[62] Deng, A., & Hooi, B. (2021, May). Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 5, pp. 4027-4035).

[63] Kravchik, M., & Shabtai, A. (2021). Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. *IEEE transactions on dependable and secure computing*, *19*(4), 2179-2197.

[64] Kim, B., Alawami, M. A., Kim, E., Oh, S., Park, J., & Kim, H. (2023). A comparative study of time series anomaly detection models for industrial control systems. *Sensors*, *23*(3), 1310.

[65] Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020, August). Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 3395-3404).

[66] Guan, S., He, Z., Ma, S., & Gao, M. (2023). Conditional normalizing flow for multivariate time series anomaly detection. *ISA transactions*, *143*, 231-243.

[67] Wu, W., Song, C., Zhao, J., & Xu, Z. (2023). Physics-informed gated recurrent graph attention unit network for anomaly detection in industrial cyber-physical systems. *Information Sciences*, *629*, 618-633.

[68] Guo, H., Zhou, Z., Zhao, D., & Gaaloul, W. (2024). EGNN: Energy-efficient anomaly detection for IoT multivariate time series data using graph neural network. *Future Generation Computer Systems*, *151*, 45-56.

[69] Lin, Q., Adepu, S., Verwer, S., & Mathur, A. (2018, May). TABOR: A graphical model-based approach for anomaly detection in industrial control systems. In *Proceedings of the 2018 on asia conference on computer and communications security* (pp. 525-536).

[70] Fährmann, D., Damer, N., Kirchbuchner, F., & Kuijper, A. (2022). Lightweight long short-term memory variational auto-encoder for multivariate time series anomaly detection in industrial control systems. *Sensors*, *22*(8), 2886.

[71] Tang, C., Xu, L., Yang, B., Tang, Y., & Zhao, D. (2023). GRU-based interpretable multivariate time series anomaly detection in industrial control system. *Computers & Security*, *127*, 103094.

[72] Sahin, A. K., Cavdar, B., Dogan, R. O., Ayas, S., Ozgenc, B., & Ayas, M. S. (2023, October). A Hybrid CNN-LSTM Framework for Unsupervised Anomaly Detection in Water Distribution Plant. In *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

[73] Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1*(1), 73-79.

[74] Yaro, A. S., Maly, F., & Prazak, P. (2023). Outlier Detection in Time-Series Receive Signal Strength Observation Using Z-Score Method with S n Scale Estimator for Indoor Localization. *Applied Sciences*, *13*(6), 3900.

[75] Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68-78.

[76] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

[77] Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018, July). Synthesizing robust adversarial examples. In *International conference on machine learning* (pp. 284-293). PMLR.

[78] Li, J., Liu, Y., Chen, T., Xiao, Z., Li, Z., & Wang, J. (2020). Adversarial attacks and defenses on cyber–physical systems: A survey. *IEEE Internet of Things Journal*, *7*(6), 5103-5115.

[79] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

ÖZGEÇMİŞ

Enis KARA, İlköğrenimini Mevlüt Selami Yardım İlköğretim Okulu'nda bitirdikten sonra lise eğitimini 2015 yılında 17 Şubat Anadolu Lisesi'nde tamamladı. 2016-2020 öğretim yılında Karadeniz Teknik Üniversitesi Mühendislik Fakültesi Elektrik-Elektronik Mühendisliği bölümünde lisans eğitimine başladı. 2020 yılında bu bölümden mezun oldu. Aynı yıl Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim dalında tezli yüksek lisans programına başladı. 2024 yılından itibaren Sakarya Üniversitesi Elektrik-Elektronik Mühendisliği Bölümü'nde araştırma görevlisi olarak görev yapmaktadır. Orta derece İngilizce bilmektedir.