KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

DOKTORA TEZİ

LTE-A KABLOSUZ AĞLARDA KÖTÜ NİYETLİ RÖLE ATAKLARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TESPİTİ

YELİZ YENGİ

KOCAELİ 2020

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

LTE-A KABLOSUZ AĞLARDA KÖTÜ NİYETLİ RÖLE ATAKLARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TESPİTİ

YELİZ YENGİ

Prof.Dr. Adnan KAVAK		
Danışman,	**Kocaeli Üniv.**	**.....**
Prof.Dr. Celal ÇEKEN		
Jüri Üyesi,	**Sakarya Üniv.**	**.....**
Doç.Dr. Kerem KÜÇÜK		
Jüri Üyesi,	**Kocaeli Üniv.**	**.....**
Doç.Dr. Sultan ALDIRMAZ ÇOLAK		
Jüri Üyesi,	**Kocaeli Üniv.**	**.....**
Doç.Dr. Ali ÇALHAN		
Jüri Üyesi,	**Düzce Üniv.**	**.....**

Tezin Savunulduğu Tarih: 04.09.2020

ÖNSÖZ VE TEŞEKKÜR

Doktora çalışmalarım boyunca profesyonel yaklaşımı ve pratik bilgisi ile doktora çalışmamı mümkün olan en iyi şekilde tamamlamama yardımcı olan; desteğini, rehberliğini hiçbir zaman benden esirgemeyen, yapıcı motivasyonu ve titiz denetimi vesilesiyle emeğine minnettar olduğum danışmanım Prof. Dr. Adnan Kavak'a içten şükranlarımı sunuyorum. Değerli destekleri, yol göstericiliği ve öngörülü yaklaşımları ile çalışmama olan katkılarından dolayı Prof. Dr. Hüseyin Arslan'a çok teşekkür ediyorum. Bu meşakkatli yolu eş danışmanım olan Doç. Dr. Kerem Küçük'ün rehberliğinde yürüdüğümden dolayı kendimi gerçekten şanslı ve ayrıcalıklı hissediyorum. Değerli yorumları ve önerileri ile tez çalışmama ciddi bilimsel katkılar sunan jüri üyelerinden, Prof. Dr. Celal Çeken'e, Doç. Dr. Sultan Aldırmaz Çolak'a ve Doç. Dr. Ali Çalhan'a çok teşekkür ediyorum. Bu bilimsel çalışmanın Türkçe'sinin de en az içeriği kadar derli toplu olması noktasında en kritik dokunuşları yapan değerli edebiyat öğretmenimiz Feyzullah Divli'ye değerli katkılarından dolayı teşekkür bir borç bilirim. Son olarak, hayat arkadaşım, akıl danışmanım Revnak Yengi'ye sonsuz sevgi ve desteği için çok teşekkür ediyorum.

Ağustos – 2020 Yeliz YENGİ

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

σ^2 : Rölede alınan sinyal gücü
c : Sınıf sayısı
f : Sinyalin özeliği
h : Karmaşık kanal katsayısı
L : Özellik sayısı
M : Veri sayısı
N : Normal dağılım
n : Toplamsal beyaz Gaussian (AWGN) gürültü sinyali
O : Karmaşıklık fonksiyonu
w : Nöron ağırlığı
x : Sembol serisi
Z : Veri matrisi
Λ : Diagonal matris
ξi : Gevşek (slack) parametre
φ : Çekirdek haritalama yöntemi
ψ : Anomali skoru
K : Çekirde haritalama işlemi
$\mathcal{P}$: Permutasyon
α : Lagrange çarpanı
λ : Köşegen faktör
ρ : Rölenin yükseltme gücü

Kısaltmalar

3G : 3. Generation (3. Nesil)
4G : 4. Genetarion (4. Nesil)
5G : 5. Generation (5. Nesil)
AF : Amplify and Forward (Yükselt ve İlet)
AUC : Ander the Area of the Curve (Eğri Kalan Alan)
DF : Decode and Forward (Çöz ve İlet)
LOF : Local Outlier Factor (Yerel Aykırılık Faktörü)
LTE-A : Long Term Evalution Advanced (Gelişmiş Uzun Dönemli Gelişim)
NN : Neural Network (Yapay Sinir Ağları)
OCSVM : One-Class Support Vector Machines (Tek-Sınıflı Destek Vektör Makineleri)
PHY : Physical Layer (Fiziksel Katman)
RB : Resource Block (Kaynak Bloğu)
SNR : Signal to Noise Ratio (Sinyal Gürültü Oranı)
SVM : Support Vector Machines (Destek Vektör Makineleri)

LTE-A KABLOSUZ AĞLARDA KÖTÜ NİYETLİ RÖLE ATAKLARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TESPİTİ

ÖZET

Uzun Vadeli Evrim Gelişmiş (LTE-A) ağlarda yapılan birçok çalışmada, röleler kullanılarak kapsama alanı ve performans artırmaya odaklanılmıştır. Kablosuz ağların doğası gereği kötü niyetli davranışlara açık olması, röle tarafından iletişimin gecikmesine ya da performansının düşmesine neden olabilmektedir. Bu nedenle son zamanlarda, iletişim performansının artırılması ve veri gizliliğinin sağlanması için, fiziksel (PHY) katman güvenliği çalışmalarının yapılması önem kazanmaktadır. Mevcut çalışmalarda PHY katman çözümlerinin yetersiz olması, yapılanların ise denetimli makine öğrenmesi tekniklerine ve istatistiksel yaklaşımlara odaklanması nedeniyle gerekli olan yüksek performanslı kötü niyetli röle tespiti sağlanamamaktadır. Denetimli makine öğrenmesi tekniklerinin veri ve donanım gereksinimleri, istatistiksel yöntemlerin ise sınırlı değişimleri tespit edebilmesi problemin çözümünde farklı yaklaşımların değerlendirilmesini gerekli kılmıştır.

Bu nedenle tez çalışmasında, işbirlikçi LTE-A ağlarındaki, kötü niyetli röle ataklarını fiziksel katmanda tespit etmek için hedef düğümde denetimsiz makine öğrenmesi yaklaşımlarının kullanılması önerilmektedir. Tek sınıflı destek vektör makinesi (OCSVM), yerel aykırı faktör (LOF) ve yalıtım ormanı (iForest), kötü amaçlı röle tespiti için uygulanmaktadır. Makine öğrenmesi algoritmalarına girdi teşkil eden özellik vektörleri, modüle edilmiş temel bant sembollerinin genlik, faz ve bağıl faz bilgileri kullanılarak oluşturulmaktadır. Makine öğrenmesi algoritmalarının performansı, kesinlik, doğruluk ve eğri altında kalan alan (AUC) ölçümleri ile değişen sinyal gürültü oranı (SNR) seviyeleri, farklı modülasyon türleri, tahsis edilen kaynak bloğu (RB) sayısı ve değişen veri boyutu eksenlerinde analiz edilmiştir. Ayrıca, önerilen denetimsiz öğrenme algoritmalarının başarımı, literatürde mevcut diğer denetimli öğrenme algoritmaları ve geleneksel istatistiksel yöntemler ile de karşılaştırılmıştır. Sonuçlar, LTE-A ağlarındaki kötü amaçlı röleleri bilhassa fiziksel katmanda tespit etmek için önerilen yaklaşımımızın etkinliğini göstermektedir.

Anahtar Kelimeler: Atak Tespiti, Denetimsiz Öğrenme, Fiziksel Katman Atak Tespiti, LTE-A Röle Güvenliği, Tek Sınıflı Öğrenme.

DETECTION OF MALICIOUS RELAY ATTACKS IN LTE-A WIRELESS NETWORKS BY MACHINE LEARNING METHODS

ABSTRACT

There are many studies in existence that focus on improving the performance of relays for Long Term Evolution Advanced (LTE-A) networks and focus on improving the performance of relays and security issues are often neglected. Due to the broadcast nature of wireless channels, relay nodes in LTE-A network may act maliciously, affect communication, reduce quality and cause delays. Recently physical (PHY) layer security has attracted researchers to provide secure communication and data privacy. The current studies are insufficient with regards to PHY malicious relay detection by focusing on supervised machine learning and statistical learning approach. Malicious relay detection requires a high level of data and hardware configurations to achieve success by using supervised learning and a statistical approach also has limitation to detect any maliciousness in relay behavior.

Therefore in this thesis we propose using an unsupervised machine learning approach at the destination node to detect malicious relay attacks in cooperative LTE-A networks based on received source signal in PHY layer. Unsupervised outlier detection algorithms are applied to detect various malicious relay behaviors. As input to these algorithm feature vectors are constructed by using amplitude, phase and relative phase information of modulated baseband symbols. The performance of the outlier detectors are evaluated with respect to precision, accuracy and under the area curve (AUC) measures for changing signal-to-noise ratio (SNR) levels, different modulation types, allocated number of resource blocks (RBs) and varying data size. The results demonstrate the effectiveness of our proposed outlier detection approach when compared to existing studies which employ supervised and conventional learning for detecting malicious relays in LTE-A networks. The results verify the contribution of this study which is the demonstration of the effectiveness of one class outlier detection approaches for detecting malicious relays in LTE-A network.

Keywords: Attack Detection, Unsupervised Learning, Physical Layer Attack Detection, LTE-A Relay Security, One Class Learning.

GİRİŞ

Radyo erişim teknikleri ve sistemleri son yıllarda büyük bir teknolojik gelişme göstermiştir. Tüketicilerin kitlesel olarak 3G'yi kullandığı bir zamanda, yüksek kaliteli telekomünikasyon hizmetlerine yönelik talebin artmasıyla, 4G hücresel ağların küresel olarak dağıtılması için Uluslararası Telekomünikasyon Birliği (ITU) tarafından belirlenen performans gereksinimlerini telekomünikasyon kurumlarınca geliştirmeye yöneltmiş ve böylece pazarın talebini karşılayan Uzun Vadeli Evrim (LTE) çözümü tanımlanmıştır. 3. Ortaklık Projesi (3GPP) tarafından kurulan LTE iki aşamada tasarlanmıştır. İlk olarak LTE standardını yani sürüm 8'i tamamlamak, daha sonra LTE'yi 4G ve uluslararası mobil telekomünikasyon gelişmiş (IMT Advanced)'in kriterleri doğrultusunda geliştirmek için LTE-advanced yani sürüm 9 ve 10'u tamamlamaktır [1].

Sürüm 10, LTE'ye işlevsellik ve performans geliştirmeleri ekleyerek daha iyi kullanıcı deneyimi sağlarken, kullanılan LTE-Advanced teknolojilerinden bazıları, taşıyıcı birleştirme (Carrier aggregation), çoklu anten geliştirmeleri ve röle entegrasyonudur. Taşıyıcı birleştirme bant genişliğini, LTE sürümü 8'den 5 kat daha fazla olan 100 MHz'e kadar arttırmaktadır[2]. Sürüm 10 ile tanımlanan röle düğümleri (RN), baz istasyonu ile kullanıcı ekipmanı arasındaki trafiğe, yüksek veri iletim hızı ile hücresel alanın içinde ve dışında kapsama alanını arttırarak çözüm getirmektedir [3]. Hücre kenarı performansı daha kritik hale geldikçe LTE-A röle teknolojisi, kolay kurulum, düşük maliyet ve yüksek performans ile seçkin bir çözüm olarak sunulmaktadır. Bununla birlikte, LTE-A ağında röle performansını artırmanın yolları üzerine birçok olağanüstü çalışma yapılmış olmasına karşın, güvenlik sorunları genellikle ihmal edilmiştir[4]. Bu nedenle güvenlik problemi günümüz çalışmalarına motivasyon kaynağı olmaktadır.

Kablosuz kanalların yapısı ve işbirlikçi sistemlerin savunmasızlığı nedeniyle, rölelerden alınan sinyaller çeşitlilik açısında yüksek kalitede olmasına rağmen, röleler kaynaktan alınan sinyalleri iletmek yerine, manipule edip ileterek kötü amaçlı

davranabilmektedirler [4]. Bu nedenle, işbirlikçi sistemlerde oluşan bu güvenlik açığı önemli bir araştırma konusu haline gelmektedir. Ağ güvenliği, bütün ağ katmanlarında geliştirilen farklı çalışmalarla desteklenmiştir ancak PHY katman güvenliği, diğer katmanlarda yapılacak bir denetimden daha fazla kanalın enerjisinin korunmasını sağlayan bir çözüm olarak sunulmaktadır. Dolayısıyla röle ile işbirliği yapmadan önce yapılacak olan bir güvenlik denetiminin sistem enerjisine ve servis kalitesine katkısının önemli olacağı düşünülmektedir.

PHY katman güvenliği, literatürde istatistiksel tabanlı geleneksel yaklaşımlarla çözülmeye çalışılmıştır ve belirli bir kapsam içerisindeki değişimlerin denetlenmesi sağlanmıştır[15-18]. Ancak PHY katmanındaki olası etkenlerin kapsamının geniş olması bu çözümleri, kapsam dışı bütün değişimler için etkisiz kılmaktadır. Makine öğrenmesi algoritmaları kullanılarak getirilmek istenen çözümlerde ise denetimli öğrenme yaklaşımı ağırlıklı olarak kullanılmış ve tanımlı atak kategorileri kapsamında sınıflandırma problemi olarak ele alınmıştır [1-7]. Bu çözümler sistemdeki belirli değişimleri tespit etmekte ve sınıflandırmakta olsa da öntanımlı modellerin dışında bir çözüm sunmamaktadır.

Bu tez çalışmasında, mevcut çalışmaların ötesine geçebilmek; daha kapsamlı, kullanışlı ve güçlü bir çözüm getirebilmek adına PHY katmanda denetimsiz makine öğrenmesi teknikleri ile atak tespit sisteminin geliştirilmesi ve röle ile işbirliği yapmadan önce atakların tespit edilebilmesi adına bir çözüm sunulmaktadır. Bu çözüm, makine öğrenmesi tekniklerini geniş bir kapsamda farklı parametreler ile inceleyerek, tekniklerin avantajlı ve dezavantajlı yönlerini, kullanışlılık ve uygulanabilirlik değerlendirmeleri ile sistemin faydalarını geniş bir perspektiften analiz etme imkanı sunmaktadır.

1. İLGİLİ ÇALIŞMALAR VE TEZİN KATKISI

Makine öğrenmesi tekniklerinin özellikle kablosuz ağ güvenliği alanında kullanımı eski bir yöntem olamamakla birlikte ilgili çalışmalar başlığı altında tartışılmaktadır. Bununla birlikte, LTE-A röle ağlar birçok kullanım senaryosuna sahip geniş bir alan olduğundan, önceki çalışmaların çoğu bu tezdeki modelden farklı kurulum senaryolarına odaklanmakta ve değişen başarı dereceleri ile tutarsız sonuçlar sunmaktadır. Ayrıca gerçek röle verilerinin elde edilmesi ve erişilmesinin zorluğundan dolayı, bu çalışmada ve literatürdeki çoğu çalışmada sadece simule edilen sistem modeli ile oluşturulan veriler kullanılmaktadır.

1.1. İlgili Çalışmalar

Önerilen çözümü sunmadan önce, bu tezde kullanılan tüm önemli bilimsel araştırma alanlarına genel bir bakış sunması amacıyla ilgili çalışmalar makine öğrenmesi teknikleri ve istatistiksel öğrenme tekniği kapsamında detaylandırılmaktadır.

1.1.1. Denetimli öğrenme çalışmaları

Denetimli öğrenme teknikleri, eğitim verilerinin hem güvenli hem de atak sınıflarına ait etiketli örnekleri içermesini ön koşul olarak sunmaktadır. Bu durumda, atak algılama modeli herhangi bir makine öğrenme sınıflandırıcısı olabilir ve problem sınıflandırma problemleri olarak düşünülebilmektedir. Bununla birlikte, normal / anormal veya güvenli / atak sınıfları arasındaki doğal olarak oluşan dengesiz oran ve veri kümesinin çoğunluğunun normal verilerden oluşması, sınıflandırıcı için bir zorluk teşkil eder ve belirli bir doğruluk performansı elde etmek için ek ön işleme adımı gerektirmektedir.

Etiketli verilerin, özellikle etiketli atak durumlarının mevcut olması nedeniyle, denetimli atak tespiti aynı zamanda atak türlerini kategorilendirebilmektedir. Verilerin derlemesinde, mümkün olduğunca uzun süre kanalın gözlemlenmesi önerilmektedir. Az miktarda derlenen güvenli verilerin bile atak olarak algılanma oranını önemli ölçüde artırabileceği gösterilmiştir [14].

J. Xing ve diğerlerinin yaptıkları çalışmalar, kanalın gizli dinlenmesini tespit edebilmek için yasal kullanıcılar ile yasal olmayanları kanal durum bilgisi (CSI) istatistiği tutularak denetimli yapay sinir ağları (NN) algoritması ile bu iki durumun tespit edilebileceğini göstermiştir [1]. Z. Deng vd. benzer şekilde NN algoritması kullanarak güvenilir röle seçimini CSI ile çok sınıflı bir model ile sağlamış, aynı zamanda geleneksel röle seçim yöntemleri ile kıyaslayarak başarımını göstermiştir [2]. X. Wang vd. karar ağaçları ile üç aşamalı denetimli bir öğrenme modeli tasarlayarak röle seçiminin daha güvenli bir şekilde gerçekleştirilebileceğinin simülasyonunu yapmışlardır. Çalışmada kanalın genel CSI bilgisi, eğitim verisi olarak kullanılmıştır [3]. A. Carreño vd. yaptıkları; aykırı değer tespiti, anomali tespiti, atak tespiti, vb. gibi kanal üzerindeki durumların tespit edilmesi için denetimli öğrenme perspektifinden gerçekleştirilen çalışmalarda, denetimli öğrenme algoritmalarının yaygın kullanım şekilleri ile birlikte başarımlarını göstermiştir [4].

S. Riyaz vd. çalışmalarında derin evrişimsel sinir ağları (CNN) algoritması ile kanaldan elde ettikleri I/Q sembollerini kullanarak oluşturdukları model ile yazılım tanımlı radyo (SDR) sistemlerinde atak tespiti yapmayı hedeflemişlerdir. Oluşturdukları radyo dalgaları parmak izi modelini SVM, lojistik regresyon (LR) gibi diğer denetimli öğrenme modelleri ile de kıyaslayarak CNN'nin başarımını göstermişlerdir.

Ayrıca diğer parmak izi üzerine yapılan çalışmalarla karşılaştırarak katkısını tartışmışlardır [5]. M. Kulin vd. uçtan uca radyo dalgalarının spektrum verileri ile CNN algoritması modeli oluşturarak sınıflandırma yapmayı hedeflemişlerdir. Spektrum verileri olarak sinyalin faz ve genlik değerlerini I/Q sembollerinden elde etmişlerdir. SNR seviyesinin radyo dalgalarını sınıflandırmadaki etkisini analiz ederek %10 luk bir etkiye sahip oluduğunu göstermişlerdir ve çalışmanın modülasyon, girişim (interference) algılama noktasında katkı sunabileceğini tartışmışlardır [6].

L. Xiao vd. kimlik sahtekarlığını önlemek için PHY katmanında radyo kanalı bilgilerini (alınan sinyalin gücü vb.) kullanarak kimlik doğrulaması yapmayı hedeflemişlerdir. Oyun teorisi yardımı ile dengeli ve dinamik öğrenme tasarlayarak, istatistiksel durumlarla tespit edilemeyen değişken durumları da kapsayan Q-learning ve Dyna-Q tabanlı kimlik doğrulama sistemini araştırmışlardır [7].

Denetimli öğrenme etkili bir yöntem olmakla birlikte, dengeli veri seti gereksinimi ve sadece eğitim verisi içerisindeki sınıflar dahilinde kategorilendirilmesi nedeniyle kapsamlı bir atak tespit sistemi, maliyetli bir yaklaşım olarak görülmektedir.

1.1.2. Yarı denetimli öğrenme çalışmaları

Yarı denetimli öğrenme algoritmalarının son on yılda popülaritesi artmakla birlikte, denetimsiz ve denetimli öğrenme algoritmaları arasında bir yaklaşıma sahiptir. Öğrenme algoritması için giriş verileri, etiketlenmiş ve etiketlenmemiş verilerin bir karışımıdır [9]. Yarı denetimli öğrenme algoritmaları, eğitim verilerinin etiketlemesi için zaman ve çabadan tasarruf etme yeteneğine sahiptir. Ek olarak, yarı-denetimli verilerin etiketlenmesi sırasında ortaya çıkabilecek insan hatasını ortadan kaldırmaktadır. Etiketlenmemiş verilerin, az miktarda etiketlenmiş verilerle birlikte kullanıldığında, öğrenme doğruluğunda kayda değer bir gelişme sağlayabileceği bulunmuştur.

Atak tespiti, E.R. Faria vd. yaptıkları çalışmalarda, kümeleme ve yarı-denetimli algoritmalar kullanılarak, örneklerin önceden tanımlanmış veri ve model üzerinde kategorize edilmesini amaçlamışlardır [8]. M. Ozay vd. yaptıkları çalışmalarda yarı denetimli öğrenme tekniği ile denetimli öğrenme tekniklerini uygulayarak, etkililiklerini kıyaslamışlar ve PHY problemlerine kısmi bir çözüm sağlanabildiğini göstermişlerdir [13].

1.1.3. Denetimsiz öğrenme çalışmaları

Denetimsiz öğrenme, denetimli öğrenmenin tam tersidir. Sistem bir öğretmen olmadan öğrenmeye çalışır, girdi verileri etiketlenmez ve dolayısıyla sınıflandırılmaz. Bu nedenle, sistem esas olarak giriş verilerindeki belirli desenlerden ibarettir. Denetimsiz öğrenme yaklaşımının bir örneği, Öklid, Jaccard ve Kosinüs mesafe metrikleri gibi uygun bir mesafe metriği ile tanımlanan benzer özelliklere dayalı olarak giriş verilerindeki yararlı kümeleri tespit etmek için kullanılan kümelenmelerdir [10].

Denetimsiz öğrenme kategorisinde yer alan aykırı örneklerin algılanması, tüm veri odaklı bilimsel disiplinlerde en önemli yaklaşımlardan biridir. Beklenen davranışla eşleşmeyen veriler genellikle bazı ilginç özelliklere sahiptir ve mevcut sorunu daha iyi

anlatmaya yardımcı olabilmektedir [11]. Denetimsiz öğrenme algoritmalarının literatürde PHY katman atak tespiti probleminde değerlendirilmemiş olmaları, bu tez çalışmasındaki önemli motivasyon kaynaklarından biridir.

1.1.4. İstatistiksel öğrenme çalışmaları

Y. Mao vd. röle düğümlerinden bozuk sinyaller göndererek iletişimin bozulmaya çalışıldığı, kod çözme ve iletme (DF) stratejisini kullanarak çoklu röle düğümleri bağlamında işbirliği iletişimindeki güvenlik sorunlarını araştırmışlar, çalışmada geleneksel PHY katman kötü niyetli röle tespitinin bu gibi durumlarda hatalı olduğu sonucunu tartışmışlardır [15]. L. Lo vd. çalışmalarında CSI bilgisi olmadan, kanalın enerji seviyesi ve sembollerin faz rotasyonlarının olasılık dağılımlarını inceleyerek işbirlikçi sistemlerde kötü niyetli röle tespitini araştırmışlardır [16].

W. Hou vd. yükselt ve ilet (AF) protokole sahip röle ağlar için, rölenin yükseltme kazancı ve gürültü enerjisi üzerinden ikili bir hipotez ile parazit ataklarına (jamming) karşı olasılık dağılımları üzerinden tespit sistemi geliştirmişlerdir. Olasılık tabanlı yaklaşımlarda genel olarak görülen bir eşik veya tespit aralığı mevcuttur. Bu aralık, dağılım dışında kalan sinyalleri ayırt etmede etkilidir [17].

T. Lv vd. kaynağın güvenilir olmayan CSI bilgisine tabi olan bir AF röle düğümü yardımıyla kaynağın bilgiyi hedefe ilettiği bir Gauss geçiş sisteminde kötü niyetli davranışın tespiti üzerinde çalışmışlardır. Özellikle, doğrudan bağlantıya sahip bir sistem için, tespit edilebilen ve tespit edilemeyen sınıfları açıkça ayırt edebilen bir tespit yaklaşımı önermişlerdir. Ayrıca saptanabilir sınıf için önerilen yaklaşımın yüksek olasılıklı kötü niyetli saldırıları tespit ettiği de bu çalışmalarda gösterilmiştir [18].

1.1.5. Takviye öğrenme çalışmaları

Herhangi bir eğitim verisi gerektirmeyen başka bir makine öğrenmesi tekniğidir ve makine deneme yanılma yöntemine göre ideal davranışı geliştirir. Markov Karar Süreçleri (MDP'ler), Qlearning, Dyna-Q ve Post-Decision State, IoT kimlik doğrulaması, kötü amaçlı yazılım tespiti ve anti-parazit iletimleri için takviye öğrenmede kullanılan algoritmadır [42,43].

1.1.6. Derin öğrenme çalışmaları

Derin öğrenme tekniğinde, birden fazla özellik seviyesi vardır ve bunlar otomatik olarak keşfedilir. Özelliklerin her seviyesi bir önceki seviyenin özelliklerinden keşfedilir. Araştırmalar derin öğrenmenin SDN bağlamında anomali tespiti için yüksek bir potansiyele sahip olduğunu göstermektedir [44]. Derin öğrenme yakşımı yüksek donanım performansına ihtiyaç duyan bir yöntemdir bu nedenle küçük donanımlarda geliştirilecek çözümler için önerilmemektedir.

1.2. Çalışmanın Katkısı

Daha önce bahsedildiği gibi bu tezin amacı LTE-A ağlarda röle ataklarının PHY katmanda denetimsiz makine öğrenmesi teknikleri ile tespit edilmesidir. Birincil araştırılacak olan: "Röle ağlarda anormallikler veya kötü niyetli davranışlar PHY katmanda denetimsiz makine öğrenmesi teknikleri ile tespit edilebilir mi?" sorusuna bu tez çalışmasında cevap aramaktır.

Bu problemin çözümü ile birlikte tez aşağıdaki araştırma sorularını cevaplamayı amaçlamaktadır:

1. Hangi modeller en iyi tahmin performansını vermektedir?
2. Bu modeller kullanılarak ne tür anormallikler daha iyi tespit edilebilmektedir?
3. Mevcut çalışmalara olan katkısı nedir?
4. Modelin güvenirliği ve dayanıklılığı farklı fiziksel koşullarda geçerli midir?
5. Sadece güvenilir tek bir kaynaktan alınan sinyalin öğrenilmesi ile atak tespiti başarılı bir şekilde yapılabilmekte midir?

Bu tez çalışmasının en büyük katkısı, LTE-A ağlarındaki kötü niyetli röle düğümlerinin PHY katmanında sadece güvenilir bir kaynağın sinyallerinin denetimsiz algoritmalar ile öğrenilerek, normal olamayan (atak) sinyallerin tespit edilebileceğinin gösterilmesidir. Ayrıca, geleneksel istatistiksel yaklaşımlar ile tespit edilmesi oldukça zor olan fiziksel katmandaki kötü amaçlı veya güvenilir rölelerin, temel bant sinyal özelliklerinin kullanılarak ML tabanlı yaklaşımlar ile ayırt edilebileceğini göstermesidir. Çalışmanın etkililiğini ve güvenirliliğini gösterebilmek için modülasyon tipi, atak tipi, kanal gürültü seviyesi, veri kümesi boyutu ve sinyal bant genişliği gibi farklı etkenlerin atak algılama performansı üzerindeki etkileri araştırılmaktadır.

1.3. Tezin Yapısı

Tez, araştırma sorusu ve motivasyonu ile başlamakta ve 1. Bölüm ilgili konuların literatür taramasının bir özetini ve önerilen çözümün temel bir çerçevesini sunmaktadır. 2. Bölüm, LTE-A röle ağların güvenlik ihtiyaçları, sistem modeli ve bu sistemde oluşabilecek olası atak modellerini açıklamaktadır. 3. Bölüm makine öğrenmesi algoritmalarının detaylandırılmasını ve 4. Bölüm makine öğrenmesi algoritmaları için eğitim verilerinin tanımlanmasını detaylandırmaktadır. 5. Bölüm sistem deneyini ve başarım parametrelerini anlatırken, 6. ve 7. Bölümler teori ve pratik deneylerden elde edilen bilgilerin değerlendirilmesini ve bu tezin sınırlamalarını, gelecekteki çalışmaları ve katkılarını içermektedir.

2. LTE AĞLARA GENEL BAKIŞ

3. Nesil Ortaklık Projesi (3GPP), 1998'de farklı telekomünikasyon dernekleri arasında bir işbirliği oluşturmak amacıyla kurulmuştur. 3GPP, evrimleşmiş paket çekirdeğini (EPC), sistem mimarisi evrimi (SAE) ve uzun vadeli evrim (LTE) olmak üzere iki ana öğe içeren yapı olarak belirlemiş ve evrensel karasal radyo erişim ağı (E-UTRAN) ile karasal radyo erişimine katkı sunmasını amaçlamıştır.

LTE, spektrum esnekliği için kapsamlı destek sağlarken, LTE FDD ve TD-LTE olarak da adlandırılan hem frekans bölmeli çoklama (FDD) hem de zaman bölmeli çoklama (TDD) modunu desteklemektedir. Sürüm 8'de tanımlanan özelliklere göre LTE, sırasıyla 300 Mbit/s ve 75 Mbit/s'ye kadar aşağı bağlantı (downlink) ve yukarı bağlantı (uplink) üst hızları ile 5 ms'den daha az tek yönlü bir radyo ağı gecikmesi ile spektrum verimliliğinde önemli bir artış sağlayabilmektedir. LTE'nin anahtar teknolojilerine örnek olarak:

- LTE FDD ve TD-LTE Ölçeklenebilir bant genişliği için 20 MHz'e kadar Dikey Frekans Bölmeli Çoklu Erişim (OFDMA) tabanlı çoklu erişim şemaları;
- Çoklu Giriş Çoklu Çıkış (MIMO) anten teknolojisi desteği;
- Yeni veri ve kontrol kanalları;
- Yeni ağ ve protokol mimarisi (iki düğüm, IP tabanlı).

Temmuz 2008'de ITU tarafından, IMT-Advanced olarak bilinen dördüncü nesil kablosuz ağların resmi özellikleri yayımlanmıştır. Gerçek bir 4G teknolojisi olabilmek için LTE, ITU tarafından verilen IMT-Advanced gereksinimlerini karşılayacak şekilde geliştirilmiştir. Gerekli iyileştirmeler 3GPP Sürüm 10'da belirtilmiş ve LTE-Advanced olarak da adlandırılmıştır. LTE-A, veri hızını daha da artırmak için LTE OFDM / MIMO mimarisini temel alır ve LTE-Advanced, downlink için üst veri hızını 1 Gbit/s'ye ve uplink için 500 Mbit / s'ye çıkarmaktadır ve aşağıdaki teknoloji bileşenleri ile desteklenmektedir:

- Taşıyıcı birleştirme;

- MIMO genişletme (Downlink: 8x8'e kadar; Uplink: 4x4'e kadar);
- Yer-uydu bağı erişim geliştirmeleri (kümelenmiş SC-FDMA ve eşzamanlı veri ve kontrol bilgileri (PUSCH ve PUCCH) iletimi;
- Hücre kenarı performansını artırma (gelişmiş hücreler arası girişim koordinasyonu (eICIC)).

3GPP Sürüm 10'da 4G LTE-Advanced standardının kullanılmasından sonra, LTE-Advanced birkaç sürümle gelişmeye devam etrmiş ve sürüm 11 başlatılmıştır, Sürüm 10 için geliştirilen temel LTE-Advanced teknolojilerine daha fazla iyileştirmeler dahil edilmiştir. Sürüm 11 sırasındaki en büyük katkılar; işbirlikçi çok noktalı iletim ve alım (CoMP), geliştirilmiş hücre içi girişim iptali (eICIC) ve mobilite yönetimi geliştirmeleridir. Sürüm 11'in 2013'ün başlarında tamamlanmasından sonra, 3GPP Sürüm-12'de standartlaştırma çalışması başlamıştır. Sürüm 12'nin birincil hedefi; mobil operatörlere kapasiteyi artırmak, pil ömrünü uzatmak, ağ düzeyinde enerji tüketimini azaltmak, maliyet verimliliğini en üst düzeye çıkarmak, çeşitli uygulamaları ve trafik türlerini desteklemek, ana taşıyıcıyı geliştirmek ve müşterilere daha zengin bir ürün sunmak için yeni seçenekler ile daha hızlı ve daha güvenilir bir ortam getirmektir.

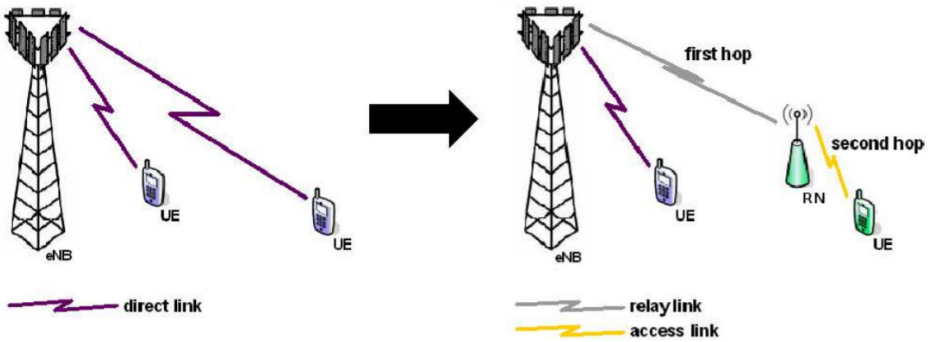
Araştırma topluluğu artık 4G'nin ötesine, gelecekteki 5G teknolojilerine doğru çalışmaktadır. ITU, son zamanlarda veri trafiğinin patlayıcı şekilde büyümesi, çok sayıda makine tipi iletişim (MTC) cihazı desteği ve güvenilir 5G sistemleri için "Vizyon" (ITU-RM.2083, 2015) konusundaki çalışmalarını tamamlamıştır. LTE evrimi ve yeni radyo erişim teknolojisinden oluşan 5G kablosuz erişim çözümüdür. LTE evrimi, 6 GHz'e kadar mevcut spektrumda geriye dönük uyumlu geliştirmelere odaklanırken 5G, LTE'nin dağıtılmadığı yeni spektruma odaklanacaktır. 5G yolunda, LTE Advanced Pro standardı olarak bilinen Sürüm 13 ve Sürüm 14 hâlihazırda Gigabit LTE sağlamak için taşıyıcı toplama, büyük MIMO, lisanslı ve lisanssız spektrum tekniklerinin eşzamanlı kullanımını gerçekleştirmektedir. Sürüm 13'ün bir diğer önemli odağı, genişletilmiş kapsama alanını da destekleyebilen MTC cihazları için maliyetin azaltılmasıdır. Yayımlanan Sürüm 14'ün odağı, mevcut massive MIMO çerçevesinin fazla sayıda antene uzatılmasını, iletim zaman aralığını 0.5ms veya daha azına düşürmeyi ve büyük MTC'yi desteklemeyi içermektedir.

2.1. LTE-A Ağlarda Röle

Tarihsel olarak, çok sekmeli (multi hop) ağ, kablolu ve kablosuz ağlarda geniş mesafe iletişimiyle ilişkilendirilmiş olup teknik olarak sinyal tekrarlayıcılardır. Geleneksel bir hücresel ağda, kaynak ve hedef (macro base station (eNB) ve user equipment (UE) veya tam tersi) doğrudan birbirleriyle iletişim kurmaktadır. Böyle bir senaryoda, eNB'den uzakta bulunan bir UE, iletimindeki yüksek yol kaybından dolayı düşük sinyal gücü ile karşılaşır. Bu nedenle, bir hücrenin büyük bir kısmı daha kötü radyo sinyal seviyesi koşullarındadır. Bu etkiye karşı koymak için iletim gücü arttırılabilir; bu da bir UE'nin iletim gücü (maksimum Etkili İzotropik olarak Yayılan Gücü (EIRP)) ile sınırlı olduğundan, güç kısıtlı iletimlere yol açabilmektedir. Ayrıca, kablosuz ağlar spektrumu yeniden kullandığından iletim gücündeki bir artış, spektrumun yeniden kullanıldığı bölgeler arasındaki paraziti artırabilmektedir. Bu ilave zorluklar, yeni nesil hücresel şebekelerde öngörülen hedef iletim oranlarına ulaşmada daha fazla komplikasyon yaratmaktadır. Daha önce de belirtildiği gibi, özellikle hücre kenarı UE'lerle ilgili olarak ağ performansı iyileştirmek için olası bir çözüm, sabit RN'leri mevcut makro eNB ağlarına entegre etmektir. RN, eNB ve UE arasında bir ara erişim noktası olarak düşünülmektedir Şekil 2.1. Mevcut hücresel ağlarla ve RN'lerle konuşlandırılan pico/femto düğümleri arasındaki fark, RN'nin onu tamamen kablosuz bir erişim noktası haline getiren bir kablosuz ana taşıyıcıya sahip olmasıdır. Bu nedenle iyi incelenmiş AF tipi RN'ler yerine DF tipi RN'ler gelecekteki ağlarda araştırma konusu olabilecektir [38]. Bir DF rölesi sadece alınan sinyalin, bilgi içeren kısmını yeniden üretip ilettiği için bir AF rölesinden farklıdır.

Bir DF rölesi, değişken kazançlı bir dijital tekrarlayıcı gibi çalışır. Alınan sinyalin yeniden iletimine izin verir ve farklı atlamalarda değişken Modülasyon ve Kodlama Şemalarına (MCS) uyarlanabilir ve iletim tekniklerinden yararlanabilir. Ayrıca, RN'ler girişimden (interference) kaçınma/hafifletme şemaları da uygulayabilir. RN'ler mobil (mobil geçici ağlarda olduğu gibi) veya sabit olabilmektedir.

Sabit RN'ler LTE-A'nın dağıtım hedeflerine daha iyi uyum sağladıkları için bu çalışmaya ilgi duyulmaktadır. Bu nedenle, sabit RN'ler kapsama alanı veya kapasite kısıtlamaları nedeniyle tek başına eNB'nin performans hedeflerini karşılayamayacağı yerlere yerleştirilebilmektedir [37,39].



Şekil 2.1. Hücresel şebeke topolojisi tekli ve çoklu atlama ile geçiş [38].

Dolayısıyla potansiyel RN uygulama senaryoları, kapsama deliklerine öyle ki hücresel alanlar gerçekte çarpışık dairesel kapsama alanlarıdır ve bu nedenle bazı ölü noktalara sahiptir. Sıcak noktalarda yoğun kullanıcı talebinin olduğu bölgelerdir ve son olarak hücre kenarlarında kapsama alanının sınır bölgelerinde uygulanabilmektedir. Bu çalışmada, hücre kenarı, sıcak noktalar ve kapsama delikleri senaryolarında performansın iyileştirilmesine odaklanılmıştır ve bu nedenle, RN'lerin bu tür konumlara çeşitliliği artırmak amaçlı yerleştirildiği varsayılmaktadır.

2.2. Röle Ağlarda Kurulum

Röle teknolojisinin kullanılmasının potansiyel olarak faydalı olduğu senaryolar 3GPP'de tartışılmıştır. Dağıtım senaryoları Tablo 2.1'de gösterilmektedir. Kapsama alanının, dağlık ve seyrek nüfuslu bölgelere (kırsal alan ve kablosuz ana taşıyıcı senaryoları) genişletilmesi, operatörler için önemli bir senaryodur. Röle teknolojisinin, sabit hatlı ana taşıyıcı bağlantılarının aksine, kapsamı ekonomik olarak genişletmek için kullanılması beklenmektedir. Röle teknolojisinin ayrıca, depremler veya diğer felaketler ya da büyük olaylar meydana geldiğinde (acil veya geçici kapsama senaryosu), yani tahsis edilmiş sabit hat ana taşıyıcı bağlantılarının konuşlandırılmasının zor olduğu durumlarda geçici kapsama sağlama konusunda etkili olması beklenmektedir. Buna ek olarak, piko ve femtoceller kentsel sıcak nokta, ölü nokta ve iç mekan sıcak nokta senaryoları için kullanılabilirken, elektrik direklerinin montajı, binaların içine kablo döşenmesi vb. gereksinimlerinden dolayı bazı ülkelerde ve bölgelerde kurulum zorlukları olabilmektedir. Ayrıca röle teknolojisinin kentsel senaryolar için de etkili olabileceği düşünülmüştür. Son olarak, hareketli mobil istasyonlardan gelen kontrol sinyallerinin hacmini azaltmak için tren ve otobüs gibi

araçlara röle istasyonlarının kurulduğu grup hareketlilik senaryosu da önerilmektedir. 3GPP'de, LTE Sürüm 10'da kapsama alanı genişletmesi için kullanılan röle teknolojisinin standartlaştırılmasına karar verilmiştir. Bu spesifikasyonlar, özellikle röle istasyonunun konumunun sabitlendiği tek duraklı röle teknolojisini ve arasındaki radyo erişim bağlantısını destekleyecektir [36].

Tablo 2.1. Röle teknolojisi uygulama senaryoları.

Senaryo	**Uygulama Alanı**	**Atlama (Hop) Sayısı**
Kentsel ve yoğun alanlar	Yüksek trafik yoğunluğuna sahip kentsel alanlarda kapsamı genişletmek ve verimliliği arttırmak için kullanılabilir	1 Atlama
Kapalı ve yoğun alanalar	İç mekan ortamlarında kapsama alanını genişletmek ve verimliliği arttırmak için kullanılabilir.	1 Atlama
Grup hareketliliği	Aktarma ve konum kayıt kontrol sinyallerini azaltmak için kamu araçlarına röle istasyonları kurulabilir.	1 Atlama
Kırsal alan	Dağlık veya seyrek nüfuslu bölgelerde kapsama alanını genişletmek için kullanılabilir.	1 Atlama
Ölü noktalar	Kapsama alanındaki boşlukları doldurmak için kullanılabilir.	1 veya daha çok atlama
Acil veya geçici kapsama alanı	Felaketler, olaylar vb. zamanlarda geçici teminat sağlamak amaçlı kullanılabilir.	1 veya daha çok atlama
Kablosuz ana taşıyıcı	Dağlık, seyrek nüfuslu bölgelerde veya uzak adalarda kapsama alanını genişletmek için kullanılabilir.	1 veya daha çok atlama

2.3. LTE-A Ağlarda Röle Seçimi ve Güvenlik

İşbirlikçi kablosuz haberleşme sistemlerinin önemli çalışma konularından biri, verimli bir geçiş protokolü tasarlamaktır. Bu tür çalışmalarda ana problem, geçiş işlevini gerçekleştirmek için oluşturulan model üzerindeki kısıtlamalara (örneğin güç, kanal, röle vb.) göre optimum tahsis şemalarını bulmaktır. Optimum röle seçimi işbirlikçi ağlarda büyük önem taşıyan sorunlardan biridir ve sistem performansını önemli ölçüde etkilemektedir. Bu alandaki çalışmalar ise PHY katman konularına odaklanarak bit hata oranı, kapasite ve kesinti davranışlarını incelemektedir [19-27]. [19]'da işbirliği kavramı için bir seçim yöntemi sunulmaktadır ve işbirlikçi röle seçiminin dağıtılmış Uzay Zaman Kodlaması (STC) programından daha iyi performans gösterdiği gösterilmiştir. [20]'de yazarlar, kısa vadeli güç kısıtı altında kesinti olasılığını en aza indiren bir röle seçim stratejisi önermişlerdir. Referans [21] birçok tek röle seçim

şemasının çeşitlilik sırasını inceler ve sonra tek röle seçimi fikrini çoklu röle seçimine genelleştirir. Çalışma [22] röle seçimini oyun teorisi bakış açısından inceler ve dağıtılmış bir alıcı/satıcı önerir. Oyun iki soruyu ele alır: Birincisi, hangi röle düğümleri işbirliğine dahil edilmelidir; ikincisi, optimum güç seviyesi nedir. [28]'de, anlık SNR'lere dayanan sınırlı ağ bilgisi gerektiren dağıtılmış bir röle seçim şeması önerilmektedir. [23]'de önerilen röle seçim politikası, sabit miktarda verinin toplam iletim süresini en aza indirmektedir. Fırsatçı röle seçim şeması [24,29- 31]'de incelenmiştir. [29]'da fırsatçı röle seçim şemasının, geleneksel şemadan çok daha iyi seçim davranışı sağladığı gösterilmiştir. [30] ve [35]'te, bir toplam güç kısıtı altında, fırsatçı röle seçim şemasının küresel olarak en uygun düzeyde olduğu gösterilmiştir. [24]'teki yazarlar, çoklu röle ağında, fırsatçı röle seçiminin geleneksel tüm katılım şemasını içeren diğer tüm stratejilerden daha iyi performans gösterdiğini göstermiştir. Ayrıca, her bir röle düğümünün kendi anlık kanal kazancına ve önceden tanımlanmış bir eşik değerine göre geçiş yapmaya veya sessiz kalmaya karar verdiği merkezi olmayan bir röle seçim algoritması [26]'da önerilmekte, bu öneride röle seçim politikası, kaynakta ve rölelerde mevcut kanal durumu bilgisi (CSI) kullanmaktadır. Seçim politikası, kaynak-röle ve röle-hedef kanal kazanımlarının maksimum anlık ölçekli harmonik ortalama fonksiyonuna sahip olan optimal bir röle seçimi gerçekleştirir. [27]'de çok kullanıcılı sistemde röle atama problemi ele alınmıştır. Önerilen görevlendirme politikası, kapsama alanını genişletmek için işbirlikçi çeşitlilik protokollerinin uygulanması için pratik bir uygulamadır. [32-35]'te kablosuz işbirlikçi ağlarının stokastik kontrolü tartışılmaktadır. [32]'de, tek kaynaklı ve tek hedefli işbirlikçi ağlar için verimi optimum ağ kontrol politikaları incelenmiştir. En uygun politika kuyruk dinamiklerini dikkate alır ve ağdaki yönlendirme, zamanlama ve kaynak tahsisini birlikte optimize eder.

[33]'teki çalışma, işbirlikçi ağlarda röle seçim problemini stokastik kontrol açısından ele almaktadır. Önerilen politika adaptif modülasyon ve kodlamanın yanı sıra ağ ömrünü en üst düzeye çıkarmak için röle seçim sürecinde kalan röle enerjisini değerlendirmektedir. [33]'teki yazarlar kablosuz işbirlikçi ağlarda röle seçim problemini, Markov karar zincirinde stokastik bir kontrol problemi olarak formüle etmektedir. [35]'teki çalışma, sınırlı işbirlikçi iletişim ağları için en uygun kaynak tahsisi politikasını araştırmakta ve hedef kesintisi olasılığına ulaşmak için dinamik

işbirliği stratejileri geliştirmektedir. [34] 'te, çok kullanıcılı işbirlikçi geçiş erişim ağlarında verim optimal röle seçimi sorunu ele alınmıştır.

İşbirliğine dayalı iletişim, önemli performans iyileştirmelerine yol açamaktadır; ancak kaynak, bir ara düğüme yani geçişe dayanmak zorunda olduğundan, güvenlik sorunları gündeme gelmektedir. İşbirlikçi iletişim tasarımı, röle düğümünün her zaman birbirlerine yardım etmeye ve seçildiklerinde etkili bir şekilde işbirliği yapmaya hazır oldukları varsayımına dayanır. Ancak, ağda bencil veya kötü niyetli niyetler için yanlış davranabilecek bir düğüm olduğunda bu varsayım geçerli olmayabilir. Bir geçiş, kaynak veya hedef olarak hizmet verirken hatalı çalışan bir düğüm, sistemin performansını bozabilir. Bu nedenle, işbirlikçi iletişimdeki güvenlik sorunları, ağdaki varlıklar arasında işbirliği gerektiren diğer senaryolardaki güvenlik endişelerine benzemektedir.

Örneğin, kötü amaçlı düğümlerin ağa katılması ve istenmeyen iletileri hedefe iletmesi ve dolayısıyla sistemin performansını düşürmesi mümkündür. Kaynak ve varış yeri arasındaki iletişimi kesintiye uğratmak için işbirliği stratejisini ihlal eden bir geçiş düğümü kendi gücünü ve kaynaklarını tüketme masrafı ile birlikte, kötü niyetli bir davranışa sahip olma karakteristiğindedir. Kötü amaçlı bir düğüm işbirliği için seçildiğinde ve iletiyi iletmesi istendiğinde, alınan paketleri iletmeyi reddedebilir, bu paketleri kuyruğunda tutabilir. Böylece iletim için bir kanal atama şansı artar ve birçok yönden, kötü niyetli bir düğüm, hatalı davranışlarla sistemin performansına etki edebilir. Ek olarak, bazı düğümler kendi enerjilerini korumak, işbirliği yapmamak, kaynağın bilgilerini diğer röle düğümleri aracılığıyla iletmesini sağlamak ve böylece işbirliğini caydırmak için bencil bir şekilde hareket edebilir. Bu sorunları tespit etmek ve çözmek için mekanizmalar olmalıdır. Aksi takdirde işbirlikçi iletişim, ciddi performans bozulmasından muzdarip olabilir.

Bencil veya kötü niyetli bir düğüm, diğer kaynaklarla veya ağ denetleyicisiyle işbirliği yapmadan ağ kaynaklarını kullanmaya çalışan veya işbirliği yaparken iletilmek istenen veriler üzerinde bir bozulmaya sebep olan bir düğüm olarak tanımlanmaktadır, ve sistemin performansını ciddi şekilde etkileyebileceği açıktır. Bu nedenle tezimizde kötü niyeti, alınan paketi iletmesi beklendiğinde alınan paketi bloke eden veya değişikliğe uğratarak ileten eylemler olarak tanımlamaktayız.

İşbirlikçi iletişimin aşağıdaki özellikleri, ataklarda savunmasızlığa neden olurken bu da ağdaki düğümlerde zayıf ve kesintili işbirliğine neden olacaktır [40].

- Olası Serbest Sürücüler: Bu bölümde belirtildiği gibi, işbirlikçi kablosuz ağın düzgün çalışmasını sağlayan temel varsayımlardan biri, düğümlerin kaynaklarını ağdaki diğer varlıklar arasında paylaşmaya istekli olmasıdır. Bu nedenle, beklemek ve seçildiklerinde diğer düğümlerle işbirliği yapmak zorundadırlar. Ancak, kuruluşların işbirliğini güvence altına alan hiçbir kontrol mekanizması yoktur. Aslında bir düğüm işbirliği yapmayı ve aktarmayı kasıtlı olarak reddedebilir. Bu bağlamda, iyi davranmış düğümler için bir teşvik mekanizması gibi düşünülebilir. İşbirliği yapmayan düğümleri engellemek için, iyi davranan düğümler, güvenilir olmayan düğümlerle işbirliği yapmayı reddedebilir.
- Merkezi Denetimin Yokluğu: Düğümler alandan bölgeye, ağdan ağa geçebilir ve kısa vadeli ilişkiler kurabilir. Bu ilişkiler hakkında eksiksiz bilgi sahibi olmak önceden mümkün değildir. Bu nedenle, kimlik doğrulama ile ağa yalnızca o meşru düğümlerin girmesine izin verilmelidir. Ayrıca, düğüm davranışının dinamik izlenmesi için bir mekanizma olmalıdır.
- Sık Topoloji Değişiklikleri: Ağ topolojilerindeki olası değişiklikler rölelerin hatalı davranışını önlemeyi zorlaştırabilir. İşbirlikçi geçiş ağlarında yer alan tüm düğümler, kaynak ve hedef arasında başarılı bir iletim sağlamak için güvenilir olmalıdır. Ağ için sabit bir topoloji şekli yoktur ve topoloji sık sık değişebilir.
- Kablosuz Ortamın Yapısı: Kablosuz ortamın doğası, herhangi bir erişime açık olması nedeniyle güvenlik açığına neden olur. Aslında, fiziksel erişimde herhangi bir sınırlama yoktur ve bu nedenle tüm düğümlerin çeşitli yönleriyle doğrulanması ve izlenmesi gerekir.
- Ölçeklenebilirlik: Herhangi bir politika veya protokol, gelen rölelerin doğal dinamik doğası ile uyumlu olmalı ve ağda genişletilebilir olmalıdır.

Bu tezde, güven tesis etme yöntemi, Bölüm 3'te önerdiğimiz atak tespiti ile herhangi bir topoloji değişikliğinde, merkezi bir kontrol olmadan kötü niyetli röleleri ayırt etmekle ilgilidir. Literatürde farklı güvenlik mekanizmaları oluşturulmuş, bu mekanizmalar uygulama modlarına göre aşağıda sıralandığı gibi kategorilendirilmiştir. Önleyici ve reaktif mekanizmalar, önleyici ve tespit tabanlı teknikler ile temelde iki tür olarak sınıflandırılmaktadırlar:

- Önleyici Mekanizma: Amaç, kimlik doğrulama, özellikle veri kaynağı kimlik doğrulaması ve şifreleme ile bütünlük koruması gibi bazı önleme tekniklerini kullanmaktır. Aslında bu tür araçlar, işbirlikçi iletişim ağlarındaki ataklara karşı savunmanın ön saflarında hareket eden güvenlik mekanizmalarıdır.
- Reaktif Mekanizma: Bu, ilk önleme mekanizmasını atlamış olan ataklara karşı ön savunma hattı olarak çalışmaktadır.
- Önleme: Ön hat savunmasıdır ve kötü amaçlı veya rakip düğümlerin ataklarını veya yetkisiz eylemlerini önlemek için tasarlanmıştır.
- Algılama: Ağa giren tüm olumsuz veya hatalı çalışan düğümlerin izlenebilmesini ve ağdan ayrılabilmesini sağlamak için tasarlanmıştır.

Bölüm 3'te önerilen yöntem, yukarıdaki önleme tekniklerini geçen ancak yine de kötü amaçlı düğümlerin, temelde güvenilir/kötü niyetli olarak ikili bir sınıflandırma ile ayırt edilerek tespit edilmesini amaçlamaktadır. Güvenli davranışın tanımı genel olarak, bir olayın beklenildiği gibi gerçekleştiği durumlardır. [54]'te belirtildiği gibi güvenen, güvenilirlik (reliability) ve güven (secure) arasındaki ilişkiyi göz ardı ederse, ilgili riskin yanlış hesaplanması söz konusu olacaktır. Güvenin yanlış değerlendirilmesi ise olası işbirliklerinin önüne geçebilir. Bu nedenle güvenirlilik ve güven arasındaki ilişkinin en dengeli şekilde kurulması gerekmektedir.

2.4. LTE-A Fiziksel Katman Kaynak Tahisisi

Paylaşılan kablosuz kanal, ağ hizmetinin en üst düzeye çıkarılması ve kullanıcıların adil kullanım kriterlerinin karşılanması için iyi tanımlanmış bir iletim mekanizması gerektirmektedir. LTE gibi yeni nesil ağlar, video gibi belirli bir kare hızı gerektiren kaynaklara yönelik, farklı bir talebi olan hizmetleri desteklemek üzere tasarlanmıştır. LTE'nin hava arayüzü - OFDM, ağır spektrumlu iletişim kanallarında yüksek spektral verimlilik ve performansa sahip gelişmiş bir iletim düzenidir.

OFDM, WLAN, HiperLAN ve IEEE 802.11 serilerinde başarılı olmuştur ve LTE'nin başarısına da büyük bir katkı sunması tasarlanmıştır. Radyo kaynak yönetimi (Radio Resource Management (RRM))'nin amacı, paraziti azaltarak, kullanıcının algıladığı servis kalitesini (Quality of Service (QoS)) iyileştirmek, spektral verimliliği arttırmak ve şebeke operatörü gelirini en üst düzeye çıkaracak mekanizmalar tasarlamaktır. Radyo kaynaklarının azlığı ve pahalılığı, performans hedeflerini karşılamaya yönelik

zorlu yayılma koşulları ile birleştiğinde RRM'yi önemli bir modül haline getirmektedir. Bir UE'nin iletişim kurmaya başlaması ve bir eNB'ye bağlanması için pil gücü ve taşıyıcı/zaman dilimi gerekmektedir. Başka bir deyişle, Güç Kontrolü (PC) ve Kaynak Tahsisi (Resource Allocation (RA)) LTE RRM'deki kilit noktalardır.

Uluslararası Mobil Telekomünikasyon Sistemi (Universal Mobile Telecommunications System (UMTS)) gibi geniş bant sistemlerinde PC, uzak bir noktadaki iletişime olan etkileri azaltmak için kullanılmaktadır.

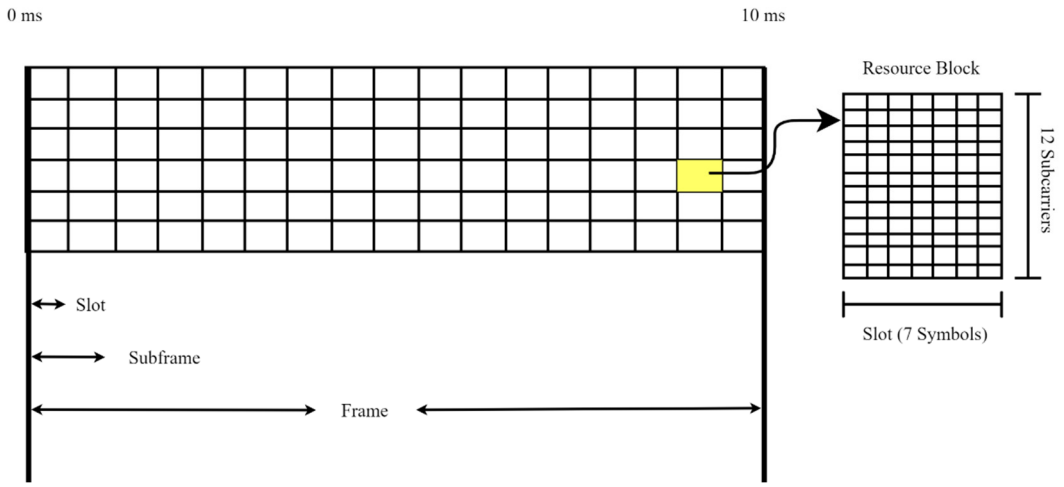
LTE geniş bant sistemlerinde, hücreler arası etkileşim teorik olarak sıfır olduğundan PC; yol kaybını, gölgelemeyi ve hücreler arası girişim (Inter-cell interference (ICI))'i telafi etmek için kullanılmaktadır. Bu çalışmada LTE-A bant genişliği ile RA ilişkisi ve fiziksel katmanda alınan kaynak bloğu miktarının sonuçlara olan etkileri incelenmiştir bu nedenle fiziksel katman çerçeve yapısı ve detayları üzerinde durulmaktadır.

LTE standardında Tip 1 ve Tip 2 olmak üzere iki tür çerçeve yapısı vardır. Tip 1, Frekans Bölmeli Çoğullama (FDD, uplink ve downlink frekansla ayrılmış) ve Tip 2 Zaman Bölmeli Çoğullama (TDD, uplink ve downlink zamanla ayrılmış) dır. Bir LTE çerçevesini tanımlamak için Tablo 2.2.' de gösterildiği üzere altı birim bulunmaktadır: çerçeve, yarım çerçeve, alt-çerçeve, yuva, sembol ve temel zaman birimi (Ts).

Tablo 2.2. LTE çerçevesini tanımlamak için kullanılan terimler.

Zaman Birimi	Değer
Çerçeve (Frame)	10 ms
Yarım Çerçeve (Half-frame)	5 ms
Alt Çerçeve (Subframe)	1 ms
Tek Yuva (Slot)	0,5 ms
Sembol (Symbol)	(0,5 ms) / 7 normal
Temel Zaman Birimi (T_s)	1 / (15000 * 2048) saniye » 32,6 ns

Bir kaynak bloğu (Resource Block (RB)), bir kullanıcıya tahsis edilebilecek en küçük kaynak birimidir. Kaynak bloğu 180 kHz genişliğinde ve 1 slot genişliğindedir. RB Frekansta 12 x 15 kHz alt taşıyıcı veya 24 x 7,5 kHz alt taşıyıcı genişliğindedir ve RB başına kullanılan alt taşıyıcı sayısı genellikle 12'dir. Frekans birimleri, alt taşıyıcıların veya kaynak bloklarının sayısıyla ifade edilebilmektedir.



Şekil 2.2. LTE FDD çerçevesinin 1.4 MHZ bant genişliği için gösterimi.

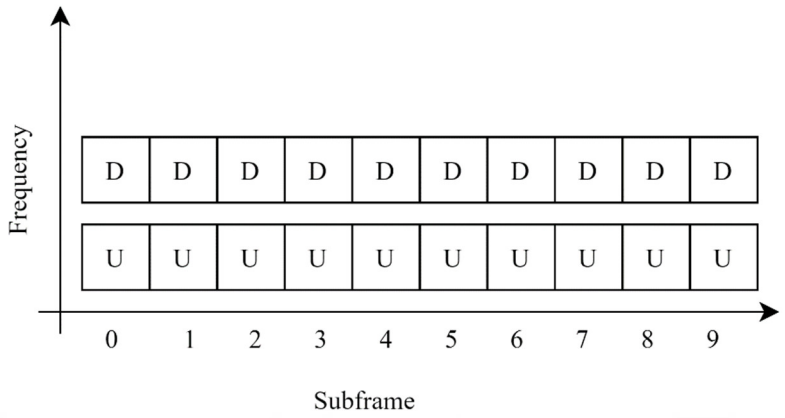
Örneğin, 5 MHz'lik bir downlink sinyali 25 RB genişliğinde veya 301 alt taşıyıcı genişliğinde olarak tanımlanabilir. Bir LTE çerçevesi için temeldeki veri taşıyıcısı, kaynak öğedir (Resource Element (RE)). 1 alt taşıyıcı x 1 sembolü olan RE, çerçevenin en küçük kısmıdır ve fiziksel bir kanal veya sinyalde verileri temsil eden tek bir karmaşık değerdir. Standart tarafından tanımlanan bant genişlikleri 1.4, 3, 5, 10, 15 ve 20 MHz'dir. Tablo 2.3. de, uplink ve downlink için her bant genişliğinde kaç alt taşıyıcı ve RB olduğunu gösterilmektedir.

Tablo 2.3. Bant genişlikleri ile kaynak blok ilişkisi.

Bant genişliği	Kaynak Blokları	Alt taşıyıcılar (downlink)	Alt taşıyıcılar (uplink)
1,4 MHz	6	73	72
3 MHz	15	181	180
5 MHz	25	301	300
10 MHz	50	601	600
15 MHz	75	901	900
20 MHz	100	1201	1200

FDD modunda, kanal uplink ve downlink çerçevelerinin her ikisi de 10 ms uzunluğundadır ve frekansta veya zamanda ayrılmaktadır. Çift yönlü FDD, uplink ve downlink çerçeveleri frekansla ayrılır ve sürekli senkronize olarak iletilir. TDD modunda, uplink ve downlink alt çerçeveleri aynı frekansta iletilir ve zaman alanında

çoğullanır. Uplink ve downlink özel alt çerçevelerin konumları, uplink-downlink yapılandırması tarafından belirlenmektedir.



Şekil 2.3. LTE FDD modunda çerçeve yapısı.

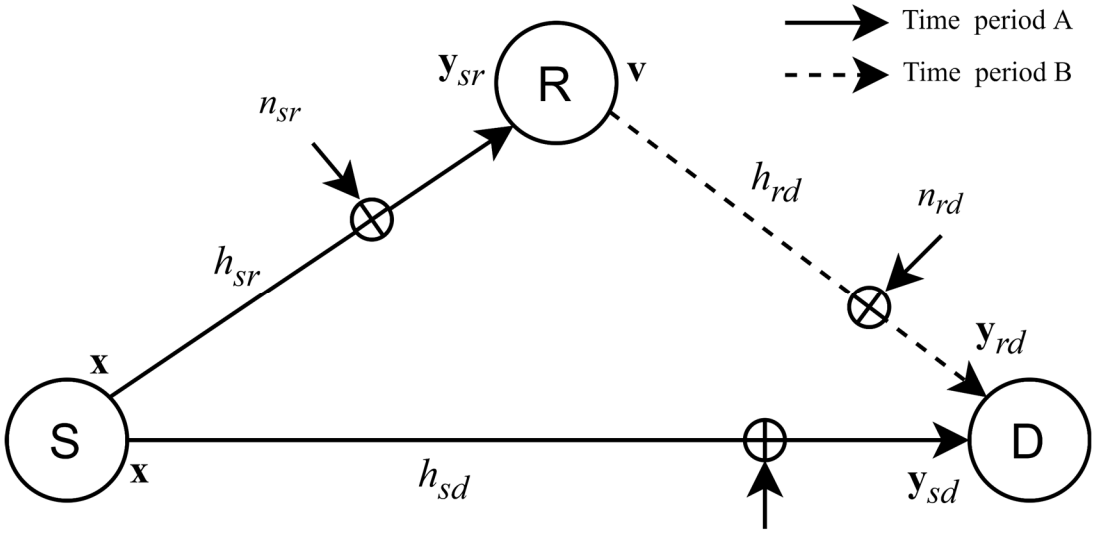
3. SİSTEM MODELİ VE PROBLEM TANIMI

3.1 Röle Ağ Sistem Modeli

Bir kaynak, bir röle ve bir hedef düğümden oluşan klasik üç düğümlü işbirliği iletişim ağı modeli tasarlanmıştır. Modelin simetrisi nedeniyle, genlik kaybı olmadan, ağdaki düğümlerin, tek seferde yalnızca tek yollu bir Rayleigh sönüm kanalı üzerinden sinyal alıp iletebilecekleri yarı çift yönlü modda çalıştığı varsayılmaktadır.

İşbirliği stratejisi iki aşamada gerçekleştirilmektedir. Faz 1'de kaynak düğüm, sinyalleri yayınlar ve hem röle hem de hedef düğüm tarafından alınır. Sönümlenen kanallar (Rayleigh fading channel) yoluyla sinyallerini gürültü oranı yüksek bir şekilde alabilmektedir. Faz 2'de röle, AF rölesi protokolünü gerçekleştirir; röle sadece aldığı sinyali yükseltir ve bir miktar güç ekleyerek hedefe iletir.

LTE-A ağ downlink iletim modeli, Şekil 3.1'de ele alınmaktadır. Burada S kaynağı eNB, R RN'yi ve D hedef düğümü temsil etmektedir ve Şekil 2.1'e atıfta bulunulduğunda UE'dir. Bu sistem modelinde, R'nin, AF geçiş protokolü kullanan katman 1 rölesi olduğu varsayılmaktadır. Kaynak düğüm (S), doğrudan iletim veya bir röle düğümü (R) aracılığıyla hedef düğüm (D) ile iletişim kurmaktadır.



Şekil 3.1. İşbirlikçi ağlarda röle sistem modeli.

LTE-A standardında yarım alt çerçeveye eşit, T1 periyodu içinde bir OFDM sinyali *x(t)* iletir. S'den yarım alt çerçevede toplam N iletilmek istenen sembol $\mathbf{x} = [x[1]\, x[2] \ldots x[N]]^T$ dir. Gönderim süresi T1 sırasında, R ve D tarafından alınan sinyal vektörleri,

$$\mathbf{y}_{sr} = h_{sr}\mathbf{x} + \mathbf{n}_{sr}, \tag{3.1}$$

$$\mathbf{y}_{sd} = h_{sd}\mathbf{x} + \mathbf{n}_{sd}. \tag{3.2}$$

T2 sırasında, röle düğümü R, alınan $\boldsymbol{y}_{sr}$ sinyalini yükseltir ve ardından, tarafına verildiği gibi D'ye yeniden iletmektedir,

$$\mathbf{v} = \alpha \mathbf{y}_{sr}, \tag{3.3}$$

$$\alpha = \frac{\rho}{P_{sr}\sigma_{sr}^2 + n_{sr}[i]}, i = 1,2, \ldots, N, \tag{3.4}$$

$$\mathbf{y}_{rd} = h_{rd}\mathbf{v} + \mathbf{n}_{rd}. \tag{3.5}$$

Yukarıdaki denklemlerde, **y** ve **n** vektörleri, aşağıdaki gibi yazılabilen örneklerden oluşmaktadır,

$$\mathbf{y} = [y[1]\, y[2] \ldots y[N]]^T, \tag{3.6}$$

$$\mathbf{n} = [n[1]\, n[2] \ldots n[N]]^T. \tag{3.7}$$

Burada röle gücü olarak $\rho = 1$ olduğunda, P_{sr}, röle düğümünde alınan sinyal gücüdür ve σ_{sr}^2, S-R kanalının varyansıdır. Denklem (3.1) ile (3.5)'de yer alan $\mathbf{n}_{sr}$, $\mathbf{n}_{sd}$ ve $\mathbf{n}_{rd}$, sırasıyla S-R, S-D ve R-D bağlantılarında σ^2 varyansı olan bağımsız ve özdeş dağılmış (i.i.d.) toplamsal beyaz Gaussian (AWGN) gürültü sinyalleridir.

Karmaşık kanal katsayıları, kablosuz kanalda yol kaybı ve sönümlenme etkilerini içeren Rayleigh sönmesi S-R, S-D ve R-D bağlantılarında sırasıyla h_{sr}, h_{sd} ve h_{rd} ile gösterilmektedir. S-R, S-D ve R-D bağlantılarındaki kablosuz kanalların yarım alt çerçeve iletimi süresinde değişmediği varsayılmaktadır.

Ek olarak hüce içi geçişlerin etkisi dikkate alınmayarak sembollerin ayrı ayrı alındığı ve kullanıcı ekipmanında ise tek anten ile iletişimin sağlandığı varsayılmaktadır.

3.2. Röle Ağlarda Atak Tanımı ve Modelleri

Geleneksel işbirlikçi ağlarda, kaynak düğümde ve röle düğümünde alınan sinyaller, sinyal kalitesini artırmak için alıcı düğümdeki farklı çeşitlilik teknikleri kullanılarak birleştirilmektedir. İşbirlikçi birleştirme teknikleri ise röle düğümlerinin kötü niyetli davranışlarını denetlemez ve sadece röle tarafından sağlanan çeşitliliği kullanmaya çalışır. Kaynak S'den hedef D'ye sinyal iletiminin doğru ve güvenilir bir şekilde alındığı göz önüne alındığında, kritik sorun, çeşitlilik için sinyalini birleştirmeden önce röleden iletimin güvenli olup olmadığını belirlemektir.

Bu çalışmanın amacı, rölenin çeşitli tipte röle atakları kapsamında güvenli olup olmadığını tespit etmektir. LTE-A röle iletişiminde, bir rölenin fiziksel katmanda oluşturabileceği olası atak türleri bir sonraki bölümde açıklanmaktadır. İletişimin doğası nedeniyle, röleden kaynaklanabilecek tüm kötü niyetli davranışları belirlemek mümkün değildir bu nedenle fiziksel katmanda meydana gelebilecek olası atak senaryolarını dikkate alarak üç farklı kötü amaçlı röle atak türü aşağıdaki gibi modellenmiştir [41].

3.2.1. Veri karıştırma atağı – A1 (Garbling attack)

Kötü amaçlı röle düğümü, sembol serisinde modüle edilmiş sembollerin yeni bir rastgele permütasyonunu üretir. Orijinal sembollerin dizini **y**, yani, $\boldsymbol{y}_{sr}[1] \ldots \boldsymbol{y}_{sr}[N]$ rastgele bir permütasyon fonksiyonu $\mathcal{P}$ ile yeniden konumlandırılmaktadır,

$$\tilde{\mathbf{y}}_{sr} = \mathcal{P}(\mathbf{y}_{sr}) \tag{3.8}$$

$\mathcal{P}$, sembol serisinde *1* ile *N* arasında eşit olarak dağıtılmış rasgele bir dizin alarak yeni bir sembol dizisi oluşturmaktadır.

3.2.2. Farklı veri iletme atağı – A2 (Regenerative attack)

Röle düğümü kasıtlı olarak orijinal **x** sinyalini yok sayar ve yeni oluşturulan sinyal $\tilde{\mathbf{x}}$ ile değiştirir, R'de üretilen yeni sinyal, orijinal sinyal x ile aynı modülasyon özelliklerine sahiptir.

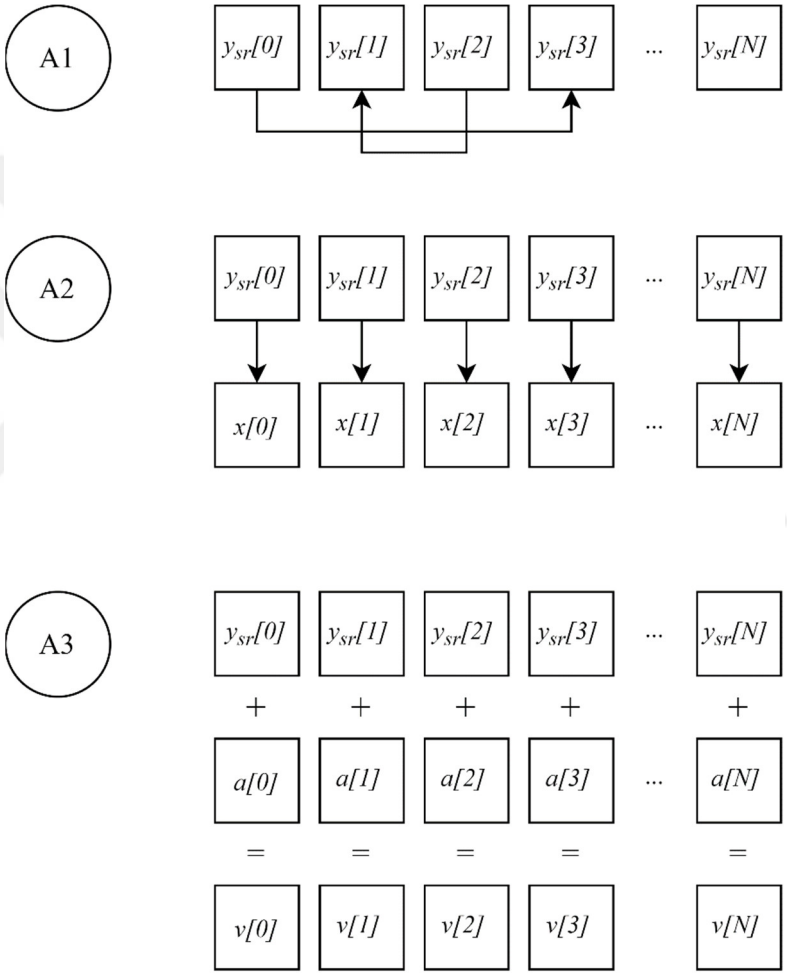
$$\tilde{\mathbf{y}}_{sr} = \tilde{\mathbf{x}}. \tag{3.9}$$

3.2.3. Veri enjeksiyon atağı – A3 (False data injection attack)

Bu atak türünde, R kasten D'ye iletilmeden önce kaynaktan alınan **x** sinyaline rastgele bir sinyal vektörü olan **a**'yı eklemektedir [13],

$$\tilde{\boldsymbol{v}} = \alpha \mathbf{y}_{sr} + \mathbf{a}, \tag{3.10}$$

$$\tilde{\mathbf{y}}_{rd} = h_{rd}\tilde{\mathbf{v}} + \mathbf{n}_{rd}. \tag{3.11}$$



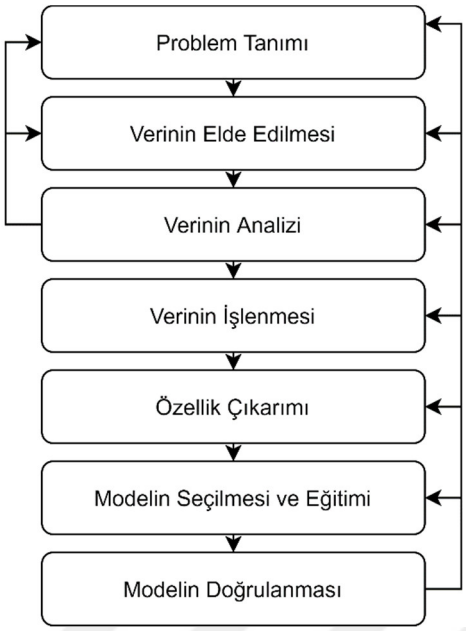
Şekil 3.2. Atak modellerinin garfiksel anlatımları.

4. MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE KÖTÜ NİYETLİ RÖLE TESPİTİ

Makine öğrenme tekniklerini kullanırken, ağ trafiğinde çerçevelerin (frame) örüntülerini analiz etmek ve sınıflandırmak için, istatistiksel temelli teknikler gibi modeller oluşturulmaktadır. Bununla birlikte, temel farklar bu metodolojinin stokastik özelliklerle sınırlı olmaması ve çerçeveleri sınıflandırmak için eşik değerleri yerine sembollerin özelliklerinin de kullanılmasıdır. Ayrıca bazı makine öğrenimi modelleri, tespit aşaması sırasında güncellenebilir ancak bu, istatistiksel tabanlı modeller için mümkün değildir [45].

Makine öğrenimi modellerinin çeşitli avantajları ve dezavantajları bulunmaktadır. Çok esnektirler çünkü tespit aşamasında güncellenebilirler. Ayrıca, çerçeveleri ve özellikleri arasındaki karmaşık korelasyonları ve bağımlılıkları da modelleyebilmektedirler. Öte yandan, bazı makine öğrenmesi teknikleri yüksek donanım özelliklerine ihtiyaç duymaktadır ve modelin tüm parametrelerini ayarlamak çok karmaşık olabilmektedir [46].

Bu bölümde, kötü niyetli röle tespit problemini çözmek için uygun öğrenme paradigmasının seçimine etki eden etkenler anlatılmaktadır. Toplamda üç öğrenme paradigması belirlenmiştir. Bunlar; denetimli, denetimsiz, ve istatistiksel öğrenmedir. Denetimli öğrenme, verilerin zaten etiketlendiğini (hangi sınıfa ait olduğu bilgisi) varsayar ve bu, temel gerçeğin önceden bilindiğini göstermektedir. Sonuç olarak, bu paradigma, sırasıyla ayrık veya sürekli sonuçları tahmin ederek hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılmaktadır. Diğer yandan, denetimsiz öğrenme yalnızca etiketlenmemiş verileri kullanmakta ve farklı sınıfları tanımlamak veya örüntüyü tanımlamak için benzer veri noktalarını kümelemektedir. Son olarak istatistiksel öğrenme, verilerin olasılık yoğunluk fonksiyonlarını dikkate alarak belli bir eşik değeri belirler ve bu eşik değerinin dışındaki değerleri aykırı olarak etiketler. Sonuç olarak, bu modeller kararların alınması veya bir planlama yapılması gerektiğinde kullanılmaktadır [47,48].



Şekil 4.1. Makine öğrenmesi modelleri tasarlanırken kullanılan genel prosedür.

Makine öğrenimesi modelleri hakkında ayrıntılar 4.4. - 4.9. alt bölümlerinde anlatılmaktadır.

4.1. Başarımın Ölçütleri

Modelin verimliliğini ölçen kayıp (loss) fonksiyonu seçilirken farklı maliyet fonksiyonları (cost function) modelin farklı yönlerini ele aldığından, söz konusu fonksiyonun seçimi önem taşımaktadır. Tablo 4.1., modelin doğruluğunu artıran üç türe bölünmüş en yaygın kullanılan performans metriklerine genel bir bakış sunmaktadır. İlk bölüm, denetlenen regresyon problemleri için konvansiyonel metrikleri açıklamaktadır. Bu durumda, Ortalama Mutlak Hata (MAE), Ortalama Kare Hata (MSE), Kök Ortalama Kare Hata (RMSE) ve Normalize Kök Ortalama Kare Hata (NRMSE). MAE ve MSE, regresyon modellerinde en çok tahmin hatalarını cezalandırmak için kullanılmaktadır ve nedeni hem MSE hem de MAE'nin uygulanmasının kolay olmasıdır. RMSE, hatanın standart sapması olarak görülebilir ve genellikle daha yorumlanabilir bir MSE formu olarak kabul edilmektedir. NRMSE, RMSE'nin normalize edilmiş şeklidir ve sonuç olarak, farklı regresyon modellerini birbirleriyle karşılaştırmak için sıklıkla kullanılmaktadır [47,48].

Tablo 4.2. Doğruluk hesaplaması için maliyet fonksiyonları.

Kayıp fonksiyonu (Loss function)	Formül
Ortalama Mutlak Hata (Mean Absolute Error (MAE))	$\frac{1}{n}\sum_{i=1}^{n}\|y_i - y_i^*\|$, burada y_i gerçek değer y_i^* ise tahmin edilen değer ve n toplam örnek sayısı.
Ortalama Karesel Hata (Mean Squared Error (MSE))	$\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$, burada y_i gerçek değer y_i^* ise tahmin edilen değer ve n toplam örnek sayısı.
Kök Ortalama Karasel Hata (Root Mean Squared Error (RMSE))	$\sqrt{MSE}$
Normalize Kök Ortalama Karesel Hata (Normalized Root Mean Square Error (NRMSE))	$\frac{RMSE}{y_{max}-y_{min}}$, burada y_{max} maksimum gerçek değer, y_{min} minimum gerçek değer.
Ortalama Doğruluk (Average Accuracy)	$\frac{1}{c}\sum_{k=1}^{c}\frac{TP_k+TN_k}{TP_k+TN_k+FP_k+FN_k}$, burada c sınıf sayısını, TP_k k sınıfı için doğru pozitif sınıflandırma sayısını, FP_k yanlış pozitif, FN_k yanlış negatif, TN_k doğru negatif sınıflandırma sayılarını temsil eder.
Kesinlik$_\mu$ (Precision$_\mu$)	$\frac{\sum_{k=1}^{c}TP_k}{\sum_{k=1}^{c}TP_k+FP_k}$
Kesinlik$_M$ (Precision$_M$)	$\frac{1}{c}\sum_{k=1}^{c}\frac{TP_k}{TP_k+FP_k}$
Hatırlama$_\mu$ (Recall$_\mu$)	$\frac{\sum_{k=1}^{c}TP_k}{\sum_{k=1}^{c}TP_k+FN_k}$
Hatırlama$_M$ (Recall$_M$)	$\frac{1}{c}\sum_{k=1}^{c}\frac{TP_k}{TP_k+FN_k}$
Alıcının çalışma karakteristik eğrisinin altındaki alan (Area under the receiver operating characteristic curve AUC-ROC)	Her ayrım eşiği için hatırlama ve yanlış pozitif oranı arasında bağlantıyı ve temsil ettiği bölgenin alanını verir.

Sınıflandırma problemleri için en sık kullanılan metrik doğruluk (accuracy) metriğidir. Bununla birlikte, veri kümesi sınıflara göre değişim gösterirse, doğruluk ölçüsü güvenilirliğini kaybetmektedir, çünkü yalnızca en yüksek yoğunluklu sınıfı tahmin eden bir model yüksek bir doğruluk elde etmektedir.

Kesinlik (precision) ve hatırlama (recall), literatürde sıkça bahsedilen metriklerdir. Bununla birlikte, yanlış negatifleri veya yanlış pozitifleri dikkate almamalarından dolayı bu değerlerin önem taşıdığı problemlerde kullanılamayacağı anlamına gelmektedir. Son olarak, alıcı işletim karakteristiği (Receiver Operating Characteristic (ROC)) ve bu karakteristik eğrinin altında kalan alan (Area Under the Curve (AUC)) skoru çok sınıflı bir bağlamda çarpık veri kümelerini açıklamak noktasında önem taşımaktadır [47-50].

4.2. Veri Analizi ve Önişleme

Veriler toplandıktan sonra, verilerdeki olası hataları tanımlamak ve modelin tasarımı ve doğrulanması sırasında ortaya çıkabilecek sorunların ilk göstergesini elde etmek için veriler analiz edilmektedir. Örneğin, hiçbir verinin eksik olmadığından, edinilmesi sırasında herhangi bir yanlışlığın bulunmadığından ve işleme sırasında veya depolamada verilerde herhangi bir hata bulunmadığından emin olmak için kontroller yapılmaktadır.

Ayrıca, verilerin sınıflara göre farklı yoğunlukta olup olmadığını ve modelde yanlış nedensel bağımlılıklar yaratabilecek verilerden belirli bilgilerin çıkarılmasının gerekip gerekmediğini kontrol etmek için bazı analizler yapılarak eğitim verisi içerisindeki olası tutarsız veriler tespit edilmelidir [48].

Üstesinden gelinmesi gereken daha karmaşık bir zorluk, veri çarpıklığı veya sınıf dengesizliği problemidir. Sınıf dengesizliği, bazı sınıfların verilerde diğerlerinden daha sık meydana görülmesi bir sınıflandırma problemini meydana getirir ve böylece makine öğrenme modelleri az yoğunluktaki sınıfları daha az doğru tahmin edebilmektedir.

Daha dengeli bir veri kümesinin oluşturulması ve kayıp fonksiyonunun yeniden ağırlıklandırılmasıdır [51,52,53]. Üç olası strateji uygulanabilmektedir: aşırı

örnekleme, yetersiz örnekleme ve sentetik örnekleme. İlk durumda, azınlık sınıfının veri sayısı, karşılık gelen veri örneklerinin çoğaltılmasıyla arttırılmaktadır. İkinci durumda, çoğunluk sınıflarının veri miktarı, veri kümesinden bir dizi temsili veri örneği seçilip geri kalanı düşürülerek azaltılabilmektedir.

Seçim, çoğunluk kümesinden rastgele örnekleme veya kümeleme teknikleri kullanılarak ve daha sonra küme başına bir dizi temsili veri örneği seçilerek gerçekleştirilebilmektedir. Üçüncü durumda, genellikle Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) algoritması [54] veya Uyarlanabilir Sentetik Örnekleme (ADASYN) algoritması [55] kullanılarak orjinal verilere dayanılarak sentetik veri örnekleri oluşturulmaktadır.

Özellik seçiminde, veri kümesinden alakasız veya gereksiz bilgiler kaldırılarak veriler azaltılmaktadır. Bunun avantajı, eğitilecek modelin aşırı donmaya karşı daha sağlam olması ve hesaplama yükünün azalmasıdır. Bunun için, ileri arama yaklaşımı, her adımda en iyi özelliğin eklendiği bir yöntem, geriye doğru arama yaklaşımı, her adımdaki veri kümesinden en kötü özelliğin kaldırıldığı bir yöntem, hibrit yaklaşımlar gibi farklı metodolojiler kullanılabilmektedir.

4.3. Modelin Seçimi ve Öğrenme Yaklaşımı

Problem için kullanılacak modeller ve ilgili öğrenme yaklaşımları belirlenirken çeşitli noktalar dikkate alınmalıdır.

İlk olarak, problemin ön koşulları belirlenmelidir. Sıklıkla dikkate alınan koşullar, modeli eğitmek için sağlanan süre, yeni veriler için tahminler yapmak için geçen süre, yapılan hatadaki izin verilen marj ve sonucu belirlemek için kullanılan yöntemin yorumlanabilirliğidir [48].

Daha sonra, modelin düzgün bir şekilde genelleştirilmesini ve karmaşıklığının azaltılmasını sağlamak için düzenlileştirme teknikleri uygun şekilde seçilmelidir. Sıklıkla kullanılan düzenlileştirmeler arasında özellik seçim algoritmaları, bırakma katmanları ve elde edilen verilere veya özelliklere yararlı gürültü eklenmesidir[56].

Üçüncüsü, eğitim yaklaşımında çeşitli seçimler yapılmalıdır. İlk karar, toplanan verileri eğitim verilerinde, test verilerinde ve doğrulama verilerinde alt bölümlere

ayırmak için kullanılan yöntemdir. Bir seçenek verileri belirli bir stratejiye göre rastgele ve sonra 3 bölüme ayırmaktır; ancak daha gelişmiş yöntemler de uygulanabilir. Bunlardan biri, veri kümesinin olduğu bir prosedür olan k-katlı çapraz doğrulamadır (k fold cross validation). Orijinal sınıf dengesizliğini koruyarak ve her parçayı tam olarak k-1 kez bir eğitime atayarak, eğitim veri seti ve test veri seti olarak verileri k eşit parçalara bölünür. Böylece eğitim veri kümesindeki her alt parça ile model oluşturup test edilebilmektedir [57]. Bununla oluşturulan modeldeki hata ve varyasyon oranı dengeli bir değere getirilerek, modelin yeni bir veri karşısındaki başarısı attırılabilmektedir.

Eğitim yaklaşımında yapılacak son adım, eğitim sırasında dinamik olarak ayarlanması gereken hiper parametrelerin seçimi ve bu parametrelerin nasıl uygulanması gerektiğidir. Bu seçim, hiper parametrenin verilerin doğruluğu üzerindeki etkisine dayanmaktadır; fakat aynı zamanda sorunun ön koşullarını da dikkate almaktadır [56]. Seçilen hiper parametreleri ayarlamak için çeşitli yaklaşımlar bulunmaktadır ve bunlardan üçü sıklıkla kullanılmaktadır: rastgele arama, ızgara arama ve bayes optimizasyonudur. İlk ikisinde, ayarlama sırasıyla hiper parametreleri rastgele bir şekilde seçerek veya en iyi model doğruluğunu sağlamak için her bir hiper parametre kombinasyonunu test ederek gerçekleştirilmektedir.

Parametrelerin seçimi için ayrıca k-katlı çapraz doğrulama yöntemide kullanılmaktadır. Tüm eğitim verisi farklı parametreler ile tekrar tekrar eğitilerek en küçük hata değerine ulaşıldığında kullanılan parametrelerin en başarılı değerler olarak nihayi modelde kullanılması tasarlanmıştır.

Son olarak, modelin kendisi ön koşullar dikkate alınarak seçilmelidir. Bu seçim sürecini desteklemek için, en sık kullanılan makine öğrenimi modellerinin kısa bir açıklaması Tablo 4.1 de verilmektedir. Tabloda detaylandırılan algoritmalar literatürde en çok denenmekte ve farklı kategorileri temsil etmektedir.

Son adımda model, önceki adımlarda yapılan seçimlere göre uygulanır. Eğitilir ve daha sonra tüm ön koşulları karşılayıp karşılamadığını kontrol etmek ve yeni veriler için tahmini doğruluğunu belirlemek için denetlenir. Bu analizlere dayanarak, modelin zayıf noktaları ve eksiklikleri tespit edilebilir ve böylece bunları çözmek için gerekli önlemler alınabilmektedir.

Tablo 4.3. Kategorilerine göre algoritmaların kısa açıklamaları ve karmaşıklıkları O(.), olarak verilmiştir.

Model	Açıklama	Eğitim Zaman Karmaşıklığı	Test Zaman Karmaşıklığı	Model Türü
Destek Vektör Makineleri (Support Vector Machine SVM)	Destek vektörleri kullanan bir sınıflandırma ve regresyon modeli, sınıflar arasındaki sınırları belirleyen (sınıflandırma) veya öngörülerin düşmesi gereken bir marj (regresyon) tanımlayan bir eğitim alt kümesi kümesi. Bu destek vektörleri daha sonra ağırlıklı özelliklerin ağırlığındaki ağırlıkları tanımlamak için kullanılır.	$O(M^3)$	$O(1)$	Denetimli Öğrenme
Yapay Sinir Ağları (Neural Network NN)	Doğrusal bir sınıflandırıcı (çok katmanlı algılayıcı) veya evrişimli bir sınıflandırıcı (evrişimli sinir ağı) kullanılarak oluşturulan birkaç düğüm katmanından oluşan bir sınıflandırma ve regresyon modeli.	*Nöron Sayısı* * $O(M)$	*Nöron Sayısı* * $O(1)$	Denetimli Öğrenme
Yerel Aykırı Faktör (Local Outlier Factor LOF)	Bir veri örneğinin yerel yoğunluğunu, yani mahalledeki veri örneklerinin sayısını tahmin eden ve kümeleri ve aykırı değerleri belirlemek için komşularının yoğunluğu ile karşılaştıran en yakın komşu model.	$O(Mk+ML)$	* $O(1)$	Denetimsiz Öğrenme
Tek Sınıflı Destek Vektör Makineleri (One Class Support Vector Machines OCSVM)	Aykırı algılama için SVM algoritmasının özel versiyonudur. OCSVM'deki hiperküre veya karar fonksiyonu, pozitif eğitim örneklerinin +1 olarak etiketlendiği küçük bir bölge için inşa edilmiştir.	$O(M^2L+M^3)$	* $O(1)$	Denetimsiz Öğrenme
İzolasyon Ormanı (Isolation Forest iForest)	İForest algoritması, veri kümesini alt ağaç topluluklarına dönüştürür. Bu rastgele ağaçların ortalama yol uzunlukları, karar fonksiyonundaki normallik ölçüm değerleridir.	$O(LMlog(M))$	* $O(1)$	Denetimsiz Öğrenme
En Küçük Kareler Aykırılık Tespiti (Least-squares Anomaly detection LSA)	En küçük kareler yöntemi, birbirine bağlı olarak değişen iki fiziksel büyüklük arasındaki matematiksel bağlantıyı, mümkün olduğunca gerçeğe uygun bir denklem olarak yazmak için kullanılan, standart bir regresyon yöntemidir.	$O(L^2M)$	* $O(1)$	İstatistiksel Öğrenme
Olasılıksal Temel Bileşen Analizi (Probabilistic Principal Component Analysis PPCA)	Verileri daha düşük boyutlu bir gizli uzay yoluyla analiz eden bir boyutluluk azaltma tekniğidir.	$O(L^2M + M^3)$	* $O(1)$	İstatistiksel Öğrenme

4.4. Veri Kümesinin İstatistiksel Özellikleri

Şekil 4.1'de D'de yani hedef düğümde alınan sinyalin genlik değerlerinin $|y_{sd}|$ ve $|y_{rd}|$'nin olasılık yoğunluk fonksiyonları (PDFs), röle atak modelleri, güvenilir röle ve kaynak sinyalleri için gösterilmektedir. Veri uzayındaki normalliği tanımlamak için elde edilen dağılım üzerinde eşik değerini kullanılarak sinyal örnekleri alınarak hesaplanmaktadır. Alınan y_{sd} ve y_{rd} sinyallerinin genlik ve faz değerleri için tamamlayıcı kümülatif dağılım fonksiyonları (CCDF's) ise Şekil 4.2'de verilmektedir.

İstatistiksel tespit yaklaşımları, alınan sinyal örneklerinin koşullu PDF'lerini $f_{y_{rd}|y_{sd}}(y_{rd}[i], y_{sd}[i])$ temel alarak veri kümesindeki normalliği bir eşik değer yardımı ile tanımlamaktadırlar [28]. CDF'lerin tersi perspektifte, aşma olasılığı (exceedance probability) olarak adlandırılan ve belirli bir seviyenin üzerindeki sinyal değerlerini temsil eden grafik, Şekil 4.2 (a)'da gösterilmektedir. Sinyal genliklerinin CCDF'lerinin, yakın aşma olasılıklarına sahip ve istatistiksel atak tespitinde kullanılmak için ayırt ediciliğinin bulunmadığı görülmektedir.

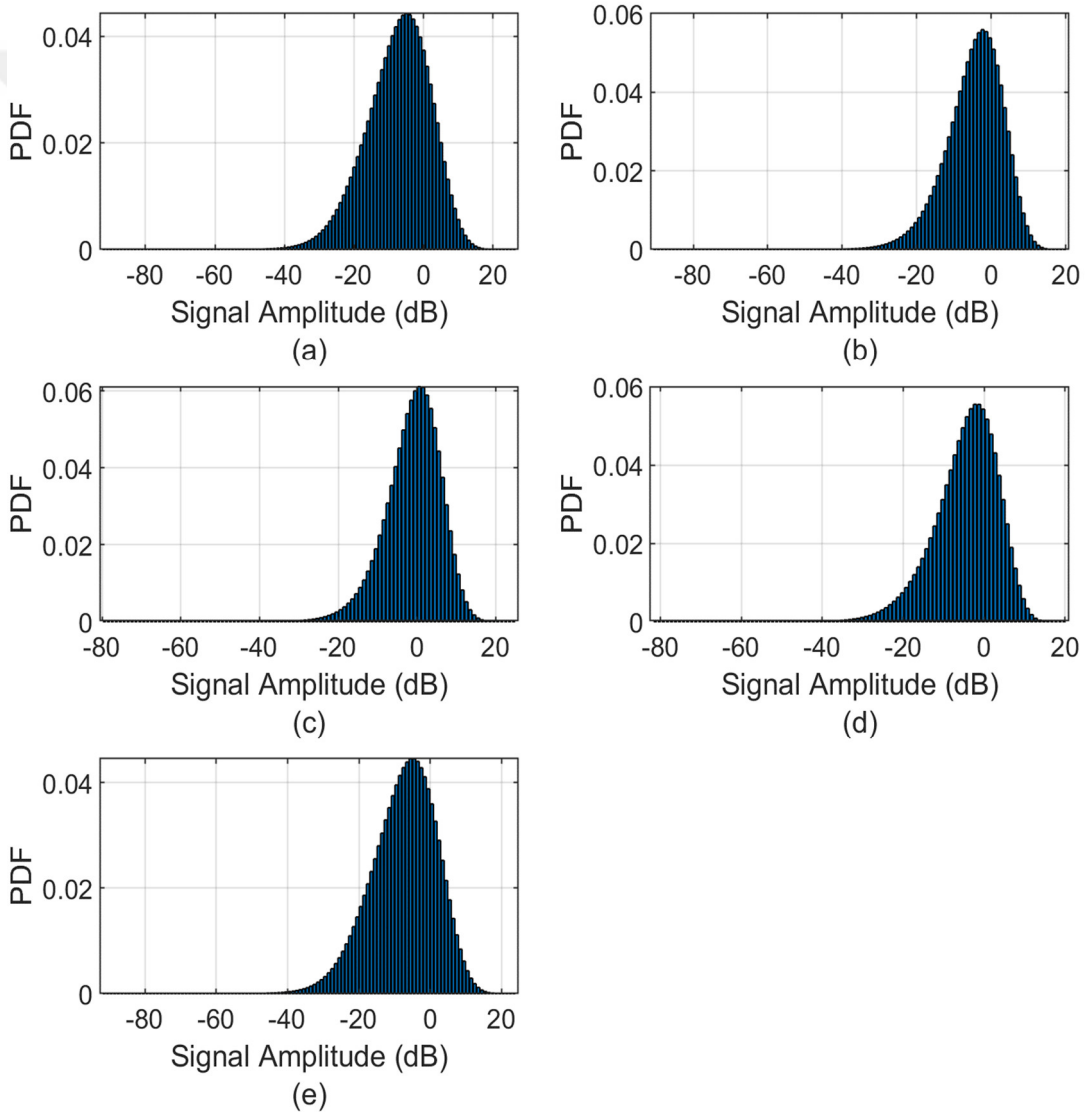
Öte yandan, sinyal fazlarının aşılma olasılıklarında bazı farklılıklar görülmektedir. Bu farklılıklar istatistiksel yöntemlerde kullanılmak üzere bir değer taşımamaktadır çünkü kaynak sinyalinin A2 ve A3 sinyallerine çok yakın olması ve güvenilir röle sinyalinin ise A1 atak modeline yakın olmasıdır. Bu durum ayırt edilebilirliği etkilemekte ve güvenilir sinyallerin atak olarak ya da atak sinyallerinin güvenilir olarak değerlendilmesi ile sonuçlanmaktadır.

Geleneksel atak tespit yaklaşımları, alınan sinyallerin istatistiksel yoğunluk fonksiyonlarını kullanmakta ve düşük yoğunluklu sinyalleri atak/aykırı değer olarak tespit edebilmesini sağlamaktadır. Bu nedenle, Şekil 4.2 ve Şekil 4.3'de gösterildiği gibi ayırt edilemeyen sinyal dağılımlarına sahip olan kötü niyetli rölelerin atak olarak tespit edilebilmesi mümkün değildir.

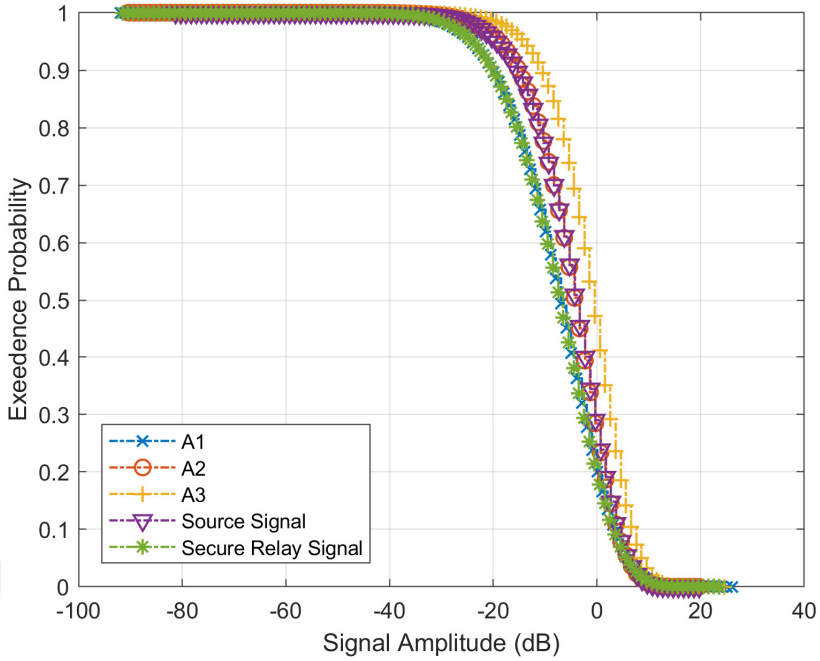
Bu tür röleleri tespit etmek için, bu tez çalışmasında, doğru tanımlanmış sinyal özelliklerine ihtiyaç duyan denetimsiz makine öğrenmesi algoritmalarının uygulanması çözüm olarak önerilmektedir. Makine öğrenmesi algoritmaları istatistiksel olarak bir birine yakın gibi görünen örnekleri dönüştürerek yani farklı

boyutlara taşıyarak ya da yeniden haritalandırarak, farklarının daha ayırt edici bir hale gelmesini sağlamaktadır.

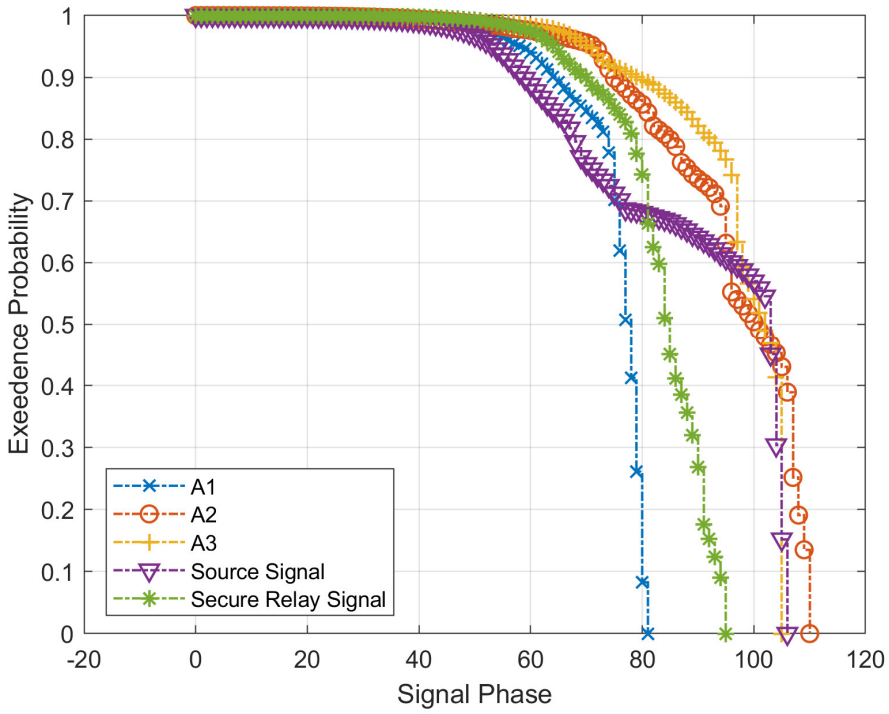
Bir sonraki bölümde, probleme çözüm getirmek adına sinyal özellikleri ve denetimsiz öğrenme algoritmaları tanımlanmaktadır. Bu modelleri geleneksel istatistiksel algılama modelleri ile karşılaştırdığımızda sunulan algılama modellerinin birçok başarısı gözlemlenmektedir. Önerilen tespit modelimizle, daha karmaşık karar fonksiyonları oluşturabilmekte ve ardışık semboller arasındaki gizli ilişkiyi araştırabilmekteyiz.



Şekil 4.2. (a) A1 atak olan RD bağlantısı, (b) A2 atak olan RD bağlantısı, (c) A3 atak olan RD bağlantısı, (d) kaynak sinyalli SD bağlantısı ve (e) güvenli olan RD bağlantısı için alınan sinyal genliklerinin PDF'leri (SNR=15 dB, Modülasyon = QPSK, Bant genişliği = 1.4 MHz).



(a)



(b)

Şekil 4.3. Sırasıyla tüm bağlantılar için hedefe alınan sinyalin genlik (a) ve faz (b) özelliğinin CCDF'leri, R-D A1 saldırısı, R-D A2 saldırısı, R-D A3 saldırısı, S-D kaynak sinyali ve R-D güvenli röle sinyali (SNR=15 dB, Modülasyon = QPSK, Bant genişliği = 1.4 MHz).

4.5. Giriş Verilerinin Özellikleri

Şekil 3.1'deki η → D bağlantısı üzerinden alınan karmaşık temel bant sembolünün n'inci örneği $y_{n,\eta\to D}$, η∈ {S, R} olsun. S kaynağından iletilen OFDM sinyalleri kablosuz kanaldan geçirilmekte ve $y_{n,\eta\to D}$ bu nedenle değiştirilmiş OFDM sinyali olarak kabul edilemektedir. S'deki OFDM sembolleri, LTE-A ağındaki QPSK, 16-QAM, 64-QAM gibi dijital olarak modüle edilmiş sembollere dayanmaktadır. Sembollerin sinyal sembol serisindeki (constellation) genlik, faz değerleri ve sembollerin faz farklılıkları veya bağıl fazları da modüle edilmiş semboller hakkında bilgi sağlamaktadır.

Modüle edilmiş semboller, seri içerisinde farklı genlik ve faz değerine sahip noktalar olarak tanımlanabilir. Bu nedenle, genlik, faz bilgisi veya $y_{n,\eta\to D}$'nin bağıl fazları, bu sinyal üzerindeki herhangi bir olağandışı etkiyi yakalamak için bazı ayırt edici bilgiler sağlayabilmektedir. Aşağıdaki özellikleri röle ataklarını algılamak için atak tespit algoritmalarına girdi olarak tanımlamaktayız:

$$f_{n,\eta\to D}^{(1)} = |y_{n,\eta\to D}|\,, \tag{4.1}$$

$$f_{n,\eta\to D}^{(2)} = \angle y_{n,\eta\to D}\,, \tag{4.2}$$

$$f_{n,\eta\to D}^{(3)} = \angle y_{n,\eta\to D} - \angle y_{n+1,\eta\to D}\,, \tag{4.3}$$

|(.)| ve ∠(.) genlik ve faz operatörleri anlamına gelmektedir. Yukarıdaki denklemlerde tanımlanan özellikler, 16-QAM modülasyonu için Şekil 4.4'de gösterilmektedir. Öyleki $y_{n,\eta\to D}$ sembolünün özellik vektörü aşağıdaki gibi tanımlanabilmektedir,

$$\mathbf{f}_n = \left[f_{n,\eta\to D}^{(1)}\ f_{n,\eta\to D}^{(2)}\ f_{n,\eta\to D}^{(3)}\right], \tag{4.4}$$

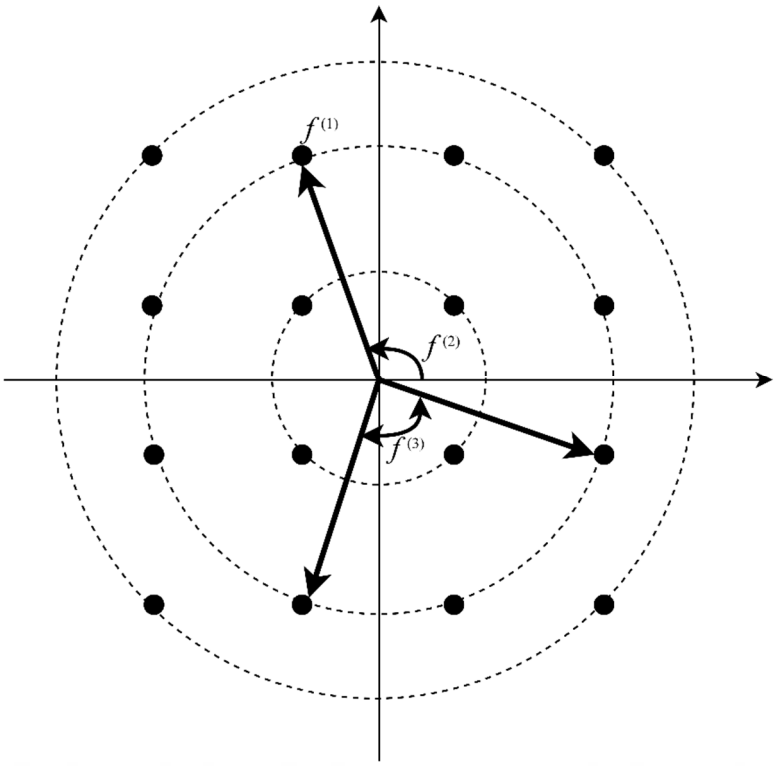
m^{th} yarım alt çerçeve, yani yuva (slot) için öğrenme aşamasında N ardışık sembol vardır ve bu nedenle özellik vektörü,

$$\mathbf{z}_m = [\mathbf{f}_1\ \mathbf{f}_2\ \ldots\ \mathbf{f}_N]^T\ \ m = 1, 2, \ldots, M. \tag{4.5}$$

M yarım alt çerçeve bulunduğundan, öğrenme aşamasındaki veri kümesi matrisi, verilen özellik vektörlerinin birleştirilmesiyle oluşturulmaktadır.

$$\mathbf{Z} = [\mathbf{z}_1\ \mathbf{z}_2\ \ldots\ \mathbf{z}_M], \tag{4.6}$$

öyleki $\mathbf{Z}$ matrisinin boyutu $L \times M$ dir ve $L=3N$.



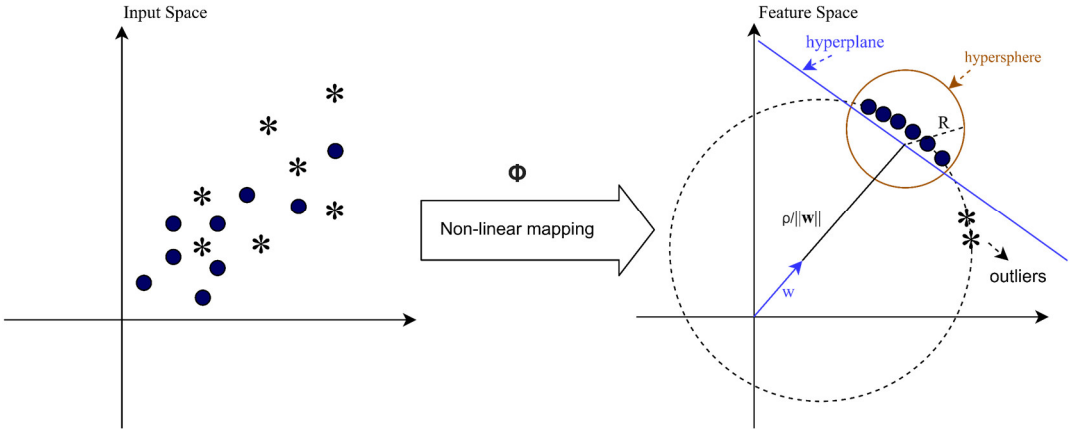
Şekil 4.4. 16-QAM modülasyonu ile alınan $y_{n,\eta \to D}$ için özelliklerin gösterimi.

4.6. Denetimsiz Öğrenme Teknikleri

4.6.1. Tek sınıflı destek vektör makineleri (OCSVM)

Tek sınıflı destek vektör makineleri (one class support vector machines OCSVM) Scholkopf vd. [68] tarafından geliştirilen aykırı algılama için SVM algoritmasının özel bir versiyonudur. OCSVM'deki hiperküre veya karar fonksiyonu, pozitif eğitim örneklerinin +1 olarak etiketlendiği küçük bir bölge inşa edilmektedir.

Karar işlevi veya sınırı, bir çekirdek işlevi tarafından eşleştirilen eğitim verilerinin çoğunu dikkate alan olabilecek en küçük bir hiper küredir. Bununla birlikte, *ζi, i = 1, ... ,M* gevşek/esnek değişkenlerini (slack variables) ekleyerek veri noktalarının sınırların dışına yerleştirilmesine de izin vermektedir.



Şekil 4.5. Radyal temelli çekirdek haritalama fonksiyonu ile OCSVM.

Sınırın dışında bulunan veri noktalarının sayısı *1/νM* ceza faktörü ile hesaplanmaktadır. Küçük bir ceza faktörü seçmek, sınırların dışında daha fazla veri noktasının bulunmasına neden olacaktır. Şekil 4.5, OCSVM'de radyal temelli fonksiyon (RBF) çekirdeğine sahip orijinal veri örneklerinin doğrusal olmayan bir haritasını göstermektedir [8]. Veri örnekleri, tüm pozitif örnekler içeren en küçük hiper küreyi bulana kadar maksimum kenar boşluğu çekirdeği işlevi ile başlangıç noktasından ayrılmaktadır. Problem tanımımız için OCSVM optimizasyon modeli aşağıdaki gibi formüle edilmiştir:

$$\min_{w,\xi_m,\rho} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu M}\sum_{m=1}^{M}\xi_m - \rho$$

$$\text{subject to} \quad \mathbf{w}^T\phi(\mathbf{z}_m) \geq \rho - \xi_m\,,\ \xi_m \geq 0,\ m = 1,2,\dots,M \tag{4.24}$$

burada ϕ (.) eğitim setindeki özellik vektörlerinin doğrusal olmayan eşleme işlevidir (çekirdek), **w** modeldeki ağırlık vektörünü temsil eder, ξ_m ağırlıkların düzenlenmesi için gevşek değişkentir, ρ sınırdan maksimum sapmayı temsil eder ve $\nu \in (0,1]$, destek vektörleri için alt limitleri belirlerken anormal değerler için bir üst limit belirlememizi sağlamaktadır. $\mathbf{z}_m$ Denklem (4.6) veri örneklerinin mesafesi ve merkezi optimizasyon modelinde ξ_m tarafından kontrol edilmektedir. Modelin test aşamasında, örnek vektör $\mathbf{z}_j$'nin, verilen bir karar fonksiyonu ile hiper düzlemin dışına çıkıp çıkmadığını belirlemeye çalışılır.

$$f(\mathbf{z}_j) = sgn\left(\sum_{m=1}^{M}\mu_m K(\mathbf{z}_m, \mathbf{z}_j) - \rho\right), \tag{4.25}$$

$$0 < \mu_m < \frac{1}{\nu M}$$

burada μ Lagrange çarpanıdır ve $K(\mathbf{z}_m, \mathbf{z}_j)$ çekirdek fonksiyonu RBF'dir,

$$K(\mathbf{z}_m, \mathbf{z}_j) = e^{-\gamma \|\mathbf{z}_m - \mathbf{z}_j\|^2}, \tag{4.26}$$

||.|| öklid norm operatörüdür ve γ çan şeklindeki eğrinin uzunluğunu ayarlamak için çekirdeğin "yayılma" parametresidir.

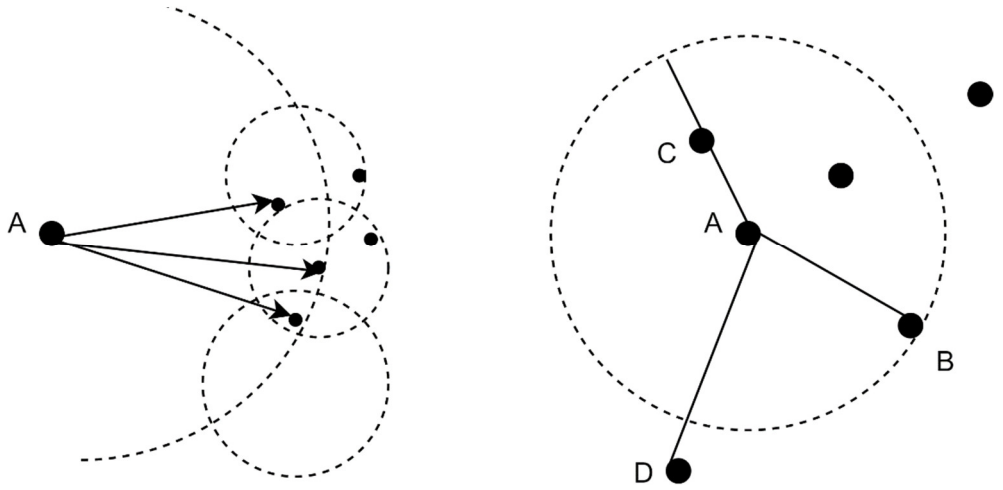
4.6.2. Yerel aykırılık faktörü (LOF)

Lokal aykırılık faktörü (local outlier factor LOF) yöntemi, belirli bir veri örneğinin komşularına göre yerel yoğunluk sapmasını tahmin eder. LOF, yoğunlukları komşularından önemli ölçüde farklı olan numunelerin aykırı olarak, yani LOF ≈1 aykırı değil, LOF ≫1 aykırı değer olarak işaretlendiği, puana dayalı bir yöntemdir. Bir örneğin yerel yoğunluğunu komşularının yerel yoğunlukları ile karşılaştırarak, benzer yoğunluktaki bölgeleri ve komşularından önemli ölçüde daha düşük yoğunluğa sahip noktaları belirler ve bunları aykırı değer olarak etiketlendirmektedir. Şekil 4.6'da yerel aykırılık tespitinin bir örneği verilmektedir, *A* noktasının *k* inci en yakın komşu ya mesafesi ve bu komşunun bir küme şeklinde kendine yakın noktaları içermesi $N_k(A)$ en yakın komşuluk olarak adlandırılmaktadır. Bu komşuluğa olan mesafe erişebilirlik mesafesi (reachability distance) olarak adlandırılır, erişebilirlik mesafesi$_k(A,B)$ = *max{k*-mesafe *(B), d(A,B)}*. Yerel erişebilirlik mesafesi ise (local reachability distance) şu şekilde tanımlanmaktadır;

$$\text{yerel erişebilirlik mesafesi}_k(A) = \frac{1}{\frac{\sum_{B \in N_k(A)} \text{erişebilirlik mesafesi}_k(A,B)}{|N_k(A)|}} \tag{4.27}$$

$$\text{yerel yoğunluk } (v_k) = \frac{\sum_{B \in N_k(A)} \text{yerel erişebilirlik mesafesi}_k(B)}{|N_k(A)| * \text{yerel erişebilirlik mesafesi}_k(A)} \tag{4.28}$$

Bir sonraki adımda karar fonksiyonu ilgili test verilerini, yerel yoğunluk değeri ve test verisinin en yakın olduğu erişebilirlik mesafesini dikkate alarak, bu verinin yeterince eğitim verilerine yakın olup olmadığını analiz ederek etiketine karar vermektedir. Şekil 4.6. da gösterilen temsile bakarak, test verisinin (A) en yakın yoğunluk merkezlerine ve genel olarak pozisyonuna bakarak hangi kümeye dahil olduğuna karar vermek olduğunu söylenebilir. İncelenen problem bazında etiketine karar vermek verinin güvenli ya da atak olarak (+1,-1) işaretlenmesi anlamına gelmektedir.



Şekil 4.6. Bir noktanın yerel yoğunluğunun komşularının yoğunluğu ile karşılaştırılması.

Yerel yoğunluk hücreleri, eğitim verisinin komşu örneklerine ulaşılarak oluşturulmaktadır, LOF'ta karar fonksiyonu [69],

$$g_{LOF}(\mathbf{z}_j) = \frac{k/M}{v_k \|\mathbf{z}_m - \mathbf{z}_j\|} \tag{4.29}$$

burada v_k, Denklem (4.28)'de verilen yerel yoğunluk hücrelerinin ortalamasıdır, $\mathbf{z}_m \in \mathbf{Z}$, $\mathbf{z}_j \in \mathbf{Z}$ Denklem (4.6)'daki eğitim verisi vektörleridir. LOF, bir noktanın komşularını yoğunluğunu bulmak için arar ve sonra diğer noktaların yoğunluğu ile karşılaştırmaktadır. Küçük bir *k* yakındaki noktalara bakarak daha yerel bir odağa sahipken, büyük bir *k* yerel aykırı değerleri tespit edememek anlamına gelmektedir. Verilerde çok fazla gürültü varsa, küçük *k* ile yanlış sonuçlar elde edilebilmektedir.

4.6.3. İzolasyon ormanı (iForest)

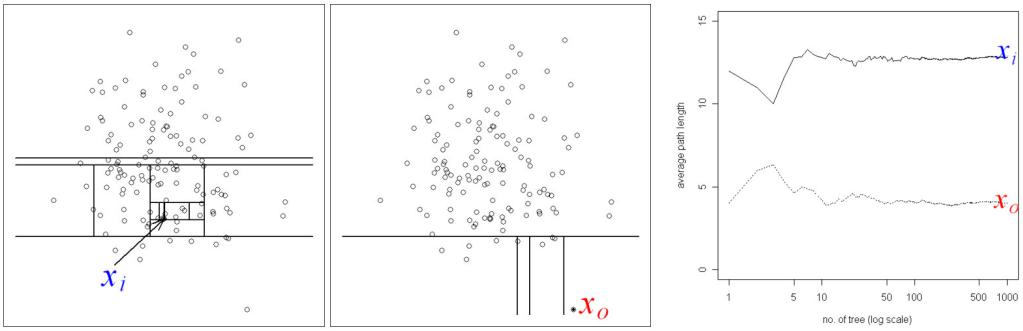
İzolasyon Ormanı (*i*Forest) adı verilen yöntem, belirli bir veri kümesi için bir ağaç (iTrees) topluluğu oluşturur; aykırılıklar iTrees üzerinde ortalama yol uzunlukları kısa olan örneklerdir. *i*Forest algoritması ikili ağaç yapıları oluşturarak rastgele tek tek örnekleri özyinelemeli olarak yalıtır. *i*Forest'in algılama doğruluğunun çok az sayıda ağaçla hızla gerçekleştirdiği çalışamalarda gösterilmektedir, yüksek verimlilikle yüksek algılama doğruluğu elde etmek için yalnızca küçük bir alt örnekleme boyutu gerektirir; ve farklı yükseklik sınırları farklı yoğunluktaki aykırı kümelerini karşılamak için kullanılabilmektedir [71,72]. Temel özelliklerinin yanı sıra derinlik, bağlantı ve model temelli olarak farklı yaklaşımlarda denenmektedir [73], ancak tüm

yaklaşımlarda izlenen genel bir yaklaşım gözlenmektedir, öyleki çalışmalar şu ortak özelliklere sahiptirler;

- İzolasyon ağaçlarının karakteristiği, mevcut örneklerde mümkün olmayan bir ölçüde alt örneklemeyi kullanmalarını sağlamaktadır.
- *i*Forest, aykırılıkları tespit etmek için mesafe veya yoğunluk ölçümü kullanmaz. Bu, tüm mesafe tabanlı ve yoğunluk tabanlı yöntemlerde kullanılan mesafe hesaplamasında önemli bir maliyeti ortadan kaldırmaktadır.
- *i*Forest, küçük bir sabit ve minimum bellek gereksinimi olan doğrusal bir zaman karmaşıklığına sahiptir; sürekli eğitim süresi ve mekan karmaşıklığı olan bir algoritmadır.
- *i*Forest, çok sayıda alakasız nitelikte çok büyük veri boyutu ve yüksek boyutlu sorunları ele almak için ölçeklendirme kapasitesine sahiptir.

Genel olarak, izolasyona dayalı bir yöntem, bireysel örnekleri ölçer ve aykırılıklar en yüksek duyarlılığa sahip olanlardır. Yalıtım fikrini gerçekleştirmek için, verileri doğal olarak izole eden bir veri yapısına yönelirilmektedir. Örneklerin özyinelemeli olarak bölümlendiği rastgele üretilen ikili ağaçlarda, anormallikler için fark edilir daha kısa yollar üretir, çünkü Şekil 4.7 (a)'da görüldüğü gibi anormalliklerin işgal ettiği bölgelerde, daha az anormallik daha az sayıda bölüme neden olur - bir ağaç yapısında daha kısa yollar ve Şekil 4.7 (b) ayırt edilebilir nitelik değerlerine sahip örneklerin bölümleme işleminin başlarında ayrılması daha olası görülmektedir. Bu nedenle, rastgele ağaçlardan oluşan bir orman toplu olarak bazı belirli noktalar için daha kısa yol uzunlukları ürettiğinde, anomaliler olma olasılığı yüksektir. Şekil 4.7`de ifade edilen anomaliler, izolasyona daha duyarlıdır ve bu nedenle kısa yol uzunluklarına sahiptir. Bir Gauss dağılımı göz önüne alındığında, (a) normal bir x_i noktası on iki rasgele bölümün izole edilmesini gerektirir; (b) bir anomali x_o sadece dört bölümün izole edilmesini gerektirir. (c) ağaç sayısı arttıkça ortalama x_i ve x_o yol uzunlukları yakınsamaktadır [70].

*i*Forest kullanarak aykırılık tespiti iki aşamalı bir işlemdir. İlk (eğitim) aşaması, verilen eğitim setinin alt örneklerini kullanarak izolasyon ağaçları oluşturulur. İkinci (değerlendirme) aşaması, her bir örnek için bir aykırılık skoru elde etmek ve test örneklerini izolasyon ağaçlarından geçirmektir.



(a) (b) (c)

Şekil 4.7. Anomalilerin izolasyonun kısa mesafelerde tespit edilmesi.

Eğitim aşamasında, iTrees, tüm örnekler izole edilene kadar bir alt örnek X' tekrar tekrar yerleştirerek oluşturulur. Eğitim aşamasının ayrıntıları Algoritmalar 1 ve 2'de bulunabilir. Her iTree, X, $X' \subset X$'in yerini almadan rastgele seçilen bir alt örnek X' kullanılarak oluşturulur. Alt örnekleme boyutu (ψ) eğitim veri boyutunu kontrol etmektedir ve ψ istenen bir değere yükseldiğinde, iForest'ın güvenilir bir şekilde algıladığını ve ψ değerini daha da artırmaya gerek olmadığı görülmektedir, çünkü algılama doğruluğunda herhangi bir kazanç olmadan işlem süresini ve bellek boyutunu artırmaktadır. Aykırılıkların "az" ve "farklı" olduğunu varsaydığımızdan, normal noktaların "çok" ve "benzer" olduğu varsayılır. Bu varsayımlar altında, iForest'in ψ aykırılıkları normal noktalardan ayırt etmesi için küçük alt örnekleme boyutu yeterli olmaktadır.

Bu algoritmanın en önemli parametresi alt ağaç sayısıdır (t) eğer alt ağaç sayısı çok yüksekse, model fazla öğrenerek yanlış alarm sayısının artmasına neden olmaktadır. Ancak, bu nedenle alt ağaç sayısı düşürülerek performans arttırımına karar verilirse, alarm sayısında azalma meydana gelecektir bu nedenle performans ve alt ağaç sayısı arasında dengeli bir nokta seçilmelidir. Dolayısıyla, alt ağaç parametresi en yüksek performans yaklaşımlarına göre seçilmektedir. Genellikle t = 100'den küçükken iyi bir şekilde maksimum performansına yakınlaştığı görülmektedir.

Eğitim sürecinin sonunda bir ağaç koleksiyonu elde edilir ve değerlendirmeye hazırdır. Değerlendirme aşamasında, tek bir yol uzunluğu h(x), kök x'den dış bir düğüme olan düğüm sayısının, hesaplanmasıyla türetilmektedir ve rastgele bir alt ağacın ortalama yol uzunluğunu tahmin edebilmektedir. Topluluğun her ağacı için h(x) elde edildiğinde, bir aykırılık skoru hesaplanmaktadır.

Herhangi bir aykırı tespit yöntemi için bir aykırı skoru gereklidir. h(x)'den böyle bir puan elde etmenin zorluğu, iTree'nin mümkün olan maksimum yükseklik ψ sırasına göre büyürken, ortalama yüksekliğin logψ'e göre artmasıdır. Farklı alt örnekleme boyutlarındaki modellerden yol uzunluklarını görselleştirmek veya karşılaştırmak gerektiğinde, h(x)'in yukarıdaki terimlerden herhangi biriyle normalleştirilmesi sınırlandırılmaz veya doğrudan karşılaştırılamaz. Bu nedenle, yukarıda belirtilen amaçlar için normalleştirilmiş bir anomali skoru gerekmektedir. c(ψ) verilen h(x)'in ortalaması olduğu için, h(x)'i normalleştirmek için kullanılmaktadır. Bir x örneğinin anomali puanı s aşağıdaki gibi tanımlanmaktadır [70],

$$f(\mathbf{x}, \psi) = 2^{-\frac{E(h(\mathbf{x}))}{c(\psi)}} \tag{4.30}$$

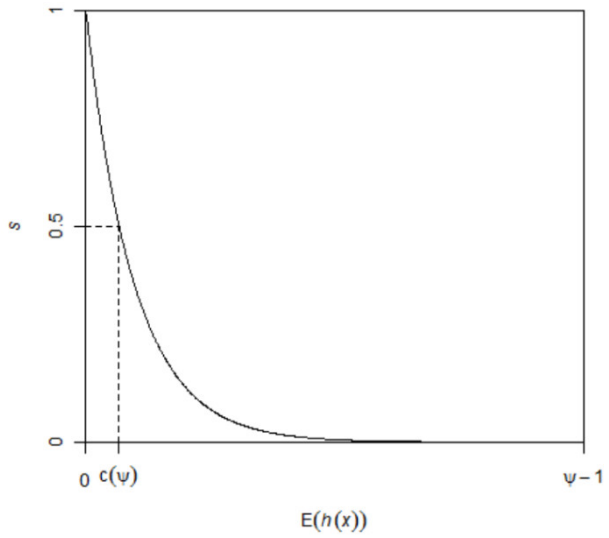
burada E(h(x)), bir iTrees koleksiyonundan alınan h(x) ortalamasıdır. Aşağıdaki koşullar aykırı puanının üç özel değerini göstermektedir,

(a) E(h(x)) → 0, s → 1;

(b) E(h(x)) → ψ − 1, s → 0; ve

(c) E(h(x)) → c(ψ), s → 0.5

Şekil 4.8, E(h(x)) ve s arasındaki ilişkiyi göstermektedir, değer aralığı 0 <s ≤ 1 ve 0 <h (x) ≤ ψ - 1'dir.



Şekil 4.8. Beklenen yol uzunluğu E(h(x)) ile anomali skorlarının ilişkisi.

4.7. Denetimli Öğrenme Teknikleri

4.7.1. Yapay sinir ağları (NN)

Denetimli öğrenme kategorisinde araştırılan sınıflandırma modellerinden ilki sinir ağıdır (perceptron). İnsan beyninin sinirsel yapısını taklit ederek zaman içinde bilgi ve deneyim kazanabilecek bir sınıflandırma ve regresyon modelidir.

Sonuç olarak bir nöral ağ, her bir nöronun ağa beslenen bilginin bir kısmını çıkardığı ve sakladığı birkaç nöron katmanından oluşmaktadır ve farklı türdeki problemlerin çözümünü en etkili şekilde ele almak için de çeşitli nöron türleri ve sinir ağı yapıları geliştirilmiştir.

Çok katmanlı algılayıcıları tanıtmak için, algılayıcı olan temel işlem birimi önce daha ayrıntılı olarak ele alınmalıdır. Şekil 4.9'da görülebileceği gibi, bir algılayıcı, ağırlıklı bir toplam ve doğrusal olmayan bir aktivasyon fonksiyonundan oluşan bir ikili sınıflandırma modelidir. Daha spesifik olarak, algılayıcı veri örneğinden veya başka bir algılayıcıdan çıktıyı $x_{i,l}$ giriş özelliği olarak alıp, bunu ilgili sinapsis ağırlığı $w_{l,j}$ ile yeniden ayarlamaktadır. Ağırlıklı girdiler daha sonra kullanılarak $t_{i,j}$'ye dönüştürülür.

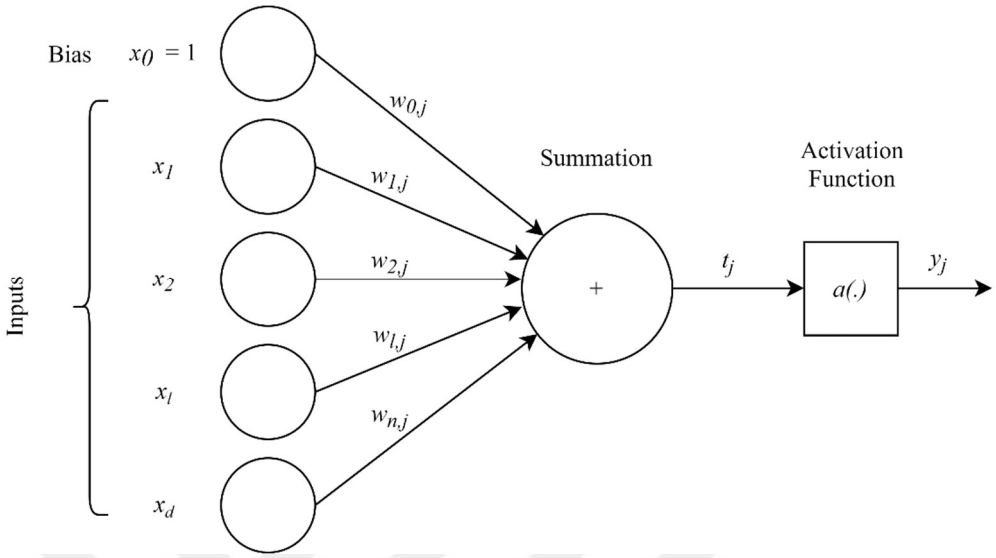
$$\forall_j = 1..c : t_{i,j} = \sum_{l=1}^{d} w_{l,j} x_{i,l} + w_{0,j} \quad (4.7)$$

burada c sınıf sayısı, d girdi sayısı ve $w_{0,j}$ algılayıcının gözlemlenen girdilerdeki örüntüleri öğrenmesine yardımcı olmak için ek bir sapma (bias) terimidir. $y_{i,j}$ çıkışı daha sonra formül 4.8'de gösterildiği gibi $t_{i,j}$ nin seçilen aktivasyon fonksiyonundan beslenmesi ile elde edilmektedir [59].

$$\forall_j = 1..c : y_{i,j} = a(t_{i,j}) \quad (4.8)$$

Yukarıda açıklanan algılayıcının en büyük kusurlarından biri, sadece ikili sınıflandırma problemlerini çözebilmesidir. c sınıflarının genel durumu ($c \geq 2$) dikkate alındığından, c her biri paralel belirli bir C_j sınıfını temsil eden ve karşılık gelen ağırlık vektörü w_j'ye sahip algılayıcılardan (perceptron) oluşmaktadır. Sonuç olarak, verilen bir örnek x_i y_i, $j = max(y_{i,k})$ [48] ise C_j sınıfı seçilerek sınıflandırılmaktadır. Algılayıcının eğitimi için prosedür, ilk olarak aktivasyon fonksiyonunun sonraki

olasılığı $Pr[C_j \mid x_i]$ döndürdüğü varsayıldığından en uygun ağırlıkları bulmak için çapraz entropi hata fonksiyonunu kullanarak tekrar ayarlama yapılmaktadır.



Şekil 4.9. Sinir ağının yapısı [59].

İkincisi, MLP'lerin tüm sinir ağları gibi toplu (mini-batch) öğrenme yaklaşımını kullandığını, yani hem kayıp fonksiyonunun hem de ağırlıkların, tüm veri seti yerine tek tek mini gruplar halinde güncellenmesidir. Sonuç olarak, çapraz entropi hatası:

$$E(w_j|\chi_k) = -\sum_{i=1}^{n_k}\sum_{j=1}^{c} b_{i,j}\log y_{i,j}$$

$$b_{i,j} = \begin{cases} 1 & x_i \in C_j \\ 0 & diğer \end{cases} \tag{4.9}$$

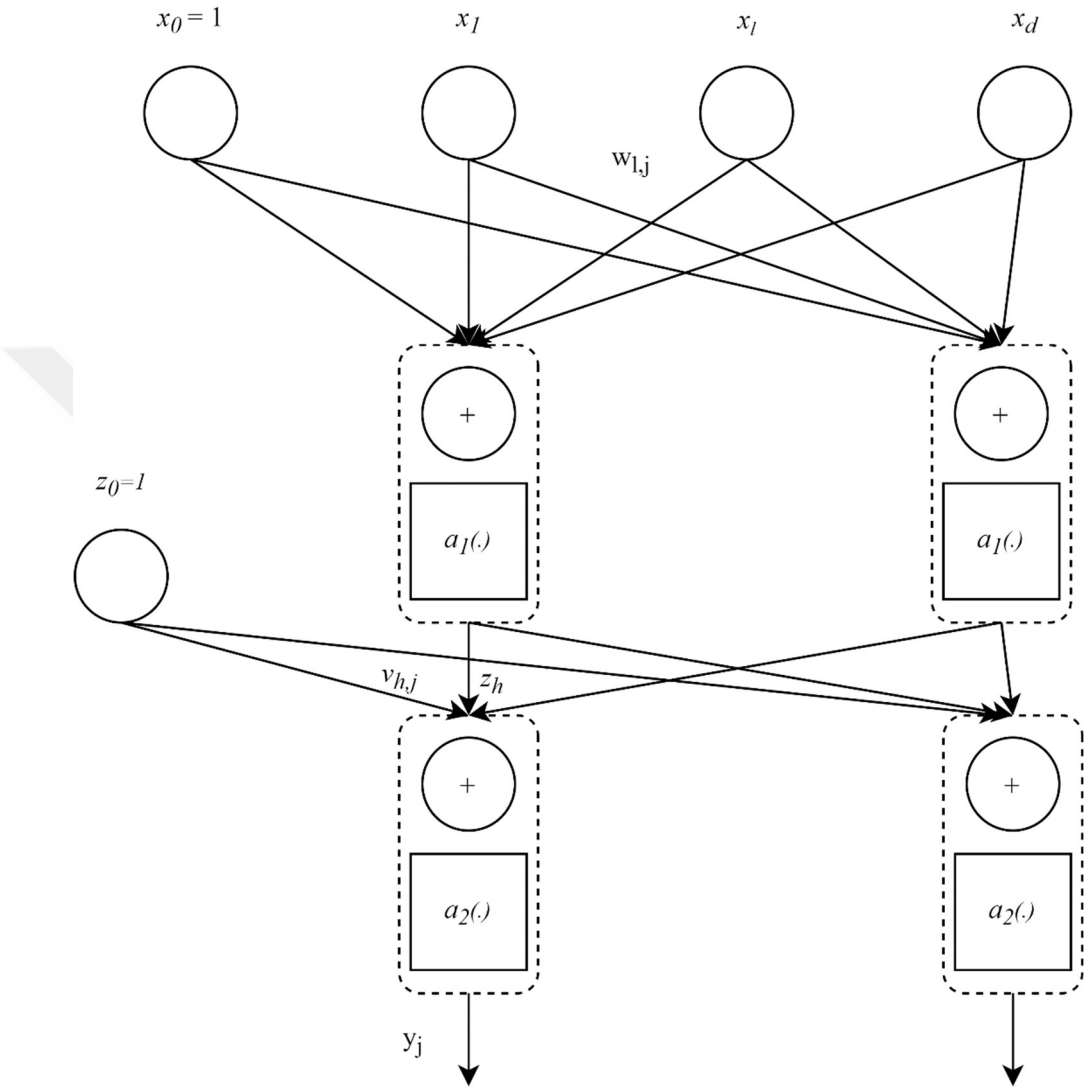
$$y_{i,j} = \Pr[C_j|x_i]$$

burada $x_i \in \chi_k$, n_k, C_j temsil edilen sınıfın veri örneklerinin sayısıdır ve w_j algılayıcının ağırlık vektörüdür.

Bir algılayıcı, yalnızca bir ağırlıklı katmandan oluştuğu için, yalnızca belirli bir girdi ve çıktı arasındaki doğrusal ilişkileri öğrenebilmektedir. Bununla birlikte, sınırlama, birkaç algılayıcının birbirine bağlanmasıyla aşılabilir, bu da giriş ve çıkış katmanı arasında ara veya gizli katmanların oluşturulması anlamına gelmektedir. Dolayısıyla, MLP (Multi Layer Perceptron), şekil 4.5'teki gibi yapılandırılmışsa karar fonksiyonu;

$$y_{i,j} = a_2\left(\sum_{h=1}^{H} v_{h,j} z_{i,h} + v_{0,j}\right) = softmax\left(\sum_{h=1}^{H} v_{h,j} a_1\left(\sum_{l=1}^{d} w_{l,h} x_{i,l} + w_{0,h}\right) + v_{0,j}\right) \tag{4.10}$$

burada H, gizli katman sayısını göstermektedir. Bu formülde, doğrusal olmayan bir aktivasyon fonksiyonuna (a_1) duyulan ihtiyaç netleşmiş ve bu işlev doğrusal veya mevcut değilse, böylece MLP tek bir algılayıcıya dönüştürülebilmektedir.



Şekil 4.10. Birden fazla katman ile tasarlanan yapay sinir ağları.

Şekil 4.10`da gösterilen birden fazla katman içeren sinaps yapısı, $x_l = 1..d$ giriş değerlerini, $z_h, h = 1..H$ gizli birimleri, z_0 gizli katman sapma değerini, y_j çıkış birimlerini, $w_{l,h}$ ilk katmanın ağırlıklarını ve $v_{h,j}$ ise ikinci katmanın ağırlıklarını, $a_1(.)$ gizli katman için aktivasyon fonksiyonu ve $a_2(.)$ Çıkış fonksiyonunu (softmax) temsil etmektedir. Aktivasyon fonksiyonu formül 4.11, a_1 aktivasyon fonksiyonu olarak kullanılır. Ayrıca, ağırlıklı toplamın deterministik karakterini olasılıklı bir tahmine dönüştürmek için softmax fonksiyonu son bir dönüşüm olarak eklenmiştir [48].

$$ReLU(r) = \max(0, r) \tag{4.11}$$

Çok katmanlı algılayıcılara uyum sağlamak için, Şekil 4.5'teki yapı göz önüne alındığında, $v_{h,j}$ algılayıcıdaki ağırlıklarla aynı şekilde hesaplanabilmektedir. Bununla birlikte, ilk katman ağırlıklarını hesaplamak için zincir kuralı kullanılmaktadır:

$$\frac{\partial E}{\partial_{w_{l,h}}} = \sum_{j=1}^{c} \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial z_h} \frac{\partial z_h}{\partial w_{l,h}} \tag{4.12}$$

Sonuç olarak, formül 4.9'un çapraz entropi kaybının kullanıldığı ve a_1 aktivasyon fonksiyonunun ReLU fonksiyonu ve a_2 softmax fonksiyonunun olduğu varsayıldığında, bir veri örneği için $v_{h,j}$ ve $w_{i,h}$ türevleri

$$\frac{\partial E}{\partial v_{h,k}} = -\sum_{i=1}^{n_k} \sum_{j=1}^{c} \frac{b_{i,j}}{y_{i,j}} y_{i,j} \left(\delta_{j,k} - y_{i,k}\right) z_{i,h}$$

$$= -\sum_{i=1}^{n_k} \left(\sum_{j=1}^{c} b_{i,j} \delta_{j,k} - y_{i,k} \sum_{j=1}^{c} b_{i,j}\right) z_{i,h} \tag{4.13}$$

$$= -\sum_{i=1}^{n_k} \left(b_{i,k} - y_{i,k}\right) z_{i,h},$$

ve

$$\frac{\partial E}{\partial y_j} = -\sum_{i=1}^{n_k} \frac{b_{i,j}}{y_{i,j}}$$

$$\frac{\partial y_j}{\partial z_h} = y_{i,j} \left(\delta_{j,h} - y_{i,h}\right) v_{h,j}$$

$$\frac{\partial z_h}{\partial w_{l,h}} = \begin{cases} x_{i,l} & \sum_{l=1}^{d} w_{l,h} x_{i,l} + w_{0,h} > 0 \\ 0 & diğer \end{cases} \tag{4.14}$$

$$\rightarrow \frac{\partial E}{\partial w_{l,h}} = -\sum_{i=1}^{n_k} x_{i,l} * H\left(\sum_{l=1}^{d} w_{l,h} x_{i,l} + w_{0,h}\right) \sum_{j=1}^{c} \left(b_{i,j} - y_{i,j}\right) v_{h,j}$$

burada H birim basamak fonksiyonudur. Son olarak, bu türevler belirlendikten sonra, ağırlıkları etkin bir şekilde güncellenmektedir.

4.7.2. Destek vektör makineleri (SVM)

SVM'ler (Cristianini ve Shawe-Taylor, 2000)[65,66] tarıfından geliştirilmiştir; ancak başlangıçta Vapnik ve iş arkadaşları tarafından tanıtılana (Boser ve diğerleri, 1992;

Vapnik, 1998)[64,67] nispeten yeni bir tür öğrenme algoritmasıdır ve bir dizi başka araştırmacı tarafından art arda genişletilmiştir. Seyrek ve gürültülü verilere göre son derece sağlam performansları, onları metin kategorizasyonundan protein fonksiyon tahminine kadar birçok uygulamada tercih edilen sistem haline getirmektedir.

Sınıflandırma için kullanıldıklarında, belirli bir ikili etiketli eğitim verisi kümesini, birbirlerinden maksimum uzak olan bir hiper-düzlemle ayırmaktadır ("maksimum marj hiper-düzlemi" olarak bilinir). Doğrusal ayrımın mümkün olmadığı durumlarda, özellik alanına otomatik olarak doğrusal olmayan bir eşleme gerçekleştiren "çekirdekler" (Kernel) teknikleri ile birlikte çalışabilmektedirler. SVM tarafından özellik alanında bulunan hiper düzlem, girdi alanındaki doğrusal olmayan karar sınırına karşılık gelir ve rastgele bir vektör düşündüğümüzde $\mathbf{x}^j$ j indexi için $\mathbf{x}^j = (x_1^j, \dots, x_2^j)$ giriş noktaları $Y^j \in \{-1, +1\}$ şeklinde etiketlenmektedir. $\Phi : I \subseteq \mathbb{R}^n \rightarrow F \subseteq \mathbb{R}^N$ şeklinde giriş uzayı $I \subseteq \mathbb{R}^n$ özellik uzayına $F \subseteq \mathbb{R}^N$ doğru haritalanmaktadır. Eğer S örneği için m adet etiketli noktamızın bulunduğunu varsayarsak $S = \{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^m, y^m)\}$. SVM öğrenme algoritması hiper düzelem ($\mathbf{w}$,b);

$$\gamma = \min_i y^i \{\langle \mathbf{w}, \Phi(\mathbf{x}^i) - b \rangle\} \tag{4.15}$$

şekilde maksimize edilir, burada $\mathbf{w}$ vektörü F ile aynı boyutlara sahiptir, $||\mathbf{w}||^2$ sabit tutulur, b gerçek bir sayıdır ve γ kenar boşluğu olarak adlandırılmaktadır. (W,φ(x_i) - b) değeri, x_i noktası ile karar sınırı arasındaki mesafeye karşılık gelir ve Y_i etiketi ile çarpıldığında, tüm doğru sınıflandırmalar için pozitif, yanlış olanlar için negatif bir değer üretmektedir. Veriler doğrusal olarak ayrılabilirse ve kenar boşluğu olarak adlandırılırsa, bu verilerin tüm veriler üzerindeki minimum değeri pozitiftir. Sınıflandırılacak yeni bir veri noktası $\mathbf{x}$ verildiğinde, bir etiket karar sınırıyla olan ilişkisine göre atanır ve ilgili karar işlevi,

$$f(x) = sign(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle - b), \tag{4.16}$$

maksimal marjı hiperdüzlemi çözmek için,

$$w = \sum_{i=1}^{m} \alpha_i y^i \Phi(\mathbf{x}^i), \tag{4.17}$$

burada α_i maksimum yapılmak istenen pozitif real sayıdır,

$$\sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j y^i y^j \langle \Phi(\mathbf{x}^i), \Phi(\mathbf{x}^j) \rangle, \tag{4.18}$$

ve bağlı fonksiyon,

$$\sum_{i=1}^{m} \alpha_i y^i = 0, \alpha_i > 0, \tag{4.19}$$

bu denklemden, eğitim noktası $\mathbf{x}_i$ ile ilişkili α_i'nin, nihai karar fonksiyonuna eklediği gücü ifade ettiğini görmek mümkündür. Bu alternatif sunumun dikkat çekici bir özelliği, genellikle sadece noktaların bir alt kümesinin sıfır olmayan α_i ile ilişkilendirilmesidir. Bu noktalara destek vektörleri denir ve ayırıcı hiper-düzleme en yakın olan noktalardır. α vektörünün seyrekliğinin çeşitli hesaplama ve öğrenmeye ilişkin sonuçları vardır.

Bir test noktası (**x**, y) için $y(\sum_{i=1}^{m} \alpha_i y_i \langle \Phi(x^i), \Phi(x) \rangle - b)$, makinenin tahmini yanlışsa miktarının negatif olduğunu ve büyük bir negatif değerin (**x**, y) noktası için algoritma tarafından eğitim verilerinden 'farklı' olarak kabul edilmekteidr. $K_{i,j} = \langle \Phi(x^i), \Phi(x) \rangle$ matrisine çekirdek (kernel) matrisi denir ve daha sonra tartışılacak olan algoritmanın uzantılarında önemli katkıları bulunmaktadır. Verilerin doğrusal olarak ayrılamaması durumunda, lineer olmayan karar sınırları sağlayan daha genel fonksiyonlar, $K_{i,j} = K(\mathbf{x}^i, \mathbf{x}^j)$ kullanılabilmektedir. İki klasik seçenek, polinom çekirdekleri $K(\mathbf{x}^i, \mathbf{x}^j) = (\langle \mathbf{x}^i, \mathbf{x}^j \rangle + 1)^d$ ve Gauss çekirdekleri $K(\mathbf{x}^i, \mathbf{x}^j) = e^{-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|}{\sigma^2}}$ dir; burada d ve σ, çekirdek parametreleridir.

Gürültü varlığında, yukarıda açıklanan standart maksimum kenar boşluğu algoritması aşırı öğrenme problemine neden olabilmekte ve daha karmaşık teknikler gerekebilmektedir. Bu sorun, maksimum kenar boşluğu algoritmasının (maximum margin algorithm) her zaman mükemmel tutarlı bir hipotez bulması, eğitim hatasına karşı esnek olmamasının bir sonucudur. Bununla birlikte, bazen, daha iyi bir tahmin gücü için bazı eğitim doğruluğundan fedakarlık yapmak gerekebilir. Eğitim hatasını tolere etme ihtiyacı, yumuşak marj ve boşluk dağılımı (margin distribution) sınıflandırıcılarının gelişmesine yol açmıştır (Cortes ve Vapnik, 1995). Bu

tekniklerden biri (Shawe-Taylor ve Cristianini, 1999) eğitim aşamasındaki çekirdek matrisini aşağıdaki gibi değiştirmektedir:

$$K \leftarrow K + \lambda 1, \tag{4.20}$$

karar aşamasında hala standart çekirdek fonksiyonunu kullanırken (4.19), λ'ya köşegen faktör denmektedir. λ ayarlanarak, eğitim hatası kontrol edilebilir ve görünmeyen noktaları yanlış sınıflandırma riskinin uygun bir λ seçimi ile azaltılabileceğini kanıtlamak mümkündür [65]. Genel eğitim hatasını kontrol etmek yerine yanlış pozitifler ve yanlış negatifler arasındaki dengeyi kontrol etmek istersek, K'yi aşağıdaki gibi değiştirmek mümkündür:

$$K \leftarrow K + \lambda D, \tag{4.21}$$

burada D, pozitif ve negatif örneklere karşılık gelen konumlarda girişleri d^+ veya d^- olan diyagonal bir matristir. Bu tekniğin α_i'nin büyüklüğünü, sınıfın büyüklüğüne bağlı olacak şekilde kontrol etmeye eşdeğer olduğunu kanıtlamak mümkündür ve daha küçük d'ye sahip sınıfta daha büyük α_i için bir sapma sağlamaktadır. Bu da asimetrik bir marja karşılık gelir; daha küçük olan d sınıfı karar sınırından daha uzakta tutulacaktır. Dengesiz veri kümeleri durumunda, $d^+ = \frac{1}{n^+}$ ve $d^- = \frac{1}{n^-}$, iki sınıfın ilgili önemlerini temel alarak, göreli önemini otomatik olarak ayarlamak için sezgisel bir yol sağlamaktadır.

4.7.3. Rastgele orman (Random Forest)

İncelenmekte olan üçüncü sınıflandırma modeli rastgele ormandır (Random Forest). Bu, sadece bir ağaç eğitimi ile karşılaştırıldığında tespit doğruluğunu önemli ölçüde artırmak amacıyla birkaç adet budanmamış karar ağacını birleştiren bir topluluk tekniğidir.

Şekil 4.6'da görülebileceği gibi, bir sınıflandırma ağacı, belirli bir girdi alanını, yerel bölgelere, her yerel bölgeyi ise özyinelemeli bölünmelerle ve girdiyi kararlar dizisi şekilde bölen deterministik bir hiyerarşik sınıflandırma modelidir. Bu, her biri belirli bir yerel bölgeyi temsil eden iki veya daha fazla alt bölgeye ve yaprak düğümlere ayırmak için her biri bir karar fonksiyonu $f_m(x)$ kullanan düğümlerden oluşan bir karar ağacı ile gerçekleştirilebilir. Bununla birlikte, rastgele bir orman karar ağacının ideal

yapısını bulmak için, NP-sert optimizasyon probleminin çözülmesi gerekir, böylece gerçekte ikili rastgele orman ağaçları oluşturmak için CART algoritmasına dayanan açgözlü bir yukarıdan aşağıya yönlü algoritma kullanılmaktadır [48, 62, 63]
Prosedür dört adımdan oluşmaktadır [62]:
1. Gini indeksi (4.22) sıfırsa, yalnızca bir sınıf içerdiğinden bir yaprak düğümü oluşturulmaktadır.
2. Aksi takdirde, veri kümesinden rastgele küçük bir özellik alt kümesi seçilir. Her özellik için en iyi bölünme, bölünmeden sonraki toplam Gini indeksinin en aza indirilmesiyle hesaplanmaktadır (4.23).
3. Düğümlerin ayrımı, Gini indeksini en aza indiren 2. adımın en iyi ayrımı seçilerek belirlenmektedir.
4. Bir iç karar düğümü oluşturulmakta ve tüm yaprak düğümleri oluşturulana kadar algoritmayı yinelemeli olarak tekrarlamaktadır,

$$Gini\left(veri_{\text{düğüm}}\right) = \sum_{i=1}^{c}\sum_{j=1,i\neq j}^{c} \Pr\left[C_i|veri_{\text{düğüm}}\right]\Pr\left[C_j|veri_{\text{düğüm}}\right]$$

$$= \frac{1}{2}\left(1 - \sum_{i=1}^{c}(\Pr\left[C_i|veri_{\text{düğüm}}\right])^2\right) \quad (4.22)$$

Denklem (4.22) bir düğümün safsızlığını belirlemek için Gini indeks formülü [48, 62].

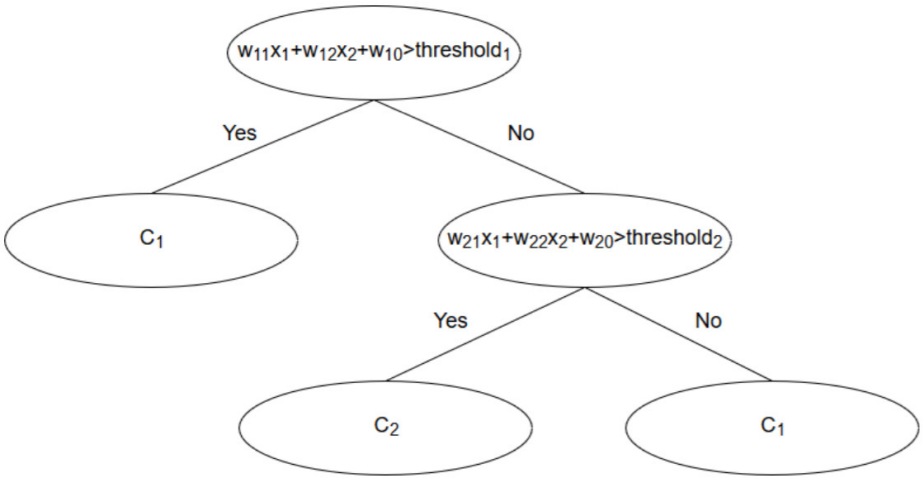
$$GiniBölünmesi(veri, x_i) = \min_{\text{böl}}(uzunluk\left(veri_{\text{böl,dol}}\right) * Gini\left(veri_{\text{böl,sol}}\right) +$$

$$uzunluk\left(veri_{\text{böl,sağ}}\right) * Gini(veri_{\text{böl,sağ}})) \quad (4.23)$$

Denkelm 4.23'de Gini bölünmesi adımı ile x_i özelliğinin bölünerek minimum değer tespit edilir ve alt dallanmalara karar verilmektedir [48].

Rasgele orman tekniğinde, doğru karar ağaçları ve minimum karşılıklı korelasyon elde etmek için üç tip rasgelelik kullanılmaktadır. Her şeyden önce, her ağaç maksimum boyutuna kadar büyütülür ve asla budanmaz, böylece her biri kullanılan veri kümesine uyum sağlamaktadır. İkinci olarak, iç düğümün veri kümesinden rastgele az sayıda özellik seçer ve daha sonra en iyi karar fonksiyonunu belirlemek için kullanılmaktadır. Tüm özelliklerin üzerinde tekrarlanmamasıyla, tekrarlamalı prosedür büyük ölçüde hızlandırılır ve farklı ağaçlar arasındaki korelasyon en aza indirilir. Son olarak, örnekleme işlemi, orijinal veri kümesinden örnekler ile değiştirilen örnekler seçilerek her bir ağaca ait sıralı veri kümesi oluşturmak için kullanılmaktadır. Sonuç olarak, her

ağaç ortalama olarak orijinal veri kümesinin %63.2'si üzerinde eğitilir ve sınıflandırma ağaçları arasındaki korelasyonu etkili bir şekilde azaltılmaktadır [48].



Şekil 4.11. İki özellikli giriş ile örnek karar ağacı [48].

4.8. İstatistiksel Öğrenme Teknikniker

4.8.1. En küçük kareler yaklaşımı (LSA)

En küçük kareler yaklaşımı (LSA) aykırılık tespitini, istatistik temelli, parametrik olmayan, ancak kare kaybı objektif fonksiyonuna dayanarak incelenmektedir. Yöntem, parametrik olmayan en küçük kareler sınıflamasında (least square classification) son çalışmaların genişletilmesinden, aykırı olmayan eğitim verileri açısından aykırılıkları modelleyen "yukarıdakilerin hiçbiri" sınıfını içermek üzere ortaya çıkmıştır. Aynı zamanda birden çok iç sınıf ve anomali arasında ayrım yapmak için de kullanılabilmektedir [74]. Etiketli eğitim verileri $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ $\mathbf{x}_i \in \mathbb{R}^d$, veri alanındaki giriş noktasıdır, burada d, giriş verilerinin boyutudur, N ise bağımsız ve aynı dağılıma sahip (i.i.d.) giriş verilerinin sayısını temsil etmektedir ve buna karşılık gelen sınıf etiketi $y_i \in Y$ veya olası sınıflar kümesidir $Y = \{1, \dots, c\}$. Verinin sınıf koşullu olasılığı $p(y|\mathbf{x})$ tahmin edilebilmekte ancak $p(y = i|\mathbf{x})'$i bütün sınıflar için $i \in Y$ tahmin etmek istediğimizde $q(y = i|\mathbf{x}, \theta_i)$ fonksiyonu kullanılmaktadır;

$$q(y = i|\mathbf{x}, \theta_i) = \theta_i^T \varphi(\mathbf{x}), \tag{4.24}$$

$$\theta^i = \left(\theta_{i,1}, \dots, \theta_{i,B}\right) B \in \mathbb{R}^B,$$

burada B parametre sayısını temsil etmektedir ve θ_i ise **x** örneğinin i sınıfına ait olma olasılığını ifade etmektedir. Bu noktaya kadar temel olarak en küçük karaler yönteminin fonksiyonunu anlatmaktadır ancak daha az maliyetli olan sınıf tabanlı öğrenme problemine dönüştürdüğümüzde, *B* boyutlu parametre vektörü *K* çekirdek fonksiyonu yardımı ile gerçek bir sınıf tabanlı olasılık hesabı yapılabilmektedir. Parametrelerin hesaplanmasında doğruluk, *J* karesel hata değerinin minimize edilmesi ile elde edilmektedir.

$$\varphi(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_1), \dots, K(\mathbf{x}, \mathbf{x}_B))^T \;\; B \in \mathrm{R}^B \tag{4.25}$$

$$J_i(\theta_i) = \frac{1}{2}\int (q(y = i|\mathbf{x}, \theta_i) - p(y = i|\mathbf{x}))^2 p(x) dx$$

En küçük kareler yaklaşımı tutarlı bir tahmin edicidir ve pratikte hesaplanması çok hızlıdır ve tek bir adımda yinelemeli parametre aramasına gerek kalmadan küresel optimumu bulmaktadır [74].

4.8.2. Olasılıksal temel bileşen analizi (PPCA)

Doğrusal boyutsallık azaltma yöntemleri, geometrik yorumlama ve hesaplama özelliklerinden dolayı yüksek boyutlu veri analizinin temel dayanağı olmuştur. Olasılıksal PCA (PPCA), Gauss latent değişken modeline dayanan PCA'in istatistiksel bir formülasyonudur ve ilk olarak 1999'da Tipping ve Bishop tarafından tanıtılmıştır [75]. PPCA modeli, p boyutlu bir gözlenen veri noktasını, $K \ll M$ olan doğrusal bir dönüşüm fonksiyonu vasıtasıyla karşılık gelen q boyutlu bir gizli değişkenle ilişkilendirerek yüksek boyutlu verilerin boyutunu azalmaktadır. PPCA'yı destekleyen istatistiksel model göz önüne alındığında, modelin genişletilmesi mümkündür ve çok sayıda istatistiksel araçla kullanılabilmektedir. PPCA gizli bir değişken modeldir ve her örnek x için üretici süreci şu şekilde tanımlanır (Şekil 4.12),

$$\mathbf{x} = \mathbf{W}_x\mathbf{z} + \boldsymbol{\mu}_x + \boldsymbol{\epsilon}_x \tag{4.26}$$

burada $\mathbf{z} \in \mathbb{R}^K$ gizli değişkenler olarak adlandırılır ve $\mathbf{W}_x$, faktör yükleri adı verilen $M \times K$ boyutlu matrisdir. İstatistiksel modelde, **z** gizli değişkenleri geleneksel olarak sıfır ortalama ve birim varyansla Gauss dağılımı olarak kabul edilir, yani $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$

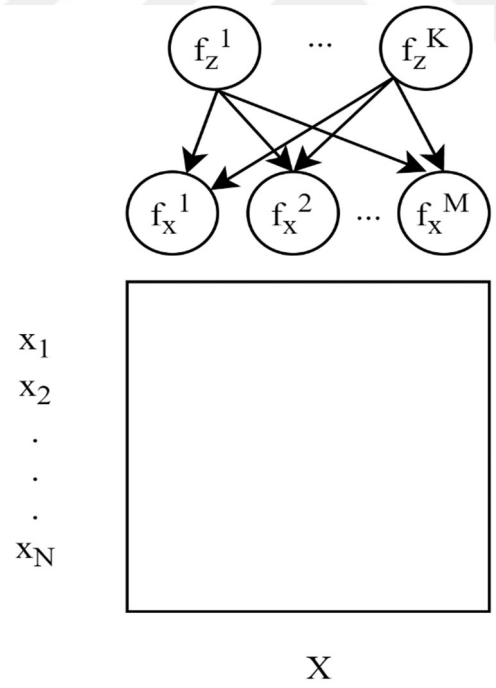
ve $\boldsymbol{\epsilon}_x$ aynı zamanda bir izotropik Gauss formu alan bir gürültü σ_x^2 seviyesi ile tanımlanmaktadır $\boldsymbol{\epsilon}_x \sim \mathrm{N}(0, \sigma_x^2\mathbf{I})$.

Ek olarak, veriler içerisindeki sıfır olmayan ortalamaya sahip örneklere izin veren $\mu_x \in \mathbb{R}^M$ parametresi bulunmaktadır.

$\mathbf{S}_x = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^{\mathrm{T}}$ veriler için örnek kovaryans matrisi $\{\mathbf{x}_n\}_{n=1}^{N}$, $\lambda_1 \geq \cdots \geq \lambda_M$ özvektörleri ile $\mathbf{u}_1, \ldots, \mathbf{u}_m$ özdeğerleri olsun, PPCA modeli K boyutlu ise, $\mathbf{W}_x$'in maksimum olabilirlik tahmini şu şekilde verilmektedir,

$$\mathbf{W}_x = \mathbf{U}_K(\boldsymbol{\Lambda}_K - \sigma_x^2\mathbf{I})^{\frac{1}{2}}\mathbf{R}, \tag{4.27}$$

burada $\boldsymbol{\Lambda}_K = diag(\lambda_1 \geq \cdots \geq \lambda_K)$, $\mathbf{U}_K = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$ ve **R,** keyfi bir K × K dikey matrisidir. Yeni **x*** girişi için ortalama projeksiyonlar **z***, aşağıdaki gibi tanımlanmaktadır,



Şekil 4.12. PPCA için X üretici değişkeninin özelliklerinin bir birleri arasında ilişkinin hesaplanması.

$$\boldsymbol{z}^* = \mathbf{R}^{\mathrm{T}}(\boldsymbol{\Lambda}_K - \sigma_x^2\mathbf{I})^{\frac{1}{2}}\boldsymbol{\Lambda}_K^{-1}\mathbf{U}_K^{\mathrm{T}}(\mathbf{x}^* - \boldsymbol{\mu}_x) \tag{4.28}$$

5. DENEYLER VE BAŞARIM ANALİZİ

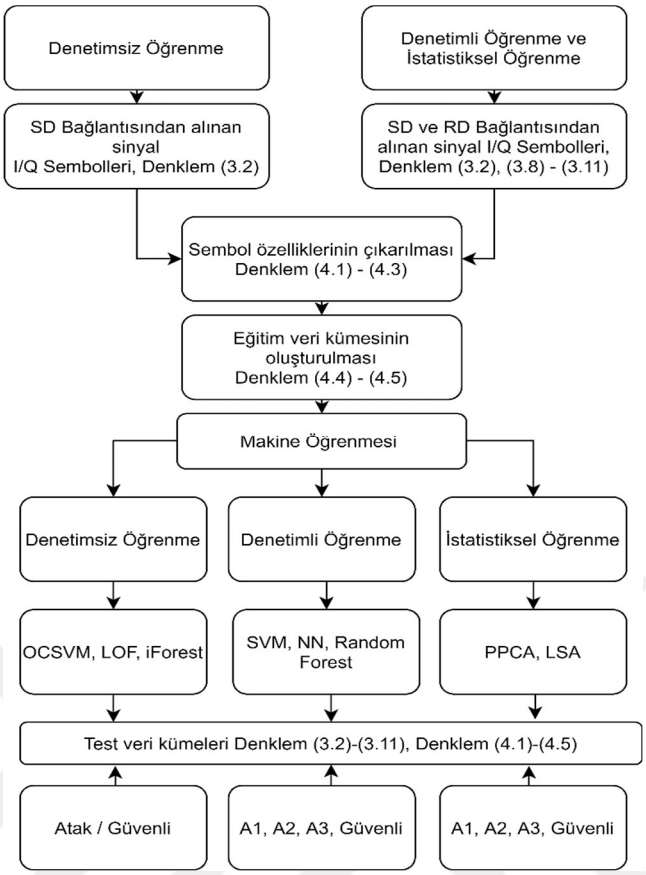
Bu bölüm, önceki bölümlerde açıklanan atak tespit yöntemlerinin performans sonuçlarını içermektedir. Deneylerde, ilk önce, denetimli, denetimsiz ve istatistiksel yöntemler için optimal parametreler bulunmaktadır. Ardından, en iyi performans gösteren modeller birbirleriyle farklı koşullarda (SNR, bant genişliği, modülasyon, vb.) gösterdikleri performanslar ile karşılaştırılmaktadır. Son olarak, sonuçlar zaman karmaşıklığı, veri boyutu, bellek kullanımı gibi çalışma zamanı karmaşıklığı perspektifinden incelenmektedir.

5.1. Uygulamalar ve Kurulum

Sinyal modeli, atak modelleri, sembollerden özelliklerin çıkarılması, test ve eğitim veri kümelerinin oluşturulmasında Matlab [76] uyugulaması kullanılmaktadır. Üretilen veri kümelerinin, öğrenme yöntemlerine girdi olarak sunulması ve test edilmesi ise “Scikit Learn” Python makine öğrenim kütüphanesi kullanılarak simüle edilmektedir [77]. Bu tezde, atak tespitine uygulanan makine öğrenimesi algoritmalarının uygulanması ve sistem modelinin oluşturulması, Intel (R) Core™ i7-7500U, CPU 2.90 GHz ve 8GB RAM donanımlı bilgisayarda gerçekleştirilmiştir.

5.2. Sistem Çalışma Modeli ve Veri Kümeleri

Denklem (3.1) ile (3.7)’de sinyal modeli oluşturduktan sonra, Denklem (4.1) ile (4.8)'de tanımlandığı gibi ardışık sembollerden özellikler çıkartılarak test ve eğitim veri kümeleri oluşturulmaktadır. Algoritmaların eğitim aşaması sırasında, LTE-A sistem modelinde 5 ms süreyle (yarım çerçeve süresi) iletilen sinyaller ile veri kümesi oluşturulmaktadır. Denklem (3.9) ile (3.11) de modellenen atak türleri, veri karıştırma (A1), farklı veri yollama (A2) ve veri enjeksiyon (A3) atak modelleri Bölüm 3'de açıklandığı gibi oluşturulmaktadır. Bölüm 4.2.’de, elde edilen semboller üzerinden çıkartılan özellikler tanımlanmıştır. Bu aşamada özellik vektörleri oluşturularak algoritmalara girdi olarak sunulmak üzere eğitim ve test veri kümeleri oluşturulmaktadır.



Şekil 5.1. Sistem simülasyonunun tüm algoritmalar için eğitim ve test aşamasında izlediği adımlar akış şemesında tanımlanmıştır.

Şekil 5.1'deki şemada belirtilen akış şemasında, sistem modeli ile sembollerin nasıl elde edildiği, veri kümelerinin oluşturulması ve oluşturulduktan sonra algoritmaların çalıştırılması adımları tanımlanmaktadır. Tüm bu süreçlerde verilerin her aşamada nasıl dönüştürüldüğünü açıklamak adına aşağıda örnek verilerle her adım detaylandırılmıştır,

1- Röleden ve kaynaktan alınan sinyallerin sembolleri,

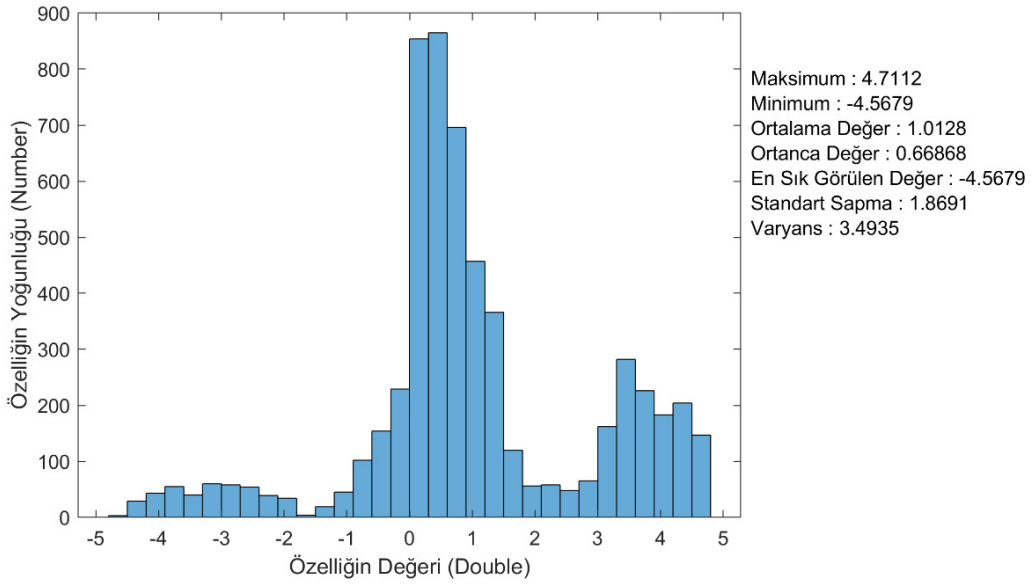
Sembol 1	Sembol 2	Sembol 3	Sembol 4	.....
+0.8594,-0.6795i	-0.3587,-0.3971i	-0.8960,+0.2202i	0.1750, +0.2047i	.....
+1.019,-0.1481i	-0.1118, - 0.5734i	-0.8159 - 0.3318i	0.0628+ 0.2336i	.....

2- Her sembolden özelliklerin çıkartılması,

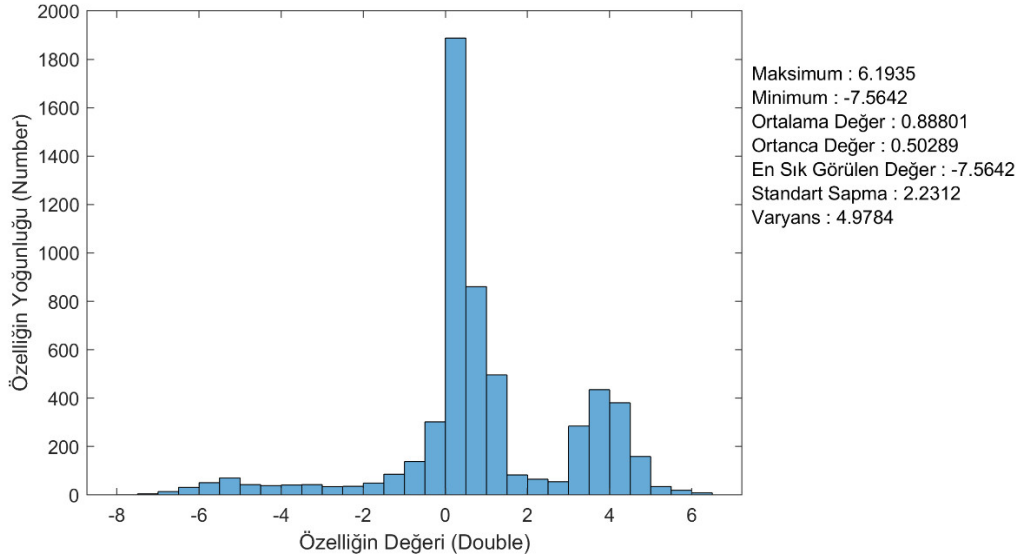
Sembol 1 (f1)	Sembol 1 (f2)	Sembol 1 (f3)	Sembol 2 (f1)	Sembol 2 (f2)	Sembol 2 (f3)	Sembol 3 (f1)	Sembol 3 (f2)	Sembol 3 (f3)	.....
1.4392	-5.1985	1.0890	0.2989	4.1384	3.1460	0.9581	-5.1332	0.5783	.....
0.4672	0.1483	4.5936	0.0893	-2.6845	4.4448	0.2413	2.5710	3.9877	.....

Tanımlanan adımlar kullanıcı ekipmanında alınan kaynak, güvenli röle ve atak röle (A1, A2, A3) sinyalleri için benzer şekilde gerçekleştirilmektedir. Her veri kümesinin

ayrı ayrı istatistiksel özellikleri ise, Şekil 5.2 de kaynakdan alınan sinyallerin özellik vektörleri elde edildikten sonraki istatistiksel özelliklerini göstermektedir. Şekil 5.3 ise güvenli röle den alınan sinyalin özellik vektörü elde edildikten sonraki istatistiksel özellikleridir. Bu iki sinyal karşılaştırıldığında kanal ve gürültü etkisinin röle sinyal üzerindeki etkileri görülebilmektedir. Bu farklılıkalar problemimizin zorluklarında birini temsil etmektedir ve yanlış alarm miktarının artmasına neden olmaktadır, bu faktörlerin etkileri başarım analizi bölümünde detaylandırılmıştır.

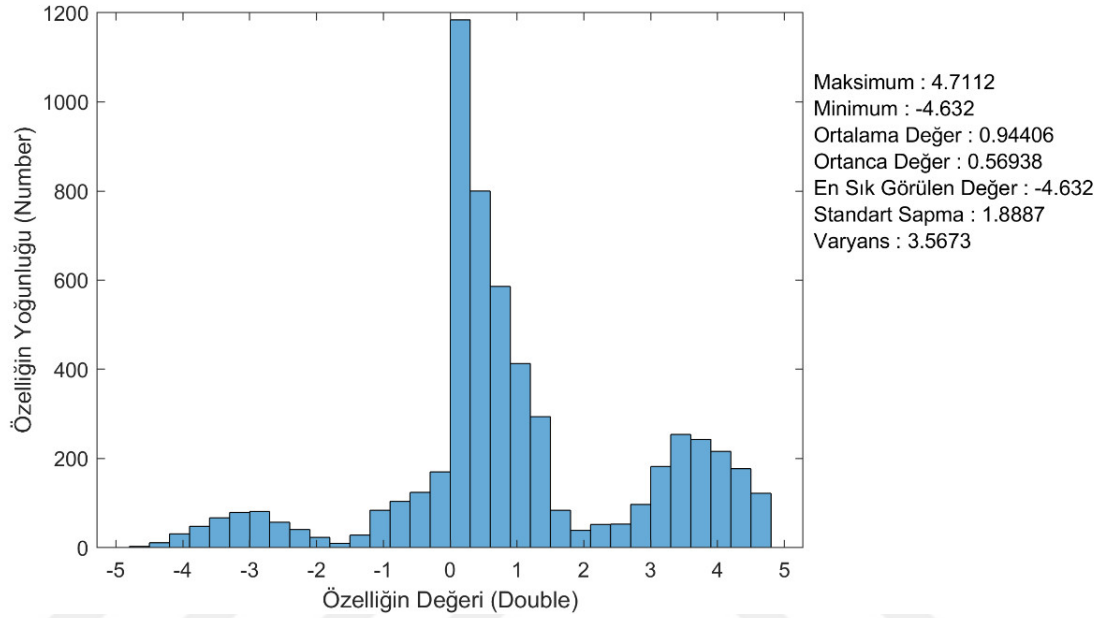


Şekil 5.2. Kaynaktan alınan sinyallerden oluşturulan eğitim veri kümesinin özellikleri.

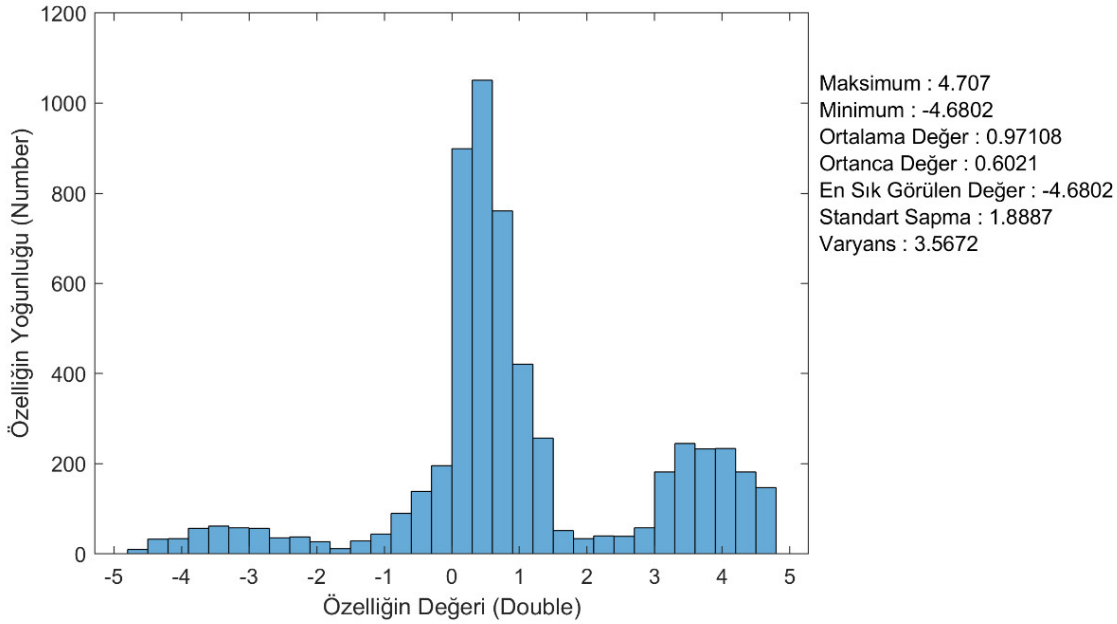


Şekil 5.3. Güvenilir röleden alınan sinyallerden oluşturulan test veri kümesinin özellikleri.

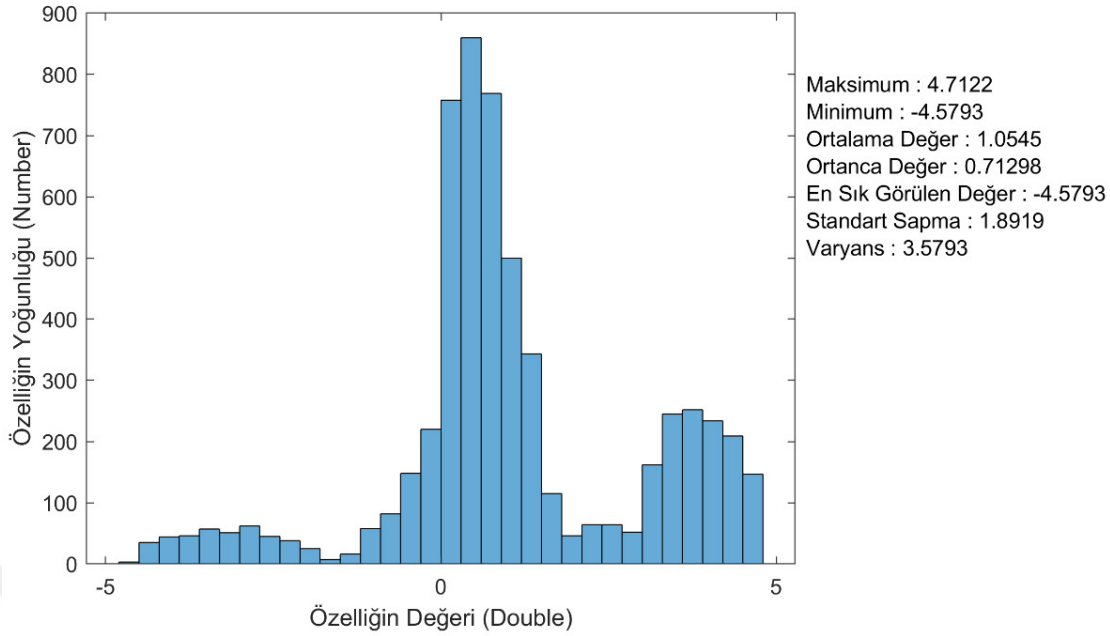
Şekil 5.4, Şekil 5.5 ve Şekil 5.6 da modellenen atakları gerçekleştiren röleden alınan sinyallerin istatistiksel özelliklerini temsil etmektedir. Atak modelleri sinyal üzerinde istatistiksel olarak algılanabilecek bir farklılığa neden olmamaktadır. Şekil 5.2 de gösterilen kaynak sinyalleri göz önüne alındığında atak modellerinin yapısal bir değişikliğe sebep olmadan veriyi manipule ettiği görülmektedir.



Şekil 5.4. A1 atak gerçekleştiren röleden alınan sinyallerden oluşturulan test veri kümesinin özellikleri.



Şekil 5.5. A2 atak gerçekleştiren röleden alınan sinyallerden oluşturulan test veri kümesinin özellikleri.



Şekil 5.6. A3 atak gerçekleştiren röleden alınan sinyallerden oluşturulan test veri kümesinin özellikleri.

5.3. Sistemin Çalışması

Simülasyonlar, Monte Carlo yaklaşımı kullanılarak sistem modeli ve algoritmalar aynı parametre değerleri ile 20 kez gerçekleştirilmiştir. Ek olarak her SNR, modülasyon türü, bant genişliği gibi alt koşulları için de 20 kez gerçekleştirilmiştir. Her çalışmada, veri kümesi rastgele oluşturulan bilgi bitlerinden ve kanal koşullarından elde edilmektedir. Oluşturulan veri seti ve algoritmaların parametreleri Tablo 6.1'de özetlenmektedir.

Denetimsiz atak tespit algoritmalarında OCSVM, LOF ve iForest algoritmalarındaki aşırı öğrenme sorunlarını önlemek için parametreler optimize edilerek seçilmektedir. OCSVM için, RBF çekirdeğini kullanılmaktadır ve ν parametresi, toplam hatalı örnek sayısına göre kenar hatalarının fraksiyonu üzerinde bir üst sınır ve destek vektörleri fraksiyonunun bir alt sınırı olan 0.3'e ayarlanmıştır. LOF için k parametresi, en derin komşuluk sayısı 60 olarak ayarlanmıştır. Alt ağaç sayısı iForest kurulumunda 20 olarak seçilmiştir. LTE-A downlink sinyali için sistem simülasyonları, sırasıyla 6, 15, 25, 50, 75, 100 kaynak bloğuna karşılık gelen 1.4, 3, 5, 10, 15 ve 20 MHz bant genişlikleri için gerçekleştirilmiştir. Denetimli algoritmalar için ise SVM algoritmasında ν parametresi 0.5, NN algoritmasında gizli katman sayısı 100, Random forest da ise maksimum ağaç derinliği 5, ve alt ağaç sayısı 10 seçilmiştir.

Tablo 5.1. Sistem simülasyonunun parametreleri.

Bant genişliği (Bandwidth (MHz))	1.4	3	5	10	15	20
Kaynak blok (Resource Block (RB))	6	15	25	50	75	100
Kaynak test ve eğitim veri sayıları	6	15	25	50	75	100
Güvenli role test ve eğitim veri sayıları	6	15	25	50	75	100
A1 atak modeli test ve eğitim veri sayıları	6	15	25	50	75	100
A2 atak modeli test ve eğitim veri sayıları	6	15	25	50	75	100
A3 atak modeli test ve eğitim veri sayıları	6	15	25	50	75	100
Veri önişleme	none					
Örneklerin veri yapısı	± Double, non zero					
Eğitim verileri min, max, mean	-74.7578, 8.9066, 1.1501					
Test verileri min, max, mean	-90.7578, 0.9066, 1.0420					
Veri boyutu (sembol sayısı) her slot için (N)	RB x 84 (Resource Block x Resource Element for 0.5 ms)					
Veri için temel bandı dinleme süresi	5 ms (Half Frame Period)					
Veri boyutu (M)	10 Slot					
Özellik vektörü boyutu (L)	3N					
Eğitim veri kümesi boyutu	LM					
OFDM alt taşıyıcı sayısı	128					
Cyclic prefix 1. Sembol için	10					
Cyclic prefix kalan için	9					
Kanal türü	Rayleigh					
Yol türü (Path)	Single Path					
Modülasyon tipleri	QPSK, 16QAM, 64QAM					
SNR seviyesi aralığı	5-30 dB					
OCSVM ve SVM algoritmalarının parametreleri	ν = 0.3 Kernel = Radial basis (RBF)					
LOF algoritma parametreleri	k = 60					
*i*Forest parametreleri	Base estimator = 20					
Random forest algoritması parametreleri	Max Depth=5, Estimators=10,					
NN algoritması parametreleri	Hidden layer = 100					

5.4. Performans Değerlendirme

Bu çalışmada tanımlanan problemin çözümü için denetimsiz makine öğrenmesi algoritmalarının performans değerlendirmesi, sadece röle ataklarının algılanmasıyla ilgili değil, aynı zamanda güvenli röle durumlarının doğru şekilde tanımlanması adınada önem taşımaktadır. Bir önceki bölümde tanımlanan veri kümelerinin istatistiksel özellikleri değerlendirildiğinde güvenli rölelerin, atak olarak tanımlanmaması atak tespit modelinin güvenirliliğinin bir göstergesi olarak düşünülebilmektedir.

Bölüm 3'de anlatılan performans metriklerini dentimli öğrenme, denetimsiz öğrneme ve istatistiksel öğrenme için kesinlik, doğruluk ve AUC'yi performans değerlendirme metrikleri olarak belirlenmiştir. Atak ve güvenli röle durumlarının/sınıflarının temsili

karışıklık matrisi, denetimsiz öğrenme için Tablo 5.1'de, denetimli öğrenme ve istatistiksel öğrenme için ise Tablo 5.2'de tanımlanmaktadır.

Tablo 5.2. Denetimsiz öğrenme yöntemleri için karmaşıklık matrisi.

		Gerçek Durum	
		Atak	Güvenli
Tahmin Edilen Durum	Atak	TP	FP
	Güvenli	FN	TN

Tablo 5.3. Denetili ve istatistiksel öğrenme yöntemleri için karmaşıklık matrisi.

		Gerçek Durum			
		Güvenli	A1	A2	A3
Tahmin Edilen Durum	Güvenli	TP	FP	FP	FP
	A1	FN	TP	FN	FN
	A2	FP	FN	TP	FN
	A3	FP	FN	FN	TP

Kesinlik, tek bir kategorinin ne kadar kesinlikte sınıflandırıldığının gösterilmesidir.

$$\text{precision} = \frac{TP}{TP+FP} \tag{5.1}$$

Bu tezde tanımlanan problemde, atak röle durumunun tespiti, denetimsiz öğrenme yaklaşımları için aykırılık tespiti olarak kabul edilmektedir. Öte yandan doğruluk, tüm sınıflar arasında gerçekten sınıflandırılmış atakları ve güvenli örnekleri ölçmektedir:

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.2}$$

Karışıklık matrisi, alıcı çalışma karakteristiği (ROC) eğrisi gibi bir sınıfın sınıflandırılmasının yeteneğini ölçmek için değerlendirme metrikleri olarak da kullanılabilmektedir. ROC eğrileri, yanlış pozitif orana (FPR) göre gerçek pozitif oranın (TPR) CDF'sinden türetilmektedir. Bir ROC eğrisi, her bir eşiğin ürettiği tüm

karışıklık matrislerini özetler ve farklı sınıflandırma eşiklerinde TPR ve FPR'yi gösterir. Sınıflandırma eşiğini düşürerek, daha fazla örnek pozitif olarak sınıflandırılabilir; böylece hem yanlış pozitifleri hem de gerçek pozitifleri arttırabilir. TPR ve FPR aşağıdaki gibi tanımlanmaktadır:

$$TPR = \frac{TP}{TP+FN} \tag{5.3}$$

$$FPR = \frac{FP}{TN+FP} \tag{5.4}$$

AUC, ROC eğrisinin altındaki iki boyutlu alanı tanımlamaktadır. Yani olası tüm sınıflandırma eşiklerinde toplam performans ölçümü sağlamaktadır. AUC un istatistiksel anlamı, modelin rastgele bir pozitif örneği rastgele bir negatif örnekten daha yüksek sıralama olasılığıdır. Ayrıca, tahminlerin mutlak değerlerinden ziyade ne kadar iyi sıralandığının ölçümünü sağlamaktadır. AUC, hangi sınıflandırma yönteminin daha iyi olduğuna karar vermemize yardımcı olmaktadır.

Bir sonraki bölümde SNR seviyesi, modülasyon tipi ve veri boyutu (M) ile ilgili hassasiyet, doğruluk ve AUC ölçümleri açısından yukarıda belirtilen OCSVM, LOF ve iForest denetimsiz algoritmalarının, SVM, NN ve Random forest denetimli algoritmaların son olarak PPCA ve LSA istatistiksel öğrenme algoritmalarının performanslarına ilişkin sonuçları tartışılmaktadır.

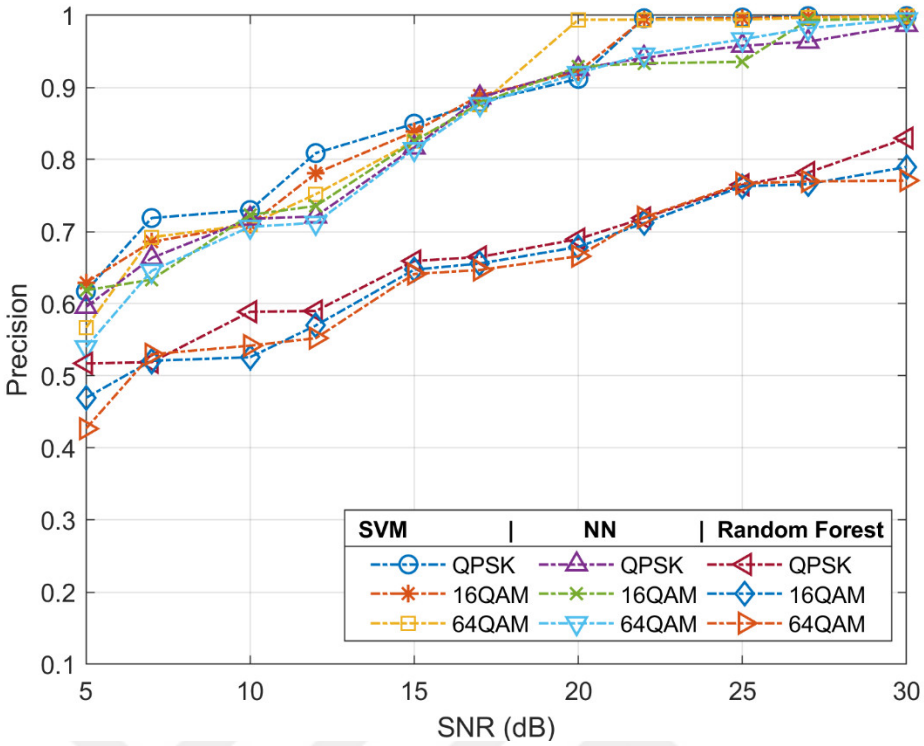
6. SONUÇLAR

6.1. Denetimli Öğrenme Yöntemleri

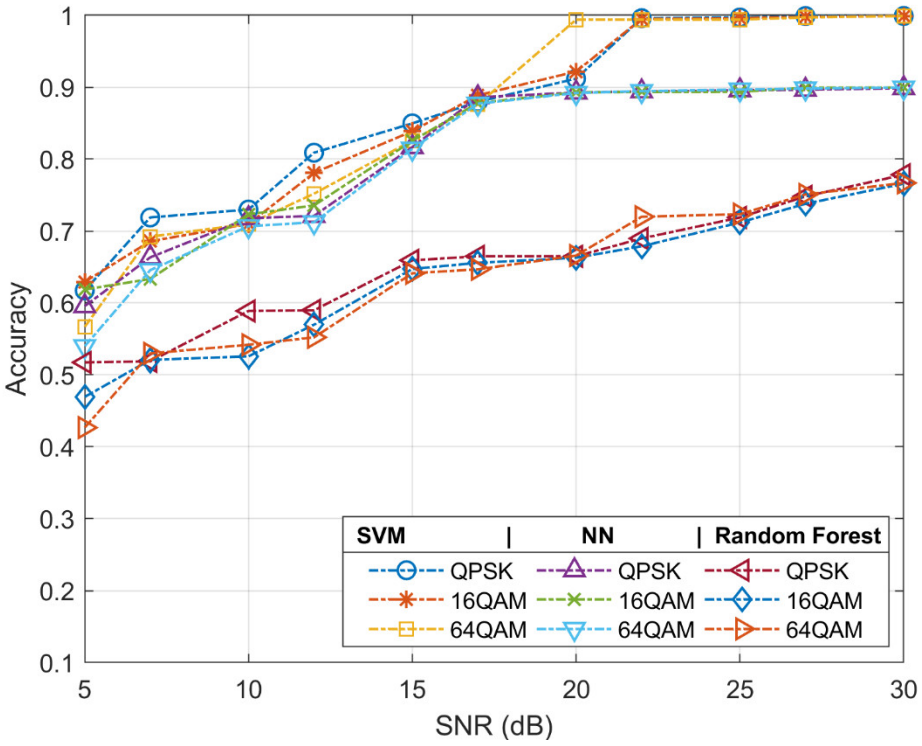
Bu bölümde denetimli öğrenme yöntemlerinin, tanımlanan atak türlerini kategorilendirmedeki deneysel sonuçları incelenmiştir. Sonuçlar SNR, modülasyon ve veri boyutu faktörlerinin etkileri grafiklerle detaylandırılmıştır. İlk olarak sunulan grafikler Şekil 6.1 ve Şekil 6.2'de güvenli röle sinyallerinin ne kadar başarılı sınıflandırılabildiği incelenmiştir. Güvenli rölelerin sınıflandırılmasında SNR seviyesi dominant bir etki olarak gözlemlenirken modülasyon tekniğindeki değişikliklerin sonuçlara etkisi ihmal edilecek kadar az olduğu görülmektedir.

SNR seviyesinin 10 dB ve aşağısında olması durumunda sınıflandırma kabiliyetinin SVM ve NN için %70'in altına, Random Forest için ise %60'ın altına indiği gözlemlenmektedir. Şekil 6.2'de doğruluk/accuracy değerlerinin de kesinlik/precision değerleri ile benzerlik göstermesi, güvenli röle sınıfı için gerçek bir değerlendirmeden bahsedilebilmesi anlamına gelmektedir. Bu karşılaştırmada Random forest algoritması en düşük performansı gösterirken SVM algoritması en yüksek doğruluk ve kesinlik değerine sahiptir. SVM algoritması 15 dB SNR seviyesinde %85 kesinlik ve doğruluktadır, NN için ise ortalama %80 ve Random forest için ortalama %65 değerindedir. SNR seviyesi 20 dB ye yükseldiğinde ise SVM algoritması ortalama %99 doğruluk ve kesinlik değerine yükselirken, NN %95 ve Random forest ise %70 değerlerine yükselmektedir.

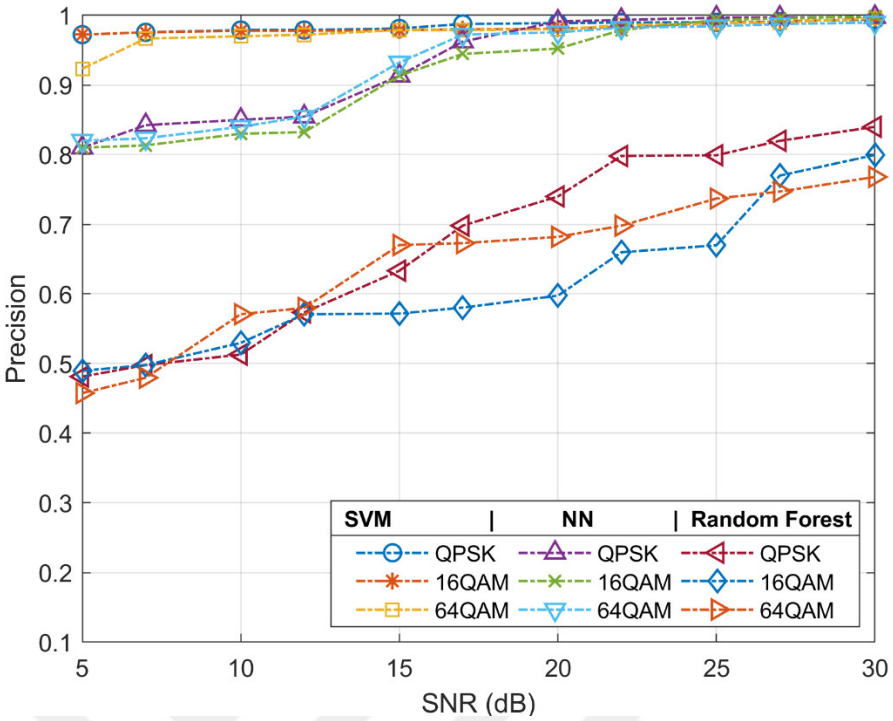
Şekil 6.3 ve 6.4'de karıştırma atağı (A1) için sınıflandırma algoritmalarının kesinlik ve doğruluk değerleri gösterilmektedir. SVM algoritması için kesinlik değeri ve doğruluk değeri en düşük SNR seviyesinden itibaren ortalama %90'ın üzerindedir. NN algoritması için kesinlik değeri %80 ile %99 arasında seyrederken doğruluk değeri %65 ile %90 arasında SNR seviyesine göre değişmektedir. Bunun nedeni algoritmanın atak olarak sınıflandırması, ancak atak türünü doğru sınıflandıramamasından kaynaklanmaktadır (örneğin A1 türü atağın A2 olarak sınıflandırılmasıdır).



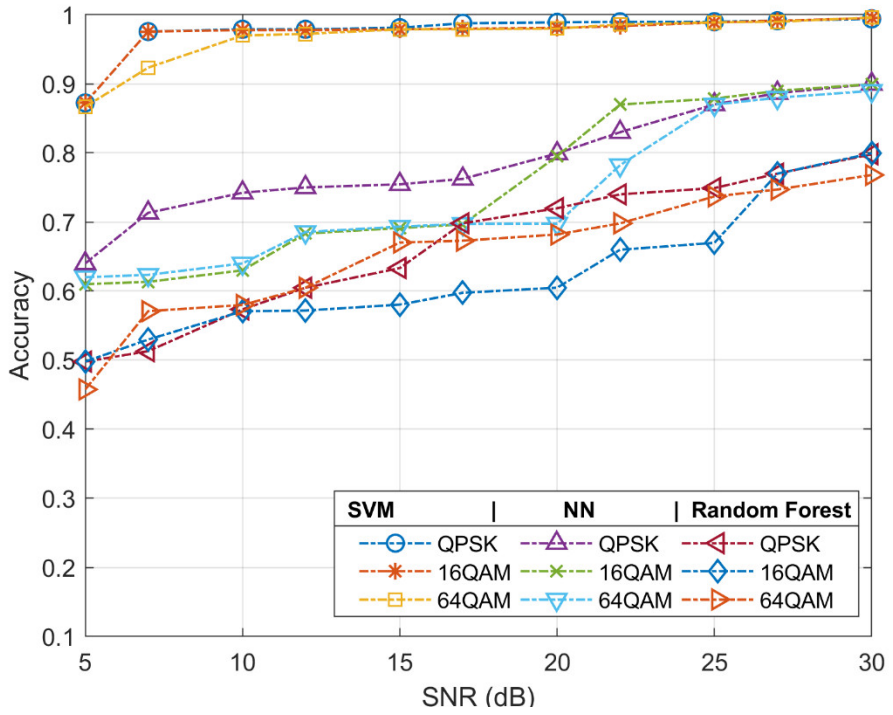
Şekil 6.1. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde güvenli röle tespitinin SNR seviyesi ekseninde kesinlik değerleri.



Şekil 6.2. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde güvenli röle tespitinin SNR seviyesi ekseninde doğruluk değerleri.

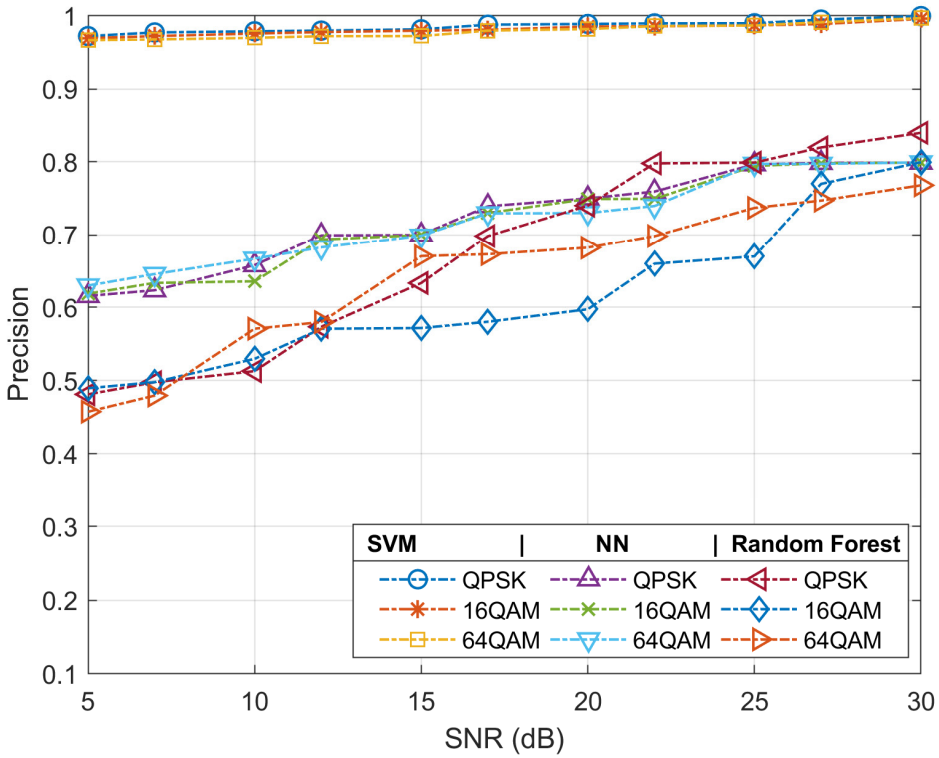


Şekil 6.3. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde karıştırma atağı (A1) tespitinin SNR seviyesi ekseninde kesinlik değerleri.

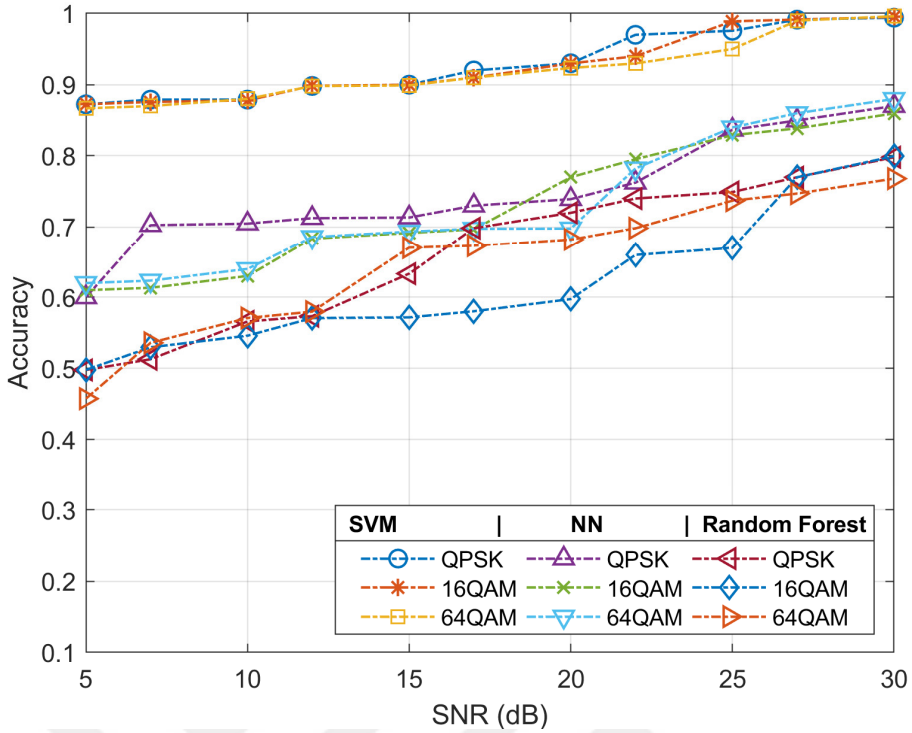


Şekil 6.4. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde karıştırma atağı (A1) tespitinin SNR seviyesi ekseninde doğruluk değerleri.

Şekil 6.5 ve Şekil 6.6'da yeni veri iletme atağı (A2) sınıflandırma sonuçları kesinlik ve doğruluk değerleri gösterilmektedir. SVM algoritması en düşük SNR seviyesinde %97'nin üzerinde kesinlik, %87'nin üzerinde ise doğruluk değerindedir. NN algoritması ve Random forest algoritması %60 ve %50 kesinlik ve doğruluk değerleri ile en düşük SNR seviyesinde en kötü sınıflandırma performansı göstermiştir. Yeni veri iletme atağı rölenin aldığı sinyali değil, farklı bir sinyali kullanıcıya iletmesi durumudur. Bu nedenle NN gibi özelliklerin ağırlıklandırılması temeline dayanan bir yaklaşım A2 türü atak modeli sınıflandırmada keskinlik ve doğruluk değerleri daha düşük olmaktadır. Random forest ve NN algoritması ancak 20 dB SNR seviyesinden sonra %70 doğruluk değerinin üzerine çıkabilmektedir ve 30 dB SNR seviyesinde NN algoritması %90, Random forest ise %80 doğruluk değerine çıkabilmektedir. Random forest gibi karar ağaçları türü algoritmarın her örneği doğru sınıflandıramaması, fazla öğrenme sorununa dayanmaktadır (ovefitting). Karar ağaçlarının ağaç derinlikleri ve alt ağaç sayılarının optimizasyonu bu noktada önem taşımaktadır ve optimum değerlerin seçilmemesi sonuçları daha fazla değiştirmektedir.



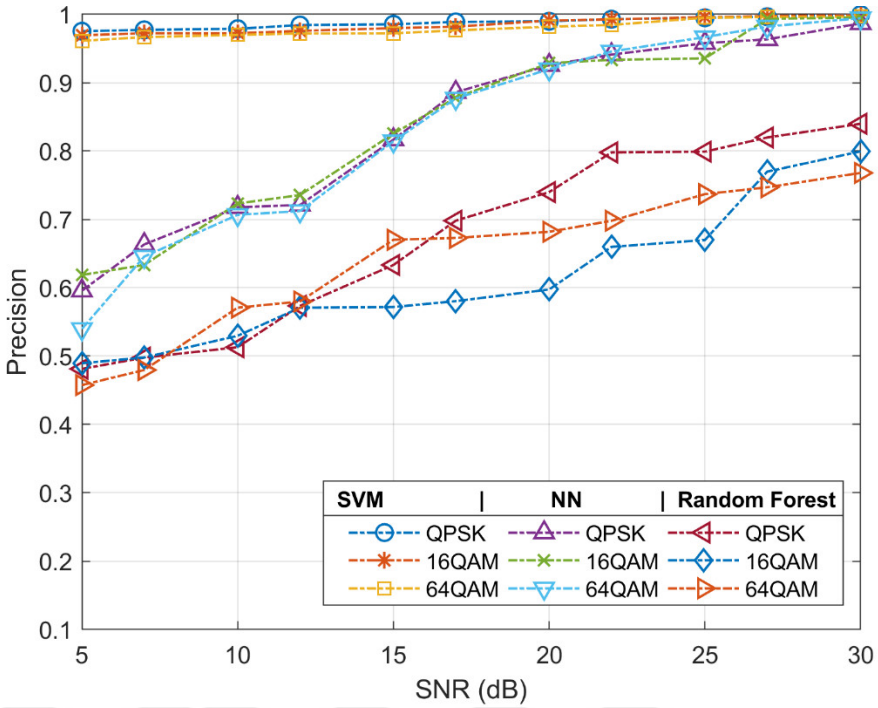
Şekil 6.5. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde farklı veri iletme atağı (A2) tespitinin SNR seviyesi ekseninde kesinlik değerleri.



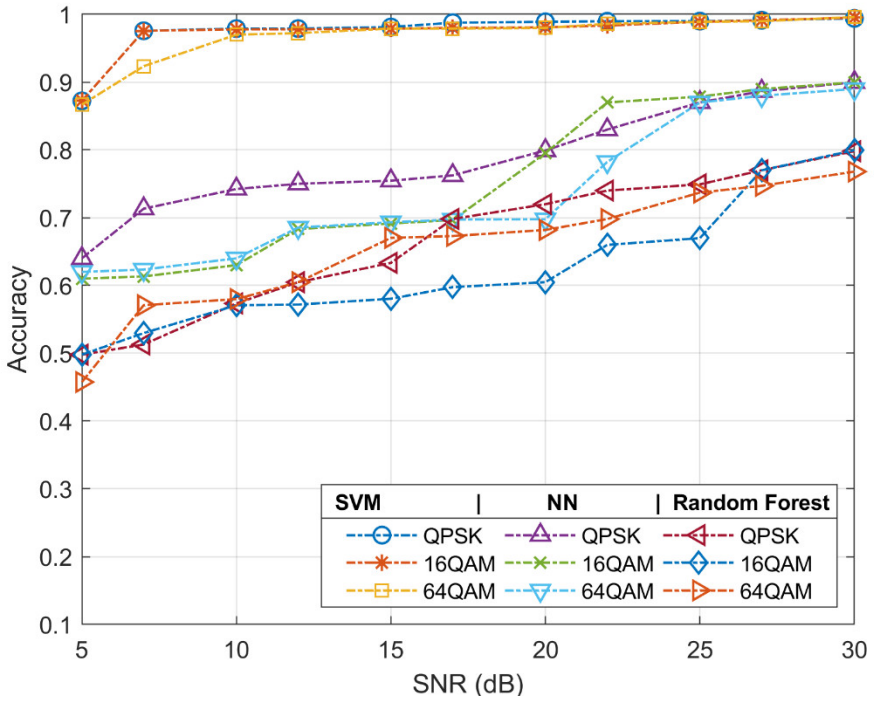
Şekil 6.6. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde farklı veri iletme atağı (A2) tespitinin SNR seviyesi ekseninde doğruluk değerleri.

Şekil 6.7 ve Şekil 6.8 de farklı veri enjeksiyon atağının (A3) tüm algoritmalar için kesinlik ve doğruluk grafikleri verilmektedir. SVM algoritması A3 türü atak modelini sınıflandırırken en yüksek performansı göstermiştir. En düşük SNR seviyesinde %98 kesinlik değeri, %88 doğruluk değerine sahiptir NN algoritması %60, Random Forest ise %50 değerinde kalmıştır. NN algoritması %60 ile %99 arası bir SNR seviyesi ile değişen kesinlik değerine, %60 ile %90 arası doğruluk değerine sahiptir. Random forest algoritması %50 ile %80 arasında kesinlik ve doğruluk değerine sahiptir. NN ve Random forest algoritmaları 15 dB SNR seviyesinden sonra A3 atak türünü sınıflandırabilmektedir.

Bu seviyede NN %80, Random forest %70 kesinlik değerine sahiptir. Doğruluk değerinde ise NN %75, Random forest %70 seviyesindedir. Farklı veri enjeksiyon atağı, alınan sinyal vektörü ile farklı bir vektörün sinyal gücü normalize edilerek eklenmiş halidir. Bu nedenle orjinal veri üzerindeki değişiklikler daha yüksektir ve yine bu nedenle NN ve Random forest bu atak türünü sınıflandırırken A2 atak türüne göre daha yüksek bir kesinliğe sahiptir.

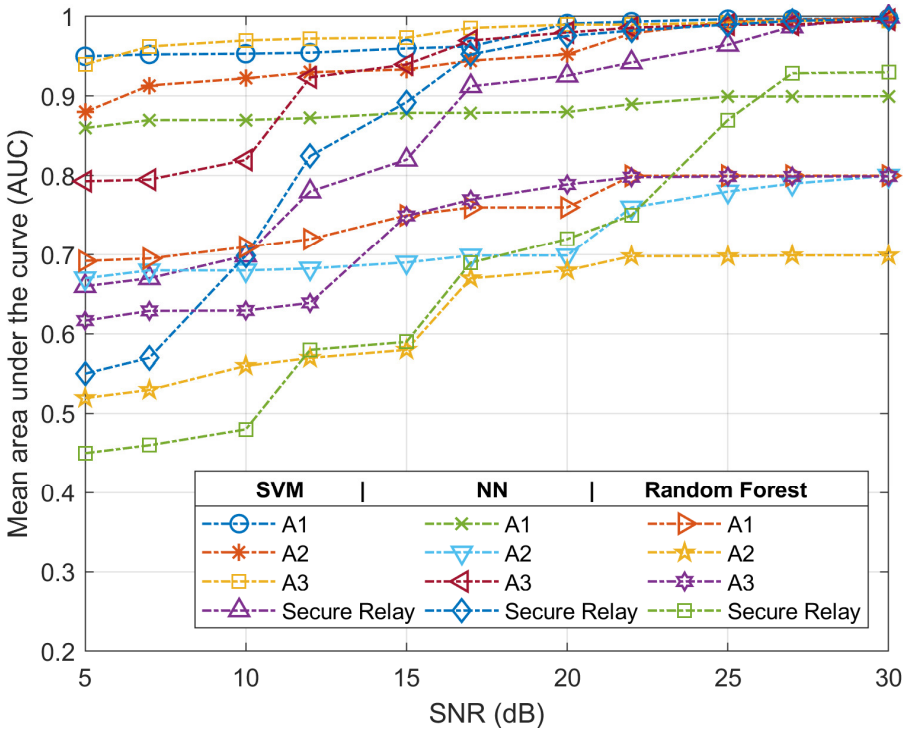


Şekil 6.7. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde veri enjeksiyon atağı (A3) tespitinin SNR seviyesi ekseninde kesinlik değerleri.



Şekil 6.8. Denetimli öğrenme algoritmalarının farklı modülasyon tiplerinde veri enjeksiyon atağı (A3) tespitinin SNR seviyesi ekseninde doğruluk değerleri.

Şekil 6.9'da tüm algoritmaların atak türlerini sınıflandırma performansları, SNR seviyesi ekseninde AUC değerleri ile gösterilmektedir. AUC, skor (sınıflandırma tahminindeki gerçek değeri) değerlerinin karakteristik grafiği altında kalan alandır. Bu alan kapsamlı olarak algoritmaların sınıflandırma tahminlerine ne kadar yakın ya da uzak olarak karara vardıklarının göstergesidir. Bu alan ne kadar genişse, skor değerleri gerçek sınıf değerine o kadar yakın olarak karar fonksiyonu tarafından hesaplandığının göstergesidir. Bu nedenle AUC değeri algoritmaların karar fonksiyonlarının etkililiğinin bir göstergesi olarak da düşünülebilmektedir.



Şekil 6.9. Tüm denetimli öğrenme algoritmalarının farklı atak türleri ve SNR seviyelerinde AUC değerlerinin karşılaştırması (modülasyon = QPSK, bant genişliği = 1.4 Mz, M=6).

SVM, 5 dB SNR seviyesinde A1, A2, A3 atak türünü en iyi sınıflandıran algoritma olarak gözlemlenmektedir. Daha sonra A1 atak modelin sınıflandırılmasında yaklaşık 0,9 değeri ile NN algoritmasının SVM (~0,95) e yakın bir performans göstermektedir. NN algoritması A3 atak modeli için 10 db SNR seviyesinden sonra ~0,95 ile SVM ~0,97 ile benzer bir perofrmans göstermektedir. Random forest algoritması karşılaştırmada en düşük performansa sahip algoritmadır. A1, A2, A3 atak modelleri için ~0,75 değerinde ortalama AUC değerine sahiptir. Güvenilir rölenin sınıflandırılmasında ise 10 dB SNR seviyesinden sonra SVM ve NN algoritması için

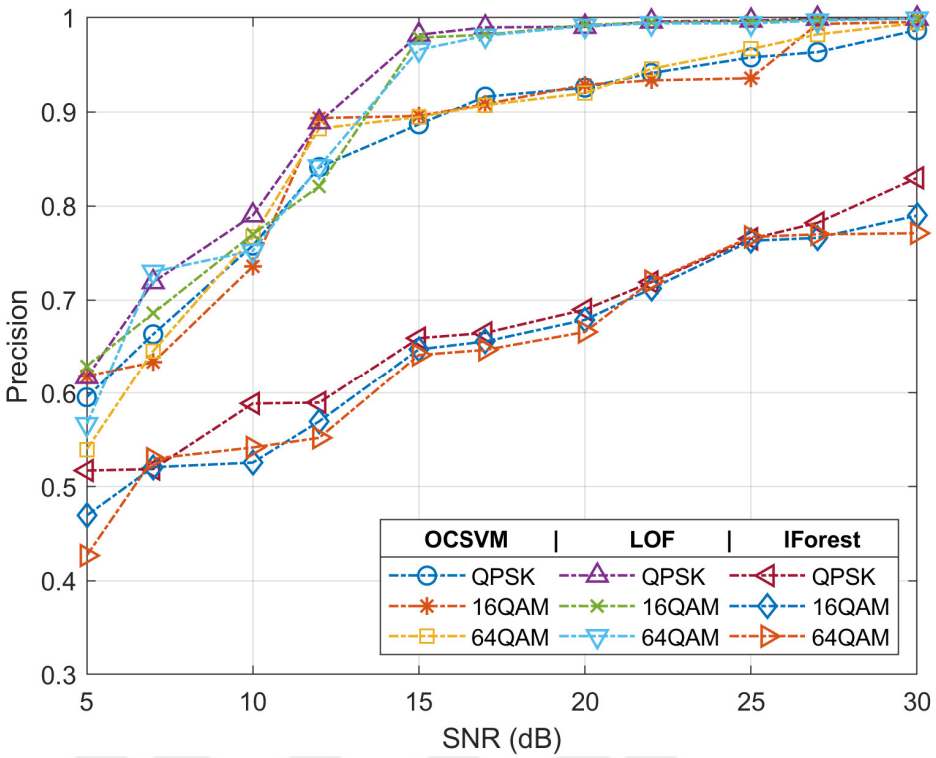
0,7 ile 0,99 değer aralığında değişmektedir. Ancak Random forest bu değer aralığına 20 dB SNR seviyesinden sonra ulaşmaktadır. Bu noktada da diğer algoritmalardan daha düşük bir performans göstermektedir.

6.2. Denetimsiz Öğrenme Yöntemleri

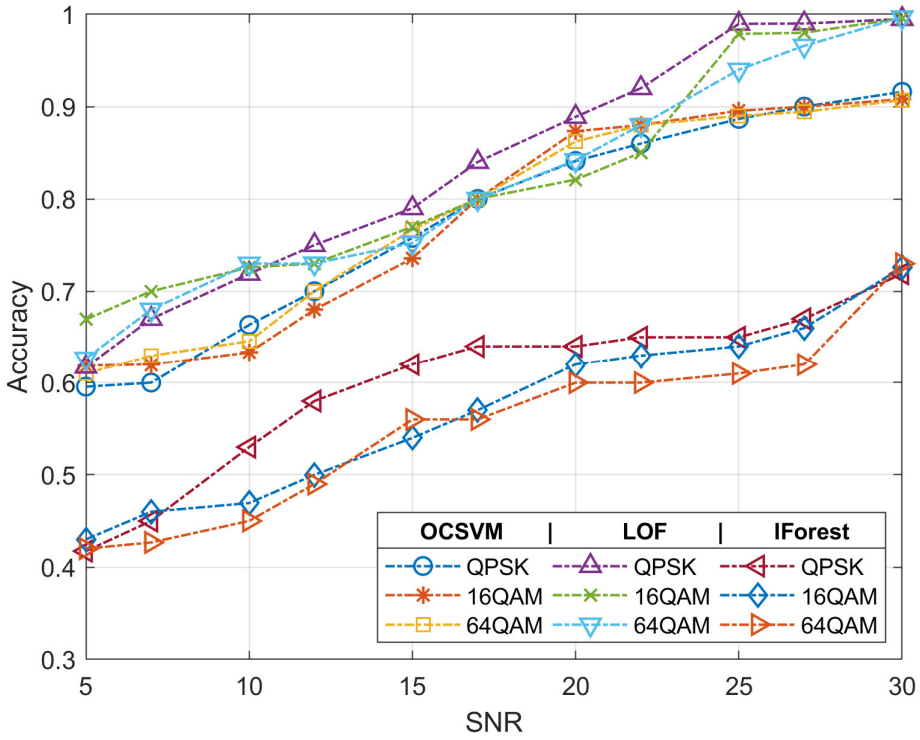
Bu bölümde denetimsiz öğrenme yöntemlerinin tanımlanan atak türleri tespitindeki performans sonuçları incelenmektedir. Algoritmaların performansı SNR, modülasyon ve veri boyutu faktörleri göz önüne alınarak grafiklerle detaylandırılmıştır. Atak türlerine göre denetimsiz algoritmaların performansı incelendiğinde, SNR düzeyindeki artışın olumlu etkisi görülmektedir. Düşük SNR seviyesinde bile, tüm algoritmaların en az %90 kesinlik seviyesine ulaştığı görülebilmektedir (Şekil 6.12, Şekil 6.14 ve Şekil 6.16). Şekil 6.10'da, iForest için hassasiyet seviyesinin %40 ile %80 arasında, OCSVM ve LOF'un %50 ile %100 arasında olduğunu gözlemlenmektedir.
Farklı SNR seviyesinde güvenli röle algılamasında, hem LOF hem de OCSVM neredeyse aynı kesinlik performansını göstermektedir ve iForest'ın kesinlik performansının bu algoritmalara göre daha düşük olduğu gözlenmektedir. Ayrıca, modülasyon tiplerinin güvenli röle tespiti için algoritmaların kesinlik seviyesi üzerindeki etkisinin önemsiz olduğu görülmektedir. OCSVM ve LOF'da %70 kesinlik seviyesine yaklaşık 10 dB SNR'da ulaşılabilirken, iForest'da 22 dB SNR seviyesinde ulaşılabileceği görülmektedir.

Kesinlik, yalnızca saldırı ile röleyi tespit etme performansını temsil eder (gerçek pozitif). Diğer yandan, doğruluk/accuracy hem röle saldırılarını hem de güvenli röleleri (gerçek pozitif ve gerçek negatif) tespit etme doğruluğu hakkında ölçüm sağlar. Bu nedenle, genel algılama performansını değerlendirmek için algoritmaların doğruluk değerleri de sunulmaktadır.

Şekil 6.11'deki güvenli röle algılamasındaki doğruluk, Şekil 6.10'daki kesinlik, yaklaşık olarak yakındır. Ancak, iForest algoritmasının doğruluğu, kesinliğinden daha kötüdür. Yani 20 dB SNR seviyesinde kesinlik ve doğruluk değerleri yaklaşık sırasıyla %70 ve %60'tır. Genel olarak, SNR varyasyonunun doğruluk ve kesinlik performansı üzerindeki etkisi, güvenli röle algılamasında dikkat çekicidir.



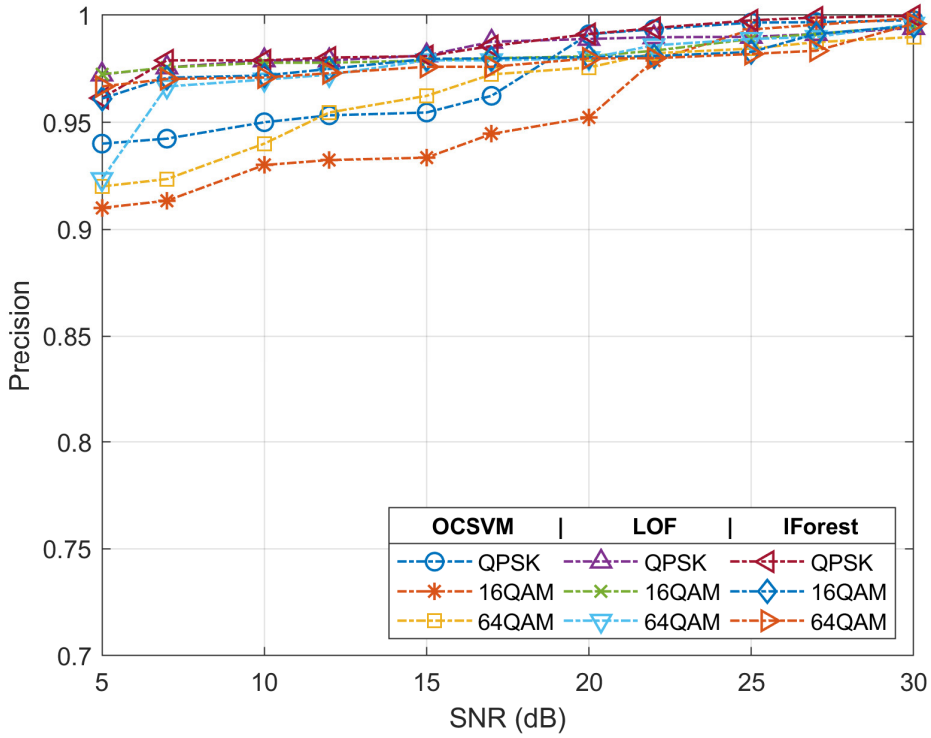
Şekil 6.10. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde güvenli röle tespitinin SNR seviyesi ekseninde kesinlik değerleri.



Şekil 6.11. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde güvenli röle tespitinin SNR seviyesi ekseninde doğruluk değerleri.

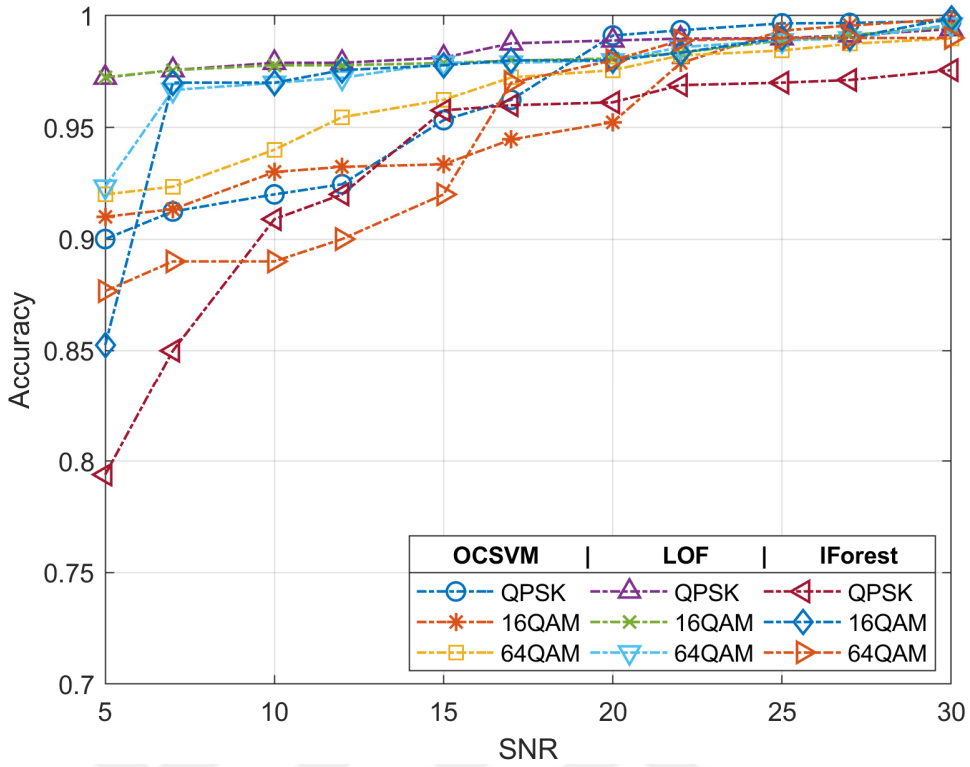
Röle saldırısı durumundaki algoritmaların performans sonuçları Şekil 6.12 ile Şekil 6.17'de gösterilmektedir. Şekil 6.12 ve Şekil 6.13'te sunulan karıştırma saldırısı modelinin (A1) tüm algoritmalar için en düşük SNR koşulları (10 dB ve altı) altında kesinlik ve doğruluk sonuçları karşılaştırıldığında, doğruluğun özellikle kesinlikten daha düşük olduğu görülmektedir.

Bunun nedeni, düşük SNR koşullarında güvenli rölenin yanlış alarm olarak değerlendirilebilmesidir ve bu da modelin algılama doğruluğunu etkilemektedir. Bununla birlikte, bu yanlış alarmlarda bile, %80 ve üzeri doğruluk elde edilir. Sadece kesinlik göz önüne alındığında, A1 saldırı modeli için tüm algoritmalarda %90'ın üzerinde performans elde edilebilmektedir.



Şekil 6.12. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde karıştırma atağı (A1) tespitinin SNR seviyesi ekseninde kesinlik değerleri.

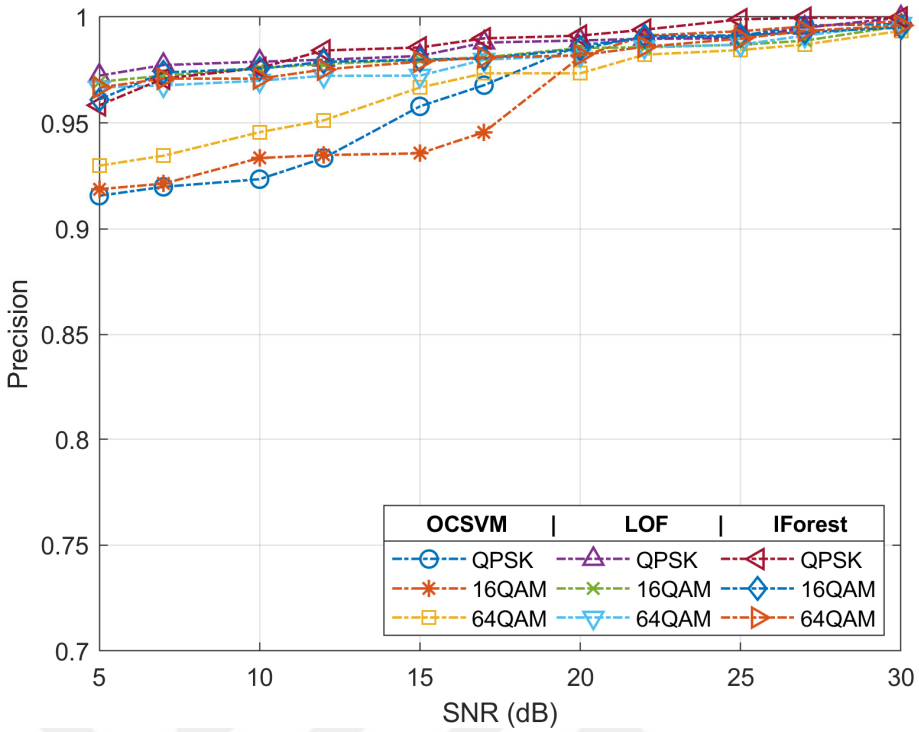
A1 ve A2 atak türlerini algılamada OCSVM, kesinlik performansı LOF ve iForest'ten 15 dB SNR seviyesine kadar daha fazla etkilenmiştir. Şekil 6.14'te gösterildiği gibi yeni veri iletme atağı (A2) tespiti, tüm SNR değerleri için LOF ve iForest algoritmaları %95'in üzerinde kesinliğe sahiptir; ancak OCSVM bunu yaklaşık 20 dB'den daha yüksek SNR seviyesinde yakalayabilmektedir.



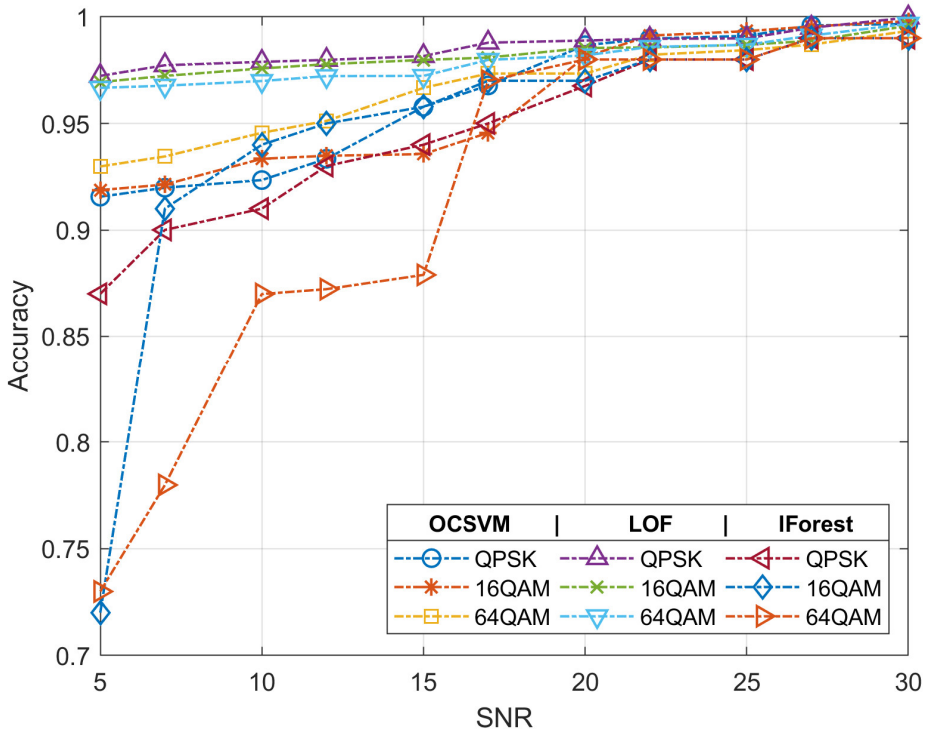
Şekil 6.13. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde karıştırma atağı (A1) tespitinin SNR seviyesi ekseninde doğruluk değerleri.

Şekil 6.15'te verilen yeni veri iletme (A2) atağı tespitinde algoritmaların doğruluğu incelendiğinde, LOF algoritması tüm SNR değerleri için %95'in üzerinde performans gösterirken, OCSVM ve iForest 20 dB'den daha yüksek SNR değerlerinde aynı performansı gösterebilmiştir. Farklı veri enjeksiyon saldırısının (A3) tespiti için algoritmaların kesinliği ve doğruluğu sırasıyla Şekil 6.16 ve Şekil 6.17'de sunulmaktadır.

Tüm algoritmalar sırasıyla %95 ve %90'ın üzerinde kesinlik ve doğruluk performansına sahiptir. Genel olarak, A1, A2 ve A3 atak türlerinin algılanmasında tüm algoritmalar %90'nın üzerinde kesinlik performansı göstermektedir. Doğruluk performansı açısından, iForest algoritmasının, özellikle A1 ve A2 ataklarını tespit etmek için düşük SNR seviyesinden, yani 10 dB ve daha düşük seviyeden etkilendiği görülmektedir. iForest algoritmasının gürültüden daha fazla etkilenmesinin nedeni, karar ağaçları temelli algoritmlarda modellerin eğitim verisine çok fazla uymasından kaynaklanmaktadır. Bu modeler eğitim verileri üzerinde yüksek başarım gösterirken farklı bir veri ile test edilmek istendiğinde doğruluk seviyeleri düşmektedir.



Şekil 6.14. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde yeni veri iletme atağı (A2) tespitinin SNR seviyesi ekseninde kesinlik değerleri.



Şekil 6.15. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde yeni veri iletme atağı (A2) tespitinin SNR seviyesi ekseninde doğruluk değerleri.

Aykırı değeri tespit oranını artırmak için; eğer LOF algoritmasında komşuluk değeri artırılırsa, OCSVM ve iForest'ten çok daha yüksek bir doğruluğa sahip olabilir. Yüksek boyutlu hiper küreyi haritalayan OCSVM için de yüksek bir doğruluk gözlenmektedir.

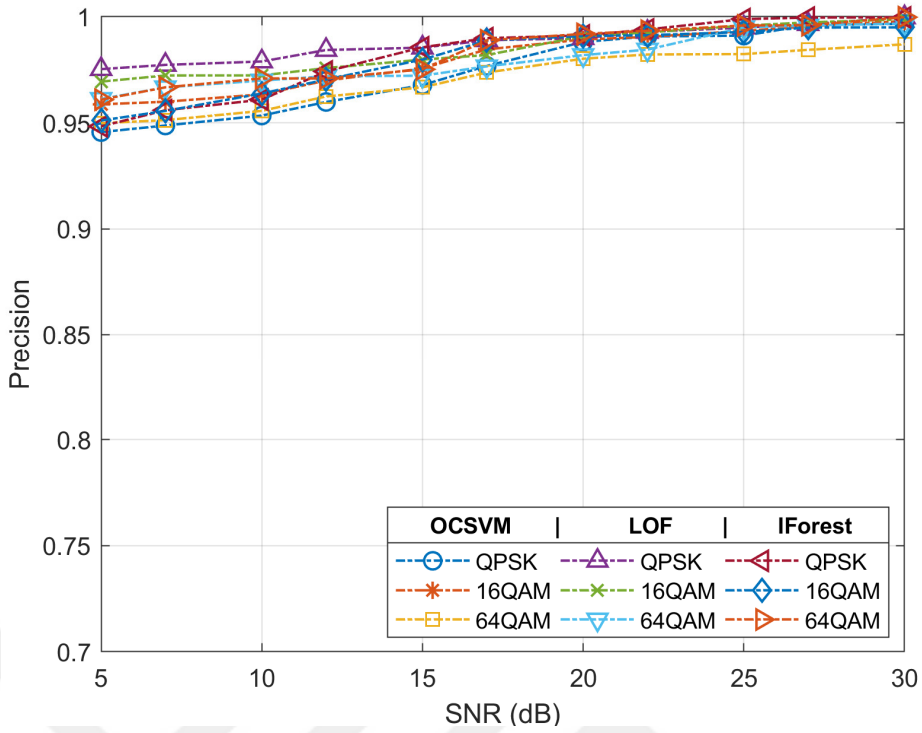
Bununla birlikte iForest, LOF ve OCSVM'den daha yüksek boyutsallıkta bile iyi bir ortalama performansa sahiptir; ancak aşırı öğrenme sorununun getirdiği yüksek yanlış alarm oranı nedeniyle daha az doğruluktadır. Güvenli röle tespiti sonucunda OCSVM, karşılaştırmamızda önemli bir ayarlama gerektirmeden az miktarda eğitim verisi ile yüksek performansa sahiptir ve pratik uygulamalar için iyi bir adaydır.

Modülasyon tiplerinin (QPSK, 16QAM, 64QAM) algoritmaların performansı üzerindeki etkisi incelendiğinde, doğruluk değerlerinin Tablo 6.2'de gösterildiği gibi, aynı atak tipi için önemli ölçüde etkilenmediği görülmektedir. Bu durum çalışmada önerdiğimiz atak tespit modelinin sağlamlığı, geometrik sintal özelliklerinin veriyi modelleme noktasındaki önemi hakkında bir bakış açısı sunmaktadır.

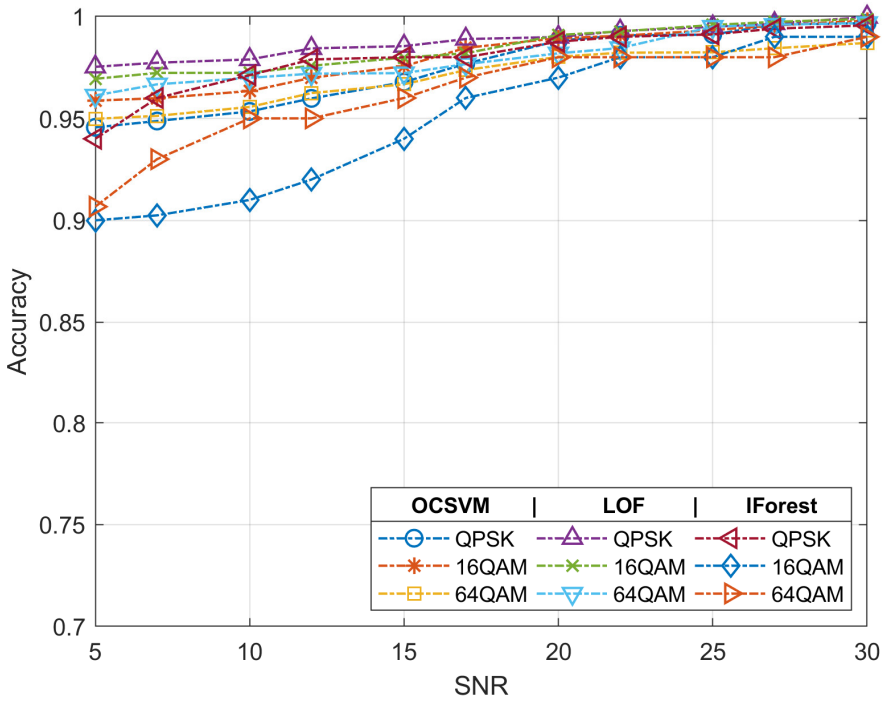
Tablo 5.1'de belirtilen parametreler algoritmların simülasyonu sırasında kullanılmıştır. Bununla birlikte, *v, k* ve taban tahmincisi (base estimator) parametrelerinin büyük değerleri, aşırı öğrenme sorununa neden olmaktadır ve bu nedenle, algoritmaların tümü için yanlış alarm oranını artırır ve saldırı algılama doğruluğunu azaltır.

Tek sınıflı aykırı değer tespit algoritmalarında kullanılan parametrelerin, en doğru sonuçları verdikleri noktada seçilmesi büyük önem taşımaktadır. Bu tez çalışmasında seçilen parametreler ise, algoritmaların bu parametreler için olası değer aralıklarında sınanmaları ile elde edilmiş en dengeli öğrenmenin sağlandığı noktalar olarak tespit edilmiştir.

Sistemin performansını değerlendirmek için kullanılan bir diğer parametre veri boyutu M'dir. Veri boyutunun etkisi için sonuçlar Tablo 6.1'de özetlenmiştir. Veri boyutunun saldırı tespiti üzerinde dikkate değer bir etkisi gözlenmemiştir. Veri boyutu özellikle güvenli röle algılama durumunda önemlidir. M değeri 6'dan 100'e yükseltildiğinde, özellikle güvenli röle algılama durumu için algoritmaların doğruluğu artmaktadır.



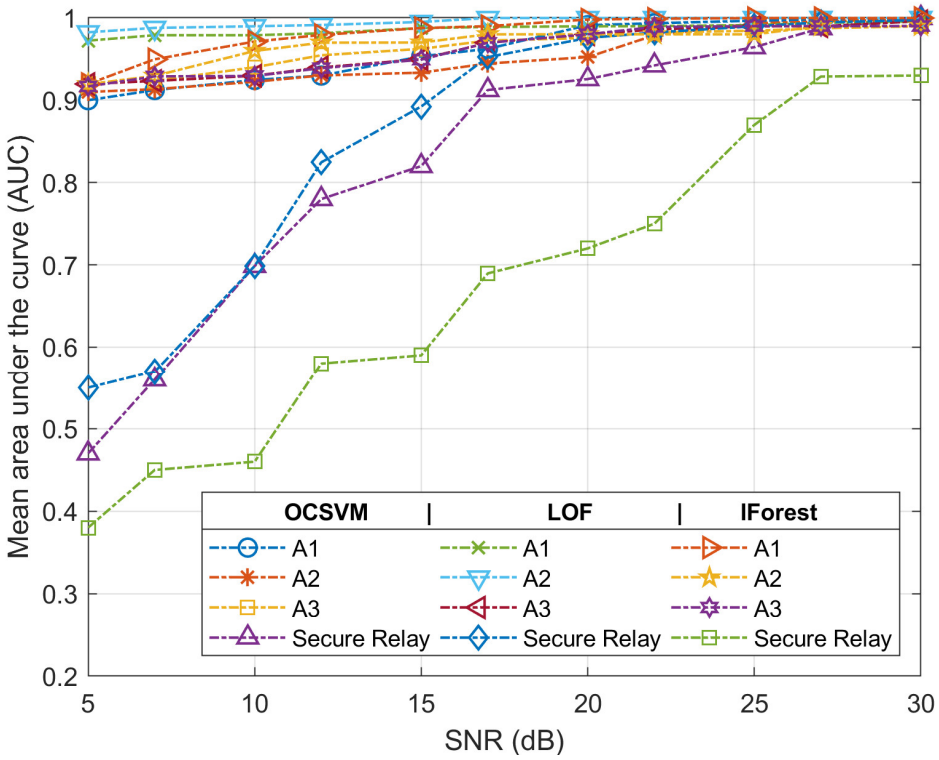
Şekil 6.16. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde veri enjeksiyon atağı (A3) tespitinin SNR seviyesi ekseninde kesinlik değerleri.



Şekil 6.17. Denetimsiz öğrenme algoritmalarının farklı modülasyon tiplerinde veri enjeksiyon atağı (A2) tespitinin SNR seviyesi ekseninde doğruluk değerleri.

Algoritmaların AUC performansı Şekil 6.18'de gösterilmektedir. Tüm algoritmaların AUC değerleri, atak tespit vakaları için 0.9'un üzerindedir. Tüm saldırı türleri için, 0.95 ve üstü AUC değerlerine sahip LOF algoritmasının performansının, diğer algoritmalardan daha üstün olduğu görülmektedir.

Güvenli röle algılama ile ilgili olarak, 15 dB SNR seviyesi için AUC değerleri incelendiğinde, LOF, OCSVM ve iForest algoritmalarının AUC değerleri sırasıyla 0.81, 0.78 ve 0.58'dir. *i*Forest algoritması güvenli röleleri algılamada bu algoritmalar kadar başarılı sonuçlar üretememiştir. LOF algoritması en büyük ortalama AUC performansına sahiptir ve daha sonra OCSVM algoritması gelmektedir. Sonuçlar; bu algoritmalar ile uygulanabilir sonuçlar elde edebilmek ve bu iki algoritmanın 10 dB ve daha düşük SNR seviyelerinde bile kötü niyetli ve güvenli röleleri tespit etmek için kullanılabileceğini göstermektedir.



Şekil 6.18. Tüm denetimsiz öğrenme algoritmalarının farklı atak türleri ve SNR seviyelerinde karşılaştırması (modülasyon = QPSK, bant genişliği = 1.4 Mz, M=6).

Yukarıda sunulan sonuçlara dayanarak, LTE-A ağındaki kötü amaçlı röle algılama sorunu hakkında iki kesin açıklama yapılabilir. İlk gözlem, denetimsiz öğrenme veya aykırı tespit yöntemlerinin kendi aralarında karşılaştırılması ile ilgilidir. OCSVM ve

LOF için yüksek doğruluk gözlenir. Bununla birlikte, yüksek yanlış alarm oranına neden olan aşırı öğrenme problemi nedeniyle, iForest'in yüksek boyutsallık ile ortalama doğruluk performansı göz önüne alınmalıdır. LOF, röle saldırılarını tespit etmede OCSVM ve iForest ile karşılaştırıldığında daha yüksek doğruluk ve kesinlik değerlerine sahiptir, ancak performansı yeterince çok sayıda komşu seçmeye bağlıdır. Bu nedenle, OCSVM, karşılaştırmamızda önemli bir ayarlama gerektirmeden, az miktarda veri performansı ile kötü amaçlı röle algılama problemi için iyi bir adaydır. İkinci gözlem, performans hakkında incelenen denetimsiz öğrenme yöntemlerinin denetimli öğrenme ve geleneksel yöntemlerle karşılaştırılmasıdır. Bu yöntemlerin ortalama tespit doğruluğu göz önüne alındığında, LOF ve OCSVM denetimli öğrenmeyi (NN, SVM, Random forest) ve geleneksel yöntemleri (Bölüm 6.3.'de detaylandırılmıştır) geride bırakmaktadır. Bunun nedeni, önerilen yaklaşımın performans karşılaştırması için önemli bir zorluk olan eğitim aşamasındaki veri esnekliğidir.

Tablo 6.1. Tüm algoritmaların modülasyon etkisinin doğruluk değerleri üzerindeki etkisi (SNR = 15 dB, Bant genişliği = 1.4 MHz, M = 6).

	QPSK			16QAM			64QAM		
	OCSVM	LOF	*i*Forest	OCSVM	LOF	*i*Forest	OCSVM	LOF	*i*Forest
Secure Relay	0.75	0.78	0.62	0.82	0.89	0.57	0.84	0.88	0.56
Garbling Attack (A1)	0.95	0.98	0.96	0.93	0.98	0.97	0.97	0.98	0.92
Regenerative Attack (A2)	0.99	0.98	0.91	0.97	0.98	0.95	0.98	0.97	0.87
False Data Injection Attack (A3)	0.99	0.98	0.98	0.98	0.98	0.94	0.98	0.98	0.97

Tablo 6.2. Tüm algoritmaların eğitim veri sayısının doğruluk değerleri üzerindeki etkisi (SNR = 15 dB, Bant genişliği = 1.4 MHz, M = 6,25,100).

	6 (1.4 MHz)			25 (5MHz)			100 (20MHz)		
	OCSVM	LOF	*i*Forest	OCSVM	LOF	*i*Forest	OCSVM	LOF	*i*Forest
Secure Relay	0.75	0.78	0.62	0.88	0.92	0.68	0.91	0.99	0.73
Garbling Attack (A1)	0.95	0.98	0.96	0.95	0.98	0.96	0.96	0.99	0.96
Regenerative Attack (A2)	0.99	0.98	0.91	0.99	0.99	0.91	0.99	0.99	0.92
False Data Injection Attack (A3)	0.99	0.98	0.98	0.99	0.99	0.98	0.99	0.99	0.98

6.3. İstatistiksel Öğrenme Yöntemleri

Bu bölümde istatistiksel öğrenme yöntemlerinin, tanımlanan atak türlerini sınıflandırmadaki performansı incelenmektedir. İstatistiksel yaklaşımın temelleri

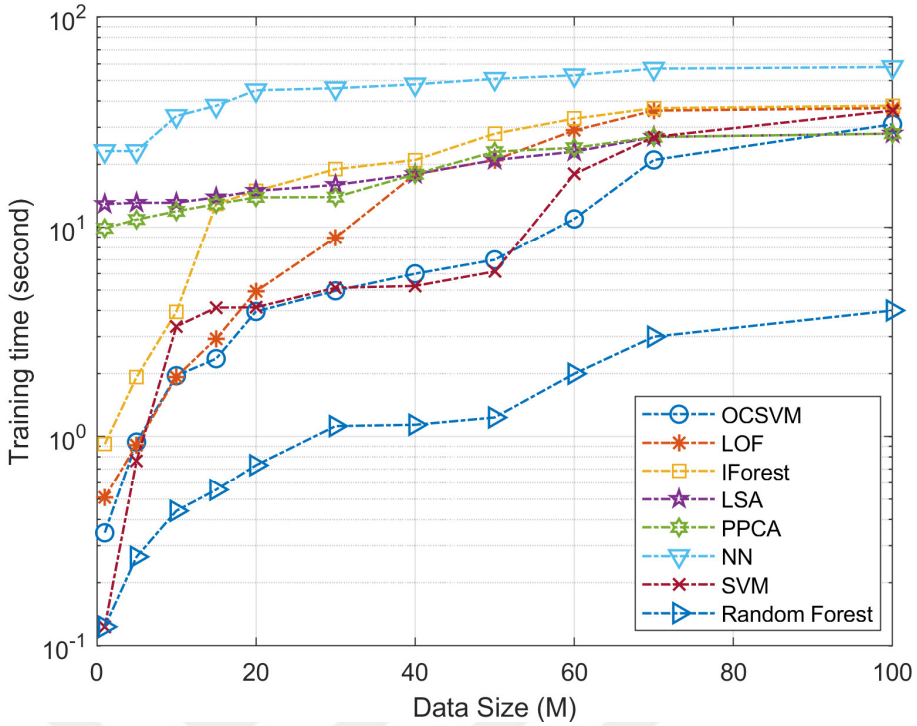
Bölüm 4.4'de veri kümesinin özellikleri üzerinden açıklanmıştır. Geleneksel olarak kullanılan bir yaklaşım olmasından dolayı bu tezde, istatistiksel yaklaşımın başarısı diğer yaklaşımlarla beraber sunulmakta, çalışmanın faydasının mevcut yaklaşımlara karşın daha yüksek performansa sahip olduğu farklı perspektiften açıklanmaktadır. PPCA yönteminin öz değerler ile olasılıksal tahminlerde bulunarak sınıflandırma sonuçları, eğitim veri kümesi içerisindeki öz değerleri ile hesaplanmaktadır. Sonuçları detaylandırdığımızda yöntemin ayrıştırma noktasında doğruluk değerinin %60'ın altında kaldığı görülmektedir. Bunun nedeni veri kümesindeki her sınıfın tek bir öz değerle ifade edilmesidir. Tek bir öz değer istatistiksel olarak sınıflandırmayı geçersiz kılmaktadır.

Benzer şekilde LSA yöntemi ise (en küçük kareler yöntemi) eğitim veri kümesi içerisinden rastgele seçilen örnekler içerisindeki mesafeyi öğrenerek sınıflandırır. Ancak sunulan atak modelleri ve sistem modeli içerisinde verilerin geometrik olarak ayrıştırılabilmesinin zorluğu hâlihazırda problemin temeli olarak sunulmaktadır. Bu nedenle bu yöntemin tüm örnekleri aykırı yani atak olarak sınıflandırması öngörüldüğü gibi gerçekleşmiştir. LSA yöntemi tüm test örneklerini aykırı olarak kategorilendirmekte ve bu noktada güvenilir röleye ait örnekleri atak olarak sınıflandırarak tezde tanımlanan sistem ve atak modelleri üzerinde bir başarımının olmadığı anlaşılmaktadır.

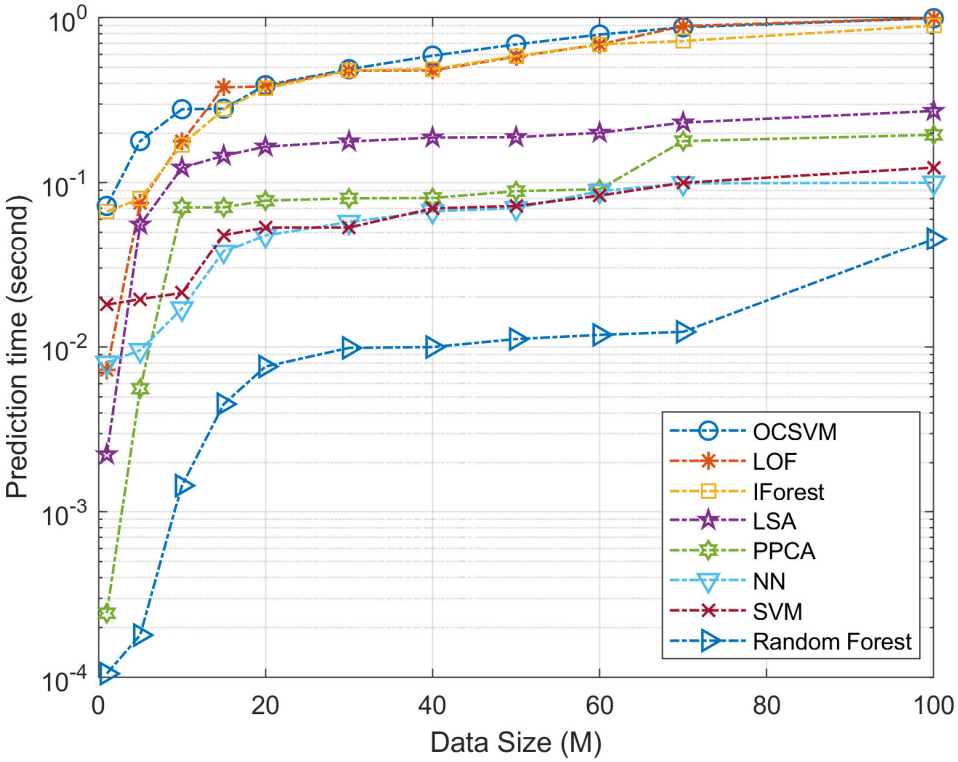
6.4. Yöntemlerin Performans ve Karmaşıklık Karşılaştırması

Algoritma uygulamalarındaki bir başka performans değerlendirmesi karmaşıklık analizidir. Bu çalışmada, atak tespitinde uygulanan makine öğrenimesi algoritmalarının karmaşıklık analizi olarak zaman ve donanım ihtiyaçları analiz edilmiştir. Simülasyonlar Intel (R) Core ™ i7-7500U, CPU 2.90 GHz ve 8GB RAM donanım özelliklerinde bilgisayarda gerçekleştirilmiştir.

Bu çalışmada uygulanan her algoritma için değişen veri boyutuna (M) karşın eğitim süresi elde edilmiş ve Şekil 6.20'de grafik olarak gösterilmiştir. Beklendiği üzere, daha büyük veri boyutu daha uzun eğitim süresine neden olmaktadır ve Random Forest, eğitim süresi açısından en hızlı algoritma olarak gözlemlenmektedir. İkinci olarak OCSVM ve SVM algoritmaları gelmektedir. Üçüncü LOF ve sırasıyla iForest, LSA, PPCA ve sonuncu olarak NN algoritması gelmektedir.



Şekil 6.19. Algoritmaların eğitim aşaması süreleri (modülasyon = QPSK, SNR = 15, bant genişliği = 1.4 MHz).

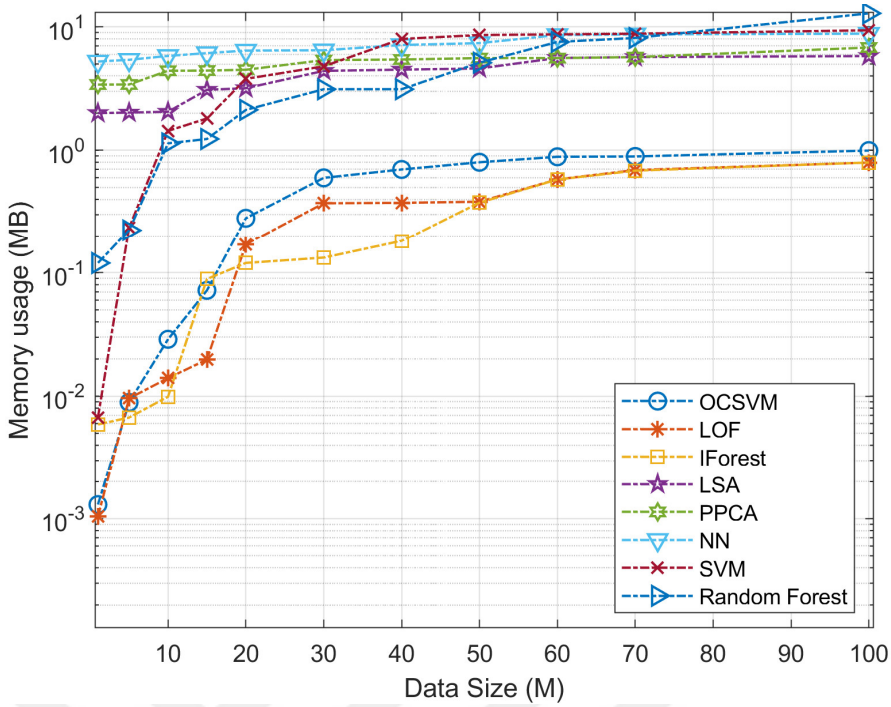


Şekil 6.20. Algoritmaların tahmin süreci süreleri (modülasyon = QPSK, SNR = 15, bant genişliği = 1.4 MHz).

Algoritmaların test fazı için zaman karmaşıklığı Şekil 6.20'de, test verilerinin atak veya güvenli olup olmadığına ya da hangi sınıfa ait olduklarına dair kararın verildiği tahmin aşamasını göstermektedir. Tahmin süresi maliyetinin denetimsiz algoritmalar için yaklaşık olarak aynı olduğu görülmektedir. LSA, PPCA, NN ve SVM ise tahmin sürelerinde birbirilerine yakınken Random Forest algoritmasının diğer algoritmalara göre en kısa süre tahminde bulunduğu görülmektedir. Eğitim aşaması sırasında algoritmaların bellek kullanımı da değişen veri boyutuna göre analizleri Şekil 6.21'de gösterilmektedir. Veri boyutu arttıkça, algoritmalar tarafından kullanılan bellek miktarı da artmaktadır; ancak en kötü durumda denetimsiz algoritmalar 1 MB bellek alanını aşmazken denetimli ve istatistiksel öğrenme yöntemlerinde 10 MB bellek kullanımları görülmektedir. Denetimsiz algoritmalar içerisinde OCSVM algoritması en büyük bellek alanını tüketmektedir. Tahmin aşamasındaki algoritmaların bellek kullanımı ise göz ardı edilecek kadar küçük olduğundan, grafiklerde gösterilmektedir.

Algoritmaların bellek tüketim maliyeti göz önüne alındığında, bu algoritmaların küçük kapasiteli cihazlarda bile uygulanmasının mümkün olabileceği gözlenmektedir. Röle ataklarını algılama sisteminin uygulanabilirliği bağlamında, algılama algoritmalarının bellek ve zaman gereksinimleri küçük kapasiteli kullanıcı ekipmanı için uygulanabilirlik göstermektedir. Şekil 6.19 ve Şekil 6.21'de gösterildiği gibi, 1.4 MHz bant genişliği ile 6 RB eğitim veri boyutu, modeli eğitmek için küçük bellek alanı ve süresi gerektirmektedir.

Algoritmaların eğitim süreleri ve tahmin süreleri açısından karmaşıklığı Şekil 6.19 ve Şekil 6.20'de verilmiştir. Algoritmaların zaman karmaşıklığında baskın bir faktör olan ve tahmin sürelerinden nispeten daha büyük olan eğitim süresi açısından OCSVM, diğer yöntemlerden daha iyi performans göstermektedir. Diğer denetimsiz yöntemler (LOF ve iForest), denetlenen yöntemlerle karşılaştırıldığında daha küçük eğitim sürelerine sahiptir. Denetimsiz yöntemler, denetimli ve istatistiksel yöntemlere kıyasla daha uzun tahmin sürelerine sahiptir. Denetimli ve istatistiksel yöntemlerin bellek kullanım performansı, Şekil 6.21'de görüldüğü gibi denetlenetimsiz yöntemlerden belirgin şekilde daha düşüktür. Bu çalışmada uygulanan algoritmalar için Random Forest, NN, SVM, LSA, PPCA, LOF, OCSVM ve iForest'in eğitim karmaşıklıkları sırasıyla O($LMlog\ (M)$), O(ML^2), O(L^3M), O(L^2M), O($L^2M + M^3$), O($Mk + ML$), O($M^2L + M^3$) ve O($LMlog\ (M)$).



Şekil 6.21. Algoritmaların eğitim aşamasında kullandıları bellek miktarı (modülasyon = QPSK, SNR = 15, bant genişliği = 1.4 MHz).

Tablo 6.2. Algoritmaların eğitim verisi ve özellik sayısı ile performans ilişkisi.

	Eğitim Süresi		Tahmin Süresi		Bellek Kullanımı	
Algoritma	$\nearrow M$	$\nearrow L$	$\nearrow M$	$\nearrow L$	$\nearrow M$	$\nearrow L$
LOF	Orta	Düşük	Orta	Düşük	Düşük	Düşük
OCSVM	Yüksek	Düşük	Yüksek	Düşük	Yüksek	Düşük
*i*Forest	Düşük	Düşük	Düşük	Düşük	Orta	Düşük
LSA	Düşük	Düşük	Orta	Düşük	Orta	Düşük
PPCA	Düşük	Yüksek	Düşük	Düşük	Düşük	Yüksek
NN	Orta	Düşük	Orta	Düşük	Orta	Yüksek
SVM	Orta	Yüksek	Düşük	Düşük	Orta	Yüksek
Random Forest	Düşük	Düşük	Düşük	Düşük	Orta	Yüksek

iForest algoritması, LOF algoritmasının karmaşıklığından daha büyüktür; fakat OCSVM algoritmasından daha az üstel karmaşıklığa sahiptir. PPCA yöntemi ise en yüksek karmaşıklığa sahiptir. LSA ve NN de benzer bir karmaşıklığa sahip olmakla birlikte aykırı tespit yöntemlerinden daha yüksektir.

Tablo 6.2, veri boyutu M ve özellik boyutu L olmak üzere algoritmaların eğitim süresini, tahmin süresini ve bellek kullanım bağımlılıklarını özetlemektedir. Öte yandan, veri boyutunun etkisi algoritmaya bağlı olarak değişmektedir.

7. SONUÇLAR VE ÖNERİLER

Bu tezde, işbirlikçi LTE-A ağlarda kötü amaçlı rölelerin fiziksel katmanda tespit edilebilmesi problemi üzerinde durulmuştur. Atakların tespiti, hedef düğümde röleden alınan sinyalin demodülasyon ve kanal tahmin işlemlerinden önce gerçekleştirilmiştir. İşbirlikçi röle sistemleri için ilk adımda, hedef düğümde alınan röle sinyali ile işbirliği yapmadan önce, atak tespiti yaparak kablosuz iletişim sistemlerinde önemli bir katkı sağlayacağı görülmektedir. Makine öğrenmesi tekniklerinin denetimli, denetimsiz ve istatistiksel öğrenme yaklaşımları kullanılarak fiziksel katmanda atak tespit problemine çözüm getirilmesi hedeflenmiştir. Daha spesifik olarak, aykırı algılama yaklaşımı altında OCSVM, LOF ve iFores algoritmaları denetimsiz bir öğrenme yöntemi olarak değerlendirilmiştir, SVM, NN ve Random forest ise denetimli sınıflandırma kategorisinde iken, son olarak PPCA, LSA istatistiksel yöntemler olarak değerlendirilmiştir. Veri karıştırma, yanlış veri enjeksiyonu ve yeni veri iletme atağı gibi röle ataklarını tespit etmek için SNR, modülasyon tipi, veri boyutu ekseninde algoritmaların kesinlik, doğruluk, AUC performansları üzerine etkileri araştırılmıştır. Bu tez çalışmasında, düşük SNR koşullarında bile yüksek kesinlik ve doğruluk performanslarına sahip tek sınıflı denetimsiz öğrenme yöntemleri kullanılarak kötü amaçlı röle ataklarının denetimli ve istatistiksel yöntemlere göre daha başarılı bir şekilde tespit edilebileceği gösterilmiştir. Denetimli algoritmaların daha fazla eğitim verisine ihtiyaç duyması, atak modellerinin eğitim verisi içerisinde olması ve kamaşıklığının daha yüksek olması nedeni ile pratikte uygulanabilirlik açısından etkili bulunmamıştır.

Tüm denetimsiz algoritmalar için AUC değerleri, atak tespit vakaları için 0.9'un üzerindedir. Tüm saldırı türleri için, LOF algoritmasının performansının OCSVM ve iForest algoritmalarından daha üstün olduğu görülmüştür. Kötü amaçlı rölelerin algılanması için algoritmaların doğruluk performansı, LTE-A ağındaki modülasyon tipi (QPSK, 16-QAM, 64-QAM) değiştirildiğinde önemli ölçüde etkilenmemiştir. Ayrıca, değişen veri boyutu M'nin saldırı tespit performansı üzerindeki etkisinin önemsiz olduğu gözlenmiştir. Ancak denetimli algoritmalar için tersi durum söz

konusudur. Eğitim veri miktarı düştüğünde doğruluk ve keskinlik değerleri de düşmektedir. Bu çalışma, eğitim aşamasında minimum kaynakların, yani LTE-A sistemlerinde en düşük bant genişliğine sahip iken ve bir çerçevede taşınabilecek en düşük RB miktarının, kötü niyetli röle ataklarını etkili bir şekilde tespit etmek için denetimsiz algoritmaların kullanılabileceğini göstermiştir. Sonuçlar, bu algoritmaların ayrıva küçük kapasiteli cihazlarda bile uygulanmasının mümkün olduğunu göstermektedir.

Bu çalışmanın en büyük katkısı, LTE-A ağındaki kötü niyetli röle düğümlerinin etkili bir şekilde algılanması için denetimsiz öğrenme yaklaşımının, aykırı algılama problemi için uygulanabilirliğinin gösterilmesidir. Ayrıca, geleneksel istatistiksel yaklaşımlar kullanılarak tespit edilmesi oldukça zor olan fiziksel katmandaki kötü amaçlı ve güvenli rölelerin temel bant sinyal özelliklerinin, ML tabanlı yaklaşımlarda, özelliklerin doğru seçilmesiyle ayırt edilebildiğinin gösterilmesidir

sistemlerinde en düşük bant genişliğine sahip iken ve bir çerçevede taşınabilecek en düşük RB miktarının, kötü niyetli röle ataklarını etkili bir şekilde tespit etmek için denetimsiz algoritmaların kullanılabileceğini göstermiştir. Sonuçlar ayrıca bu algoritmaların küçük kapasiteli cihazlarda bile uygulanmasının mümkün olduğunu göstermektedir.

Bu çalışmanın en büyük katkısı, LTE-A ağındaki kötü niyetli röle düğümlerinin etkili bir şekilde algılanması için denetimsiz öğrenme yaklaşımının, aykırı algılama problemi için uygulanabilirliğinin gösterilmesidir. Ayrıca, geleneksel istatistiksel yaklaşımlar kullanılarak tespit edilmesi oldukça zor olan fiziksel katmandaki kötü amaçlı ve güvenli rölelerin temel bant sinyal özelliklerinin, ML tabanlı yaklaşımlarda, özelliklerin doğru seçilmesiyle ayırt edilebildiğini gösterilmiştir.

KAYNAKLAR

[1] Xing J., Lv T., Zhang X., Cooperative Relay Based on Machine Learning for Enhancing Physical Layer Security, *IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Istanbul, Turkey, 1-6, 2019.

[2] Deng Z., Sang Q., Gao Y., Cai C., Optimal Relay Selection for Wireless Relay Channel with External Eavesdropper: a NN-based Approach, *IEEE/CIC International Conference on Communications in China (ICCC)*, Beijing, China, 515-519, 2018.

[3] Wang X., Liu F., Data-Driven Relay Selection for Physical-Layer Security: A Decision Tree Approach, *IEEE Access*, 2020, 8, 12105-12116.

[4] Carreño A., Inza I., Lozano J. A., Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework, *Artificial Intelligence Review*, 2019, 3575–3594.

[5] Riyaz S., Sankhe K., Ioannidis S., Chowdhury K., Deep Learning Convolutional Neural Networks for Radio Identification, *IEEE Communications Magazine*, 2018, **56**(9), 146-152.

[6] Kulin M., Kazaz T., Moerman I., De Poorter E., End-to-End Learning From Spectrum Data: A Deep Learning Approach for Wireless Signal Identification in Spectrum Monitoring Applications, *IEEE Access*, 2018, 6, 18484-18501.

[7] Xiao L., Li Y., Han G., Liu G., Zhuang W., PHY-Layer Spoofing Detection With Reinforcement Learning in Wireless Networks, *IEEE Transactions on Vehicular Technology*, 2016, **65**(12), 10037-10047.

[8] Faria E.R., de Leon P., Ferreira Carvalho A.C., Gama J., MINAS: multiclass learning algorithm for novelty detection in data streams, *Data Min Knowl Disc*, 2016, **30**(3), 640–680.

[9] Géron A., *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, USA: O'Reilly, 2017.

[10] Zhu, X., & Goldberg, A. B., Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 2009, **3**(1), 1-130.

[11] Agrawal S., Agrawal J., Survey on anomaly detection using data mining techniques, *Procedia Computer Science*, 2015, 60, 708-713.

[12] Domingues R., Filippone M., Michiardi P., Zouaoui J., A comparative evaluation of outlier detection algorithms: Experiments and analyses, *Pattern Recognition*, 2018, 74, 406-421.

[13] Ozay, M., Esnaola, I., Vural, F. T. Y., Kulkarni, S. R., & Poor, H. V. Machine learning methods for attack detection in the smart grid. IEEE transactions on neural networks and learning systems, 2015, **27**(8), 1773-1786.

[14] Aggarwal, C. C., Outlier analysis. In Data mining, *Springer*, Cham, 2015, 237-263.

[15] Mao Y., Wu M., Tracing Malicious Relays in Cooperative Wireless Communications, *IEEE Transactions on Information Forensics and Security*, 2007, **2**(2), 198-212.

[16] Lo L., Huang W., Misbehavior Detection Without Channel Information in Cooperative Networks, *IEEE Vehicular Technology Conference (VTC Fall)*, San Francisco, CA, 1-5, 2011.

[17] Hou W., Wang X., Refaey A., Misbehavior detection in amplify-and-forward cooperative OFDM systems, *IEEE International Conference on Communications (ICC)*, Budapest, 5345-5349, 2013.

[18] Lv. T., Yin Y., Lu Y., Yang S., Liu E., Clapworthy G., Physical Detection of Misbehavior in Relay Systems With Unreliable Channel State Information, *IEEE Journal on Selected Areas in Communications*, 2018, **36**(7), 1517-1530.

[19] Beres E., Adve R., Selection cooperation in multi-source cooperative networks, *IEEE Trans. Wireless Commun*., 2008, 7, 118-127.

[20] Zhang Y., Xu Y., Cai Y., Relay selection utilizing power control for decodeand-forward wireless relay networks, *Proc. 2nd International Conference on Signal Processing and Communication Systems (ICSPCS 2008)*, Gold Coast,Australia, Dec. 2008.

[21] Jing Y., Jafarkhani H., Single and multiple relay selection schemes and their achievable diversity orders, *IEEE Trans. Wireless Commun*., 2009, 8, 1414-1423.

[22] Wang B., Han Z., Liu K. J. R., Distributed relay selection and power control for multiuser cooperative communication networks using stackelberg game, *IEEE Trans. Mobile Comput*., 2009, 8, 975-990.

[23] Nam S., Vu M., Tarokh V., Relay selection methods for wireless cooperative communications, *Proc. Conf. on Inform. Sciences and Systems (CISS)*, March 2008.

[24] Wu N. E., Huang W. C., Li H. J., A novel relay selection algorithm for relaying networks, *Proc. IEEE Vehicular Technology Conference (Fall)*, Anchorage, Alaska, Sep. 2009.

[25] Yi Z., Kim I. M., Diversity order analysis of the decode and-forward cooperative networks with relay selection, *IEEE Trans. Wireless Commun*., 2008, **7**(5), 1792-1799.

[26] Ibrahim A., Sadek A. K., Su W., Liu K. J. R., Cooperative communications with relay selection: When to cooperate and whom to cooperate with?, *IEEE Trans. Wireless Commun*., 2008, 7, 2814-2827.

[27] Sadek A. K., Han Z., Liu K. J. R., An efficient cooperation protocol to extend coverage area in cellular networks, *Proc. IEEE Wireless Communications and Networking Conference*, Las Vegas, NV, Apr. 2008.

[28] Bletsas A., Lippman A., Reed D. P., A simple distributed method for relay selection in cooperative diversity wireless networks, based on reciprocity and channel measurements, *61st IEEE Vehicular Technology Conference (VTC'05)*, Stockholm, Sweden, May 2005.

[29] Zhao Y., Adve R. S., Lim T. J., Improving amplify-and-forward relay networks: optimal power allocation versus selection, *IEEE Trans, on Wireless Commun*., 2007, 6, 3114-3123.

[30] Bletsas A., Shin H., Win M. Z., Outage probability at arbitrary SNR with cooperative diversity, *IEEE Commun. Letters*, 11, 261-263.

[31] Bletsas A., Shin H., Win M. Z., Cooperative communications with outageoptimal opportunistic relaying, *IEEE Trans, on Wireless Commun*., 2007, 6, 3450-3460.

[32] Yeh E., Berry R., Throughput optimal control of cooperative relaying networks, *IEEE Trans. Inform*. Theory, special issue on Models, Theory and Codes for Relaying and Cooperation in Communication Networks, 2007, 53, 3827-3832.

[33] Wei Y., Yu F. R., Song M., Leung V. C. M., Energy efficient distributed relay selection in wireless cooperative networks with finite state markov channels, *Proc. IEEE Globecom*, Honolulu, Hawaii, USA, Dec. 2009.

[34] Halabian H., Lambadaris I., Lung, Srinivasan A., Throughputoptimal relay selection in multiuser cooperative relaying networks, *IEEE MILCOM 2010*, San Jose, CA, USA, Nov. 2010.

[35] Urgaonkar R., Neely M. J., Delay-limited cooperative communication with reliability constraints in wireless networks, *Proc. IEEE INFOCOM 2009*, Rio De Janeiro, Brazil, Nov. 2010.

[36] Iwamura M., Takahashi H., Nagata S., Relay Technology in LTE-Advanced, *NTT DoCoMo Technical Journal*, 2010, **12**(2), 29-36.

[37] Nourizadeh H., Nourizadeh S., Tafazolli R., Performance Evaluation of Cellular Networks with Mobile and Fixed Relay Station, *IEEE Vehicular Technology Conference VTC*, Montreal, Canada, September 2006.

[38] Shrestha S., Chang K., Analysis of Outage Capacity Performance for Cooperative DF and AF Relaying in Dissimilar Rayleigh Fading Channels, *IEEE International Summit on Information Theory ISIT*, Toronto, Canada, July 2008.

[39] Walke B. H., Wijaya H., Schultz D. C., Layer-2 Relays in Cellular Mobile Radio Networks. *IEEE Vehicular Technology Conference VTC*, Melbourne, Australia, May 2006.

[40] Djenouri D., Khelladi L., Badache N., A survey of security issues in mobile ad hoc and sensor networks, *IEEE Commun. Surveys & Tutorials*, 2005, **7**(4), 2-28.

[41] Yengi Y., Kavak A., Arslan H., Küçük K., Yiğit H., Malicious Relay Node Detection with Unsupervised Learning in Amplify-Forward Cooperative Networks, *2019 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sakhier, Bahrain, 1-5, 2019.

[42] Xiao L., Wan X., Xiaozhen l., Yanyong Z., DiWu, IoT Security Techniques Based on Machine Learning, *arXiv preprint*, 2018, **1**(1), 1-20.

[43] Maini V., Machine Learning for Human, 19 August 2017. [Online]. Available: https://medium.com/machine-learning-for-humans/reinforcement-learning-6eacf258b265. (Ziyaret tarihi: 14 Temmuz 2020).

[44] Tang T. A., Mhamdi L., McLernon D., Zaidi S. A. R., Ghogho M., Deep learning approach for network intrusion detection in software defined networking, *IEEE*, Fez, 2016.

[45] Garc´ıa-Teodoro P., D´ıaz-Verdejo J., Maci´a-Fern´andez G., V´azquez E., Anomalybased network intrusion detection: Techniques, systems and challenges, *Computers and Security*, 2009, **28**(1-2), 18–28.

[46] Atli B. G., Miche Y., Kalliola A., Oliver I., Holtmanns S., Lendasse A., Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space, *Cognitive Computation*, 2018, **10**(5), 848–863.

[47] Boutaba R., Salahuddin M. A., Limam N., Ayoubi S., Shahriar N., Estrada-Solano F., Caicedo O. M., A comprehensive survey on machine learning for networking: evolution, applications and research opportunities, *Journal of Internet Services and Applications*, 2018.

[48] Alpaydin E., *Introduction to Machine Learning*, 3rd ed. MIT Press, 2014.

[49] Loss Functions, https://mlcheatsheet.readthedocs.io/en/latest/loss functions.html (Ziyaret tarihi: 1 Ağustos 2020).

[50] Branco P., Torgo L., Ribeiro R. P., Relevance-based evaluation metrics for multi-class imbalanced domains, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Cham*, 698–710, 2017.

[51] Karagod V., How to Handle Imbalanced Data: An Overview, https://www.datascience.com/blog/imbalanced-data (Ziyaret tarihi: 13 Temmuz 2020).

[52] Seif G., Handling Imbalanced Datasets in Deep Learning, https://towardsdatascience.com/handling-imbalanced-datasets-in-deep-learning-f48407a0e758 (Ziyaret tarihi: 1 Temmuz 2020).

[53] Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., SMOTE: Synthetic Minority Over-sampling Technique, *Tech. Rep*., 2002.

[54] He H., Bai Y., Garcia E. A., Shutao L., ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).* Hong Kong, China: IEEE, 1322–1328, 2008.

[55] Boutaba R., Salahuddin M. A., Limam N., Ayoubi S., Shahriar N., Estrada-Solano F., Caicedo O. M., A comprehensive survey on machine learning for networking: evolution, applications and research opportunities, *Journal of Internet Services and Applications*, 2018.

[56] Dambre J., *Lecture 5: Machine learning in practice*, University of Ghent, Belgium, 2017.

[57] Drakos G., Cross-Validation, https://towardsdatascience.com/cross-validation-70289113a072 (Ziyaret tarihi: 4 Temmuz 2020).

[58] Burlutskiy N., Petridis M., Fish A., Chernov A., Ali N., An Investigation on Online Versus Batch Learning in Predicting User Behaviour, *Research and Development in Intelligent Systems XXXIII. Springer International Publishing*, 2016, 11, 135–149.

[59] Dias L., Cerqueira J. J. F., Assis K. D. R., Almeida R. C., Using artificial neural network in intrusion detection systems to computer networks, *2017 9th Computer Science and Electronic Engineering Conference (CEEC)*, 2017, 145–150.

[60] Kingma D. P., Ba J. L., Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations*, 2014.

[61] Reddi S. J., Kale S., Kumar S., On the convergence of Adam and Beyond, *ICLR*, 2018.

[62] Breiman L., Friedman J., Stone C. J., Olshen R., Classification and Regression Trees. *Taylor & Francis*, 1984.

[63] Dhaene T., Decision trees and Random Forests, *University of Ghent*, Belgium, 2017.

[64] Vapnik V., *Statistical Learning Theory*. Wiley, New York, 1998.

[65] Cristianini. N., Shawe-Taylor. J. An Introduction to Support Vector Machines, *Cambridge University Press*, Cambridge, 2000.

[66] Shawe-Taylor J., Cristianini N., Further results on the margin distribution. *In Proceedings of the 12th Annual Conference on Computational Learning Theory ACM Press*, New York, 1999.

[67] Boser B.E., Guyon I.M., Vapnik V.N. A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory ACM Press*, Pittsburgh, PA, 144–152, 1992.

[68] Schölkopf B., Williamson R.C., Smola A.J., Shawe-Taylor J., Platt J.C., Support vector method for novelty detection, *Advances in neural information processing systems*, 582-588, 2000.

[69] Breunig M.M., Kriegel, Ng R.T., Sander J., LOF: identifying density-based local outliers, *ACM sigmod record,* 2000, **29**(2), 93–104, 2000.

[70] Liu F.T., Ting K.M., Zhou Z.H., Isolation-based anomaly detection, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2013, **6**(1), 1-39.

[71] Knorr E. M., Ng R. T., Tucakov V., Distance-based outliers: algorithms and applications. *The VLDB Journal,* 2000, 8, 3-4, 237–253.

[72] Ghoting A., Otey M. E., Parthasarathy S. Loaded: Link-based outlier andanomaly detection in evolving data sets. *ICDM'04: Proceedings of the Fourth IEEE International Conference on Data Mining. IEEE Computer Society*, Washington, DC, USA, 387–390, 2004.

[73] Abe N., Zadrozny B., Langford, J. Outlier detection by active learning. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, Philadelphia, PA, USA, 504–509, 2006.

[74] Quinn J. A., Sugiyama M. A least-squares approach to anomaly detection in static and sequential data. *Pattern Recognition Letters*, 2014, 40, 36-40.

[75] Tipping M. E., Bishop C. M., Probabilistic principal component analysis. *Journal of the Royal Statisitical Scoiety*, **B**(61),611–622, 1999.

[76] The Mathworks, Inc., MATLAB:2017b, MATLAB version 9.3.0.713579 (R2017b), Natick, Massachusetts, 2017.

[77] Pedregosa F., Varoquaux F., Michel A., Thirion B., Grisel O., Vanderplas J., Scikit-learn: Machine learning in Python, *Journal of machine learning research,* 12, 2825-2830, 2011.

KİŞİSEL YAYIN VE ESERLER

[1] **Yengi Y**., Kavak A., Arslan H., "Physical Layer Detection of Malicious Relays in LTE-A Network Using Unsupervised Learning," *IEEE Access*, 8, 154713-154726, 2020.

[2] **Yengi Y**., Kavak A., Arslan H., Küçük K., Yiğit H., Malicious Relay Node Detection with Unsupervised Learning in Amplify-Forward Cooperative Networks, *2019 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sakhier, Bahrain, 1-5, 2019.

[3] **Yengi, Y**., Khan S.A, Küçük K., Design and performance analysis of information centric network for Internet of Things, *25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey, 1-4, 2017.

[4] **Yengi Y**. K. Küçük, Context aware internet of things for large scale data analytics, *International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkey, 702-706, 2017.

ÖZGEÇMİŞ

İlk, orta ve lise öğrenimini İstanbul'da tamamladı. 2006 yılında girdiği Sakarya Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü'nden 2010 yılında Bilgisayar Mühendisi olarak mezun oldu. 20014-2016 yılları arasında, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı'nda Yüksek Lisans öğrenimini tamamladı. 20016-2020 yılları arasında, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı'nda Doktora öğrenimini tamamladı.