Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1828346

Open Access Journal
e-ISSN: 2619 – 8991

Araştırma Makalesi (Research Article)
Cilt 9 - Sayı 2: 608-615 / Mart 2026
(Volume 9 - Issue 2: 608-615 / March 2026)

TİROİT KANSERİ NÜKS TAHMİNİ: SENTETİK VE GERÇEK VERİ KARŞILAŞTIRMASI

Emrullah GAZİOĞLU[1]*

[1]*Şırnak Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 73000, Şırnak, Türkiye*

Özet: Tıbbi araştırmalar başta olmak üzere diğer birçok araştırma dallarında veri kıtlığı ve/veya gizlilik kısıtlamaları, çeşitli çalışmalar için yapay zekâ (makine öğrenmesi, derin öğrenme) modellerinin geliştirilmesinde engeller oluşturmaktadır. Bu çalışmada, sentetik veri üretiminin küçük veri setlerini büyütmede nasıl bir çözüm üretebileceği gösterilmiştir. UCI Makine Öğrenmesi Deposu'ndan elde edilen Farklılaşmış Tiroit Kanseri Nüks Veri Seti kullanılmıştır. Orijinal verinin istatistiksel özelliklerini koruyan büyük ölçekli sentetik veri üretilmiş ve sınıf dengesizliği iyileştirilmiştir. XGBoost, LightGBM, k-En Yakın Komşu (kEYK) ve Karar Ağacı (KA) algoritmaları ile performans değerlendirmesi yapılmıştır. Sonuçlara bakıldığında, sentetik veri ile eğitilen modellerin orijinal veri ile karşılaştırılabilir veya daha iyi performans gösterdiği ortaya çıkmıştır. Topluluk yöntemlerinde tutarlı performans, basit modellerde ise kayda değer iyileşmeler elde edilmiştir. Kararlılık analizi, XGBoost ve LightGBM'in en tutarlı modeller olduğunu göstermiştir.

Anahtar Kelimeler: Sentetik veri, Tiroit kanseri, Nüks tahmini, Makine öğrenmesi, Gizlilik koruma, Tıbbi yapay zekâ

Thyroid Cancer Recurrence Prediction: A Comparison of Synthetic and Real Data

Abstract: Data scarcity and/or privacy constraints in medical research and many other research fields pose obstacles to the development of artificial intelligence (machine learning, deep learning) models for various studies. In this study, it has been demonstrated how synthetic data generation can provide a solution for augmenting small datasets. The Differentiated Thyroid Cancer Recurrence Dataset obtained from the UCI Machine Learning Repository was used. Large-scale synthetic data preserving the statistical characteristics of the original data was generated and class imbalance was improved. Performance evaluation was conducted with XGBoost, LightGBM, k-Nearest Neighbors (kNN) and Decision Tree (DT) algorithms. Looking at the results, it was revealed that models trained with synthetic data showed comparable or better performance than original data. Consistent performance was obtained in ensemble methods, while significant improvements were achieved in simple models. Stability analysis showed that XGBoost and LightGBM were the most consistent models.

Keywords: Synthetic data, Thyroid cancer, Recurrence prediction, Machine learning, Privacy preservation, Medical artificial intelligence

***Sorumlu yazar (Corresponding author):** Şırnak Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 73000, Şırnak, Türkiye
E mail: gazioglu@sirnak.edu.tr (E. GAZİOĞLU)
Emrullah GAZİOĞLU https://orcid.org/0000-0002-7615-305X

Gönderi: 21 Kasım 2025 **Received:** November 21, 2025
Kabul: 03 Şubat 2026 **Accepted:** February 03, 2026
Yayınlanma: 15 Mart 2026 **Published:** March 15, 2026

Cite as: Gazioğlu, E. (2026). Thyroid cancer recurrence prediction: A comparison of synthetic and real data. *Black Sea Journal of Engineering and Science*, *9*(2), 608–615.

1. Giriş

Tiroit kanseri, endokrin sistem kanserlerinin en yaygın türüdür ve dünya genelinde artan bir görülme sıklığı göstermektedir (Siegel vd., 2023). Farklılaşmış tiroit kanseri (papiller ve foliküler tiroit kanserleri), tüm tiroit kanserlerinin yaklaşık %90'ını oluşturur ve genel olarak iyi prognoza (seyir) sahip olmasına rağmen, hastaların %10-30'unda hastalığın nüks ettiği görülebilmektedir (Haugen vd., 2016). Nüks riski yüksek olan hastaların erken dönemde belirlenmesi, takip stratejilerinin optimize edilmesi ve gereksiz müdahalelerin azaltılması açısından kritik öneme sahiptir.

Yapay zekâ ve makine öğrenmesi yöntemleri, tıbbi verilerin analizi ve hastalık tahmininde giderek daha fazla kullanılmaktadır (Rajkomar vd., 2019). Ancak, tıbbi verilerin hassasiyeti nedeniyle veri paylaşımı ve model geliştirilmesi, ciddi etik ve yasal kısıtlamalarla karşı karşıyadır. Kişisel Verilerin Korunması Kanunu (KVKK) ve Genel Veri Koruma Tüzüğü (General Data Protection Regulation) (GDPR) gibi düzenlemeler, gerçek hasta verilerinin paylaşımını ve kullanımını sınırlamaktadır (Voigt ve Von dem Bussche, 2017). Bu kısıtlamalar, özellikle nadir hastalıklar veya küçük veri setleri söz konusu olduğunda, model geliştirme süreçlerini zorlaştırmaktadır.

Sentetik veri üretimi, bu sorunlara alternatif bir çözüm olarak öne çıkmaktadır. Sentetik veri, orijinal verinin istatistiksel özelliklerini ve dağılımını korurken, gerçek bireylere ait bilgi içermeyen yapay veri setidir (El Emam vd., 2020a). Bu yaklaşım, mahremiyet korunurken veri paylaşımına olanak tanıması, sınıf dengesizliği problemlerinin giderilmesi ve küçük veri setlerinin büyütülmesi açısından önemli avantajlar sağlamaktadır.

Bu çalışmanın amacı, tiroit kanseri nüks tahmini bağlamında sentetik veri kullanımının etkinliğini değerlendirmek ve küçük bir tıbbi veri setinin sentetik olarak büyütülmesinin model performansına etkisini kapsamlı bir şekilde analiz etmektir.

Çalışmanın ana katkıları şu şekilde özetlenebilir: (i) Küçük bir tıbbi veri setinin (306 örnek) 49 kat büyütülerek (15.000 örnek) sentetik veri üretimi ve istatistiksel özelliklerinin korunması, (ii) Orijinal ve sentetik veri ile eğitilen dört farklı makine öğrenmesi algoritmasının (XGBoost, LightGBM, kEYK, KA) performans karşılaştırması, (iii) Model kararlılık analizi ile sentetik verinin farklı algoritma türleri üzerindeki etkisinin incelenmesi, (iv) Sentetik verinin tıbbi araştırmalarda veri kıtlığı ve gizlilik engellerini aşmada pratik bir çözüm olarak uygulanabilirliğinin gösterilmesi.

Çalışma, sentetik veri üretiminin sadece veri büyütme aracı olarak değil, aynı zamanda gizlilik koruyucu ve model performansını iyileştirici bir yaklaşım olarak tıbbi yapay zekâ uygulamalarında kullanılabileceğini ortaya koymaktadır.

1.1. Literatür Taraması

1.1.1. Tiroit kanseri nüks tahmini

Tiroit kanseri nüks tahmininde makine öğrenmesi yöntemlerinin kullanımı son yıllarda artış göstermektedir. Borzooei vd., (2024) çalışmalarında, Farklılaşmış Tiroit Kanseri Nüks veri setini kullanarak nüks tahmininde makine öğrenmesi modellerinin etkinliğini göstermişlerdir. Çalışmada, duyarlılık (sensitivity) ve özgüllük (specificity) metrikleri üzerinden değerlendirme yapılmış ve yüksek tahmin doğruluğu elde edilmiştir.

Habchi vd. (2023) sistematik derlemesinde, tiroit kanseri tanısında derin öğrenme, yapay sinir ağları ve topluluk yöntemlerinin yaygın olarak kullanıldığını ortaya koymuştur. Özellikle topluluk yöntemlerinin (ensemble methods), tek başına algoritmalara göre daha yüksek performans gösterdiği belirtilmiştir.

1.1.2. Tıbbi verilerde sentetik veri kullanımı

Sentetik veri üretimi, tıbbi makine öğrenmesi çalışmalarında giderek artan bir ilgi görmektedir. Son dönemde, CTGAN (Xu vd., 2019) ve TVAE gibi özel sentetik veri üretim yöntemleri tabular veri için geliştirilmiştir. Bu yöntemler, koşullu üretim ve varyasyonel öğrenme ile yüksek kaliteli sentetik veri üretebilmektedir.

Yale vd. (2020), çalışmalarında, elektronik sağlık kayıtlarından sentetik hasta verileri üretmek için Değişken Otokodlayıcı (Variational Autoencoder) (VOK) ve Koşullu ÜÇA (Conditional GAN) (KÜÇA) yöntemlerini karşılaştırmışlar ve sentetik verilerin gerçek verilerle yüksek istatistiksel benzerlik gösterdiğini ortaya koymuşlardır. Ancak, nadir klinik olayların sentetik veride yeterince temsil edilmemesi bir sınırlama olarak çalışmalarında belirtilmiştir.

Chen vd. (2021) tarafından yapılan sistematik derlemede, Üretken Çekişmeli Ağlar (Generative Adversarial Networks) (ÜÇA) tabanlı sentetik veri üretim yöntemlerinin tıbbi görüntüleme, elektronik sağlık kayıtları ve genomik verilerde başarılı uygulamaları incelenmiştir. Çalışmada, sentetik verinin veri büyütme (data augmentation), gizlilik koruma ve nadir durumların modellemesi açısından önemli faydalar sağladığı vurgulanmıştır.

Büyük Dil Modelleri (Large Language Model) (BDM) tabanlı yaklaşımlar, son dönemde sentetik tablo verisi üretiminde yeni bir paradigma oluşturmuştur. Borisov vd. (2022) çalışmalarında, dil modellerinin yapılandırılmış tablo verisi üretimindeki potansiyelini göstermişler ve geleneksel ÜÇA tabanlı yöntemlere alternatif sunmuşlardır. BDM tabanlı sentetik veri üretimi, kategorik değişkenlerin doğal dil benzeri yapısını daha iyi modelleyebilmekte ve nadir kombinasyonları üretebilmektedir. Bu çalışmada kullanılan LLM tabanlı yaklaşım, bu yöntemlere alternatif bir paradigma sunmaktadır.

1.1.3 Veri gizliliği ve mahremiyet

Tıbbi verilerde gizlilik ve mahremiyet koruma, yapay zekâ uygulamalarının en kritik sorunlarından biridir. Sağlık bilgi sistemlerinde tutulan kişisel sağlık verileri, bireylerin en hassas bilgilerini içermekte ve bu verilerin yetkisiz erişime, kötüye kullanıma veya ifşaya karşı korunması hem etik hem de yasal bir zorunluluktur. Sağlık verilerinde gizlilik ve mahremiyet kavramlarının kapsamlı tanımlanması ve korunması gerekliliği ilk olarak Donaldson ve Lohr (1994) tarafından detaylı şekilde ele alınmış ve bu çalışma alandaki temel referanslardan biri haline gelmiştir. Kişisel Verilerin Korunması Kanunu (KVKK) ve Genel Veri Koruma Tüzüğü (General Data Protection Regulation) (GDPR) gibi düzenlemeler, gerçek hasta verilerinin paylaşımını ve kullanımını ciddi şekilde sınırlamaktadır (Voigt ve Von dem Bussche, 2017). Lobato de Faria ve Cordeiro (2014), sağlık verisi gizliliğinin modern toplumda yaşadığı krizi vurgulayarak, geleneksel yasal araçların bu verileri korumada yetersiz kaldığını ve yeni teknolojik çözümlere ihtiyaç duyulduğunu belirtmiştir.

Veri gizliliğini korumak için geleneksel olarak k-anonimlik (k-anonymity), farklılıksal gizlilik (differential privacy) ve veri maskeleme gibi yöntemler kullanılmaktadır. Ancak bu yaklaşımlar, veri kalitesinde önemli kayıplara yol açabilmekte ve istatistiksel analizlerin doğruluğunu etkileyebilmektedir. Bender vd. (2020), istatistiksel kurumların hassas veritabanlarından doğru sonuçlar hesaplama, sonuçları geniş çapta paylaşma ve bilimsel inceleme ile tekrarlanabilirliği kolaylaştırma ihtiyacı ile bireysel gizlilik ve veri mahremiyeti çıkarları arasında denge kurması gerektiğini vurgulamıştır. Bu bağlamda sentetik veri, veri erişimi sağlama yöntemlerinin evriminde önemli bir alternatif olarak öne çıkmaktadır. El Mestari vd. (2024) tarafından yapılan sistematik incelemede, makine öğrenmesi sistemlerinde veri korumaya yönelik gizlilik artırıcı teknolojilerin (Privacy-Enhancing Technologies - PETs) tek başına yeterli olmadığı, veri sahiplerinin makine öğrenmesi yaşam döngüsü boyunca karşılaştığı

risklerin kapsamlı bir şekilde ele alınması gerektiği belirtilmiştir.

El Emam vd. (2020b) tarafından yapılan sistematik derlemede, sentetik verinin yeniden tanımlama (re-identification) riskini azaltma ve k-anonimlik, farklılıksal gizlilik gibi geleneksel gizlilik koruma yöntemlerine üstünlükleri tartışılmıştır. Sentetik verinin, istatistiksel yararı (utility) korurken gizlilik sağlama konusunda daha etkin olduğu gösterilmiştir. Gerçek hasta verilerinden türetilen sentetik kayıtlar, orijinal bireylere ait bilgi içermediği için yeniden tanımlama riski pratikte ortadan kalkmakta, bu da KVKK ve GDPR gibi düzenlemelere uyumu kolaylaştırmaktadır.

Jordon vd. (2022) tarafından yapılan çalışmada, sentetik veri kalitesini değerlendirmek için benzerlik (fidelity), kullanışlılık (utility) ve gizlilik (privacy) boyutlarından oluşan bir çerçeve (framework) önerilmiştir. Bu çerçeve, sentetik verinin hem istatistiksel özellikleri koruması hem de gizlilik garantileri sağlaması gerektiğini vurgulamaktadır. Sentetik veri üretimi, tıbbi araştırmalarda veri paylaşımını mümkün kılarken, hasta mahremiyetini koruma ve yasal düzenlemelere uyum sağlama açısından pratik bir çözüm sunmaktadır.

2. Materyal ve Metot

2.1. Veri Seti Tanımı

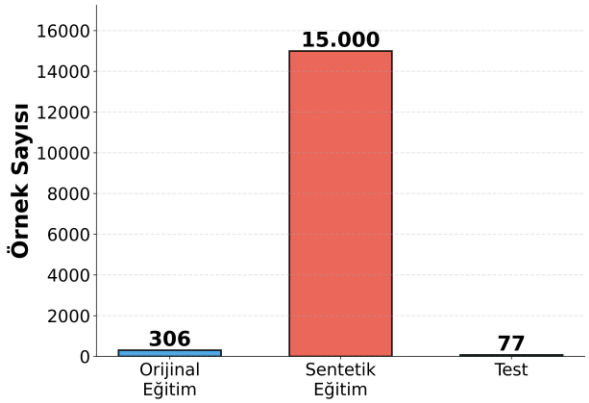
Bu çalışmada UCI Makine Öğrenmesi Deposu'ndan elde edilen Farklılaşmış Tiroit Kanseri Nüks Veri Seti kullanılmıştır (Borzooei vd., 2024). Veri seti, farklılaşmış tiroit kanseri tanısı almış 383 hasta kaydından oluşmaktadır. Hedef değişken, hastalığın nüks durumunu (ikili: Evet/Hayır) ifade etmektedir; 275 hastada (%71,8) nüks gözlemlenmezken, 108 hastada (%28,2) nüks görülmüştür. Bu durum, 2.55:1 oranında orta düzeyde bir sınıf dengesizliği oluşturmaktadır.

Veri seti 16 tahmin edici değişken içermektedir: Sayısal değişken olarak Yaş (tanı anındaki hasta yaşı) bulunmaktadır. Kategorik değişkenler; cinsiyet, sigara öyküsü, radyoterapi maruziyeti, tiroit fonksiyon kategorileri, fizik muayene bulguları, adenopati, patoloji tipi, fokalite ve tedaviye yanıt bilgilerini kapsamaktadır. Sıralı değişkenler (ordinal) ise klinik evreleme ve risk seviyelerini temsil etmekte olup Risk Seviyesi, T Evresi, N Evresi, M Evresi ve TNM Evresi değişkenlerinden oluşmaktadır. Bu özellikler, hastalık ilerlemesi ve tedavi sonuçlarını klinik olarak yorumlanabilir şekilde tanımlamaktadır.

Şekil 1'de veri setinin boyutları gösterilmektedir. Orijinal eğitim verisi 306 örnekten oluşurken, sentetik veri ile bu sayı 15000'e çıkarılmıştır. Test veri seti ise 77 örnekten oluşmaktadır.

3.2. Veri Ön İşleme

Model geliştirmeden önce, veri seti kapsamlı bir temizleme sürecinden geçirilmiştir. Eksik değer içeren kayıtlar, kalan gözlemlerin güvenilirliğini korumak amacıyla liste bazında silme (listwise deletion) yöntemiyle çıkarılmıştır. Hedef değişken, orijinal olarak "Recurred" etiketi altında "Yes" ve "No" değerleriyle kodlanmışken, betikler (script) arası tutarlılık sağlamak için küçük harfe çevrilmiştir.



Şekil 1. Veri seti boyutlarının karşılaştırması. Sentetik veri ile orijinal veriye göre 49 kat artış sağlanmıştır.

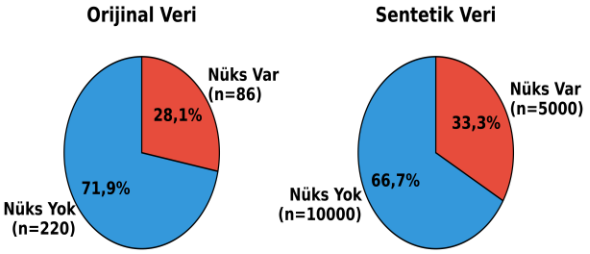
Veri seti, %80-%20 oranında eğitim (train) ve test alt kümelerine bölünmüştür. Orijinal sınıf dağılımını korumak için tabakalı örnekleme (stratified sampling) uygulanmış, bu işlem sonucunda 306 eğitim ve 77 sınama örneği elde edilmiştir. Tekrarlanabilirliği sağlamak amacıyla sabit rastgele tohum (random seed) (42) kullanılmıştır. Bölme işlemi, her iki alt kümede de yaklaşık %72 "Nüks Yok" ve %28 "Nüks Var" dağılımını korumuştur.

Kategorik ve sıralı değişkenler makine tarafından okunabilir formatlara dönüştürülmüştür. Sıralı değişkenler (Risk Seviyesi, T/N/M Evreleri) klinik hiyerarşilerine göre sayısal olarak, nominal kategorik değişkenler tek-sıcak kodlama (one-hot encoding) ile kodlanmıştır. İkili hedef değişken ise 0 ("Hayır") ve 1 ("Evet") olarak kodlanmıştır.

Sentetik veri üretimi, veri sızıntısını (data leakage) önlemek amacıyla yalnızca eğitim alt kümesindeki 306 kayıt kullanılarak gerçekleştirilmiştir. Test kümesi (77 kayıt), sentetik veri üretim sürecine hiçbir şekilde dahil edilmemiştir. Bu yaklaşım, modellerin gerçek dünya performansının tarafsız değerlendirilmesini sağlamaktadır.

Ağaç tabanlı modellerin (XGBoost, LightGBM, KA) doğal olarak ölçek bağımsız olması nedeniyle öznitelik ölçeklendirme (feature scaling) uygulanmamıştır. Kodlanmış öznitelik yapısı, kEYK gibi mesafe tabanlı algoritmaların da ek normalizasyon olmadan yeterli performans göstermesini sağlamıştır.

Şekil 2'de orijinal ve sentetik verilerin sınıf dağılımlarına ait karşılaştırma verilmiştir. Orijinal veride %71,8 - %28,2 oranında dengesizlik bulunurken, sentetik veride bu oran %66,7 - %33,3'e iyileştirilmiştir.

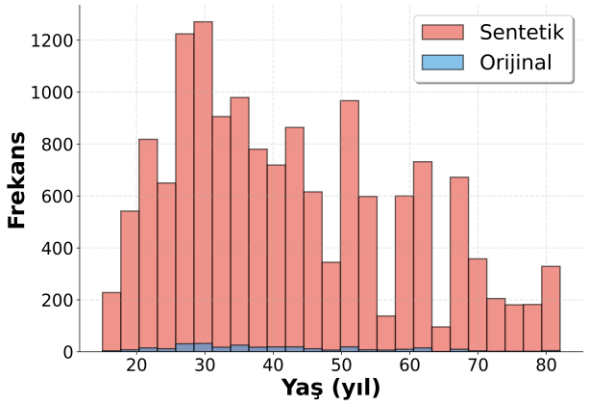


Şekil 2. Orijinal ve sentetik verilerde sınıf dağılımı karşılaştırması. Sentetik veride sınıf dengesizliği iyileştirilmiştir.

3.3. Sentetik Veri Üretimi

Sentetik veri, DataXid platformu (DataXID, 2024) kullanılarak üretilmiştir. DataXid, BDM tabanlı tabular veri üretim yöntemi kullanan, yapay zekâ destekli bir platformdur. Platform, orijinal veriden istatistiksel dağılımı ve ilişkileri öğrenerek, gerçek kayıtlara ait bilgi içermeyen yeni örnekler oluşturmaya imkân vermektedir. Sayısal değişkenler (yaş) sürekli değer aralıklarında, kategorik değişkenler doğal dil birimleri (token) olarak işlenmiştir. Model, orijinal verinin öznitelik korelasyonlarını ve koşullu olasılık dağılımlarını öğrenerek, klinik olarak tutarlı hasta profilleri üretmiştir. Sentetik veri üretiminde çeşitlilik ve gerçekçilik dengesi gözetilmiş ve toplamda 15.000 örnek üretilmiştir: 10.000 "nüks yok", 5.000 "nüks var" (2:1 oranı). Bu oran ile, orijinal verideki sınıf dengesizliği (2.55:1) iyileştirilmek amaçlanmıştır.

Şekil 3'te orijinal ve sentetik verilerin yaş dağılımları gösterilmiştir. Sentetik verinin, orijinal verinin yaş dağılımını başarılı bir şekilde taklit ettiği görülmektedir.



Şekil 3. Orijinal ve sentetik verilerde yaş dağılımı karşılaştırması. Sentetik veri, orijinal verinin istatistiksel özelliklerini korumaktadır.

3.4. Makine Öğrenmesi Modelleri

Bu çalışmada, farklı öğrenme paradigmalarını ve model karmaşıklık seviyelerini temsil eden dört makine öğrenmesi algoritması kullanılmıştır: XGBoost, LightGBM, k-En Yakın Komşu (kEYK) ve Karar Ağacı (KA). Bu algoritmaların seçiminde, küçük tıbbi veri setlerinde performans gösterebilme yetenekleri, sınıf dengesizliği ile başa çıkma kapasiteleri ve yorumlanabilirlik özellikleri dikkate alınmıştır. Topluluk yöntemleri (XGBoost, LightGBM) karmaşık öznitelik etkileşimlerini yakalayabilirken, basit modeller (kEYK, KA) sentetik veri büyütmenin etkisini değerlendirmek için referans noktası oluşturmaktadır.

XGBoost (Extreme Gradient Boosting), Chen ve Guestrin (2016) tarafından geliştirilen, gradyan güçlendirme algoritmasının optimize edilmiş bir uygulamasıdır. XGBoost, zayıf öğrenicileri (tipik olarak karar ağaçları) sıralı bir şekilde birleştirerek, her yeni ağacın önceki ağaçların hatalarını düzeltmesini sağlar. Algoritma, düzenlileştirme (regularization) teknikleri ile aşırı uyumu önlerken, ikinci dereceden gradyan bilgisi kullanarak hızlı ve doğru öğrenme gerçekleştirir. XGBoost'un sınıf dengesizliği problemlerinde etkin olması ve küçük veri setlerinde bile yüksek performans gösterebilmesi, bu çalışma için önemli avantajlar sağlamaktadır. Ayrıca, öznitelik önem skorları sayesinde hangi klinik değişkenlerin nüks tahmininde daha etkili olduğu yorumlanabilir.

LightGBM (Light Gradient Boosting Machine), Ke vd. (2017) tarafından geliştirilen, XGBoost'a alternatif bir gradyan güçlendirme yöntemidir. LightGBM, yaprak-odaklı (leaf-wise) ağaç büyütme stratejisi kullanarak, seviye-odaklı (level-wise) büyütmeye göre daha derin ve spesifik karar sınırları oluşturabilir. Algoritma, gradyan-tabanlı tek-taraflı örnekleme (Gradient-based One-Side Sampling - GOSS) ve özel öznitelik paketleme (Exclusive Feature Bundling - EFB) teknikleri ile hem hızlı hem de bellek-verimli çalışır. LightGBM'in XGBoost'tan temel farkı, daha hızlı eğitim süresi ve büyük veri setlerinde daha iyi ölçeklenebilirlik sağlamasıdır. Bu çalışmada, sentetik verinin 15.000 örneğe çıkarılması nedeniyle LightGBM'in veri büyüklüğü ile başa çıkma kapasitesi değerlendirilmiştir.

k-En Yakın Komşu (kEYK) algoritması, parametresiz ve örnek-tabanlı bir sınıflandırma yöntemidir. Algoritma, test örneğinin öznitelik uzayında en yakın k komşusunu bularak, bu komşuların sınıf etiketlerine göre çoğunluk oylaması ile tahmin yapar. kEYK'nin en önemli avantajı, herhangi bir model eğitimi gerektirmemesi ve verinin yerel yapısını doğrudan kullanmasıdır. Ancak, yüksek boyutlu verilerde ve büyük veri setlerinde hesaplama maliyeti artmaktadır. Bu çalışmada kEYK, sentetik veri büyütmenin basit, mesafe-tabanlı algoritmaların performansına etkisini değerlendirmek amacıyla dahil edilmiştir. Özellikle, 306 örnekten 15.000 örneğe çıkışın öznitelik uzayında daha yoğun bir kaplama sağlayarak kEYK'nin yerel örüntü öğrenme kapasitesini nasıl etkilediği incelenmiştir.

Karar Ağacı (KA) modeli, hiyerarşik karar kuralları kullanarak veriyi alt gruplara bölen, yorumlanabilir bir algoritmadır. Her düğümde, veriyi en iyi ayıran öznitelik ve eşik değeri seçilir, böylece yaprak düğümlerde homojen sınıf dağılımları elde edilir. KA'nın en büyük avantajı, klinik karar süreçlerine benzer "eğer-o zaman" kuralları üretmesi ve tıbbi personel tarafından kolayca

anlaşılabilir olmasıdır. Ancak, KA modelleri küçük veri setlerinde aşırı uyum (overfitting) riskine sahiptir ve yüksek varyans gösterebilir. Bu çalışmada KA, sentetik veri büyütmenin basit, tek ağaç modellerinde aşırı uyum riskini nasıl azalttığını ve genelleme kapasitesini nasıl iyileştirdiğini göstermek için kullanılmıştır. 306 örneklik küçük veri setinde eğitilen KA'nın aşırı uyum yapma eğilimi, 15.000 örneklik sentetik veri ile eğitildiğinde daha dengeli ve sağlam karar sınırları öğrenmesi beklenmiştir.

Her model standart hiperparametrelerle eğitilmiştir: gradyan güçlendirme modelleri (XGBoost, LightGBM) için 100 tahmin edici (n_estimators=100), KA için maksimum derinlik 10 (max_depth=10), kEYK için komşu sayısı 5 (n_neighbors=5) kullanılmıştır. Standart hiperparametre kullanımı bilinçli bir tercihtir ve bunun üç temel gerekçesi bulunmaktadır: Birincisi, Chen ve Guestrin (2016) çalışmasında belirtildiği gibi, XGBoost ve LightGBM gibi topluluk yöntemlerinin varsayılan parametreleri çoğu veri setinde yüksek performans göstermektedir. İkincisi, küçük test setlerinde (77 örnek) kapsamlı hiperparametre optimizasyonu, aşırı uyum riskini artırarak sonuçların genellenebilirliğini azaltabilir. Üçüncüsü, bu çalışmanın birincil odağı sentetik veri kalitesinin değerlendirilmesidir; dolayısıyla model karşılaştırmalarında adil bir zemin oluşturmak için tüm modellerin standart konfigürasyonlarla eğitilmesi tercih edilmiştir. Modeller, kodlanmış eğitim verisi üzerinde eğitilmiş ve ayrılan test seti üzerinde değerlendirilmiştir.

3.5. Değerlendirme Ölçütleri

Model performansı, hem sınıflandırma eşiğine bağlı (doğruluk, kesinlik, duyarlılık, F1) hem de eşikten bağımsız (AUC-ROC) metrikler kullanılarak değerlendirilmiştir. F1 skoru, birincil eniyileme (optimization) ölçütü olarak kullanılmış, kesinlik ve duyarlılık dengesini sağlayarak sınıf dengesizliğine karşı sağlamlık göstermiştir. AUC-ROC, modellerin karar eşiklerinden bağımsız olarak genel ayırt etme yeteneğini ölçmüştür.

İkincil ölçütler olarak doğruluk, kesinlik ve duyarlılık kullanılmıştır. Bu ölçütler toplu olarak farklı klinik dengeleri (trade-off) yakalar. Nüks tahmininde, yanlış negatif (false negative) (kaçırılan nüks vakaları) minimizasyonu özellikle önemlidir, çünkü bu tür hatalar ikincil tedavi müdahalelerini geciktirebilir.

3.6. Deneysel Kurulum

Yapılan testlerde iki farklı veri yapılandırması karşılaştırılmıştır:

Orijinal Veri Seti: 306 eğitim örneği, doğal sınıf dengesizliği (%72 Hayır, %28 Evet)

Sentetik Veri Seti: 15.000 sentetik üretilmiş örnek (10.000 Hayır ve 5.000 Evet), azınlık sınıfı temsilini iyileştirmek ve gizlilik koruyan model geliştirmeyi sağlamak amacıyla üretilmiştir.

Tüm modeller, orijinal ve sentetik veri setleri üzerinde ayrı ayrı eğitilmiş, ardından aynı 77 örneklik gerçek sınama seti üzerinde değerlendirilmiştir. Karşılaştırmalı performans analizi, mutlak ve göreceli ölçüt farklılıkları kullanılarak yapılmıştır.

3.7. İstatistiksel Analiz

Gerçek ve sentetik veri üzerinde eğitilen modeller arasındaki karşılaştırma için iki farklı analiz yöntemi uygulanmıştır: göreceli değişim analizi ve model kararlılık analizi.

3.7.1. Göreceli değişim analizi

Her model için orijinal ve sentetik veri arasındaki göreceli yüzde değişim (η) şu şekilde hesaplanmıştır (eşitlik 1):

$$\eta\,(\%) = \frac{(\text{Metrik}_{sent.} - \text{Metrik}_{orij.})}{\text{Metrik}_{orij.}} \times 100 \qquad (1)$$

Bu analiz sonuçları Tablo 1'de her metrik için sunulmuştur. Pozitif yüzde değerleri sentetik verinin o metrikte iyileşme sağladığını, negatif değerler ise performans kaybını göstermektedir.

3.7.2. Model kararlılık analizi

Her modelin veri tipleri arasındaki tutarlılığını ölçmek için kararlılık skoru hesaplanmıştır. Tüm metrikler genelinde ortalama varyans hesaplanarak, kararlılık skoru $1 - Ort.Varyans$ formülü ile elde edilmiştir. Yüksek kararlılık skoru (1'e yakın), modelin farklı veri kaynaklarında tutarlı performans gösterdiğini ifade etmektedir. Analiz sonuçları Tablo 3'te gösterilmiştir: Ortalama varyans modelin genel tutarlılığını, maksimum varyans en büyük performans farklılığını, kararlılık skoru ise genel sağlamlığı göstermektedir.

4. Bulgular

4.1. Model Performans Karşılaştırması

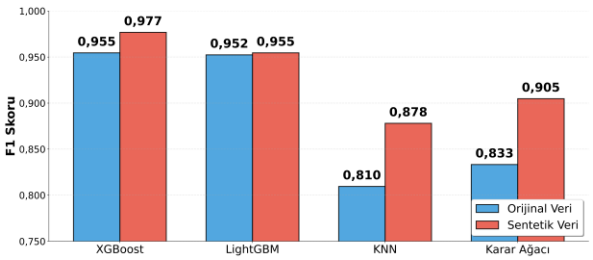
Tablo 1'de tüm modellerin orijinal ve sentetik veri üzerindeki performans ölçütleri sunulmaktadır. Sonuçlar, sentetik veri ile eğitilen modellerin, orijinal veri ile eğitilen modellerle karşılaştırılabilir performans gösterdiğini ortaya koymaktadır.

Şekil 4'te modellerin F1 skorları karşılaştırılmaktadır. Tüm modellerde sentetik veri ile eğitim, orijinal veriye çok yakın veya daha iyi performans göstermektedir.

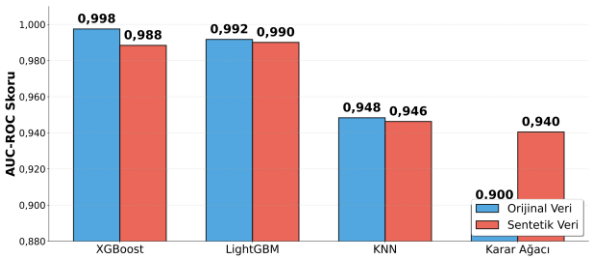
Şekil 5'te modellerin AUC-ROC skorları gösterilmektedir. Topluluk yöntemleri (XGBoost ve LightGBM) hem orijinal hem de sentetik veride 0,98'in üzerinde AUC-ROC skorları elde etmiş, bu da mükemmel ayırt etme yeteneği göstermektedir.

Tablo 1. Model performans karşılaştırması: Orijinal ve sentetik veri üzerinde eğitilen modellerin sınama seti üzerindeki performansı

Metrik	XGBoost	LightGBM	kEYK	KA
Orijinal Veri				
Doğruluk	0,974	0,974	0,896	0,896
Kesinlik	0,955	1,000	0,850	0,769
Duyarlılık	0,955	0,909	0,773	0,909
F1	0,955	0,952	0,810	0,833
AUC-ROC	0,998	0,992	0,948	0,900
Sentetik Veri				
Doğruluk	0,987	0,974	0,935	0,948
Kesinlik	1,000	0,955	0,947	0,950
Duyarlılık	0,955	0,955	0,818	0,864
F1	0,977	0,955	0,878	0,905
AUC-ROC	0,988	0,990	0,946	0,940
Göreceli Değişim (%)				
Doğruluk	+1,33	0,00	+4,35	+5,80
Kesinlik	+4,76	-4,55	+11,46	+23,50
Duyarlılık	0,00	+5,00	+5,88	-5,00
F1	+2,33	+0,23	+8,46	+8,57
AUC-ROC	-0,91	-0,17	-0,22	+4,50



Şekil 4. F1 skoru karşılaştırması. LightGBM modelinde tam eşitlik, basit modellerde iyileşme gözlemlenmiştir.



Şekil 5. AUC-ROC skoru karşılaştırması. Topluluk yöntemleri her iki veri tipinde de yüksek performans göstermiştir.

4.2. Model Kararlılık ve Performans Sıralaması

Model kararlılık analizi, XGBoost ve LightGBM'in en tutarlı modeller olduğunu ortaya koymuştur. Bu iki topluluk yöntemi, veri tipleri arasında neredeyse identik performans sergilemiştir. kEYK ve KA daha düşük kararlılık skorları göstermiş, bu da sentetik verinin bu modellerin performansını daha fazla etkilediğini ortaya koymuştur. F1 skoruna göre model sıralamasında, XGBoost-Sentetik en yüksek performansı göstermiştir (Tablo 2). Dikkat çekici bir bulgu, sentetik veri ile eğitilen modellerin ilk iki sırada yer almasıdır. 8 model-veri kombinasyonu içinde sentetik veri ile eğitilen 4 modelden 3'ü (XGBoost, LightGBM, KA) üst sıralarda yer almış, bu da sentetik verinin bazı durumlarda orijinal veriden daha iyi performans sağladığını göstermektedir.

Tablo 2. Model sıralaması: F1 skoru ve AUC-ROC'ye göre

Sıra	Model	Veri Tipi	F1 Skoru	AUC-ROC
1	XGBoost	Sentetik	0,977	0,988
2	LightGBM	Sentetik	0,955	0,990
3	XGBoost	Orijinal	0,955	0,998
4	LightGBM	Orijinal	0,952	0,992
5	KA	Sentetik	0,905	0,940
6	kEYK	Sentetik	0,878	0,946
7	KA	Orijinal	0,833	0,900
8	kEYK	Orijinal	0,810	0,948

Tablo 3. Model kararlılık analizi: Veri tipleri arası tutarlılık

Model	Ort, Varyans	Maks, Varyans	Kararlılık Skoru
XGBoost	0,0179	0,0455	0,9821
LightGBM	0,0189	0,0455	0,9811
kEYK	0,0505	0,0974	0,9495
KA	0,0780	0,1808	0,9220

5. Tartışma

5.1. Sentetik Veri ile Gerçek Veri Performans Eşitliği

Bu çalışmanın en önemli bulgusu, sentetik veri ile eğitilen modellerin orijinal veri ile eğitilen modellerle karşılaştırılabilir performans göstermesidir. Özellikle LightGBM ve XGBoost gibi topluluk yöntemlerinde, sentetik veri kullanımı model performansını korumuş ve bazı durumlarda iyileştirmiştir. Bu sonuç, sentetik veri üretim sürecinin orijinal verinin temel istatistiksel özelliklerini, korelasyon yapısını ve sınıf ilişkilerini başarılı şekilde koruduğunu göstermiştir.

LightGBM'de gözlemlenen minimal değişimler (F1 skorunda %0,23 artış: 0,952'den 0,955'e; AUC-ROC'de %0,17 azalma: 0,992'den 0,990'a), modelin sentetik veride kesinlik-duyarlılık dengesini yeniden ayarladığını ancak genel ayırt etme kapasitesini koruduğunu göstermektedir.

XGBoost'ta ise F1 skorunda %2,33 iyileşme gözlemlenmiştir (0,955'ten 0,977'ye), ancak AUC-ROC'deki %0,91'lik azalma (0,998'den 0,988'e) kabul edilebilir düzeydedir. Rajkomar vd. (2019) tarafından yapılan çalışmada belirtildiği gibi, tıbbi tahmin modellerinde küçük metrik değişimleri, gizlilik koruma ve veri paylaşımı dezavantajları göz önüne alındığında kabul edilebilir dengeler olarak değerlendirilmektedir.

Bu bulgular, sentetik verinin sadece veri büyütme değil, aynı zamanda model kalitesini koruyarak gizlilik sağlayan bir çözüm sunduğunu ortaya koymaktadır.

5.2. Basit Modellerde Performans İyileşmesi

kEYK ve KA modellerinde sentetik veri kullanımının performansı önemli ölçüde artırması, ilgi çekici bir bulgudur. kEYK'de F1 skoru %8,46 artış gösterirken

(0,810'dan 0,878'e), KA modelinde %8,57'lik bir iyileşme kaydedilmiştir (0,833'ten 0,905'e). AUC-ROC açısından ise KA'da %4,50 artış gözlemlenmiştir (0,900'den 0,940'a). Bu iyileşme, sentetik verinin iki temel avantajını ortaya koymaktadır:

Veri Büyütme Etkisi: 306 örnekten 15.000 örneğe çıkış, öznitelik uzayında daha yoğun bir kaplama sağlamış ve bu basit modellerin öğrenme kapasitesini artırmıştır. Shorten ve Khoshgoftaar'in (2019) yaptıkları çalışmada da belirttikleri gibi, veri büyütme küçük veri setlerinde özellikle etkilidir.

Aşırı Uyum Azaltma: Orijinal verinin küçük boyutu, KA gibi yüksek değişkenlik (variance) gösteren modellerde aşırı uyum riskini artırmaktadır. Sentetik verinin büyük boyutu, daha sağlam ve genellenebilir karar sınırları öğrenilmesini sağlamıştır.

5.3. Topluluk Yöntemlerinin Sağlamlığı

XGBoost ve LightGBM gibi topluluk yöntemlerinin hem küçük hem de büyük veri setlerinde yüksek performans göstermesi, gradyan güçlendirme algoritmalarının farklı veri boyutlarında sağlamlığını doğrulamaktadır. Kararlılık analizi, XGBoost (kararlılık skoru: 0,9821) ve LightGBM'in (kararlılık skoru: 0,9811) veri tipleri arasında en tutarlı performansı sergilediğini ortaya koymuştur. Yüksek kararlılık skorları, sentetik veri üretim sürecinin bu modellerin öğrendiği temel örüntüleri başarılı şekilde koruduğunu göstermiştir.

Orijinal veride eğitilen XGBoost modelinin çok yüksek AUC-ROC skoru, küçük veri setlerinde aşırı uyum riskine işaret etmektedir. Ancak test seti performansının korunması, modelin genelleme kapasitesini sürdürdüğünü göstermektedir. Sentetik veri ile eğitilen modellerde gözlemlenen daha dengeli metrikler, aşırı uyum riskinin azaldığını ve modelin daha sağlam öğrenme gerçekleştirdiğini ortaya koymuştur.

5.4. Klinik Çıkarımlar ve Veri Kıtlığı Problemine Çözüm

Tiroit kanseri nüks tahmininde sentetik veri kullanımının başarılı sonuçlar vermesi, tıbbi yapay zekâ uygulamalarına ilişkin önemli çıkarımlar elde edilmesini sağlamıştır. Bu çalışmada gösterilen 49 kat veri büyütme ve model performansının korunması, veri kıtlığının ciddi engel oluşturduğu birçok alanda (tıbbi veya değil) sentetik veri üretiminin uygulanabilir bir çözüm sunduğunu göstermiştir.

Sentetik veri, KVKK ve GDPR gibi düzenlemelere uygun şekilde veri paylaşımına olanak tanır ve araştırmacıların gerçek hasta verilerini paylaşmadan model geliştirmesini sağlar. Bu özellik, nadir hastalıklar, pediatrik vakalar veya hasta sayısının doğal olarak kısıtlı olduğu alanlarda yapay zekâ modellerinin geliştirilmesini mümkün kılar. Ayrıca, azınlık sınıfı örneklerinin sayısını artırarak sınıf dengesizliği problemine çözüm sunar ve modellerin kritik vakaları daha iyi öğrenmesini sağlar.

Çok merkezli çalışmalarda, farklı hastanelerin kendi verilerinden sentetik veri üretip paylaşması veri siloları problemini aşabilir ve model genelleme kapasitesini artırabilir. Bu yaklaşım, büyük veri setlerine erişimi olmayan araştırma merkezlerinin yapay zekâ çalışmalarına katılmasını sağlayarak tıbbi yapay zekâ araştırmalarında demokratikleştirici bir etki yaratır. Bu çalışmada da elde edilen yüksek performans skorları, bu modellerin klinik karar destek sistemlerinde kullanılabileceğini göstermiştir.

5.5. Sentetik Veri Kalitesi

Sentetik verinin yaş dağılımı orijinal veriyle yüksek benzerlik göstermekte ve dağılımsal tutarlılık (distributional fidelity) sağlamaktadır. Jordon vd. (2022) tarafından önerilen çerçeveye göre, başarılı bir sentetik veri üretimi benzerlik, kullanışlılık ve gizlilik boyutlarında dengeli olmalıdır. Bu çalışmadaki sonuçlar, üretilen sentetik verinin benzerlik ve kullanışlılık boyutlarında başarılı olduğunu göstermiştir.

Sentetik verinin önemli bir avantajı, sınıf dengesini kontrollü şekilde iyileştirme yeteneğidir. Orijinal veride 2,55:1 oranında (%71,8 Hayır - %28,2 Evet) bulunan sınıf dengesizliği, sentetik veride 2:1 oranına (%66,7 Hayır - %33,3 Evet) iyileştirilmiştir. Geleneksel aşırı örnekleme (oversampling) yöntemleri aynı örneklerin tekrarlanması nedeniyle aşırı uyum riskini artırırken, sentetik veri üretimi her bir örneğin benzersiz olmasını sağlar. SMOTE (Chawla vd., 2002) gibi ara değer üretimine dayalı yöntemler, mevcut örnekler arasında doğrusal bir geçiş (enterpolasyon) yaptıkları için, bu çalışmadaki ölçek artışında (306'dan 15.000'e) gerçekçi sonuçlar üretmeyebilir. Gelecek çalışmalarda, orta ölçekli veri setlerinde karşılaştırmalı analizler önerilmektedir.

5.6. Kısıtlılıklar

Bu çalışmanın bazı kısıtlılıkları bulunmaktadır. Çalışma tek bir tiroit kanseri veri seti üzerinde gerçekleştirilmiş olduğundan, sonuçların genellenebilirliği farklı kanser türleri ve tıbbi alanlar üzerinde doğrulama gerektirmektedir. Test kümesinin boyutu istatistiksel güç açısından yetersiz olduğundan sonuçları daha sağlam hâle getirmek için, tamamen ayrı bir insan grubu üzerinde (bağımsız bir kohortta) tekrar doğrulama yapılması gerekmektedir.

Katkı Oranı Beyanı

Yazarın katkı yüzdeleri aşağıda verilmiştir. Yazar makaleyi incelemiş ve onaylamıştır.

	E.G.
K	100
T	100
Y	100
VTI	100
VAY	100
KT	100
YZ	100
KI	100
GR	100

K= kavram, T= tasarım, Y= yönetim, VTI= veri toplama ve/veya işleme, VAY= veri analizi ve/veya yorumlama, KT= kaynak tarama, YZ= yazım, KI= kritik inceleme, GR= gönderim ve revizyon.

Çatışma Beyanı

Yazar bu çalışmada hiçbir çıkar ilişkisi olmadığını beyan etmektedirler.

Etik Onay Beyanı

Bu çalışmada kullanılan veri seti, UCI Makine Öğrenmesi Deposu'ndan halka açık olarak erişilebilen, anonim hasta verilerinden oluştuğundan etik kurul onayı gerekmemektedir.

Destek ve Teşekkür Beyanı

Bu çalışmada kullanılan Farklılaşmış Tiroit Kanseri Nüks veri setini halka açık olarak sunan araştırmacılara teşekkür ederiz. Orijinal tiroit kanseri veri seti UCI Makine Öğrenmesi Deposu'nda halka açıktır (https://archive.ics.uci.edu). Sentetik veri, DataXid platformu ile üretilmiştir ve gizlilik nedeniyle halka açık değildir. Model eğitim kodları, makul talep üzerine sorumlu yazardan temin edilebilir.

Kaynaklar

Bender, S., Jarmin, R. S., Kreuter, F., & Lane, J. (2020). Privacy and confidentiality. In *Big data and social science* (pp. 313–331). Chapman and Hall/CRC.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2022). Language models are realistic tabular data generators. *arXiv*. https://doi.org/10.48550/arXiv.2210.06280

Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhian, A. (2024). Machine learning for risk stratification of thyroid cancer patients: A 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2095–2104. https://doi.org/10.1007/s00405-023-08766-2

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497. https://doi.org/10.1038/s41551-021-00751-8

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

DataXID. (2024). *DataXID: Blockchain-powered synthetic data platform*. https://dataxid.com

Donaldson, M. S., & Lohr, K. N. (Eds.). (1994). *Health data in the information age: Use, disclosure, and privacy*. National Academies Press.

El Emam, K., Mosquera, L., & Bass, J. (2020b). Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research*, *22*(11), e23139. https://doi.org/10.2196/23139

El Emam, K., Mosquera, L., & Hoptroff, R. (2020a). *Practical synthetic data generation: Balancing privacy and the broad availability of data*. O'Reilly Media.

El Mestari, S. Z., Lenzini, G., & Demirci, H. (2024). Preserving data privacy in machine learning systems. *Computers & Security*, *137*, 103605. https://doi.org/10.1016/j.cose.2023.103605

Habchi, Y., Himeur, Y., Kheddar, H., Boukabou, A., Atalla, S., Chaker, D., & Mansoor, W. (2023). AI in thyroid cancer diagnosis: Techniques, trends, and future directions. *Systems*, *11*(10), 519. https://doi.org/10.3390/systems11100519

Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E., Pacini, F., Randolph, G. W., Sawka, A. M., Schlumberger, M., Schuff, K. G., Sherman, S. I., Sosa, J. A., Steward, D. L., Tuttle, R. M., & Wartofsky, L. (2016). 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. *Thyroid*, *26*(1), 1–133. https://doi.org/10.1089/thy.2015.0020

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data: What, why and how? *arXiv*. https://doi.org/10.48550/arXiv.2205.03257

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146–3154). Curran Associates.

Lobato de Faria, P., & Cordeiro, J. V. (2014). Health data privacy and confidentiality rights: Crisis or redemption? *Revista Portuguesa de Saúde Pública*, *32*(2), 123–133. https://doi.org/10.1016/j.rpsp.2014.10.001

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(1), 17–48. https://doi.org/10.3322/caac.21763

Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems 32* (pp. 7335–7345). Curran Associates.

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244–255. https://doi.org/10.1016/j.neucom.2020.07.134.