DERİN ÖĞRENMEDE DİFERANSİYEL MAHREMİYET

Yavuz CANBAY[1] ve Şeref SAĞIROĞLU[2]
[1]Kahramanmaraş Sütçü İmam Üniversitesi, Bilgisayar Mühendisliği Bölümü, Kahramanmaraş
[2]Gazi Üniversitesi, MF, Bilgisayar Mühendisliği Bölümü, Ankara
yavuzcanbay@ksu.edu.tr, ss@gazi.edu.tr

ÖZET

Verinin boyut ve çeşitlilik olarak arttığı, kişisel verilerin kolaylıkla paylaşıldığı ve ihlallerinin sayısının hızla yükseldiği günümüzde veri mahremiyeti, üzerinde çokça çalışılan ve önlemler geliştirilen konuların başında gelmektedir. Kişisel verileri kullanan, depolayan veya işleyen her türlü uygulama, ürün veya sistem, veri mahremiyetini sağlamak, korumak ve doğru bir şekilde uygulandığını göstermek zorundadır. Son yıllarda veri mahremiyeti kapsamında pek çok yeni çözümler geliştirilse de teknolojik gelişmeler, yapay zekâdaki ilerlemeler, derin öğrenme yaklaşımlarının uygulama başarısı, bu yaklaşımların pek çok alanda kullanılmaya başlanması ve yapısı itibariyle kara-kutu çözüm sağlaması, veri mahremiyeti açısından yeni endişeleri de beraberinde getirmiştir. Bu çalışmada, günümüzün önemli yapay zekâ teknolojilerinden biri olan derin öğrenmede, kişisel bilgi içeren verilerin analiz edilmesi sürecinde mahremiyet koruyucu çeşitli önlemler incelenmiş, bu önlemlerden en çok kullanılanı olan diferansiyel mahremiyet açıklanmış ve derin öğrenmedeki uygulamaları karşılaştırılmıştır. Çalışmamızın, kişisel verileri işleyen derin öğrenme tabanlı uygulamalarda, oluşabilecek ihlallerin önlenmesine, karşılaşılabilecek risklerin doğru belirlenmesine ve gereken önlemlerin daha sağlıklı alınmasına katkı sağlayacağı değerlendirilmektedir.

***Anahtar Kelimeler*—**Derin öğrenme, Mahremiyet, Diferansiyel mahremiyet, İnceleme

Differential Privacy in Deep Learning

ABSTRACT

Today, where data increases in size and diversity, personal data is easily shared and privacy violations increase rapidly and personal data is protected by laws, data privacy is one of the topics that have been studied and measures have been developed. Any application, product or system that uses, stores or processes personal data must ensure and protect data privacy and demonstrate that it has been implemented properly. Although many new solutions have been developed within the scope of data privacy in recent years, technological developments, advances in artificial intelligence, application success of deep learning approaches, the usage of these approaches in many areas, having black-box structure have brought new concerns in terms of data privacy. In this study, various measures to protect privacy in deep learning, which is one of the most important artificial intelligence technologies of today, are examined; differential privacy, which is the most preferred one among these measures, is explained and its applications in deep learning are compared. It is evaluated that our study will contribute to the prevention of violations that may occur in the deep learning-based applications that process personal data, to determine the risks that may be encountered and to take the necessary precautions more accurate.

***Keywords*—** Deep learning, Privacy, Differential privacy, Review

I. GİRİŞ (INTRODUCTION)

Dijitalleşen dünya, insan odaklı ve kişilere özel çok miktarda veri üretilmesine imkân tanımaktadır. Sosyal medya ortamlarından elektronik alışverişlere, sağlık uygulamalarından akıllı şehirlere kadar pek çok teknoloji, geleneksel teknolojilere kıyasla daha fazla kişisel veri üretmeye ve toplamaya olanak sağlamaktadır. Böylesi bir dijital dünyada insanların bıraktığı her ayak izi, onları daha çok tanımayı ve tanımlamayı sağlamakla beraber daha fazla kişisel veri toplamaya da imkân sağlamaktadır. Örneğin, bir sosyal medya platformunda yapılan beğeniler,

takip edilen kişiler, akrabalık ilişkileri, yüklenen video ve resimler bireyi daha fazla tanımlamayı sağlayan pek çok özellikten bazılarıdır. Gelişen bu teknolojilerle beraber üretilen, depolanan ve işlenen verinin miktarı artmaktadır. Özelikle insanın da bir parçası olduğu her türlü dijital ortamda üretilen veriler, içerisinde çok sayıda ve çeşitte kişisel bilgi barındırabilmektedir. Verinin üretilmesinden depolanmasına kadar geçen süreçte güvenlik önlemleri önemli bir rol oynarken, verinin işlenmesi ve analiz edilmesinde ise veri mahremiyeti önlemleri önemli bir unsur olarak karşımıza çıkmaktadır.

Bireyi doğrudan veya dolaylı olarak tanımlayabilecek her türlü veri kişisel veri olarak kabul edilmektedir. Sağlık verileri, ticari veriler, demografik veriler, konum verileri, maaş verileri, epostalar, fotoğraflar ve videolar gibi pek çok veri çeşidi kişisel verilere örnek olarak verilebilir [1].

Bir veri üretici olarak insan, ürettiği kişisel verilerini açığa vurma veya paylaşma konusunda bir iradeye sahiptir. Bu irade ise doğrudan kişisel veri mahremiyetini belirler. Bu kapsamda veri mahremiyeti, "bireylerin bilgilerinin doğru kullanımı ve bireyin hangi bilgisinin kiminle ve ne derecede paylaşılmasına karar verme mekanizması" olarak literatürde tanımlanmıştır [2].

Veri mahremiyeti saldırganlar tarafından ifşa edilme riski taşıdığı için koruma gerektiren bir unsur olarak karşımıza çıkmaktadır. Kişisel verilerin ifşa edilmesi halinde ait olduğu bireyin toplumdan dışlanması, ayrımcılığa uğraması, itibar kaybı gibi pek çok sorunla karşılaşma ihtimalini de ortaya çıkarır. Bu ihtimalleri minimize etmek ve mahremiyet saldırılarını mümkün olabildiği kadar bertaraf etmek gerekir [3].

Bu amaçla literatürde önerilmiş çeşitli yaklaşımlar ve koruma modelleri mevcuttur. Veri mahremiyeti genel olarak aşağıda sunulan alt konularda çalışılan geniş bir konudur;

- Mahremiyet Korumalı Veri Yayınlama [4, 5],
- Mahremiyet Korumalı Veri Madenciliği [6].

Mahremiyet korumalı veri yayınlama, kişisel veri barındıran veri kümelerinin topluma veya araştırmacılara paylaşılması amacıyla verinin mahremiyetini koruyarak yayınlaması sürecini ifade eder. Buradaki husus, veriden daha fazla değer üretilebilmesi amacıyla verinin mahremiyet tedbirleri çerçevesinde üçüncü kişilerle paylaşılmasıdır. Mahremiyet korumalı veri madenciliği ise veri mahremiyetini dikkate alarak kişisel veri içeren veri kümesi üzerinde veri madenciliği işlemlerinin yapılmasını ifade eder.

Literatürde hem mahremiyet korumalı veri yayınlama hem de mahremiyet korumalı veri madenciliği için geliştirilmiş çok sayıda çözüm vardır. Gerek gelişen ve değişen teknolojiler gerekse de evrilen problemler ile beraber veri mahremiyeti kapsamında yeni çözümlerin sürekli olarak geliştirildiği görülmektedir. Geliştirilen bu çözümlerin hitap ettiği teknolojiyi de dikkate aldığı bir gerçektir. Örneğin büyük veri teknolojisi için geliştirilen bir mahremiyet koruma çözümünün, yapay zekâ gibi bir teknoloji için geliştirilen çözüme göre farklı bir yapıya sahip olacağı kesindir.

Kişisel verilerin işlenmesinde KVKK [7] ve GDPR [8] gibi yasal zorunlulukların olması, verinin temin edilmesinden analiz edilmesine kadar geçen süreçte tüm tarafların bu yasa ve diğer çeşitli direktiflere uyması gerekliliğini ortaya koymaktadır.

Zeki sistemlerin hukuki sorumlulukları, karşılaşılan etik problemler, doğacak olumsuzluklardaki yasal boşluklar son dönemde üzerinde çalışılan ve çözüm üretilmeye çalışılan konuların başındadır. Yapay zekânın bir alt dalı olan derin öğrenme, yapay sinir ağları mimarisine dayalı bir makine öğrenmesi yaklaşımıdır. Genel olarak bir girdi katmanı, çok sayıda gizli katman ve bir tane çıktı katmanından oluşur. Günümüzde pek çok alanda kullanılmakla beraber özellikle görüntü işleme, ses işleme, doğal dil işleme, makine çevirisi gibi alanlarda sıklıkla tercih edilmektedir [9].

Derin öğrenme yaklaşımları, açık kaynak olarak bulunan veya sunulan çeşitli veriler üzerinde zeki çözümler sunabilirken, kurumsal veya kişisel veriler kullanarak çok farklı ve şaşırtıcı derecede yeni çözümlerde sunabilmektedirler. Böylesi bir durumda, ortaya çıkabilecek muhtemel mahremiyet ihlallerini de dikkate alarak ve verilerin işlenmesi, saklanması veya taşınması gerekmektedir [10].

Mahremiyet korumalı veri yayınlamada *k*-Anonimlik [11], *l*-Çeşitlilik [12] ve *t*-Yakınlık [13] gibi temel mahremiyet koruma modelleri

geliştirilmiş ve çeşitli saldırılara karşı birer çözüm olarak uygulanmıştır. Diferansiyel mahremiyet [14] ise son yıllarda popüler olan önemli bir başka mahremiyet koruma modelidir. Veri yayınlarken ve analiz ederken bu yöntem önemli bir üstünlük sunmaktadır. Özellikle diğer koruma modellerinin arka plan bilgisi saldırısına karşı korumasız olmasından dolayı, bu sorunu gidermek amacıyla geliştirilmiştir.

Bu çalışmada, veri mahremiyeti, mahremiyet saldırı ve koruma modelleri hakkında bilgi verilmiş, diferansiyel mahremiyet modeli detaylıca açıklanmış, derin öğrenmede geliştirilen diferansiyel mahremiyet yaklaşımları ve uygulamaları gözden geçirilmiş ve sonuçta bu alanda karşılaşılan problemler ile olası çözüm önerileri sunulmuştur.

Veri mahremiyetini korumada kullanılan ve en basit yöntem olarak kabul edilen kimliksizleştirme; tekil-tanımlayıcı olarak nitelendirilen ve bireyi doğrudan tanımlayan özniteliklerin (kimlik numarası, pasaport numarası vb.) yayınlanan veriden çıkarılması olarak tanımlanır. Kimliksizleştirme, her ne kadar mahremiyet koruyucu bir yaklaşım olsa da yeterli seviyede bir koruma sağlayamadığı Sweeney tarafından yapılan araştırmada [15] gösterilmiştir. Sweeney yaptığı bu çalışmada, kimliksiz sağlık verilerini oy verileri ile cinsiyet, posta kodu ve doğum tarihi öznitelikleri üzerinde eşleştirerek Amerika nüfusunun %87'sinin kimlik bilgilerinin ifşa edilebileceği göstermiştir. Benzer başka bir olayda ise 2006 yılında AOL firması 650 bin kullanıcısına ait 20 milyon arama sorgusu verisini kullanıcı kimliği ve IP numarası bilgilerini silip kimliksizleştirerek yayınlamış, ancak birkaç gün içerisinde bu sorguların kimlere ait olduğu araştırmacılar tarafından tespit edilmiştir [12, 16].

Yukarıda belirtilen örnek durumlar dikkate alındığında, kimliksizleştirme yönteminin veri mahremiyetini tam olarak koruyamadığı belirlenmiş ve bundan dolayı mahremiyet koruyucu yeni yaklaşımlara ihtiyaç duyulduğu raporlanmıştır. Bu ihtiyacı gidermek amacıyla literatürde *k*-Anonimlik [11], *l*-Çeşitlilik [12] ve *t*-Yakınlık [13] gibi daha güçlü mahremiyet koruma modelleri geliştirilmiştir. Ancak özellikle gelişen teknoloji, artan bilgi ve veri kaynakları ve verinin elde edilme yöntemlerinin güçlenmesiyle beraber, kişiler hakkında elde edilebilecek arka plan bilgilerinin de elde edilmesi kolay hale gelmiştir. Böylesi bir durumun getirdiği en önemli dezavantaj ise kişisel verilerin ifşa riskinin artmasıdır.

Kişiler hakkında çok miktarda arka plan bilgisine sahip olan bir saldırganın, kişiyi ifşa etme ihtimali de artmaktadır. Böylesi bir saldırıyı engellemede yukarıda belirtilen üç koruma modelinin zayıf kaldığını tespit eden Dwork ve ark. [14], diferansiyel mahremiyet modelini önermişlerdir.

Bu çalışmanın ikinci bölümünde veri mahremiyetini hedef alan saldırılar açıklanmış, üçüncü bölümünde veri mahremiyetini koruma modelleri sunulmuş, dördüncü bölümünde diferansiyel mahremiyet kavramı açıklanmış, beşinci bölümde derin öğrenmede diferansiyel mahremiyeti uygulayan çalışmalar özetlenerek karşılaştırılmış, altıncı bölümde derin öğrenmede diferansiyel mahremiyeti uygulama türleri sunulmuş ve son bölümde ise sonuç ve değerlendirmelere yer verilmiştir.

II. VERİ MAHREMİYETİNİ HEDEF ALAN SALDIRILAR VE VERİ İFŞALARI (DATA PRIVACY ATTACKS AND DISCLOSURES)

Veriden değer üretmek her nekadar önemli olsa da, verinin mahremiyetini sağlamak da o derecede kıymetlidir. Mahremiyet seviyesi ile veri faydası arasındaki dengeyi bulmak veya bu dengeyi gözeterek işlemleri yapmak veya çalışmaları yürütmek gerekir. Diğer bir ifadeyle, veri mahremiyetinin tamamen karşılandığı bir durumda veri faydasından, veri faydasının tamamen karşılandığı bir durumda ise veri mahremiyetinden söz edilemez. Veri mahremiyeti ile veri faydası arasında bir denge kurmak ise oldukça zor bir problemdir [17, 18].

Literatürde veri mahremiyetine yönelik çok sayıda tehdit, saldırı veya olumsuz durum mevcuttur. Bunların başında arka plan bilgileri ile veri bağlama (eşleştirme) yöntemleri kullanılarak yapılan veri ifşası [4, 19, 20] gelir. Arka plan bilgisi ile yayınlanan veriler arasında kayıt, hassas öznitelik veya tablo düzeyinde bağlantı kurarak çeşitli saldırılar düzenleyebilir [21, 22]. Bu saldırılar sonucunda kimlik ifşası [23-25], hassas öznitelik ifşası [26, 27] ve üyelik ifşası [28] gerçekleşebilir. Kimlik bilgileri içeren veri kümeleri ile yayınlanan kimliksiz verileri yarı tanımlayıcılar düzeyinde eşleştirerek kimlik ifşası gerçekleştirilirken [29], yayınlanan verideki hassas özniteliklerin homojen dağılımına bağlı olarak kurbanın hassas bilgilerini açığa çıkarılarak hassas

öznitelik ifşası yaşanır [12]. Ayrıca, saldırgan paylaşılan veri kümesinde kurbana ait verilerin olduğunu öğrendiğinde yayınlanan veriye göre çıkarımlar yapılarak kurbanın veriyi yayınlayanla ilişkisini ortaya koyarak üyelik ifşası gerçekleştirebilir [28].

Yukarıda kısaca özetlenen bu temel saldırı türleri ve ifşalar literatürde bilinen genel saldırı türleridir. Bunlara ek olarak, özellikle bu makaleye konu olan derin öğrenme kapsamındaki saldırı türleri Mayıs 2020'de yayımlanan en güncel inceleme çalışmasında [30] çok net olarak ifade edilmiş olup, doğrudan veya dolaylı olarak yapılan saldırılarla bilgi çıkarımını ve yukarıda açıklanan literatürün gruplanması Şekil 1'de gösterilmiştir.

Bu bölümde özetlenen bilgi ve kimlik ifşaları ile yapılan saldırıları önlemek için bu saldırıları bertaraf edecek çeşitli koruma modellerine ihtiyaç vardır. Bu modeller üçüncü bölümde açıklanmıştır.

III. VERİ MAHREMİYETİ KORUMA MODELLERİ (PRIVACY PRESERVING MODELS)

Derin öğrenme yaklaşımlarında mahremiyet korumanın uygulandığı aşamalar ve bu aşamalarda kullanılan yaklaşımlar, kısa bir süre önce yayımlanan bir çalışmada özetlenmiş [30] olup bu yaklaşımlar Şekil 2'de verilmiştir. Verilen bu şekil, yapay zekâda, veri toplama ve hazırlama aşamasında kullanılabilecek yaklaşımları, eğitim aşamasında kullanılabilecek çözümleri ve sonuç çıkarımında oluşabilecek riskleri önlemek için kullanılabilecek çözümleri içermektedir.

Şekil 2'de gösterilen sınıflandırma incelendiğinde;

- diferansiyel mahremiyetin her üç aşamada da kullanılabileceği,
- veri hazırlama aşamasında, veri sahibinin kimliğinin tespit edilmesini zorlaştırmak amacıyla, veri üzerinde genelleştirme ve baskılama gibi teknikler kullanarak anonimleştirilen verilerin ifşa saldırılarına karşı korunabileceği [31],
- anonimleştirmede mahremiyet koruma gereksinimlerini karşılayan *k*-Anonimlik, *l*-Çeşitlilik, *t*-Yakınlık, *δ*-Mevcudiyet ve *ε*-Diferansiyel Mahremiyet gibi literatürde bilinen ve yaygın olarak kullanılan temel mahremiyet koruma modellerinden faydalanarak koruma işlemleri gerçekleştirilebileceği görülmektedir.

Yukarıda belirtilen ve önemli gördüğümüz bazı yaklaşımlar aşağıda kısaca özetlenmiştir.

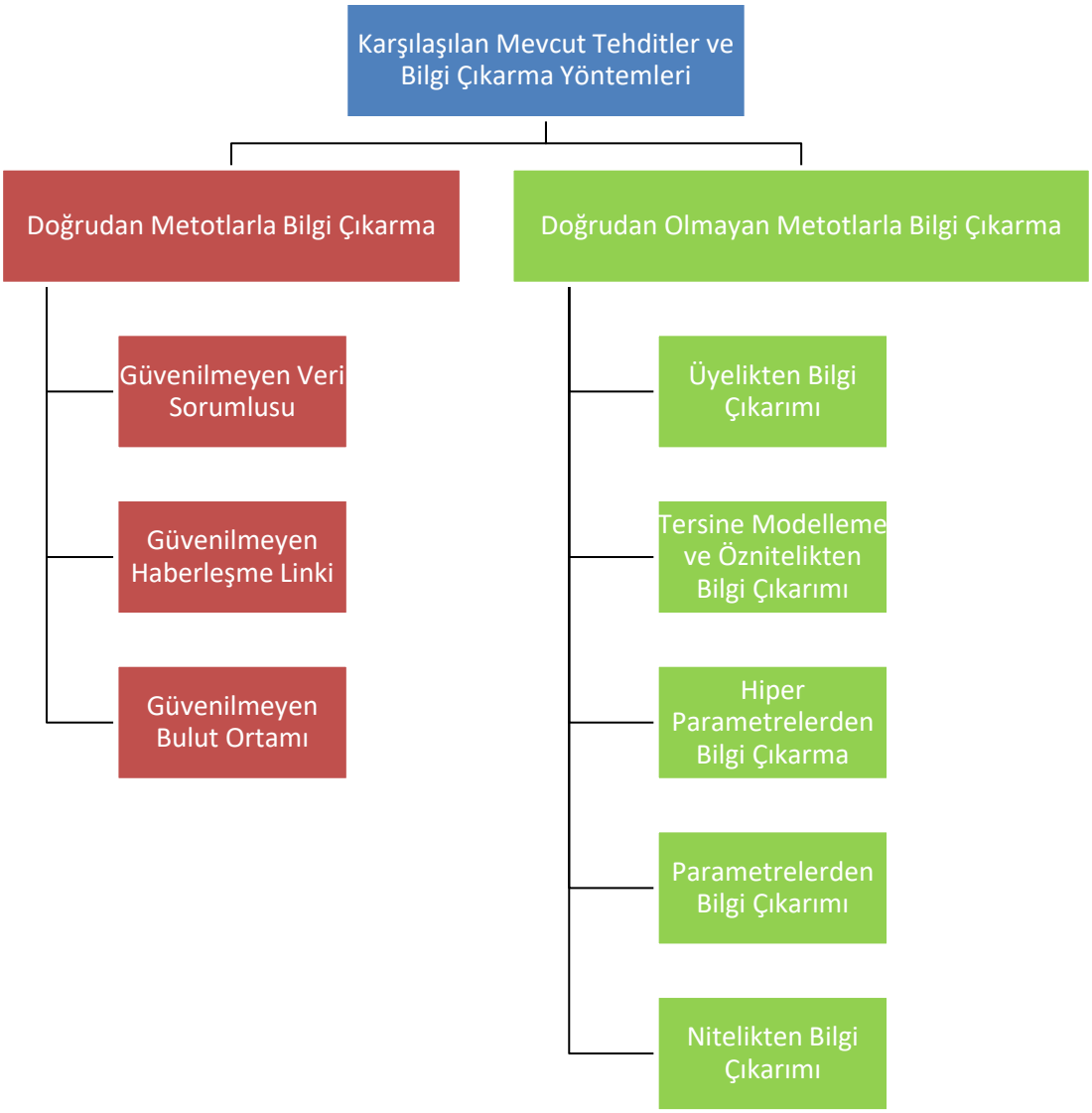
k-Anonimlik; bir verinin en az *k-1* tane veriden ayırt edilememesini sağlayan mahremiyet koruma modelidir. Kimliği ifşa edilmek istenen kişinin yani kurbanın yarı tanımlayıcılarının değeri bilinse bile, o kişinin kaydı diğer *k-1* tane kayıttan ayırt edilemez [11]. Bu şekilde kayıt bağlama saldırısı olarak da bilinen kimlik saldırısı engellenmiş olur. İlk bakışta basit bir problem olarak görünmesine rağmen optimum *k*-Anonimliği sağlamanın NP-Zor bir problem olduğu çeşitli çalışmalarda [17, 32-34] ispatlanmış ve gerek optimal gerekse de yakın optimal çözümler geliştirilmeye çalışılmıştır.

l-Çeşitlilik; *k*-Anonimlik, kimlik ifşası saldırılarına karşı koruma sağlarken hassas verilerin ifşasına karşı bir koruma sağlayamaz. Machaanavajjhala ve ark. [12], *k*-Anonimlik modelinin bu sorununu vurgulayarak hassas özniteliklerin de mahremiyetini koruyan *l*-Çeşitlilik modelini önermiştir. Bu model, hassas verilerin ifşa edilmesini mümkün olabildiği kadar engellemek amacıyla hassas verilerin çeşitliliğinin en az *l* sayıda olmasını garanti eder.

t-Yakınlık; *l*-Çeşitlilik, güçlü bir mahremiyet modeli olmasına rağmen, Li ve ark. [35] çarpık veri dağılımına sahip veri kümelerinde *l*-Çeşitlilik modelinin yetersiz olduğunu göstermiş ve *t*-Yakınlık modelini önermiştir.

l-Çeşitlilik, hassas değerler arasındaki anlamsal yakınlıklara ve hassas değerlerin dağılımının genel dağılımdan önemli ölçüde farklı olmasına bağlı olarak yapılabilecek muhtemel çarpıklık saldırılarına karşı mahremiyet korumasında yetersiz kalır. Örneğin, bir hassas verinin tüm tablodaki oranı %5 iken, bir eşlenik sınıfı içerisindeki oranı %50 ise bu durumda ciddi bir mahremiyet ihlali ortaya çıkabilir. *t*-Yakınlık yöntemi, yarı-tanımlayıcılar üzerindeki herhangi bir gruptaki bir hassas özniteliğin dağılımını tüm tablodaki özniteliklerin dağılımına yakın olmasını gerektirir [5, 35, 36].

δ-Mevcudiyet; arka plan bilgisine sahip bir saldırgan, yayınlanan anonim veride kurbana ait verilerin de olduğunu bilmesi durumunda önemli bir mahremiyet ihlali gerçekleşebilir.

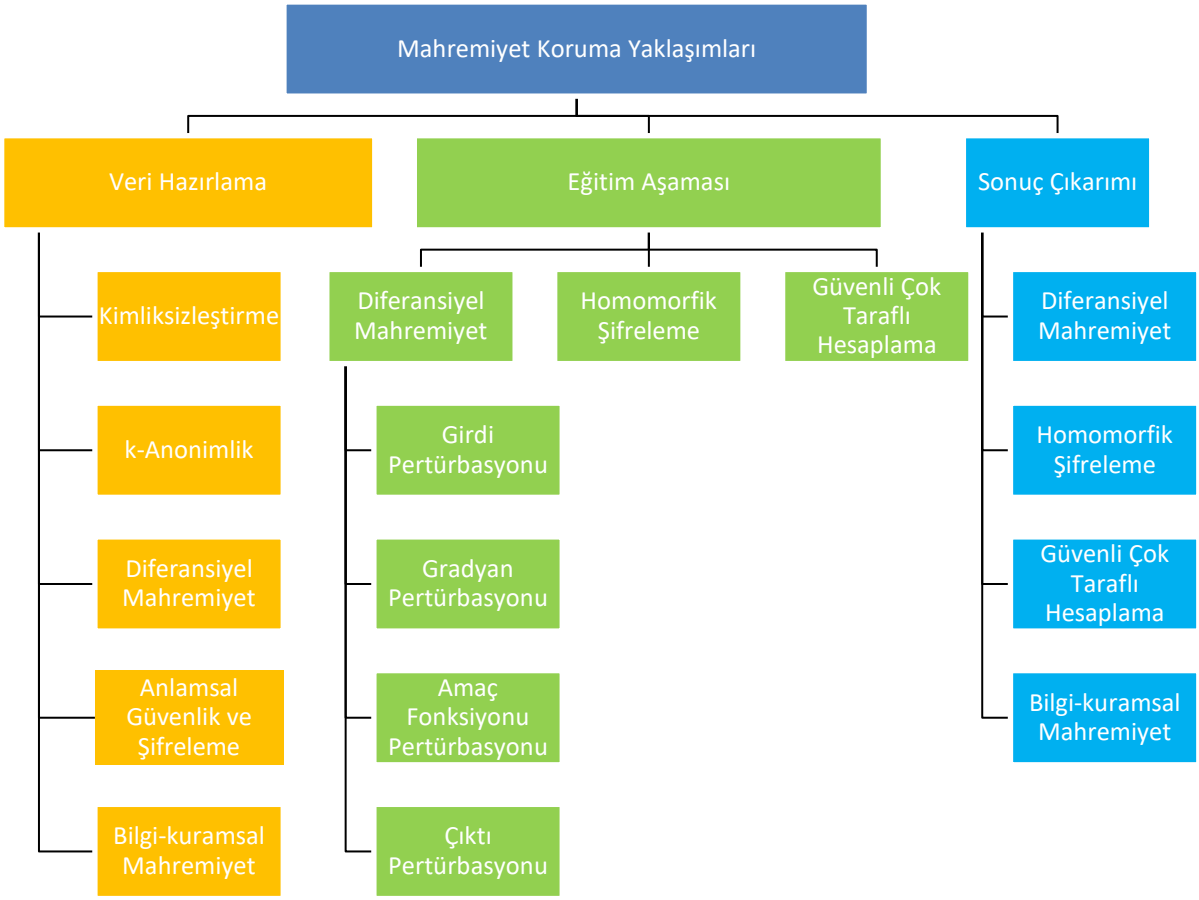


Şekil 1. Derin öğrenme yaklaşımlarında karşılaşılan tehditlerin sınıflandırılması [30]

Arka plan bilgisine sahip bir saldırgan üyelik bilgisini de kullanarak veri bağlama yöntemleriyle yapacağı saldırılar ile kimlik ifşası gerçekleştirebilir. *k*-Anonimlik, *l*-Çeşitlilik ve *t*-Yakınlık modelleri kimlik ve öznitelik ifşalarına karşı koruma sağlarken üyelik ifşalarına karşı koruma sağlayamaz. Bu problemi çözmek adına Nergiz ve ark. [28], δ-Mevcudiyet modelini önermiştir. Önerilen bu modeldeki temel yaklaşım, yayınlanan veri kümesinin saldırganın arka plan bilgisini temsil eden genel veri kümesinin alt kümesi olarak modellenebilmesidir.

ε-Diferansiyel Mahremiyet; arka plan bilgisi saldırılarını engellemek amacıyla 2006 yılında Dwork ve ark. [14] tarafından önerilmiştir. Saldırganların, arka plan bilgilerini kullanarak yapabileceği her türlü saldırıyı minimize etmek amacıyla çeşitli mekanizmaları kullanarak veriye gürültü ekleme prensibine dayanır. Bu model ileriki beşinci bölümde detaylı olarak açıklanmıştır.

Şekil 2'den de net olarak görülebileceği gibi diferansiyel mahremiyet her üç grupta da öne çıkan bir koruma yöntemi olarak karşımıza çıkmaktadır. Bundan dolayı, diferansiyel mahremiyet yaklaşımı daha detaylı olarak incelenmiş ve Bölüm 4'de açıklanmıştır.



Şekil 2. Derin öğrenmede kullanılan mahremiyet koruma yaklaşımlarının sınıflandırılması [30]

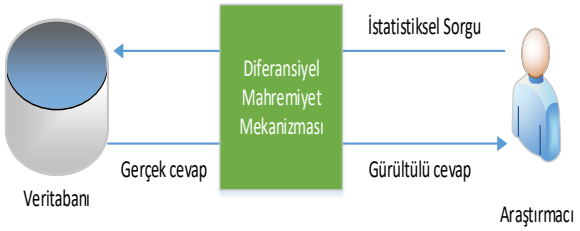
IV. DİFERANSİYEL MAHREMİYET (DIFFERENTIAL PRIVACY)

Dwork ve ark. [14] tarafından önerilen ve günümüzde sıklıkla kullanılan bir mahremiyet koruma modelidir. İstatistiksel veri tabanları üzerine yapılan analizlerde gürültü ekleme metoduna dayanarak, girdi, model ve çıktı seviyesinde koruma sağlar.

k-Anonimlik, *l*-Çeşitlilik, *t*-Yakınlık ve δ-Mevcudiyet gibi geleneksel yaklaşımların en önemli zafiyeti arka plan bilgisi saldırılarına karşı tam bir koruma sağlayamamalarıdır. Ancak diferansiyel mahremiyet modeli bu problemi gidererek yüksek mahremiyet garantisi sunar.

Anonimleştirme, bir veri kümesine veya bir sorgu sonucuna çeşitli dönüşümler uygulayarak orijinaline mümkün olduğu kadar yakın yeni bir veri kümesi veya sorgu sonucu elde edilmesini sağlar [5]. Genelleştirme, baskılama, gürültü ekleme vb. gibi yaklaşımlar, gerek veri kümesi seviyesinde gerekse de sorgu sonucu seviyesinde anonimleştirmeyi sağlayan çeşitli yapılardır. Bu yapılar kullanılarak anonimleştirilen bir veri kümesi veya sorgu sonucu üzerinde geliştirilen bir model veya analizin doğruluğu, orijinal veri üzerinde geliştirilecek bir model veya analizin doğruluğuna göre daha düşük seviyede kalacaktır. Dolayısıyla, bir veri kümesine veya sorguya diferansiyel mahremiyetin uygulanması ile elde edilecek modelin başarımı da orijinal veri üzerinde geliştirilecek modelin başarımına göre düşük olacaktır.

Şekil 3'de diferansiyel mahremiyet koruma süreci gösterilmiştir. Bu süreçte veri üzerinde analiz işlemi yapmak isteyen araştırmacı, veri tabanına istatistiksel bir sorgu ("Sum", "Count", "Mean" vb.) gönderir. Bu sorgu veri tabanında çalıştırılır ve dönecek gerçek cevap diferansiyel mahremiyet mekanizmasına gönderilir. Bu mekanizma kullanılarak gerçek cevaba gürültü eklenir ve gürültülü cevap araştırmacıya gönderilir. Bu şekilde araştırmacının gerçek veriye ulaşması yerine yaklaşık cevaplara ulaşması sağlanır.



Şekil 3. Diferansiyel mahremiyet süreci

U bir veri uzayı ve $|U|$ ise bu uzayın büyüklüğü olsun. Tablo formundaki bir veri kümesinin d boyutlu r kayıtlarını içerdiği kabul edilsin. D_1 ve D_2 sadece bir kayıtta farklılık gösteren ve komşuluk veri kümeleri olarak adlandırılan veri kümeleri olsun. f sorgu fonksiyonu, veri kümelerinden çeşitli sorgu sonuçlarını döndürsün ve sorgu fonksiyonları grubu F ile temsil edilsin.

Diferansiyel mahremiyetin temel amacı D_1 ve D_2 veri kümeleri arasında f sorgusunun sonuçları arasındaki farkı minimize etmektir. f sorgusunun sonuçlarındaki maksimum fark hassasiyet, Δf, olarak isimlendirilir. Diferansiyel mahremiyet, D veri kümesine uygulanan random bir M mekanizması ile sağlanır [37].

ε-Diferansiyel Mahremiyet; bir M mekanizması, her S çıktı kümesi için; herhangi D_1 ve D_2 komşu veri kümeleri için aşağıdaki şartı sağlaması halinde ε-diferansiyel mahremiyeti sağlar;

$$Olasılık[M(D_1) \in S] \leq \exp(e)\, Olasılık[M(D_2 \in S]$$

(ε, δ)-Diferansiyel Mahremiyet; bir M mekanizması her S çıktı kümesi için; herhangi D_1 ve D_2 komşu veri kümeleri için aşağıdaki şartı sağlaması halinde (ε, δ)-diferansiyel mahremiyeti sağlar;

$$Olasılık[M(D_1) \in S] \leq \exp(e)\, Olasılık[M(D_2 \in S] + \delta$$

Bu tanımlardan yola çıkarak; ε-Diferansiyel Mahremiyet temel diferansiyel mahremiyet olarak tanımlanırken, (ε, δ)-Diferansiyel Mahremiyet ise $\delta > 0$ için yaklaşık diferansiyel mahremiyet olarak tanımlanır [38].

Bir sorgunun hassasiyeti, mekanizmanın yapacağı pertürbasyonun miktarını belirler. Eğer bir f sorgusu D veri kümesine uygulanırsa, f sorgusunun sonucuna eklenecek olan gürültünün miktarını hassasiyet miktarı belirler. Literatürde, global ve lokal olmak üzere iki farklı hassasiyet çeşidi diferansiyel mahremiyet kapsamında kullanılmaktadır. Global hassasiyet; komşu veri kümelerine uygulanan sorgular arasındaki maksimum farkı verir ve eklenecek gürültünün miktarını belirler. Uygulanan sorgunun tüm veri kümeleri arasındaki maksimum farkını dikkate aldığı için, veri kümelerinden ziyade sorguya bağlı bir yapıdır. Sorgu hassasiyetinin düşük olması halinde kullanımı için iyi bir seçim olarak değerlendirilmektedir. Örneğin; "Count" sorgusunun hassasiyeti 1 olarak kabul edilmektedir.

Lokal hassasiyet; global hassasiyet bazı sorgularda doğru sonucun üretilmesine engel olamamaktadır. Bu durumda kullanılacak olan veri kümelerinin kısıtlanması gerekir. Bu aşamada önceden belirlenen bir eşik değerini sağlayan komşu veri kümelerinin belirlenerek hassasiyet ölçümü yapılması gerekebilir. Bu amaçla geliştirilen lokal hassasiyet, "Median" gibi sorgularda tercih edilir. "Count" ve "Range" gibi sorgular için lokal hassasiyet global hassasiyet ile aynıdır.

Literatürde diferansiyel mahremiyeti sağlamak amacıyla kullanılan çeşitli mekanizmalar mevcuttur. Bu mekanizmalar aşağıda açıklanmıştır.

i. Laplace Mekanizması

Laplace mekanizması kullanılarak gerçekleştirilen ε-diferansiyel mahremiyet için bir tanım aşağıda sunulmuştur.

Rasgeleleştirilmiş bir K fonksiyonu, sadece bir kaydın farklı olduğu D_1 ve D_2 komşu veri kümeleri için ε-diferansiyel mahremiyeti sağlar, öyle ki her $S \subseteq Range(K)$ için;

$$Olasılık[K(D_1) \in S] \leq \exp(e)\, x\, Olasılık[K(D_2 \in S]$$

Bir f sorgu fonksiyonu için, bu fonksiyonun hassasiyeti aşağıdaki gibi tanımlanır;

$f: D \rightarrow R^d$ olmak üzere, D_1 ve D_2 komşu veri kümeleri için f sorgusunun L_1-*hassasiyet*'i;

$$\Delta f = \max_{D_1 D_2} \|f(D_1) - f(D_2)\|_1$$

olarak kabul edilir. Sonuç olarak; ε diferansiyel mahremiyet parametresini temsil etmek üzere ilgili mekanizma şu şekilde oluşturulur;

$$K(D) = f(D) + Lap(0, \Delta f/\varepsilon)$$

Yukarıda verilen bilgiler dikkate alındığında, f sorgu fonksiyonunu, $Lap(0, \Delta f/\varepsilon)$ ise Laplace dağılımını (ortalama: 0, ölçek: $\Delta f/\varepsilon$, standart sapma: $\sqrt{2}\Delta f/\varepsilon$) ve son olarak Δf ise $\|f(D_1) - f(D_2)\|_1$ için maksimum değeri gösteren *L_1-hassasiyeti* temsil etmektedir. Ayrıca ε değerinin, mahremiyet koruma ile ters ve sorgu fonksiyonunun hassasiyeti ile doğru orantılı olduğu unutulmamalıdır [39].

ii. Gauss Mekanizması

Gauss mekanizmasının diferansiyel mahremiyette kullanılması ile uygulanan (ε, δ)-diferansiyel mahremiyet tanımı aşağıda sunulmuştur [40, 41].

Rasgeleleştirilmiş bir K fonksiyonu, sadece bir kaydın farklı olduğu D_1 ve D_2 komşu veri kümeleri için (ε, δ)-diferansiyel mahremiyeti sağlar, öyle ki her $S \subseteq Range(K)$ için;

$$Olasılık[K(D_1) \in S] \leq \exp(e)\ x\ Olasılık[K(D_2 \in S] + \delta$$

Gauss mekanizması, ε-diferansiyel mahremiyet için bir esneklik yapısı olarak δ parametresini kullanır.

Bir f sorgu fonksiyonu için, bu fonksiyonun hassasiyeti aşağıdaki gibi tanımlanır;

$f: D \to R^d$ olmak üzere, D_1 ve D_2 komşu veri kümeleri için, f sorgusunun *L_2-hassasiyet*'i;

$$\Delta f^2 = \max_{D_1 D_2} \|f(D_1) - f(D_2)\|_2$$

olarak kabul edilir.

Sonuç olarak, N normal dağılımı temsil etmek üzere ilgili mekanizma şu şekilde oluşturulur [10];

$$K(D) = f(D) + N(0, \Delta f^2 x \sigma^2)$$

iii. Eksponansiyel Mekanizma

Sayısal olmayan sorgular için diferansiyel mahremiyeti sağlamak amacıyla kullanılan bir mekanizmadır. Bir skor fonksiyonu $q(D, \phi)$, D veri kümesi için elde edilen $\phi \in \Phi$ çıktısını değerlendirmede kullanılır. Bu skor fonksiyonu uygulamadan uygulamaya göre değişiklik gösterebilir [41].

Δq, q skor fonksiyonunun hassasiyeti olsun. Bir K mekanizması, aşağıdaki durum için ε-diferansiyel mahremiyeti sağlar;

$$K(D) = \{\exp\left(\frac{\varepsilon x q(D, \phi)}{2\Delta q}\right)$$
$$ile\ doğru\ orantılı\ bir\ olasılıkta\ \phi'ı\ geri\ dönder\ \}$$

V. DERİN ÖĞRENMEDE DİFERANSİYEL MAHREMİYETİ UYGULAYAN ÇALIŞMALAR (DIFFERENTIAL PRIVACY APPLICATIONS IN DEEP LEARNING)

Bu bölümde literatürde diferansiyel mahremiyeti uygulayan derin öğrenme çalışmaları gözden geçirilmiş ve aşağıda özetlenmiştir;

- Sun ve ark. [10], derin sinir ağları (deep neural networks-DNN) için diferansiyel mahremiyeti (DM) sağlanmış gradyan inişi algoritması geliştirmiştir. Önerilen algoritmanın değerlendirilmesi için Diyabet veri kümesi kullanılmış ve yapılan deneysel çalışmalarda önerilen algoritmanın diferansiyel mahremiyeti sağlanmış karar ağacından daha iyi sonuç verdiği gözlemlenmiştir.
- Arachchige ve ark. [42], dağıtık derin öğrenme mimarisinde eğitim için lokal DM'ye dayalı yeni bir algoritma önerdiler. DM'yi uygulamak için evrişimsel sinir ağı (convolutional neural networks – CNN) mimarisine bir gürültü ekleme modülü uygulanmıştır. MNIST ve CIFAR-10 veri kümeleri kullanılarak yapılan deneylerde başarılı sonuçlar elde edildiği gözlemlenmiştir.
- Sei ve ark. [43], diferansiyel mahremiyetin uygulandığı derin sinir ağları yaklaşımları önerdiler. İlk yaklaşımda, diferansiyel mahremiyeti orijinal veriye uygulamış ve sonra elde edilen anonim veriye makine öğrenmesi uyguladılar. İkinci yaklaşımda ise önce orijinal veri kümesine makine öğrenmesini uygulayıp daha sonra makine öğrenmesi algoritmasından elde ettikleri modeli anonimleştirdiler. Son olarak ise, eğitim aşamasında makine öğrenmesinin parametrelerinin her bir değeri bir algoritma kullanılarak anonimleştirilir. Deneysel çalışmalarda Adult veri kümesi kullanılmış ve başarılı sonuçlar elde edilmiştir.
- Papernot ve ark. yaptıkları çalışmada [44], derin öğrenme modelinde eğitim aşamasında diferansiyel mahremiyeti sağlamak için bir model önerdiler. Önerilen model veri kümeleri üzerinde mahremiyeti sağlanmış

modellerin eğitilmesine imkân tanımaktadır. MNIST ve SHVN veri kümelerinin kullanıldığı test çalışmalarında başarılı sonuçlar elde edildiği gözlemlenmiştir.

- Abdi ve ark. [45], mahremiyet korumalı derin öğrenme için yeni bir algoritma geliştirdiler. Geliştirdikleri bu algoritmada diferansiyel mahremiyeti stokatsik gradyan iniş algoritmasına uyguladılar. Önerilen bu algoritmayı test etmek için MNIST ve CIFAR-10 veri kümelerini kullandılar ve başarılı sonuçlar elde ettiler.
- Gong ve ark. yaptıkları çalışmada [46], diferansiyel mahremiyet ve homomorfik şifrelemeyi birleştiren çok taraflı bir sistem geliştirdiler. Mahremiyeti garantilemek için diferansiyel mahremiyeti, kullanıcı ve sunucu tarafındaki veri sızıntılarını engellemek için de homomorfik şifrelemeyi kullandılar. Önerilen sistemi test etmek için MNIST ve SHVN veri kümelerinden faydalanılmış ve başarılı sonuçlar elde edilmiştir.
- Zao ve ark. tarafından yapılan bir başka çalışmada ise [47], derin öğrenmede diferansiyel mahremiyeti sağlama üzerine çeşitli çalışmalar gözden geçirilmiş, zorluklar, fırsatlar ve çözüm önerilerinde bulunulmuştur. Derin öğrenmede diferansiyel mahremiyetin girdi katmanında, gizli katmanlarda ve son olarak çıktı katmanında sağlanabileceği belirtilmiştir.
- Yang ve ark. [48], uç bilişimde mahremiyeti dikkate alan derin öğrenme modellerinin oluşturulmasında diferansiyel mahremiyeti uygulayan modeller önerdi. Derin öğrenme modeli eğitilirken diferansiyel mahremiyet uygulandı ve başarılı sonuçlar elde edildi. MNIST veri kümesi kullanılarak yapılan deneylerde, veri faydası ile veri mahremiyeti arasında bir dengenin sağlandığı raporlanmıştır.
- Hao ve ark. yaptıkları çalışmada [49] federe derin öğrenmenin mahremiyet sorunu üzerine yoğunlaşarak homomorfik şifrelemeyi diferansiyel mahremiyet ile entegre eden yeni bir federe derin öğrenme protokolü geliştirmiştir. Her bir kullanıcının yerelde oluşturduğu gradyana önce gürültü eklenmekte ardından ise şifrelenerek ana sunucuya iletilmektedir. MNIST veri kümesi kullanılarak yapılan deneylerde başarılı sonuçlar elde edilmiştir.
- Zhao ve ark. [50], doğrudan veri paylaşımı yapmaksızın kullanıcıların bir işbirliği çerçevesinde işbirlikçi derin öğrenme modeli geliştirmelerine imkân tanıyan bir sistem önerdi. Bu sistemde, her bir katılımcı lokaldeki kendi veri üzerinde lokal modelini eğitir ve sadece bu modelin parametrelerini paylaşır. Model parametrelerinin paylaşımda mahremiyeti sağlamak için derin öğrenmenin eğitim sürecinde amaç fonksiyonuna gürültü eklenir. Deneysel çalışmalarda başarılı sonuçların elde edildiği bildirilmiştir.
- Rahman ve ark. [51], diferansiyel mahremiyeti uygulayan derin öğrenme modeli için üyelik çıkarım saldırısı üzerine yoğunlaştı. Diferansiyel mahremiyeti sağlanmış derin öğrenme modeli için modelin eğitim verisi hakkında çıkarımlarda bulunabilme üzerine araştırma yaptı. MNIST ve CIFAR-10 veri kümelerinin kullanıldığı deneysel çalışmalarda, diferansiyel mahremiyeti sağlanmış derin öğrenme modelinin güçlü saldırılara karşı veri faydasından kabul edilebilir seviyede ödün vererek bir koruma sağladığı vurgulanmıştır.
- Xu ve ark. [52] çekişmeli üretici ağlar (generative adversarial networks - GAN) için eğitim aşamasında ortaya çıkması muhtemel mahremiyet sorunlarını gidermek adına gradyanlara gürültü ekleme prensibine dayalı olarak diferansiyel mahremiyeti sağlanmış yeni bir çekişmeli üretici ağ önerdi. MNIST, LSUN ve CelebA veri kümelerinin kullanıldığı deneysel çalışmalarda, önerilen modelin başarılı sonuçlar verdiği ve mahremiyeti sağladığı belirtilmiştir.
- Abay ve ark. yaptıkları çalışmada [53], diferansiyel mahremiyet ve derin öğrenmeyi kullanarak sentetik veri üretmek için yeni bir yaklaşım geliştirdi. Gradyanlara gürültü ekleme prensibinin uygulandığı çalışmada, dokuz farklı veri kümesi kullanılarak önerilen modelin doğruluğu ortaya konulmuştur.
- Phan ve ark. [54] derin sinir ağlarında diferansiyel mahremiyeti sağlamak için yeni bir yaklaşım önerdi. Veri kümesindeki özelliklere ve kayıp fonksiyonunda katsayılara gürültü ekleme yaklaşımının uygulandığı çalışmada, MNIST ve CIFAR-10 veri kümeleri kullanılarak yapılan deneylerde başarılı sonuçlar elde edildiği raporlanmıştır.
- Yu ve ark. [55] derin öğrenme için eğitim

aşamasında veri mahremiyetini sağlamak amacıyla diferansiyel mahremiyeti uygulayan bir yaklaşım önerdi. Diferansiyel mahremiyetin sunduğu veri faydasını arttırmak için bu yaklaşımı genişleterek uyguladılar. MNIST, CIFAR ve Cancer veri kümeleri kullanılarak yapılan çalışmada başarılı sonuçların elde edildiği gözlemlenmiştir.

- Liu ve ark. [56] çekişmeli üretici ağlarda veri mahremiyetini sağlamak için diferansiyel mahremiyeti kullanarak eğitim aşamasında gradyanlara gürültü ekleme yaklaşımını kullandılar. Bu yaklaşımı kullanarak önerdikleri modeli MNIST veri kümesi ile yaptıkları testte başarılı sonuçlar elde edildiği raporlanmıştır.
- Huang ve ark. [57] evrişimsel sinir ağlarında eğitim veri üzerinde model eğitme aşamasında ortaya çıkabilecek mahremiyet ihlallerini gidermek amacıyla, diferansiyel mahremiyeti dikkate alan yeni bir algoritma önerdi. MNIST ve CIFAR-10 veri kümeleri kullanılarak yapılan deneysel çalışmalarda başarılı sonuçlar elde edildiği belirtilmiştir.
- Soykan ve ark. [58] akıllı şebekelerde tüketim tahminlemesi yaparken kullanıcı tüketim verilerinin herhangi bir ihlale maruz kalmasını engellemek amacıyla diferansiyel mahremiyeti uygulayan uzun kısa süreli bellek (long short term memory – LSTM) mimarisine dayalı yeni bir model önermiştir. Açık olarak yayınlanan bir veri kümesi üzerinde yapılan deneylerde başarılı sonuçlar alındığı belirtilmiştir. Önceki kısımlarda sunulan literatür taramasındaki çalışmalar, çeşitli kriterler dikkate alınarak karşılaştırılmıştır.

Kullanılan yaklaşımlar ile mahremiyeti sağlama veya artırmaya yönelik yaklaşımlar özetlenerek Tablo 1'de verilmiştir. Bu kapsamda sunulan çalışmalar Bölüm 3, 4 ve 5'de sunulan hususlar dikkate alınarak ilgili bu çalışmalar, derin öğrenmede kullanılan mimariler, kullanılan veri kümeleri, diferansiyel mahremiyetin uygulama türleri ve diferansiyel mahremiyet mekanizmaları açısından değerlendirilmiştir.

Tablo 1. Derin öğrenme çalışmalarında diferansiyel mahremiyet kullanımlarının karşılaştırılması

Çalışma	Mimari	Veri Kümesi	Uygulama Türü	Mekanizma
[10]	DNN	Diyabet	Modele gürültü ekleme	Laplace
[42]	CNN	MNIST ve CIFAR-10	Modele gürültü ekleme	Laplace
[43]	DNN	Adult	Modele gürültü ekleme	Laplace
[44]	GAN	MNIST ve SHVN	Sentetik veri üretme	Laplace
[45]	CNN	MNIST ve CIFAR-10	Modele gürültü ekleme	Gauss
[46]	CNN	MNIST ve SHVN	Modele gürültü ekleme	Laplace
[48]	CNN	MNIST	Modele gürültü ekleme	Gauss
[49]	CNN	MNIST	Modele gürültü ekleme	Laplace
[50]	CNN	MNIST ve SHVN	Kayıp fonksiyonu pertürbasyonu	Laplace
[51]	CNN	MNIST ve CIFAR-10	Modele gürültü ekleme	Gauss
[52]	GAN	MNIST, LSUN ve CelebA	Sentetik veri üretme	Gauss
[53]	Autoencoder	9 farklı veri kümesi	Sentetik veri üretme	Gauss
[54]	CNN	MNIST ve CIFAR-10	Kayıp fonksiyonu pertürbasyonu	Laplace
[55]	CNN	MNIST, CIFAR-10, Cancer	Modele gürültü ekleme	Gauss
[56]	GAN	MNIST	Modele gürültü ekleme	Gauss
[57]	CNN	MNIST ve CIFAR-10	Modele gürültü ekleme	Gauss
[58]	LSTM	Elektrik tüketim verisi	Modele gürültü ekleme	Gauss

Tablo 1'de sunulan bilgiler dikkate alındığında diferansiyel mahremiyeti uygulayan tüm çalışmalarda özellikle 3 farklı aşamada diferansiyel mahremiyetin derin öğrenmede uygulandığı görülmektedir.

Bu türler için;

- sentetik veri üretmede GAN ve Autoencoder mimarilerinin kullanıldığı,
- modelin mahremiyetinin sağlanmasında DNN, CNN, GAN ve LSTM mimarilerinin tercih edildiği,
- kayıp fonksiyonu pertürbasyonunda ise CNN mimarisinin uygulandığı,
- MNIST ve CIFAR-10 veri kümeleri de deneysel çalışmalarda sıklıkla kullanıldığı,
- Şekil 1'de gösterilen doğrudan olmayan metotlarla bilgi çıkarımının kapsadığı tehditlerin engellenmesinde diferansiyel mahremiyetin kilit bir rolünün olduğu ve uygulanması gerektiği ve
- Şekil 2'de derin öğrenme aşamaları temel alınarak gruplandırılan mahremiyet koruma modelleri için diferansiyel mahremiyetin her aşamada kullanılabilen tek mahremiyet koruma modeli olduğu görülmektedir.

VI. DERİN ÖĞRENMEDE DİFERANSİYEL MAHREMİYETİ UYGULAMA TÜRLERİ (TYPES OF APPLICATION OF DIFFERENTIAL PRIVACY IN DEEP LEARNING)

Diferansiyel mahremiyet, derin öğrenmede 3 farklı türde uygulanmaktadır [47]. Bunlar, giriş katmanında mahremiyetini korumak için sentetik veri üretme, gizli katmanlarda (modelin) mahremiyeti sağlamak için modele gürültü ekleme ve son olarak çıkış katmanında mahremiyeti sağlamak için kayıp fonksiyonu pertürbasyonudur. Bu üç farklı uygulama türü derin öğrenme yapısına ayrı ayrı uygulanabileceği gibi birlikte de uygulanabilir.

Giriş katmanında mahremiyet, diferansiyel mahremiyeti sağlanan çekişmeli üretici ağlar kullanılarak sentetik veriler üretmeyle sağlanır [52]. Gizli katmanların ya da başka bir ifadeyle eğitim verisi üzerinde oluşturulacak modelin mahremiyetini sağlamak için de model parametrelerine gürültü ekleme işlemi yapılır. Burada genellikle tercih edilen yaklaşım; diferansiyel mahremiyet mekanizmaları kullanılarak gradyanlara gürültü eklenmesidir [42]. Son olarak, çıkış katmanında mahremiyeti sağlamak için kayıp fonksiyonu pertürbe edilerek modelin uygulanmasıyla elde edilecek çıktıların mahremiyeti sağlanır.

Diferansiyel mahremiyetin derin öğrenmedeki uygulama türleri Şekil 4'de gösterilmiş olup; şeklin sol tarafındaki yapı bir derin öğrenme sürecini temsil ederken, sağ tarafındaki kutular ise ilgili katmanlarda diferansiyel mahremiyeti sağlamada uygulanan yöntemleri göstermektedir.

Verilen şekilde, giriş katmanında sentetik veri üretimi, gizli katmanlarda gradyanlara gürültü eklenmesi ve çıkış katmanında ise kayıp fonksiyonunun pertürbasyonu ile diferansiyel mahremiyet derin öğrenme aşamalarında uygulanmaktadır.

Burada verilen bilgiler ışığında, derin öğrenmenin her aşamasında diferansiyel mahremiyetin farklı türlerde uygulanmasıyla veri mahremiyetinin sağlanabildiği görülmektedir.

VII. SONUÇ VE DEĞERLENDİRMELER (CONCLUSION AND EVALUATIONS)

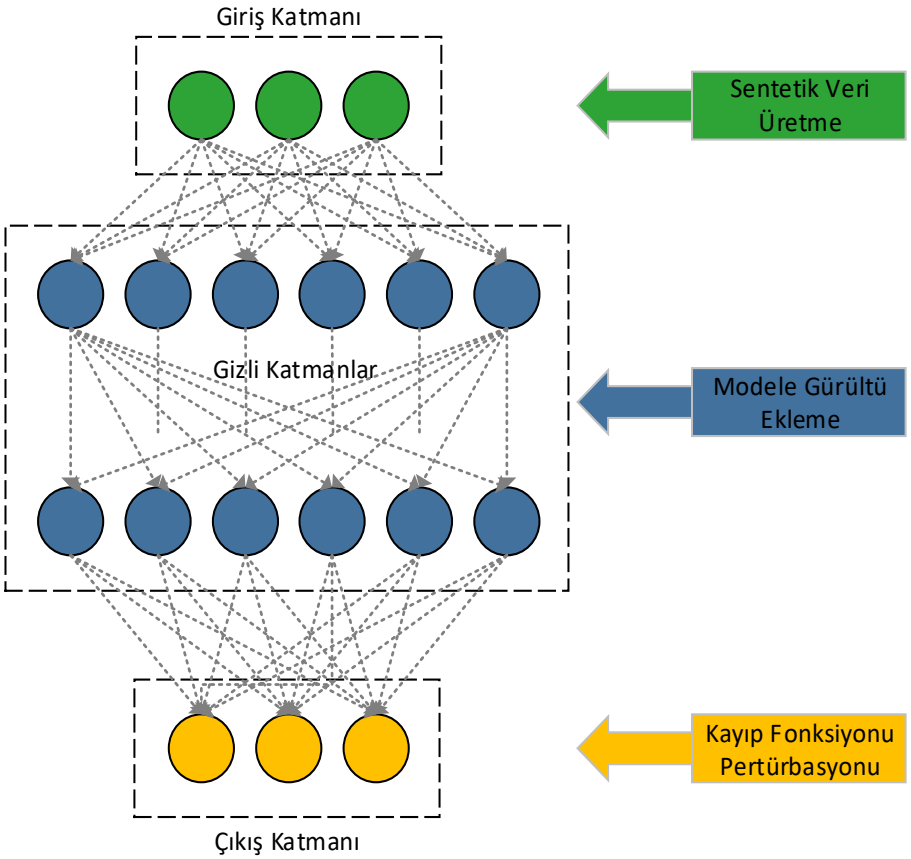
Bu çalışmada, diferansiyel mahremiyeti uygulayan derin öğrenme çalışmaları ilk defa detaylıca gözden geçirilerek tablo halinde karşılaştırılmış, derin öğrenme modellerinin geliştirilmesinde oluşabilecek mahremiyet ihlallerinin neler olduğu araştırılmış ve bunların giderilmesine yönelik olarak kullanılabilecek çözümler özetlenmiş, mahremiyet koruyucu yöntemler araştırılmış ve derin öğrenmede diferansiyel mahremiyetin kullanılması değerlendirilerek bu alandaki korkuların boyutu kapsamlı olarak tartışılmıştır.

Derin öğrenme yaklaşımlarında diferansiyel mahremiyeti sağlamaya yönelik olarak incelenen çalışmalar irdelendiğinde;

- Derin öğrenmenin günümüzde popüler bir alan olması ile beraber, özellikle kişisel verilerin işlenmesinde mahremiyet koruyucu çeşitli önlemlere ihtiyaç duyduğu açıktır. Literatürdeki çalışmalar dikkate alındığında diferansiyel mahremiyetin derin öğrenme yöntemlerine başarılı bir şekilde uygulandığı, diferansiyel mahremiyet çözümlerinin derin öğrenme yaklaşımlarında mahremiyeti

sağlamada en önemli çözüm olduğu görülmüştür.

- Tablo 1'de görüldüğü üzere DNN, CNN, LSTM ve GAN gibi mimarilere diferansiyel mahremiyetin sıklıkla uygulandığı belirlenmiştir. Ayrıca;
  - Giriş katmanının mahremiyetini korumada diferansiyel mahremiyeti uygulayan GAN ve Autoencoder mimarilerinin kullanılarak sentetik veri üretildiği,
  - Eğitim verisi üzerinde geliştirilecek modelin mahremiyetini sağlamada gradyanlara diferansiyel mahremiyet uygulanarak gürültü eklendiği,
  - Çıktı katmanının mahremiyetini sağlamada ise kayıp fonksiyonu pertürbasyonu yönteminin tercih edildiği görülmüştür.
- Derin öğrenme modellerinde uygulanan mahremiyet çözümlerinde daha çok açık sunulan benchmark veri kümelerinin kullanıldığı, bu veri kümelerinin MNIST, CIFAR-10 ve SHVN gibi standart olarak kullanılan veri kümeleri olduğu ve yeni veri kümelerine ihtiyaç olduğu, özellikle de kişisel veri mahremiyetini yüksek oranda sağladığını gösterir daha çok çalışmaya ihtiyaç olduğu,
- Hem yüksek başarımlı ve farklı çözümlerin geliştirilmesi hem de mahremiyetin sağlanmasına yönelik olarak yeni ve kapsamlı çalışmaların yapılmasında, bu çözümlerin uygulamalarda kullanılmasında ve bu örneklerin paylaşılmasında fayda olacağı değerlendirilmektedir.



Şekil 4. Derin öğrenmede diferansiyel mahremiyeti uygulama türleri [47]

Ayrıca aşağıdaki hususlara da odaklanılmasında fayda olacaktır. Bunlar;

- Derin öğrenmede kullanılan diğer mimariler için de diferansiyel mahremiyeti uygulayan çalışmalara ihtiyaç vardır,
- Derin öğrenme uygulamalarında, verinin

alınmasından çıktı üretilmesine kadar geçen süreçte mahremiyetin doğru bir şekilde anlaşılması ve uygulanması gereklidir,

- [30] numaralı inceleme çalışmasında vurgulandığı gibi, derin öğrenme tabanlı kullanıcı mahremiyetini sağlamaya yönelik çözümler incelendiğinde, derin öğrenmede mahremiyeti sağlayan çalışmaların sayısında aşama temelli (Şekil 2) bir dengesizlik olduğu, özellikle çıkarım aşamasında mahremiyetin sağlanmasına yönelik olarak ilginin düşük olduğu belirtilmiştir. Yine aynı çalışmada, Federe Öğrenme (Federated Learning) ve Vanilla ile Boomerang gibi Ayrıştırılmış Öğrenme (Split Learning) yaklaşımları kullanarak mahremiyeti arttırıcı çözümlerin mutlaka geliştirilmesi gerektiği vurgulanmıştır,
- Kişisel veriler üzerinde derin öğrenme uygulamaları geliştiren kişi, kurum ve kuruluşların KVKK ve GDPR gibi yasal mevzuatları dikkate almaları ve bunun için mahremiyet koruyucu yaklaşımları kullanarak çözümler üretilmesi gereklidir,
- Ülkemizde makine öğrenmesi alanında yapılan tüm çalışmalarda, kişisel verilerin işlenmesi halinde mahremiyete en üst seviyede odaklanarak, katma değeri yüksek ve aynı zamanda kişilerin hak ve özgürlüklerine saygı duyan uygulamalar, çözümler, ürünler, algoritmalar ve platformlar geliştirilmeli ve yaygınlaştırılmalıdır,
- Kritik verilerin paylaşılmadan değerlendirilmesi ve yeni çözümler geliştirilmesi için Federe Öğrenme (Federated Learning) gibi yapılar kurulmalı ve veri mahremiyetine saygı göstererek verilerden değer elde edilebilecek çözümlerin geliştirilmesine yönelik platformlar geliştirilmelidir ve
- Bu çalışmada incelenen derin ağ yapılarına ilave olarak Siamese Ağ ve Triplet Ağ [59] gibi ağlarda da diferansiyel mahremiyet çalışmalarının yapılması gerekmektedir.

Bu çalışmanın, oluşabilecek olan ihlallerin önlenmesine, karşılaşılabilecek risklerin doğru belirlenmesi ve gereken önlemlerin daha sağlıklı alınmasına katkı sağlayacağı değerlendirilmektedir.

TEŞEKKÜR

Bu makaleye verdiği desteklerden dolayı, Kahramanmaraş Sütçü İmam Üniversitesi DataVision Laboratuvarı (DevLab) ile Gazi Üniversitesi Büyük Veri ve Bilgi Güvenliği Laboratuvarına (BIDISEC) teşekkür ederiz.

KAYNAKLAR

[1] S. De Capitani Di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data Privacy: Definitions and Techniques," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, pp. 793-817, 2012.

[2] P. Jain, M. Gyanchandani, and N. Khare, "Big Data Privacy: A Technological Perspective and Review," *Journal of Big Data*, vol. 3, p. 25, 2016.

[3] Y. Canbay, "Aykırı Veri Yönelimli Fayda Temelli Büyük Veri Anonimleştirme Modeli," Doktora Tezi, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, 2019.

[4] B. Fung, K. Wang, R. Chen, and P. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *Computing Surveys*, vol. 42, p. 14, 2010.

[5] B. C. Fung, K. Wang, A. W. Fu, and S. Y. Philip, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. USA: CRC Press, 2010.

[6] C. C. Aggarwal and S. Y. Philip, *Privacy-Preserving Data Mining: Models and Algorithms*. USA: Springer Science & Business Media, 2008.

[7] (11.03.2020). *Kişisel Verilerin Korunması Kanunu*. İnternet Sayfası: http://www.resmigazete.gov.tr/eskiler/2016/04/20160407-8.pdf

[8] (12.03.2020). *General Data Protection Regulation*. İnternet Sayfası: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[9] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375-9389, 2017.

[10] Z. Sun, Y. Wang, M. Shu, R. Liu, and H. Zhao, "Differential Privacy for Data and Model Publishing of Medical Data," *IEEE Access*, vol. 7, pp. 152103-152114, 2019.

[11] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557-570, 2002.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity," International Conference on Data Engineering, Atlanta, USA, 2006.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *IEEE International Conference on Data Engineering*, Istanbul, Turkey, 2007, pp. 106-115.

[14] C. Dwork, "Differential Privacy," *International Colloquium on Automata, Languages and Programming*, Berlin, Heidelberg, 2006, pp. 1-12.

[15] L. Sweeney. (19.02.2018). *Simple Demographics Often Identify People Uniquely*. İnternet Sayfası: https://dataprivacylab.org

[16] R. Motwani and S. Nabar, "Anonymizing Unstructured Data," *arXiv:0810.5582*, 2008.

[17] A. Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Paris, France, 2004, pp. 223-228.

[18] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," *International Conference on Very Large Data Bases*, Trondheim, Norway, 2005, pp. 901-909.

[19] B. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," *International Conference on Very Large Data Bases*, Vienna, Austria, 2007, pp. 770-781.

[20] Y. Canbay, Y. Vural, and Ş. Sağıroğlu, "OAN: aykırı kayıt yönelimli fayda temelli mahremiyet koruma modeli," *Journal of the Faculty of Engineering & Architecture of Gazi University*, vol. 35, 2020.

[21] R. C. Wong, A. W. Fu, K. Wang, and J. Pei, "Minimality Attack In Privacy Preserving Data Publishing," *International Conference on Very Large Data Bases*, Vienna, Austria, 2007, pp. 543-554.

[22] G. Duncan and D. Lambert, "The risk of disclosure for microdata," *Journal of Business & Economic Statistics*, vol. 7, pp. 207-217, 1989.

[23] C. Skinner and D. J. Holmes, "Estimating the Re-Identification Risk per Record in Microdata," *Journal of Official Statistics*, vol. 14, pp. 361-372, 1998.

[24] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the Re-Identification Risk of Clinical Data Sets," *Bmc Medical Informatics and Decision Making*, vol. 12, p. 66, 2012.

[25] W. Winkler, "Masking and Re-Identification Methods for Public-Use Microdata: Overview and Research Problems," *International Workshop on Privacy in Statistical Databases*, Barcelona, Spain, 2004, pp. 231-246.

[26] J. Domingo-Ferrer and V. Torra, "A Critique of k-Anonymity and Some of Its Enhancements," *International Conference on Availability, Reliability and Security*, Barcelona, Spain, 2008, pp. 990-993.

[27] X. Sun, L. Sun, and H. Wang, "Extended k-Anonymity Models Against Sensitive Attribute Disclosure," *Computer Communications*, vol. 34, pp. 526-535, 2011.

[28] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Databases," *ACM SIGMOD International Conference on Management of Data*, Beijing, China, 2007, pp. 665-676.

[29] L. Sweeney, "Computational Disclosure Control: A Primer on Data Privacy Protection," Ph. D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA, 2001.

[30] F. Mirshghallah, M. Taram, P. Vepakomma, A. Singh, R. Raskar, and H. Esmaeilzadeh, "Privacy in Deep Learning: A Survey," *arXiv:2004.12254*, 2020.

[31] X. Zhang, L. T. Yang, C. Liu, and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Mapreduce on Cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, pp. 363-373, 2014.

[32] B. Kenig and T. Tassa, "A practical approximation algorithm for optimal k-anonymity," *Data Mining and Knowledge Discovery*, vol. 25, pp. 134-168, 2012.

[33] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, *et al.*, "Approximation Algorithms for k-Anonymity," *Journal of Privacy Technology*, pp. 1-18, 2005.

[34] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, *et al.*, "Anonymizing Tables," *International Conference on Database Theory*, Edinburgh, UK, 2005, pp. 246-258.

[35] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 943-956, 2010.

[36] A. Gkoulalas Divanis and G. Loukides, *Medical Data Privacy Handbook*. Switzerland: Springer, 2015.

[37] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, "Differentially private data publishing and analysis: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, pp. 1619-1638, 2017.

[38] A. Beimel, K. Nissim, and U. Stemmer, "Private learning and sanitization: Pure vs. approximate differential privacy," *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, ed: Springer, 2013, pp. 363-378.

[39] A. Alnemari, C. J. Romanowski, and R. K. Raj, "An adaptive differential privacy algorithm for range queries over healthcare data," *IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 397-402.

[40] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2006, pp. 486-503.

[41] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, pp. 211-407, 2014.

[42] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "Local Differential Privacy for Deep Learning," *IEEE Internet of Things Journal*, 2019.

[43] Y. Sei, H. Okumura, and A. Ohsuga, "Privacy-Preserving Publication of Deep Neural Networks," *IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems*, 2016, pp. 1418-1425.

[44] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv:1610.05755*, 2016.

[45] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, *et al.*, "Deep learning with differential privacy," *ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308-318.

[46] M. Gong, J. Feng, and Y. Xie, "Privacy-enhanced multi-party deep learning," *Neural Networks*, vol. 121, pp. 484-496, 2020.

[47] J. Zhao, Y. Chen, and W. Zhang, "Differential privacy preservation in deep learning: Challenges, opportunities and solutions," *IEEE Access*, vol. 7, pp. 48901-48911, 2019.

[48] Y. Yan, Q. Pei, and H. Li, "Privacy-Preserving Compressive Model for Enhanced Deep-Learning-Based Service Provision System in Edge Computing," *IEEE Access*, vol. 7, pp. 92921-92937, 2019.

[49] M. Hao, H. Li, G. Xu, S. Liu, and H. Yang, "Towards Efficient and Privacy-Preserving Federated Deep Learning," *IEEE International Conference on Communications*, 2019, pp. 1-6.

[50] L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, "Privacy-preserving collaborative deep learning with unreliable participants," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1486-1500, 2019.

[51] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership Inference Attack against Differentially Private Deep Learning Model," *Transactions on Data Privacy*, vol. 11, pp. 61-79, 2018.

[52] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "GANobfuscator: Mitigating information leakage under GAN via differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 2358-2371, 2019.

[53] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018, pp. 510-526.

[54] N. Phan, X. Wu, H. Hu, and D. Dou, "Adaptive laplace mechanism: Differential privacy preservation in deep learning," *IEEE International Conference on Data Mining*, 2017, pp. 385-394.

[55] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, "Differentially private model publishing for deep learning," *IEEE Symposium on Security and Privacy*, 2019, pp. 332-349.

[56] Y. Liu, J. Peng, J. J. Yu, and Y. Wu, "Ppgan: Privacy-preserving generative adversarial network," *arXiv:1910.02007*, 2019.

[57] X. Huang, J. Guan, B. Zhang, S. Qi, X. Wang, and Q. Liao, "Differentially Private Convolutional Neural Networks with Adaptive Gradient Descent," *IEEE Fourth International Conference on Data Science in Cyberspace*, 2019, pp. 642-648.

[58] E. U. Soykan, Z. Bilgin, M. A. Ersoy, and E. Tomur, "Differentially Private Deep Learning for Load Forecasting on Smart Grid," *IEEE Globecom Workshops*, 2019, pp. 1-6.

[59] M. Kaya and H. Ş. Bilge, "Deep metric learning: a survey," *Symmetry*, vol. 11, p. 1066, 2019.