Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Derleme Makalesi

Derin Öğrenme Modellerinde Mahremiyet ve Güvenlik Üzerine Bir Derleme Çalışması

Gülsüm YİĞİT [a,*], Ayşe KALE [b]

[a] *Bilgisayar Mühendisliği, Mühendislik ve Doğa Bilimleri Fakültesi, Kadir Has Üniversitesi, İstanbul, Türkiye*
[b] *Bilgisayar Mühendisliği, Mühendislik-Mimarlık Fakültesi, Beykent Üniversitesi, İstanbul, Türkiye*
** Sorumlu yazarın e-posta adresi: gulsum.yigit@khas.edu.tr*
DOI: 10.29130/dubited.864635

ÖZ

Son dönemlerde derin öğrenmedeki devrim niteliğindeki gelişmeler ile birlikte yapay zekaya yönelik beklentiler gün geçtikçe artmaktadır. Konuşma tanıma, doğal dil işleme (NLP), görüntü işleme gibi birçok alanda etkin bir şekilde uygulanabilen bir araştırma alanı olan derin öğrenme klasik makine öğrenmesi ile karşılaştırıldığında daha yüksek başarı göstermektedir. Derin öğrenme ile geliştirilen modellerde eğitim ve tahminleme sırasında büyük miktarda veri kullanılmakta ve kullanılan veriler kişisel verilerden oluşabilmektedir. Bu verilerin işlenmesi sırasında kişisel verilerin korunması kanununa (KVKK) aykırı olmaması oldukça önemlidir. Bu nedenle verilerin gizliliği ve güvenliğinin sağlanması oldukça önemli bir husustur. Bu çalışmada, derin öğrenme modelleri geliştirilirken yaygın kullanılan mimariler verilmiştir. Verilerin gizliliği ve güvenliğini artırmak için literatürde yaygın olarak karşılaşılan güvenli çok partili hesaplama, diferansiyel mahremiyet, garbled devre protokolü ve homomorfik şifreleme araçları özetlenmiştir. Çeşitli sistem tasarımlarında kullanılan bu araçların yer aldığı güncel çalışmalar taranmıştır. Bu çalışmalar, derin öğrenme modelinin eğitim ve tahminleme aşamasında olmak üzere iki kategoride incelenmiştir. Literatürdeki çeşitli modeller üzerinde uygulanabilen güncel saldırılar ve bu saldırılardan korunmak amacıyla geliştirilen yöntemler verilmiştir. Ayrıca, güncel araştırma alanları belirlenmiştir. Buna göre, gelecekteki araştırma yönü kriptografik temelli yöntemlerin karmaşıklığının azaltılması ve geliştirilen modelin güvenilirliğini belirlemek için çeşitli ölçme ve değerlendirme yöntemlerinin geliştirilmesi yönünde olabilir.

Anahtar Kelimeler: *Derin Öğrenme, Güvenli Çok Partili Hesaplama, Diferansiyel Mahremiyet, Homomorfik Şifreleme, Garbled Devreler Protokolü.*

A Review Study on Privacy and Security in Deep Learning Models

ABSTRACT

With the advanced progress in deep learning in recent times, expectations of artificial intelligence are increasing day by day. Deep Learning, a research area, applied effectively in many areas such as speech recognition, Natural Language Processing (NLP), and image processing, shows higher success than classical machine learning algorithms. Large amounts of data are employed during the training and prediction stages of deep learning models, in which the data may consist of the user's sensitive data. Therefore, it mustn't contradict the principle of protecting personal data while training and predicting. Consequently, it is crucial to ensure the privacy and security of data. In this study, various deep learning architectures are described. The widely used cryptographic tools applied to preserve the privacy and security of data are summarized. These are homomorphic encryption, differential privacy, secure multi-party computation, and garbled circuits. Recent works considering the privacy and security of deep learning models are examined under two different categories: train stage and

Geliş: 19/01/2021, Düzeltme: 12/07/2021, Kabul: 28/07/2021

prediction stage. In addition, the attacks against deep learning models and techniques for protection against these attacks are reviewed. Finally, open research areas are determined. Reducing the complexity of cryptographic-based models and developing evaluation methods to determine the model's privacy and security stand out as future research areas.

Keywords: *Deep learning, Secure multiparty computation, Differential privacy, Homomorphic encryption, Garbled circuits protocol.*

I. GİRİŞ

Teknolojinin gün geçtikçe ilerlemesi ve artan internet kullanımı ile birlikte her gün çok büyük miktarda ses, görüntü, metin gibi farklı türlerde veriler üretilmektedir. Bu veriler üzerinde son dönemlerde devrim niteliğinde iyileşmeler kaydeden derin öğrenme yöntemleri yaygın olarak kullanılmaktadır. Ses, görüntü ve metin verilerini sınıflandırma, görsel ve metin soru cevaplama sistemleri, oyun oynama, makine çeviri yöntemleri de dahil olmak ve bu alanlarla sınırlı olmamak üzere çeşitli önemli ve sık karşılaşılan problemlerde çözüm olarak kullanılmaktadır. Derin öğrenme modellerinin doğruluk başarısı verinin büyüklüğü ile doğru orantılıdır.

Derin öğrenmenin birçok alanda kullanımının artması ile birlikte klasik makine öğrenmesine göre daha yüksek performanslı sonuçlara ulaşılmıştır. Derin öğrenme modelinin eğitiminde kullanılan eğitim verileri kişilerin konum bilgileri, sosyal medyadan toplanan ses, görüntü, video verileri vb. olmak üzere kişisel bilgilerden oluşmaktadır [1-4].

Derin öğrenme modellerinin eğitiminde hassas veri kullanıldığında çeşitli gizlilik sorunları ile karşılaşılabilmektedir. Örneğin, sosyal medyadaki milyonlarca kullanıcının görüntü, mesaj ve video verileri gizlilik problemleri ile karşı karşıyadır. Bu verileri tutan şirketlerin verileri ne kadar süre ile tuttukları bilinmediğinden sosyal medya kullanıcıları verilerin silindiğinden hiçbir zaman emin olamamaktadır. Ayrıca, bu verilerin ne amaçla kullanılacağı kullanıcılar tarafından kontrol edilememektedir.

Bu çalışmada, Bölüm 2'de, literatürde yaygın bir şekilde kullanılan derin öğrenme mimarileri verilmiştir. Derin öğrenme temelli modellerde en çok kullanılan kriptografik araçlar Bölüm 3'te özetlenmiştir. Bölüm 4'te, derin öğrenme mimarileri kullanılarak geliştirilen modellerde güvenliği ve gizliliği sağlamak üzerine literatürde yer alan çalışmalar eğitim ve tahminleme sırasında olmak üzere iki kategoride özetlenmiştir. Bölüm 5'te ise derin öğrenme modelleri üzerinde uygulanan saldırılar incelenmiş ve bu saldırılara yönelik tasarlanan literatürdeki korunma yöntemleri verilmiştir. Gelecekteki araştırma yönleri Bölüm 6'da incelenmiş ve çalışmanın sonuç bölümü Bölüm 7'de bulunmaktadır.

II. DERİN ÖĞRENME

Derin öğrenme mimarilerinin yapısı temel olarak girdi ve çıktı katmanları ve bu katmanların arasında bulunan gizli katmanlardan oluşmaktadır. Çeşitli veri türleri üzerinde uygulanabilen bu mimarilerin temel çalışma mekanizması bir katmandan elde edilen çıktı vektörünün takip eden katmanın girdisi olarak işlenmesidir.

A. KONVOLÜSYONEL SİNİR AĞLARI (CNN)

Konvolüsyonel sinir ağları (CNN) yüz tanıma, resim sınıflandırma, nesne belirleme gibi bilgisayarla görü problemleri başta olmak üzere çeşitli problemler üzerinde uygulanan bir derin öğrenme

mimarisidir [1, 2, 5]. Tek katmanlı sinir ağları modellerine kıyasla CNN mimarisi ile oluşturulan modeller daha yüksek performans değerlerine ulaşmaktadır. CNN ağı genel olarak üç farklı bölümden oluşmaktadır.

Birinci bölümde bir veya daha fazla konvolüsyon işlemi gerçekleştirilir. Konvolüsyon katmanı, konvolüsyonel sinir ağlarının ilk katmanıdır. Konvolüsyon işleminin ilk değişkeni $\boldsymbol{x}$ girdisi ve ikinci değişkeni $\boldsymbol{w}$ kernel matrisidir. Kernel matrisi, girdi üzerinde konvolüsyon işlemi yapılarak dolaşır ve sonuçlar bir öznitelik haritasını oluşturur. Farklı kernel matrisleri ile konvolüsyon işlemi tekrarlanır ve her oluşturulan öznitelik haritası bir özellik türünü tespit eder. Bu aşamada öznitelik haritasının gerçek giriş verisine göre boyutunu korumak için piksel ekleme yöntemleri uygulanabilir ve sonrasında doğrusal olmayan bir aktivasyon fonksiyonu (ReLu, Tanh, Sigmoid vb.) ile bir aktivasyon işlemi gerçekleştirilir. Bu aşamada amaç doğrusallık problemini ortadan kaldırarak sistemin tek bir perceptron gibi davranmasına engel olmaktır.

İkinci bölümde ise bir havuzlama fonksiyonu ile önceki aşamanın çıktısı değiştirilir. Havuzlama fonksiyonu, çıktının yakınlarındaki bilgilerin bir özeti olacak şekilde (maksimum havuzlama, ortalama havuzlama, L2- Norm havuzlama vb.) değiştirilmesini temel alır. Farklı boyutlardaki girdileri işleyebilmek için önemli bir aşamadır ve girdi boyutundan bağımsız olarak aynı boyutta bir özet haritası elde edilir. Temelde havuzlama katmanı bir örnekleme işlemidir.

Üçüncü aşama ise tam bağlantı katmanı olarak adlandırılır ve önceki bölümlerin çıktıları, her bir nöronun tüm nöronlara bağlı olduğu tam bağlantılı bir sinir ağına tabi tutulur. Tam bağlantı katmanı konvolüsyonel sinir ağlarının sonuna eklenen bir katman olup öğrenme işleminin yapıldığı önceki aşamaların skor sonuçlarını optimize etmek için kullanılır.

B. ÖZYİNELEMELİ SİNİR AĞLARI (RNN)

Okunan bir yazı kendisinden önce gelen her kelimenin anlaşılmasıyla bir anlam ifade etmektedir. Yapay sinir ağları kendisinden önce gelen bilgilerden çıkarım yapma kabiliyetine sahip değildir. Yapay sinir ağlarında karşılaşılan bu probleme Özyinelemeli Sinir Ağları (RNN) ile cevap bulunabilmektedir. Son dönemlerde RNN ile geliştirilen uygulamalar robotik, el yazısı tanıma, ses tanıma, müzik oluşturma ve diğer birçok temel görev için kullanılmaktadır [3,4,6,7]. Ayrıca, DNA verileri üzerinde gerçekleştirilen biyomedikal uygulamalarda [8,9], makine çevirisi sistemlerinde [10], metin sınıflandırma [11] gibi önemli NLP problemlerinde kullanılabilmektedir.

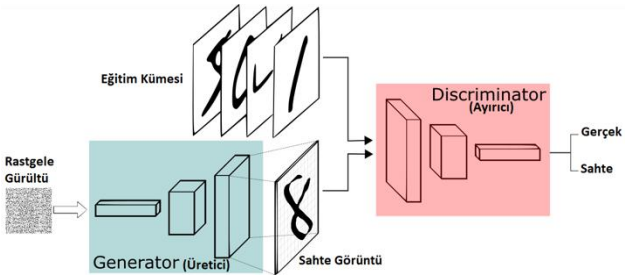
RNN ağlarında karşılaşılan en önemli problemlerden biri kaybolan gradyan problemidir. Genellikle çok eskiye dönük verilerden çıkarım yapmak gereken görevlerde ortaya çıkmaktadır. Ağın üzerinde ileri yönlü bir geçiş ile bir tahmin üretilir. Kayıp fonksiyonu ile elde edilen tahmin değeri karşılaştırılarak bu değer için bir hata değeri hesaplanır. Geriye dönük yayılımda hesaplanan bu hata değerinin kullanılmasıyla nöronların gradyan hesaplaması yapılır. Geri yayılım sırasında gradyan değerinin küçülüp 0'a yakınsaması durumunda modelin öğrenme gerçekleştirmesi mümkün değildir. Karşılaşılan bu problemi ele alan uzun-kısa vadeli bellek ağları (LSTM) [12] ve kapılı tekrarlayan hücreler (GRU) geliştirilmiştir. LSTM ve GRU ağları bilgi akışını düzenleyebilmek için RNN'lerden farklı olarak kapı mekanizmalarına sahiptirler. LSTM unutma, girdi ve çıktı kapıları olmak üzere üç kapıya sahiptir. GRU ise sıfırlama ve güncelleme kapıları ile RNN'deki gradyan problemlerine çözüm sağlamaktadır.

C. ÇEKİŞMELİ ÜRETİCİ AĞLAR (GAN)

Derin öğrenme alanında son dönemlerde yaygın kullanılan derin öğrenme mimarilerinden bir diğeri de Çekişmeli Üreticili Ağlar (GAN)'dır [13]. Bu ağlar, resimden resim elde edilmesi ve insan yüzü üretimi [14,15], insan-insan etkileşimi modellenmesi [16], yara dokularının görselleştirilmesi [17] ve yapay tıbbi veri üretilmesi [18] gibi birçok alanda uygulanabilmektedir. GAN ağları oyun senaryosu ile temeli atılmış bir derin öğrenme tekniğidir. Bu mimaride, Şekil 1'de görselleştirildiği üzere bir üretici ağ ve bir üretici ağın rakibi, ayırıcı ağ bulunur. Üretici tıpkı bir oto kodlayıcı gibi çalışarak veri

kümesinin dağılımına benzer ancak rastgele bir vektör ile sahte veri üretir. Ayırıcı ağ ise eğitim kümesinin örnekleri ve üretici ağın ürettiği örnekleri karşılaştırır ve bir olasılık değeri hesaplar. Bu olasılık değeri üretici ağın ürettiği örneğin eğitim setinden olma olasılığıdır.

GAN mimarisi ile öğrenme bir ödül teorisi ile gerçekleşir. Bu teoride, bir υ fonksiyonu ayırıcı ağın ödül fonksiyonudur. Üretici ağın ödül fonksiyonu ise -υ 'dir. Her iki taraf da alacağı ödülü büyütmeye odaklıdır.



Şekil 1.*Çekişmeli Üretici Ağ (GAN) modeli [25]*

GAN ağında eğitime rastgele bir gürültü ile başlanır ve üretici ayırıcıyı ürettiği verinin gerçek olduğuna ikna etmeye çalışır. Ayırıcı, gelen örnekleri sahte veya gerçek olarak sınıflandırarak üreticiyi veri setine uygun örnekler üretmesi için eğitir. Üretici de ayırıcıyı seçicilik konusunda eğitir ve her bir aşamada üretici ve ayırıcı birlikte eğitilmiş olur. Bu süreçte veri setinden küçük örnekler alınarak ayırıcı ağda stokastik gradyan azaltma uygulanarak geri yayılım algoritması ile ağırlıklar eğitilir. Bir yakınsama noktasına ulaşana kadar oyun devam eder. Yakınsama noktasına yaklaştıkça üreticinin örneklerinin de gerçek veriye benzerliği artar. Böylece, öğrenme süreci gerçekleştirilmiş olur. Tüm bu süreçler boyunca üretici ve ayırıcı, üretip ayırt etme işlemini kuralsız olarak gerçekleştirir.

III. KRİPTOGRAFİK ARAÇLAR

Derin öğrenme modellerinde kullanılan girdi ve elde edilen tahmin verilerinin güvenliğini ve gizliliğini güvence altına almak için çeşitli kriptografik temelli yöntemler yaygın olarak kullanılmaktadır. Kriptografi, bir metni bir anahtar ile şifreleyerek metni anlaşılmaz hale getiren ve şifrelenmiş metni tekrar şifresiz bir forma dönüştürülebilen yöntemlerin incelendiği bir bilim dalıdır. Simetrik şifreleme ve asimetrik şifreleme olmak üzere iki temel şifreleme tekniği bulunmaktadır.

Simetrik Şifreleme: Metni şifreleyen gönderici ve şifreli metni tekrar şifresiz metne dönüştürebilen alıcının aynı anahtara sahip olması durumudur. Hem alıcı hem de göndericinin sahip olduğu anahtar hem şifreleme hem de şifre çözümleme sırasında kullanılmaktadır.

Asimetrik Şifreleme: Simetrik şifrelemenin aksine metnin şifrelenmesi ve şifre çözümleme süreçlerinde alıcı ve göndericinin farklı anahtarlara sahip olması durumudur. Şifrelemede <*açık anahtar*> ve şifre çözümlemede <*gizli anahtar*>adı verilen (<*açık anahtar, gizli anahtar*>) anahtar çifti kullanılmaktadır.

Çalışmanın bu bölümünde, derin öğrenme modellerinde veri güvenliği ve gizliliğini sağlamak için literatürde yaygın olarak uygulanan kriptografik temelli yöntemler hakkında bilgi verilmiştir.

A. HOMOMORFİK ŞİFRELEME

Homomorfik şifreleme algoritması, şifreleme ve şifre çözümleme sırasında farklı anahtar kullanımına sahip asimetrik şifreleme temelli bir yöntemdir. Bu işlemler sırasında <açık anahtar (pk), gizli anahtar

(sk)> çifti kullanılmaktadır [26]. Şifreli metinlerin üzerinde şifre çözümleme işlemini yapmadan çeşitli hesaplamalar yapabilen çeşitli problemler üzerinde yaygın kullanılan bir yöntemdir.

Homomorfik şifreleme üç olasılıksal polinom zamanlı algoritmadan oluşmaktadır. Bunlar, (1) anahtar üretme, (2) şifreleme ve (3) şifre çözümleme işlemleridir.

Anahtar üretme: Gen (Üretici) güvenlik parametresi 1^k olan genel <açık anahtar (pk), gizli anahtar (sk)> çifti (pk, sk) ∈ K × K'dır.

Şifreleme: Şifreli metin C, E : K × M → C olduğu gibi açık anahtar K ve bir mesaj M ile elde edilir.

Şifre çözümleme: Şifreli metin C ve gizli anahtar K kullanılarak şifresiz metin M elde edilir.
D : K × C → M.

Şifreleme E_{pk}(M)=C, şifre çözümleme D_{sk}(C)=M ile gösterilebilir. Bu teknik ile (pk,sk) ← Gen(1^k) olmak üzere toplama (+) ve çarpma (·) işlemleri üzerinde bütün k, m_1, m_2 değerleri için aşağıdaki işlemler yapılabilir.

$$E_{pk}(m_1) \cdot E_{pk}(m_2) = E_{pk}(m_1 + m_2) \quad (1)$$

Homomorfik şifreleme, şifreli metinler üzerindeki hesaplama yapabilmesinden dolayı çeşitli Güvenli Çok Partili Hesaplama (SMPC) protokolünün tasarımında, şifrelenmiş veri tabanlarındaki sorgularda, anonim veri toplama gibi önemli uygulama alanlarında kullanılabilmektedir.

B. OBLİVİOUS TRANSFERİ PROTOKOLÜ

Oblivious Transferi Protokolü, 1981'de geliştirilen iki taraflı bir şifreleme protokolüdür [19]. Tablo 1'de görülen işlevselliğe sahip olan Oblivious Transferi, gönderen ve alıcı olmak üzere iki taraftan oluşmaktadır. Gönderenin iki girdisi x_0, x_1 ve alıcının ise girdi verisi olarak σ biti bulunmaktadır. Bu protokolünün çıkışında alıcı x_σ'yı öğrenmektedir, ancak hangi mesajı aldığına dair olan σ biti hakkında hiçbir bir veriye ulaşamamakta ve gönderici ise hiçbir bilgi öğrenememektedir.

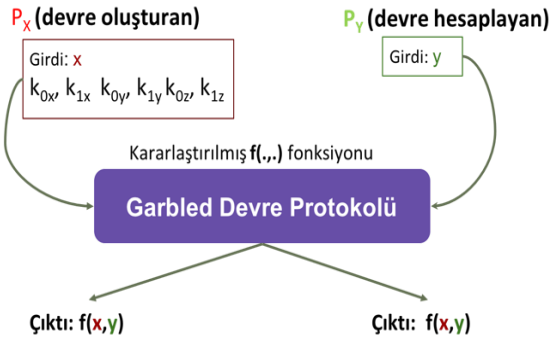
Tablo 1. *Oblivious transferi protokolü*

	Gönderen	**Alıcı**
Giriş	x_0, x_1	σ
Çıkış	⊥	x_σ

Oblivious transferi protokolü Yao milyonerler problemi, küme eşitlik testleri (DNA dizilimleri ile ebeveynlik testi), SMPC içeren protokoller ve garbled devreler protokolü gibi çeşitli kriptografik yöntemlerin geliştirilmesinde temel bir araç olarak kullanılmaktadır [20-23] .

C. GARBLED DEVRELER PROTOKOLÜ

Garbled devreler protokolü, iki taraflı güvenli hesaplama protokolleri geliştirilirken her problem üzerinde uygulanabilen matematiksel genel bir yaklaşımdır [24]. Güncel literatürdeki derin öğrenme temelli modellerde verilerin gizliliği ve güvenliğini artırmak amacıyla kullanılan kriptografik araçlardan biridir.



Şekil 2.*Garbled devreler protokolü temel çalışma mekanizması.*

Garbled devreler protokolü, x ve y girişleri olan sırasıyla P_X, P_Y tarafları üzerinden yürütülmektedir. Şekil 2'de Garbled devreler protokolü mekanizmasının temel prensibi verilmiştir. Buna göre, giriş-çıkış verileri olarak {0,1} değerlerine sahip NAND, AND, XOR-NOR, OR, NOT kapılarından oluşan bir devre olarak oluşturulabilen f(.) fonksiyonunu hesaplayabilmektir. Taraflardan biri buradaki işlevin hesaplanması için devreyi oluştururken, diğer taraf ise oluşturulan devreyi hesaplamaktadır.

Şekil 2'de olduğu gibi genel olarak, P_X kararlaştırılmış fonksiyon olan f'in şifreli halini kullanır. P_Y ise hesaplama veya P_X girdisi hakkında herhangi bir şey öğrenmeyen devrenin hesaplayıcısı olarak tanımlanabilir. Bu hesaplamada her girdi kapısı için bir Oblivious Transferi protokolü uygulanır.

Bu protokolde takip edilen adımlar aşağıda özetlenmiştir.

Taraflar: P_X (garbled devreyi oluşturan taraf) ve P_Y (garbled devreyi hesaplayan taraf)

Girdi: P_X= $\{x_0, x_1, \dots, x_n\}$ ve P_Y=$\{y_0, y_1, \dots, y_n\}$ verilerine sahiptir.
Çıkış kapısı her iki tarafın girdilerini herhangi bir üçüncü kişi veya birbirlerine sızdırmadan f(x_i, y_i) fonksiyonunu değerlendirir. Burada i = {1,2,...n}'dir.

Protokol:

- Tarafların veriler üzerinde hesaplaması için kararlaştırdıkları f fonksiyonu ilk olarak bir mantıksal devreye dönüştürülür ve her bir kapıda şifrelemede kullanılmak üzere rastgele anahtarlar üretilir.
- Üretilen rastgele anahtarların mantıksal ifadesi ile şifreleme dizgesi değiştirilerek bir doğruluk tablosu oluşturulur.
- P_X oluşturduğu doğruluk tablosunun karıştırılması ile elde edilen tabloyu ve kendi şifrelenmiş girdisini P_Y'e gönderir.
- P_Y, P_X'in karıştırılmış girdilerini ve garbled devreyi alır. Gönderilen veriler karıştırılmış bir şekilde ve şifreli olmasından dolayı P_Y devredeki giriş değerlerinin hangi girdi bit değerlerine karşılık geldiğini bilememektedir.
- Devredeki doğru girişi seçmek için P_Y girdisinin her bir değeri için P_X ile Oblivious Transferi protokolü yürütülür.

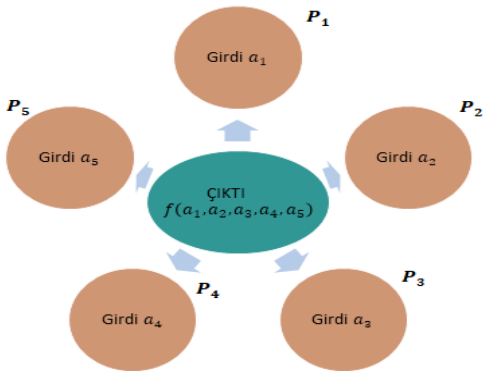
Sonuç:

- Yürütülen her bir Oblivious Transfer Protokolü sırasında P_X, P_Y'nin gönderdiği her bir bit değerine karşılık gelen rastgele bir dizesini gönderir. Oblivious Transferinin çalışma prensibinden dolayı taraflar arasında veri sızdırılması mümkün değildir. Dolayısıyla, P_X, P_Y' nin herhangi bir girdi değerini öğrenemez.

- P_Y, girdi bit değerlere karşılık gelen rastgele dizisini almayı bitirdiği zaman P_Y tarafı devreyi hesaplayabilmek için gereken bütün verilere sahiptir. Böylece, P_Y tarafı devreyi hesaplayarak elde ettiği hesaplama sonucunu P_X tarafı ile paylaşır.

D. GÜVENLİ ÇOK PARTİLİ HESAPLAMA (SMPC)

Derin öğrenme modelleri eğitim verisine sahip çoklu katılımcıların olması durumunda her bir katılımcının verisinin başka herhangi bir katılımcı ile paylaşılmaksızın genel bir derin öğrenme modeli eğitilebilir [25-27]. Bu tür eğitim süreçlerinde çok partili derin öğrenme modellerinde gizlilik sorununun azaltılmasına yönelik önemli bir ilerlemenin olduğu görülmektedir.

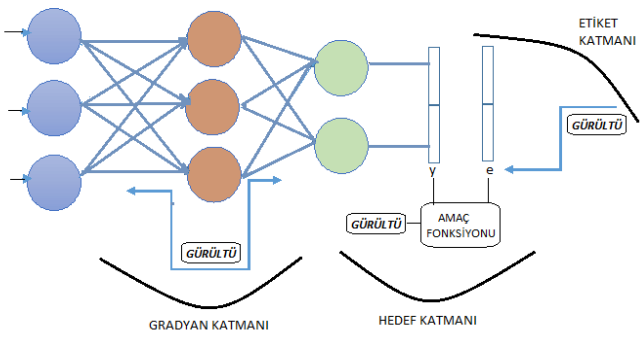


Şekil 3.*SMPC çalışma mekanizması*

Güvenli Çok Partili Hesaplama (SMPC), güvenilir merkezi bir otoriteye ihtiyaç duyulmadan katılımcılar arasında verilerin gizlilik ve güvenliğinin sağlanması ile ortak kararlaştırılmış bir fonksiyon hesaplama protokollerini konu almaktadır (Şekil 3). Bu yaklaşımın arkasında yatan temel prensip katılımcıların kişisel verileri üzerinde ortaklaştırılmış fonksiyonları hesaplamaktır. SMPC protokolünde fonksiyon çıktılarını katılımcılara bildirirken aynı zamanda çıktı üzerinden ve kendi özel girdilerinden başka katılımcı(ların) girdilerini hesaplayamayacakları ve çıktı dışında hiçbir bilginin sızmaması prensibine dayanmaktadır [28,29].

E. DİFERANSİYEL MAHREMİYET

Derin öğrenme mimarileri kullanılarak tasarlanan modeller üzerinde diferansiyel mahremiyet uygulanarak eğitim aşamasında kullanılan veriler çeşitli saldırılara karşı korunabilmektedir. Diferansiyel mahremiyet, veriye doğrudan erişimi engelleyerek gizliliği artırmaya yönelik literatürde yaygın bir şekilde karşılaşılan popüler metotlardan biridir. Bu metot ile gerçek verilere gürültü eklenerek veri maskelenir [30]. Derin öğrenme modelleri üzerinde diferansiyel mahremiyetin uygulandığı temel aşamalar Şekil 4'te verilmiştir.



Şekil 4. *Derin öğrenmede diferansiyel mahremiyete genel bakış*

Örneğin, rastgele bir fonksiyon *f* ve *D* hastanedeki hastaların verilerinden oluşan veri kümesi olmak üzere *f(D)* fonksiyonunu güvenli bir şekilde hesaplanması gerekmektedir. *f(D)* fonksiyonu hesaplanırken hasta sayısı ve hastaların kişisel verilerini korumak için fonksiyona gürültü eklenir. Eklenen bu gürültü miktarı fonksiyonda kullanılan verinin hassasiyeti ile ilgili olarak artırılıp azaltılabilir. *f* ve $||.||$ sırasıyla sorgu fonksiyonu ve norm fonksiyonu olmak üzere verilerin hassasiyet değeri $s(f, ||.||)$ aşağıdaki gibi hesaplanmaktadır.

$$s(f, \|.\|) = \| f(D) - f(D') \| \quad (2)$$

D′ veri kümesindeki örnekler D veri kümesinden bir örnek farklılık göstermekle birlikte D'nin kardeş veri kümesidir. Bu kardeş veri kümeleri için L1 veya L2 normundaki en büyük uzaklık hassasiyet değerini vermektedir.

IV. DERİN ÖĞRENME MODELLERİ ÜZERİNDE GİZLİLİĞİ VE GÜVENLİĞİ ARTIRMAYA YÖNELİK ÇALIŞMALAR

Literatürdeki çalışmalarda derin öğrenme modelleri için verilerin güvenliği ve gizliliğini artırmak amacıyla yapılan çalışmalar eğitim veya tahminleme aşamasında uygulanmaktadır.

A. EĞİTİM AŞAMASINDA

Bu bölümde, derin öğrenme modellerinde verilerin güvenliği ve gizliliğini artırmak için modelin eğitim aşamasında uygulanan literatürdeki güncel çalışmalar özetlenmiştir.

Shokri ve arkadaşları (2015), çok sayıda katılımcıdan oluşan büyük bir grubun özel verilerini herhangi bir katılımcıya sızdırmadan bir sinir ağı modeli oluşturarak katılımcıların verilerinden birlikte öğrenmeyi sağlayan genel bir sistem tasarlamışlardır [25]. "Dağıtık Stokastik Gradyan İniş" optimizasyon yöntemini kullanarak gruptaki her bir katılımcının kendi özel verilerini birbirinden bağımsız olarak eğitme yaklaşımı kullanılmaktadır. Tasarlanan yöntem ile MNIST verisi üzerinde %99.14 ve SVHN verisi üzerinde ise %93.12 başarı performansları elde edilmiştir. Verilerin güvenliğini artırmak ve sızıntı risklerinin en aza indirmek için parametreler güncellenerek (gürültü eklenerek) diferansiyel mahremiyet yaklaşımı uygulanmıştır.

Abadi ve arkadaşları (2016), stokastik gradyan iniş yöntemi ve diferansiyel mahremiyetin yaklaşımları ile yeni bir yöntem geliştirmişlerdir [31]. Veriyi korumak amacıyla her eğitim örneğinin hassasiyeti sınırlandırılmış ağ parametrelerini güncellemeden önce gradyana gürültü eklenmiştir. Özellik seçiminde ilk olarak Temel Bileşen Analizi'ne [32] diferansiyel mahremiyetin uygulanması yaklaşımı kullanılmıştır. CIFAR-10 verisi üzerinde %73, MNIST verisinde ise %97 başarı oranı elde edilmiştir.

Chase ve arkadaşları (2017), yapay sinir ağları ile oluşturdukları model üzerinde eğitimde kullanılan verilerdeki her bir örneğin gizliliğini korumak için SMPC ve diferansiyel mahremiyet yöntemlerini birlikte kullanarak yeni bir yöntem tasarlamışlardır [26]. Brett ve arkadaşları (2019), diferansiyel mahremiyeti biyomedikal veriler üzerinde geliştirilen bir modele entegre ederek verilerin güvenliği ve gizliliğini korumayı amaçlamışlardır [33]. Eğitim aşamasında rastgele seçilen bir örneğin model üzerindeki en yüksek olabilecek etkisinin sınırlandırılması ve rastgele gürültü eklenmesi yaklaşımı kullanılmıştır.

Birden fazla katılımcıdan oluşan derin öğrenme modellerinin eğitilmesinde katılımcılar hassas verilerini paylaşmadan kararlaştırılan ortak bir amaç doğrultusunda geliştirilen derin öğrenme modeli üzerinde verilerini kullanarak eğitim gerçekleştirirler. Gong ve arkadaşları (2020), merkezi bir sunucu ile verilerin gizlilik riskini önlemek için homomorfik şifreleme ve diferansiyel mahremiyet

yöntemlerinin birlikte kullanıldığı çoklu katılımcıyı destekleyen bir derin öğrenme sistemi tasarlamıştır [27] . Her bir katılımcı ilk olarak derin öğrenme kullanarak oluşturdukları modelleri kişisel verileri ile bağımsız bir şekilde eğitilmesinin sağlanmasından sonra homomorfik şifreleme uygulayıp gradyan değerlerini merkezi bir sunucuya gönderirler. Böylece katılımcıların kişisel verilerinin sızıntısı önlenirken oluşturulan genel modele katkıda bulunmak amaçlanmaktadır.

Chaudhuri ve arkadaşları (2011), eğitim verilerinin hassas bilgilerden oluştuğu bir derin sinir ağı modeli üzerinde sınıflandırma problemine odaklanmıştır [34]. Deneysel risk minimizasyonu kullanılarak sınıflandırıcılarda gizliliği korumayı amaçlayan bir yöntem tasarlamışlardır. Bu çalışmada, sınıflandırıcı tarafından her bir eğitim örneği için tahmin değeri belirlenirken eğitim verisi üzerindeki ortalama tahmin hatasının minimize edilmesi yaklaşımı kullanılmıştır. Ayrıca, gizliliği korumak için hassas veriler üzerinde diferansiyel mahremiyet yöntemini kullanmışlardır. Boura ve arkadaşları (2019), CNN ağları ile oluşturulan modele homomorfik şifreleme algoritmasını entegre ederek yeni bir yöntem geliştirmişlerdir [35].

Nandakumar ve arkadaşları (2019), yapay sinir ağları ile geliştirdikleri modellerin şifrelenmiş eğitim verileri ile eğitimi gerçekleştirmek için homomorfik şifreleme algoritmasını kullandıkları bir yöntem geliştirmişlerdir [36]. Her bir katılımcı kendi gizli anahtarını kullanarak kendi özel verilerini şifreleyebilmekte ve şifrelenmiş verilerini ortak servis sağlayıcısı ile paylaşmaktadır. Servis sağlayıcı ise katılımcıların şifresiz verilerini görmeden ortak kararlaştırılmış modeli şifrelenmiş veriler ile eğitmektedir. Eğitimin sonucunda ortaya çıkan model şifrelenmiş olmakta ve böylece servis sağlayıcı katılımcıların şifresiz verileri veya eğitim sonucunda öğrenilen model parametreleri hakkında hiçbir bir çıkarımda bulunamaz. Ayrıca, şifrelenmiş metinleri paketlemenin etkili bir şekilde uygulanması ile homomorfik şifreleme tekniğinde yapılan hesaplamaların hızlandırılması üzerinde çalışılmıştır. Şifreli metnin paketlenmesi homomorfik şifreleme tekniğinde yaygın kullanılan bir tekniktir. Gerekli önyüklemede yapılan işlemlerin sayısının en aza indirilmesi ve her bir nöron üzerindeki hesaplamaların paralelleştirilmesi yaklaşımları ile hesaplamaların karmaşıklığının azaltılması amaçlanmıştır. MNIST veri kümesi kullanılarak yapılan deneyler sonucunda şifrelenmiş veri ile modelin eğitilmesinin şifresiz veri ile modelin eğitilmesinden yaklaşık olarak 4-5 derece daha yavaş olduğu görülmüştür.

Tran ve arkadaşları (2021), derin öğrenme modellerinde verilerin gizliliğini korumak amacıyla Merkezi Olmayan Güvenli Çerçeve (SDTF) ismini verdikleri yeni bir yöntem tasarlamışlardır [37]. Herhangi bir üçüncü taraf sunucusuna ihtiyaç duyulmadan merkezi olmayan bir ağ üzerinde paralel eğitim sürecinin desteklenmesi ile verilerin mahremiyetinin korunması amaçlanmaktadır. Büyük bir gruptaki katılımcıların girdilerin toplamını güvenli bir şekilde hesaplayan Güvenli Toplam Protokolü tasarlanmıştır. Model paylaşım sürecini sağlamak için yerel modelleri dürüst ama meraklı taraflardan 2'si gizli olsa bile korumak için randomizasyon teknikleri ve Güvenli Toplam Protokolü birleştirilmiştir. Güvenli Paylaşım Protokol ismini verdikleri protokol ile gruptaki katılımcıların yerel ara parametreler ve eğitim girdileri hakkında bilgi sızdırılmadan küresel bir model eğitimi amaçlanmıştır. MNIST ve UCI SMS spam veri kümesi üzerinde yapılan deneyler sonucunda önerilen yöntem ile oluşturulan modelin yüksek başarı oranı ve verimliliği elde edilmiştir.

Syed ve arkadaşları (2020), homomorfik şifreleme kullanarak derin öğrenme modellerinde ve klasik makine öğrenmesi modellerinde verilerin güvenliğini ve gizliliğini artırmak amacıyla yeni bir yöntem geliştirmiştir [38]. Model geliştiricilerin verilere doğrudan erişememesi amaçlanmıştır. Önerilen yöntem akıllı şebekelerde arıza tanımlama, yerelleştirme ve yük tahmini uygulamaları için test edilmiştir. Şifrelenmiş veriler üzerinde eğitilen modelin doğruluğu şifrelenmemiş veriler üzerinde eğitilen modele çok yakın doğruluk değerlerine sahiptir.

B. TAHMİNLEME AŞAMASINDA

Bu bölümde, literatürde yer alan derin öğrenme temelli mimarilerde güvenliği ve gizliliği korumak amacıyla modelin tahminleme aşamasında uygulanan yöntemlerin yer aldığı çalışmalar incelenmiştir.

DeepSecure derin öğrenme ile geliştirilen modeller üzerinde verilerin gizliliği ve güvenliğini artırmayı amaçlayan bir sistemdir [39]. Gizliliği ve güvenliği artırmak için Garbled Devreler protokolünün derin öğrenme temelli modellerde tahminleme aşamalarına entegre edilmesi yaklaşımı kullanılmıştır. Garbled Devreler protokolünün uygulanmasından önce verinin boyutu küçültülür. Veri üzerinde uygulanan ön işleme aşaması şifreleme protokolünden bağımsız çalışmaktadır.

Dowlin ve arkadaşları (2016), CryptoNets olarak adlandırılan Microsoft'un katmanlı homomorfik şifreleme yöntemi kullanılarak sinir ağlarını şifreleyen bir yöntem tasarlanmıştır [40]. İlk aşamada, veri sahibinin verisinin şifreli olarak bir bulut hizmetine gönderilmesi sağlanmaktadır. Bulut sunucusunda şifreli veriyi çözmek için gereken anahtar bulunmadığından veriler sunucuda şifreli bir şekilde tutulmaktadır. Sunucu, sinir ağları ile eğitilen modele şifrelenmiş veriyi iletmektedir. Böylece, şifrelenmiş tahminleme veri sahibine ulaşır. Sonuç olarak, tahminleme sonucunu şifre anahtarına sahip olan veri sahibinin öğrenmesi sağlanmaktadır.

Xie ve arkadaşları (2019), BAYHENN olarak adlandırdıkları derin sinir ağlarının güvenliğini artırmak için Bayesçi derin öğrenme ve homomorfik şifreleme algoritmasını birleştirerek yeni bir sistem geliştirmişlerdir [41]. Kullanıcı ve sunucu arasındaki geliştirilen protokolde kullanıcı ilk olarak kişisel verilerini şifremekte ve şifrelenmiş verilerini sunucuya modelin girdisi olarak göndermektedir. Sunucu ise tasarlanan protokoldeki aşamaları takip ederek şifreli girdideki doğrusal hesaplamaları yapmaktan sorumludur. Doğrusal hesaplamalar yapılırken homomorfik şifreleme kullanılmıştır. Sunucu tarafından elde edilen çıktı kullanıcıya iletilir. Kullanıcı ise sunucudan aldığı çıktıyı yeniden şifreleyerek bir sonraki gizli katmandaki hesaplamaların yapılması için sunucuya tekrar gönderir. Bu işlemler tekrarlanıp tamamlandıktan sonra kullanıcı son modelden elde edilen sonuca güvenli bir şekilde şifre çözümlemesi yaparak ulaşabilmektedir.

Riazi ve arkadaşları [2019], XONN olarak adlandırılan derin sinir ağları ile geliştirilen modeller üzerinde kötü niyetli taraflara karşı güvenliği ve gizliliği artırmak amacıyla Yao'nun Garbled Devreler protokolünün kullanıldığı bir sistem tasarlamıştır [42]. XONN sisteminde temel amaç derin sinir ağları ile oluşturulan modelin otomatik bir şekilde eğitilmesi ve kullanıcının girdisinin ya da eğitilen modelden elde edilen çıktının orjinal halinin sunucunun ulaşamayacağı şekilde tutulması ile verinin gizliliğinin sağlanmasıdır. XONN, ikili yapay sinir ağları ile Garbled Devreler protokolünün bir arada kullanılması ile tasarlanmıştır. Bu sistemde, derin öğrenme kullanılarak oluşturulan modelin karmaşıklığı yüksek olan matris işlemleri Garbled Devreler protokolünde XNOR ile değiştirilmiştir. XONN sistemi meme kanseri, diyabet, karaciğer hastalığı ve sıtma ile ilgili veriler üzerinde değerlendirilmiştir.

Bittner ve arkadaşları (2020), derin öğrenme tabanlı ses sınıflandırması için gizliliği artırmayı amaçlayan literatürdeki ilk çözümü sunmuşlardır [43]. "MPC-friendly CNN" modeli olarak tanıttıkları SMPC'e dayanan yaklaşımdır. Bir tarafın (Alice) konuşma sinyalini başka bir tarafın (Bob), derin bir sinir ağı kullanılarak şifreli bir şekilde sınıflandırmasına olanak tanır. İki konvolüsyonel bloğa sahip bir CNN mimarisi kullanmışlardır. Önerilen yaklaşım, RAVDESS veri kümesindeki ses dosyalarının kullanılmasıyla sesten duygu tanıma görevi için değerlendirilmiştir. Elde edilen sonuçlara göre, yarı dürüst bir durumda konuşma sinyali 0.3 saniyenin altında sınıflandırılabilmiştir. Kötü niyetli durumda ise bu süre yaklaşık 1.6 saniyedir. Her iki durumda da bilgi sızıntısının olmaması amaçlanmıştır.

Tablo2. *Literatürdeki modellerde kullanılan kriptografik araçlara göre karşılaştırılması*

Model	Homomorfik Şifreleme	Garbled Devreler	SMPC	Diferansiyel Mahremiyet
Shokri ve arkadaşları (2015) [25]	+		+	
Orlandi ve arkadaşları (2007) [44]	+			
Chase ve arkadaşları (2017) [26]			+	+
Chaudhuri ve arkadaşları (2011) [34]				+
Dowlin ve arkadaşları (2016) [40]	+			
Abadi ve arkadaşları (2016) [31]				+
Rouhani ve arkadaşları (2018) [39]		+		
Gong ve arkadaşları (2020) [27]	+		+	
Xie ve arkadaşları (2019) [41]	+			
Riazi ve arkadaşları (2019) [42]		+		
Tran ve arkadaşları (2021) [37]			+	
Syed ve arkadaşları (2020) [38]	+			
Bittner ve arkadaşları (2020) [43]			+	

Tablo 2'de literatürdeki derin öğrenme temelli modellerin güvenliği ve gizliliğini korumak için geliştirilen sistemlerin Bölüm 2'de verilen kriptografik araçları kullanıp kullanmamasına göre karşılaştırılması yapılmıştır. Buna göre, [35, 37]'deki çalışmalarda homomorfik şifreleme ve SMPC'nin birlikte kullanıldığı modeller yer almışken, [48,51]'deki yer alan sistem tasarımlarında Garbled Devreler Protokolü kullanılmıştır. Ayrıca, diferansiyel mahremiyetin derin öğrenme mimarilerine de uygulandığı örnek çalışmalar da bulunmaktadır [26,34].

V. DERİN ÖĞRENME MODELLERİ ÜZERİNDE GERÇEKLEŞTİRİLEN SALDIRI TÜRLERİ VE KORUNMA TEKNİKLERİ

Bu bölümde, derin öğrenme temelli modeller üzerinde uygulanabilen saldırı türleri incelenmiş olup olası saldırıların riskini azaltmak amacıyla literatürdeki çalışmalarda yer alan korunma yöntemleri incelenmiştir.

A. SALDIRI TÜRLERİ

Derin öğrenme ile oluşturulan modeller üzerinde uygulanabilen iki büyük saldırı çeşidi bulunmaktadır. Bunlar, zehirlenme ve kaçınma saldırılarıdır. Zehirlenme saldırıları modelin eğitim aşamasında uygulanan saldırılardır. Kaçınma saldırıları ise uygun olmayan test örneklerinin test aşamasında verilmesi ile modelin yanlış tahmin değerleri üretmesine sebep olan saldırı çeşididir. Bu bölümde, literatürde yer alan saldırı çeşitlerine odaklanılmış olup güncel çalışmalar özetlenmiştir.

A.1. Zehirlenme Saldırıları

Bir modelin eğitim aşamasında kullanılan veri kümesine kötü niyetli bir örneğin entegre edilmesi durumudur.

Shafahi ve arkadaşları (2018), eğitim örneklerindeki özelliklerin gösterim şeklinin değiştirilmesi ile zehirlenme saldırısı yapmışlardır [45]. Uçtan uca öğrenme ve öğrenmenin aktarımı ile saldırı denemişlerdir. Öğrenme aktarımında eğitilmiş modelin en son katmanı olan softmax katmanına odaklanılmıştır. Modelin bu aşamasında uygulanan zehirlenme saldırıların oldukça büyük etkiler oluşturduğu gözlemlenmiştir. "Tek atışla zehirleme" olarak isimlendirilen saldırı çeşidini öğrenme

aktarımı ile birlikte uygulamışlardır. Eğitim örneklerine bir kötücül örnek eklenmesiyle tahminlemenin %100 başarısız olduğu görülmüştür. Uçtan uca eğitim modelinde ise öğrenmenin aktarımında eğitilen son katmanın aksine tüm katmanlar eğitilmektedir. Bu da uçtan uca eğitim modelinin daha çok vakit aldığını göstermektedir.

A.2. Kaçınma Saldırıları

Test aşamasında gerçekleştirilen saldırı çeşididir. Eğitim aşamasındaki sinir ağı modeline odaklanmamakta fakat modelin yanlış tahminlemeler üretmesine sebep olmaktadır. Önceden eğitilen model üzerinde gerçekleştirilen literatürdeki en çok bilinen saldırı çeşididir. Beyaz kutu ve kara kutu saldırıları olmak üzere iki grupta incelenebilir.

A.2.1. Beyaz Kutu Saldırıları

Önceden eğitilmiş derin öğrenme modelinde kullanılan tüm eğitim örnekleri, modelin parametreleri, katman sayısı bilgisi, aktivasyon fonksiyonu, bias veya ağırlıklar gibi model ile ilgili bütün verilerin bilindiği saldırı çeşididir. Bu gruptaki saldırılar eğitilen modelin gradyanları üzerinden gerçekleştirilir.

Baluja ve arkadaşları (2017), Saldırı Tabanlı Dönüşüm Mimarisi (ATN) olarak adlandırdıkları kendini denetleyen ve sinir ağı mimarilerini belirlenen bir hedef ağına ya da ağ kümesine karşı çeşitli kötücül örnekler üretmek üzere eğiten bir mimari tasarlamışlardır [46]. Bu mimari gerçek örnekler üzerinde yapılan küçük değişiklikler ile gerçek örneklerden ayırt edilmesi zor olan örnek verilerin üretilmesi amaçlanmıştır. Deneyler, MNIST ve ImageNet veri kümeleri üzerinde yapılmıştır.

A.2.2. Kara Kutu Saldırıları

Saldırganın eğitilen model ile ilgili sadece belirli girdi örneklerine karşılık tahminleme bilgisinin olduğu saldırı çeşididir. Örneğin, x örneği için saldırgan M modeli üzerinde sadece f(x;M) bilgisine sahiptir. Beyaz kutu saldırılarının aksine M modelindeki öğrenilmiş ağırlık, bias değerleri saldırganın bilgisi dahilinde değildir.

Gragnaniello ve arkadaşları (2019), derin öğrenme modelleri üzerinde tahminleme aşamasında yeni bir kaçınma saldırı çeşidi tasarlamışlardır [47]. Bu çalışmada, kaçınma saldırılarından kara kutu saldırısına odaklanılmıştır. Eğitilmiş model üzerinde saldırı yapılırken kullanılan görsel istenen sınıflandırmaya erişinceye kadar tekrarlanarak değiştirilir. Ancak, yapılan değişikliklerin görseldeki sapmaların kalitesi üzerindeki etkisi sınırlandırılmıştır. Ayrıca, tahminleme aşamasındaki etkisinin yüksek olduğu görseller düşük gradyan kısmında olacak şekilde sınırlandırılmıştır. Bunun sonucunda, kalitesi yüksek olan kötücül örnek girdiler oldukça hızlı şekilde oluşturulmuştur.

Li ve arkadaşları (2019), Vanilya Derin Sinir Ağları ve çeşitli saldırıdan korunma yöntemleri ile elde edilmiş modeller için kaçınma saldırısı tasarlamıştır [48]. Hedeflenen bir sinir ağına herhangi "optimum" bir kötücül örnek atamak yerine ilk olarak girdi örneklerinin merkezindeki ufak bir bölgede olasılık yoğunluk dağılımı bulunur. Böylece, sinir ağları ile elde edilen modelin içerisindeki katmanları ve öğrenilmiş ağırlık ve bias değerleri korunmuş olur. Bu şekilde geliştirilen yöntemler üzerinde yapılan testler sonucunda birçok test örneği için en gelişmiş kaçınma saldırı türünden daha yüksek performans elde edilmiştir.

B. SALDIRI KORUNMA TEKNİKLERİ

Bu bölümde, derin öğrenme temelli modellerin kötücül saldırı türlerine karşılık korunması için literatürde yer alan çeşitli çalışmalar incelenmiştir.

Steinhardt ve arkadaşları (2017), modelin eğitimi sırasında gerçekleştirilen zehirlenme saldırılarından korunmak amacıyla çeşitli yöntemler tasarlamıştır [49]. İkili sınıflandırma sırasında pozitif ve negatif

sınıflandırıcıların orta noktası bulunup orta noktadan uzakta olan örnekler eğitim kümesinden kaldırılır. Orta noktaların elde edilmesi iki şekilde yapılmıştır. Birincisi, literatürde küre korunma metodu olarak geçmektedir. Kürenin yarıçapından dışarıda kalan örneklerin kaldırılması yaklaşımıdır. İkinci yöntem ise tabaka savunması metodu olarak bilinmektedir. Buna göre bir çizgiden uzakta olan noktaların eğitim kümesinden çıkarılması yaklaşımı takip edilir.

Paudice ve arkadaşları (2018), zehirlenme saldırılarına karşı etkili bir savunma yöntemi tasarlamışlardır [50]. Saldırı yapan taraf, saldırdığı taraf (savunucu) üzerinde büyük etkili zehirlenme saldırısı yapmayı hedeflemektedir. Saldırının etkisini azaltmak amacıyla güvenilir veri kümesi olan D, pozitif D+ ve negatif D- olmak üzere farklı sınıflara ayrılır. Aykırı saptama yöntemi ile gerçek veri kümesinde yer alan her bir örnek için aykırı puan hesaplanır. Her örnek için elde edilmiş aykırı puanı ölçme için Destek Vektör Makinesi'ni kullanmanın en etkili yol olduğu savunulmuştur. Eğitim örnekleri üzerinde toplamsal dağılım fonksiyonu ile aykırı değerlerin tespitindeki eşik değeri hesaplanır. Elde edilen ters örnekler gerçekteki noktalarından uzaktadır. Bu ters örneklerin eğitim kümesinden kaldırılmasıyla yeni veri kümeleri elde edilir. Savunucu yeni elde edilen verileri kullanarak modelini eğitir. Bu çalışmada, bu şekilde elde edilen ters örneklerin etkin bir şekilde tespit edildiği ve filtrelendiği görülmektedir.

Tramer ve arkadaşları (2017), farklı modellerden üretilen kötücül örneklerin kullanılmasıyla eğitimdeki örnek sayısını artırarak kaçınma saldırılarından olan kara kutu saldırılarına karşı etkili olan bir korunma yöntemi tasarlamışlardır [51]. Bunun sonucunda, veri kümesinde yer alan kötücül örnek türlerinin sayısı arttırılmıştır. Inception v3 ve Inception ResNet v2 öğrenmenin aktarımı modellerinin ImageNet verisini eğiterek diğer modellerden aktarılan kötücül örneklere karşı modelin dayanıklılığın arttırılarak kaçınma saldırılarından korunmaya yönelik daha etkin modeller elde edilmiştir.

Buckman ve arkadaşları (2018), Termometre Kodlayıcı olarak adlandırdıkları sinir ağı modelleri üzerinde yapılan değişiklikler ile ağın olumsuz örneklere karşı dayanıklılığını önemli miktarda artıran bir yöntem tasarlamıştır [52]. Termometre kodlayıcısı One-Hot Kodlamanın çalışma mekanizmasına benzer olup büyüklüğü kodlamanın başka bir yoludur. Ayrıca, Termometre Kodlayıcı yöntemi ile sinir ağı modellerinin saldırılara karşı dayanıklığını artırıp zayıflığını azaltabileceği savunulmuştur. MNIST, CIFAR-10/100 ve SVHN veri kümeleri üzerinde kötücül örneklere karşı dayanıklılığı değerlendirme üzere çeşitli deneyler uygulanmıştır. Eğitim sırasında değil ancak tahminleme sırasında uygulanabilen beyaz kutu saldırısı olarak doğruluk MNIST'de %94,30 ve CIFAR-10'da %79,16 civarında olup en yüksek doğruluğa ulaştığı görülmektedir.

VI. AÇIK PROBLEMLER

Derin öğrenme ile geliştirilen modeller üzerine güvenliği ve gizlilik riskini azaltmak için literatürdeki yöntemlerden homomorfik şifreleme yüksek bir başarı skorlarına ulaşsa da derin modeller için bu yöntem uygun değildir. Bunun nedeni, derin modeller için homomorfik şifreleme ile hesaplama karmaşıklığı oldukça yüksektir. Daha yüksek hesaplama karmaşıklığı modeller üzerinde daha yavaş performans elde edilmesi ile sonuçlanmaktadır. Bundan dolayı, sinir ağının polinomlarla mümkün olduğu en düşük dereceye sahip olması oldukça önemlidir.

Homomorfik şifrelemenin yanı sıra derin öğrenme modellerinde gizliliği korumak amacıyla diferansiyel mahremiyetin modeller üzerinde uygulandığı çalışmalar mevcuttur. Bu da gradyan değerlerine ve/veya amaç fonksiyonlarına gürültü değerlerinin eklenmesi ile sağlanmaktadır. Böylece, modelin gizlilik sınırları çizilmesi amaçlanır. Ancak, bu çizilen sınırların verilerin gizliliğini korumak için ne derece yeterli olduğu bilinmemektedir. Oluşturulan derin öğrenme temelli modellerin ne denli güvenli ve saldırılara karşı dayanıklı olduğunu belirleyebilmek için çeşitli ölçütlere ihtiyaç duyulmaktadır. Gelecekteki araştırma yönlerinden biri olarak bu ölçütlerin oluşturulması ve değerlendirme yöntemleri üzerinde çalışılması gösterilebilir.

Derin öğrenme modelleri üzerinde gerçekleştirilen zehirlenme saldırılarını önlemek için literatürde yer alan yaklaşımlar, şüpheli kötücül örnekleri veri kümesinden kaldırmak veya bu örneklerin tekrar etiketlenmesi yaklaşımı kullanmaktadır [50]. Fakat, veri kümesindeki bazı örneklerin (noktaların) ortadan kaldırılması veya tekrar etiketlenmesi, modelin karar sınırı üzerinde önemli miktarda değişiklik gösterebilir. Sonuç olarak modelin elde edilen başarı performansının değişmesi veya modeli diğer zehirlenme saldırılarına karşı zayıf hale getirebilir. Modelin gizliliği ve güvenliğini belirlemek için farklı deney ve değerlendirme metotlarına gereksinim bulunmaktadır.

VII. SONUÇ

Son dönemdeki derin öğrenmenin çeşitli alanlarda getirdiği gelişmeler ile devrim niteliğinde olabilecek çalışmalar yapılmakta ve beklenti gün geçtikçe artmaktadır. Derin öğrenmenin en yaygın kullanıldığı alanlar konuşma tanıma, ses tanıma, doğal dil işleme, görüntü işleme vs. gösterilebilir. Derin öğrenme yöntemleri kullanılarak çeşitli modeller eğitilerek tahmin değerleri üretilir. Bu veriler hastane, banka veri tabanları gibi kişilerin hassas verilerinden oluşabilmektedir. Bu verilerin derin öğrenme modellerinde kullanılması veri sahipleri için güvenlik ve gizlilik riskleri doğurmaktadır.

Bu çalışmada, derin öğrenme veya sinir ağı modellerinde kullanılan verilerin gizliliğini ve güvenliğini artırmak için uygulanan araçlar verilmiştir. Bu araçlar, kriptografik temelli homomorfik şifreleme, garbled devreler, güvenli çok partili hesaplama ve diferansiyel mahremiyettir. Ayrıca, literatürde bu kriptografik araçlar kullanılarak veri güvenliği ve gizliliğini artırmak amacıyla oluşturulan derin öğrenme temelli mimariler incelenmiş, bu araçların modeller üzerinde hangi aşamada ve nasıl uygulandığına dair bilgiler verilmiştir. Literatürde yer alan derin öğrenme temelli modellerin üzerinde çeşitli saldırı tekniklerinin uygulandığı gözlemlenmiştir. Ayrıca, derin öğrenme modellerine uygulanan güncel literatürde yer alan saldırı ve korunma yöntem ve çeşitleri verilmiştir. Derin öğrenme modelleri üzerindeki gizlilik ve güvenliği artırmaya yönelik yaklaşımlar için karmaşıklığın azaltılması; modelin güvenilir ve dayanıklı olup olmadığını belirlemek için değerlendirme kriterlerinin geliştirilmesi gelecekteki araştırma yönlerinden biri olarak ifade edilebilir.

VIII. KAYNAKLAR

[1] Y. Kim, "Convolutional neural networks for sentence classification," *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, ss.1746–1751.

[2] O. Ronneberger, P. Fischer ve T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing And Computer-Assisted Intervention*, 2015, ss. 234-241.

[3] P. Pan, Z. Xu, Y. Yang, F. Wu ve Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, ss. 1029-1038.

[4] G. Parascandolo, H. Huttunen, ve T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings,", *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016, ss. 6440-6444.

[5] Z. Cai, Q. Fan, R. S. Feris, ve N. Vasconcelos, "A unified multi-scale deep concutional neural network for fast object detection," n *European conference on computer vision*, 2016, ss. 354-370.

[6] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, ve C. Pal, "Recurrent neural networks for emotion recognition in video", *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 2015, ss. 467-474.

[7] T. Hughes ve K. Mierle, "Recurrent neural networks for voice activity detection", *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, ss. 7378-7382.

[8] B. Alipanahi, A. Delong, M. T. Weirauch ve B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, c. 33, s. 8, ss. 831-838, 2015.

[9] R. Xu, D. C. Wunsch ve R. L. Frank, "Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, c. 4, s. 4, ss. 681-692, 2007.

[10] M. Auli, M. Galley, C. Quirk, ve G. Zweig, "Joint language and translation modeling with recurrent neural networks" *Conference on Empirical Methods in Natural Language Processing*, 2013, ss.1044-1054.

[11] S. Lai, L. Xu, K. Liu ve J. Zhao, "Recurrent concutional neural networks for text classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, c. 29, s. 1, ss. 2267–2273, 2015.

[12] S. Hochreiter ve J. Schmidhuber, "Long Short-Term Memory," *Neural Computation, c. 9, s. 8, ss. 1735-1780*, 1997.

[13] I. J. Goodfellow, J.P. Abadie, M. Mirza, B. Xu, D.W.Farley, S. Ozair, A. Courville ve Y. Bengio., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, 2014, ss. 2672– 2680.

[14] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, ve J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, ss. 8789-8797.

[15] A. Ghosh, V. Kulharia, V. Namboodiri, P. H. S. Torr, ve P. K. Dokania, "Multi-agent Diverse Generative Adversarial Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, ss. 8513-8521.

[16] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, ve A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, ss. 2255-2264.

[17] F. Lau, T. Hendriks, J. Lieman-Sifry, B. Norman, S. Sall, ve D. Golden, "ScarGAN: Chained generative adversarial networks to simulate pathological tissue on cardiovascular MR scans," *Deep Learning In Medical Image Analysis And Multimodal Learning For Clinical Decision Support*, ss. 343-350, 2018.

[18] A. Beers, J. Brown, K. Chang, J. P. Campbell, S. Ostmo, M. F. Chiang ve J. Kalpathy-Cramer, "High-resolution medical image synthesis using progressively grown generative adversarial networks," *arXiv: 1805.03144*, 2018.

[19] M. O. Rabin, "How to exchange secrets with oblivious transfer", *IACR Cryptol. ePrint Arch.*, c. 2005, s. 187, 2005.

[20] F. Bruekers, S. Katzenbeisser, K. Kursawe, ve P. Tuyls, "Privacy-Preserving Matching of DNA Profiles.," *IACR Cryptol. ePrint Arch.*, c. 2008, s. 203, 2008.

[21] A. C. C. Yao, "How to Generate and Exchange Secrets.," *Annual Symposium on Foundations of Computer Science*, 1986, ss. 162-167.

[22] V. Kolesnikov, A. R. Sadeghi, ve T. Schneider, "Improved garbled circuit building blocks and applications to auctions and computing minima,", *International Conference on Cryptology and Network Security*, Berlin, Heidelberg, 2009, ss. 1-20.

[23] S. Jarecki ve V. Shmatikov, "Efficient two-party secure computation on committed inputs," *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Berlin, Heidelberg, 2007, ss. 97-114.

[24] A. C. Yao, "Protocols for secure computations", *23rd Annual Symposium on Foundations of Computer Science*, 1982, ss. 160-164.

[25] R. Shokri ve V. Shmatikov, "Privacy-preserving deep learning," *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, ss. 1310-1321.

[26] M. Chase, R. Gilad-Bachrach, K. Laine, K. Lauter ve P. Rindal, "Private collaborative neural network learning," *IACR Cryptol. ePrint Arch.*, c. 2017, s. 762, 2017.

[27] M. Gong, J. Feng ve Y. Xie, "Privacy-enhanced multi-party deep learning," *Neural Networks*, c. 121, ss. 484-496, 2020.

[28] Y. Lindell, "Secure multiparty computation for privacy preserving data mining," *Encyclopedia of Data Warehousing and Mining*, ss. 1005-1009, 2011.

[29] M. Ben-Or, S. Goldwasser ve A. Wigderson, "Completeness theorems for non-cryptographic fault-tolerant distributed computation," *Proceedings of the Annual ACM Symposium on Theory of Computing*, ss. 351-371, 1988.

[30] C. Dwork, "Differential privacy: A survey of results," *International conference on theory and applications of models of computation*, Berlin, Heidelberg, 2006, ss. 1-19.

[31] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar ve L. Zhang "Deep learning with differential privacy," *Proceedings of SIGSAC conference on computer and communications security*, ss. 308-318, 2016.

[32] C. Dwork, K. Talwar, A. Thakurta ve L. Zhang, "Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis," *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, ss. 11-20, 2014.

[33] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd ve C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, c. 12, s. 7, 2019.

[34] K. Chaudhuri, C. Monteleoni, ve A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, c. 12, s. 3, 2011.

[35] C. Boura, N. Gama, M. Georgieva, ve D. Jetchev, "Simulating homomorphic evaluation of deep learning predictions", *International Symposium on Cyber Security Cryptography and Machine Learning*, Cham, 2019, ss. 212-230.

[36] K. Nandakumar, N. Ratha, S. Pankanti, ve S. Halevi, "Towards deep neural network training on encrypted data," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[37] A. Tran, T. Luong, J. Karnjana ve V. Huynh, "Neurocomputing An efficient assroach for privacy preserving decentralized deep learning models based on secure multi-party computation," *Neurocomputing*, c. 422, ss. 245-262, 2021.

[38] D. Syed ve S. S. Refaat, "Privacy Preservation of Data-Driven Models in Smart Grids Using Homomorphic Encryption", *Information*, c. 11, s. 7, ss. 1–17, 2020.

[39] B. D. Rouhani, M. S. Riazi ve F. Koushanfar, "DeepSecure: Scalable provably-secure deep learning," *Proceedings of the 55th Annual Design Automation Conference*, 2018, ss. 1-6.

[40] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig ve J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy,", *33rd International Conference on Machine Learning, ICML*, 2016, ss. 201-210.

[41] P. Xie, B. Wu ve G. Sun, "Bayhenn: Combining Bayesian deep learning and homomorphic encryption for secure DNN inference,", *IJCAI International Joint Conference on Artificial Intelligence*, 2019.

[42] M. S. Riazi, M. Samragh, H. Chen, K. Laine, K. Lauter ve F. Koushanfar, "XONN: XNOR-based oblivious deep neural network inference,"*28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, ss. 1501-1518.

[43] K. Bittner, M. De Cock ve R. Dowsley, "Private Speech Characterization with Secure Multiparty Computation," *arXiv:2007.00253*, ss. 1–40, 2020.

[44] C. Orlandi, A. Piva ve M. Barni "Oblivious neural network computing via homomorphic encryption", *EURASIP Journal on Information Security*, ss. 1-11, *2007.*

[45] A. Shafahi, W.R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras ve T. Goldstein, "Poison frogs! Targeted clean-label poisoning attacks on neural networks," *Advances in Neural Information Processing Systems*, 2018.

[46] S. Baluja ve I. Fischer, "Adversarial transformation networks: Learning to generate adversarial examples," *arXiv: 1703.09387*, 2017.

[47] D. Gragnaniello, F. Marra, G. Poggi ve L. Verdoliva, "Perceptual quality-preserving black-box attack against deep learning image classifiers," *arXiv: 1902.07776*, 2019.

[48] Y. Li, L. Li, L. Wang, T. Zhang ve B. Gong, "N Attack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," *36th International Conference on Machine Learning*, 2019, ss. 3866-3876.

[49] J. Steinhardt, P. W. Koh ve P. Liang, "Certified defenses for data poisoning attacks" *Advances in Neural Information Processing Systems*, 2017.

[50] A. Paudice, L. Muñoz-González, A. György ve E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," *arXiv*: *1802.03041*, 2018.

[51] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh ve P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv*: *1705.07204*, 2017.

[52] J. Buckman, A. Roy, C. Raffel ve I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," *6th International Conference on Learning Representations*, 2018.