Sentetik Veri Üretici Değerlendirme Görselleştirme Aracı

Synthetic Data Generator Evaluative Visualization Tool

Ahmet Yasin Aytar, Selim Balcısoy
Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey 34956
ahmet.aytar@sabanciuniv.edu, selim.balcisoy@sabanciuniv.edu

***Özetçe*—Hassas alanlarda makine öğrenimi kullanımının artması, gizlilik koruması ve veri dengesi için sentetik veri üretimini zorunlu kılmıştır. Bu çalışma, sentetik veri oluşturma sürecini optimize eden yenilikçi bir Değerlendirme Görselleştirme Aracı sunmaktadır. Araç, GAN modelinin kayıplarını ve performans metriklerini gerçek zamanlı görselleştirerek eğitim sürecinin aktif izlenmesini sağlar. Veri kümesinin dinamik örneklemeleriyle GAN eğitimini aşamalı genişleten sistem, kullanıcıların tüm veri setini kullanmadan veri yakınsamasını iyileştirmelerine ve hiperparametreleri etkili ayarlamalarına olanak tanır. Çalışma, aracın etkinliğini doğrulamakta ve potansiyel iyileştirmeleri vurgulamaktadır. Bu araştırma, üretim sürecini basitleştiren ölçeklenebilir bir çözüm sunarak sentetik veri gerektiren alanlara önemli katkı sağlamaktadır.**

Anahtar Kelimeler — Sentetik Veri; GANs; Görsel Analitik

***Abstract*—The increasing use of machine learning in sensitive domains has necessitated synthetic data generation to ensure privacy protection and data balance. This study introduces an innovative Synthetic Data Generator Evaluation Visualization Tool to address the complexities of creating and evaluating synthetic data. Our tool visualizes GAN model losses and performance metrics in real-time, enabling users to actively monitor and evaluate the training process. By progressively expanding GAN training with dynamic sampling of the dataset, our tool allows users to iteratively improve data convergence and efficiently adjust hyperparameters without using the entire dataset. The study validates the tool's effectiveness through rigorous testing and highlights potential improvements to extend its applicability. This research advances synthetic data studies by providing a scalable solution that simplifies the generation process, making a significant contribution to data-driven fields requiring synthetic datasets.**

Keywords — Synthetic Data; GANs; Visual Analytics

I. GİRİŞ

Veri bilimi ve makine öğrenimi alanında, gerçek verilerin gizlilik endişeleri, erişilebilirlik sorunları veya kalite kısıtlamaları nedeniyle sınırlı olduğu durumlarda, sentetik veri üretimi kritik bir odak alanı olarak ortaya çıkmıştır. Otantik kümelerin özelliklerini yansıtacak şekilde yapay olarak oluşan sentetik veriler, özellikle veri hassasiyetinin büyük önem taşıdığı ve veri dengesizliğinin yaşandığı alanlarda merkezi bir role sahiptir. Bu kapsamda ortaya çıkan Çekişmeli Üretici Ağlar (GAN'lar), 2014 yılında tanıtılan ve sentetik veri üretiminde devrim yaratan derin öğrenme modellerinin öncüsüdür [1].

GAN'lar etkili araçlar olarak kabul edilmiş olsa da, mod çöküşü ve yakınsama başarısızlığı gibi önemli zorluklar içermektedir. Mod çöküşü, GAN'ların çeşitlilik eksikliği yaşadığı ve üretici modelin sınırlı bir örnek alt kümesi ürettiği durumu ifade ederken; yakınsama başarısızlığı, üretici ve diskriminatör arasındaki dengesiz eğitim dinamiklerinden kaynaklanmaktadır. Mevcut yaklaşımlar bu sorunları ele alsa da, özellikle büyük veri kümeleri söz konusu olduğunda eğitim sürecinin yoğun hesaplama gereksinimleri, bu sorunların etkin bir şekilde izlenmesini zorlaştırmaktadır. GAN'ların performansını değerlendirmede, sentetik ve gerçek veri dağılımları arasındaki farkı ölçen Kullback-Leibler (KL) sapması, sentetik veri analizinde en yaygın kullanılan benzerlik metriği olarak öne çıkmaktadır [2]. KL sapması, iki olasılık dağılımı arasındaki farklılığı nicel olarak ölçerek, üretilen sentetik verilerin gerçek verilere ne kadar yakınsadığını değerlendirmede kritik bir rol oynamaktadır.

Önerdiğimiz Sentetik Veri Üretici Değerlendirme Görselleştirme Aracı, TensorFlow ekosisteminde yaygın kullanılan TensorBoard [3] gibi mevcut görselleştirme araçlarının ötesine geçerek özgün katkılar sunmaktadır. TensorBoard, eğitim sürecinde kayıp ve metrik değerlerinin pasif izlenmesine olanak tanırken, bizim aracımız veri kümesinin dinamik örneklemeleri ile GAN eğitimini aşamalı şekilde genişleterek hesaplama yükünü azaltır ve kullanıcıların eğitim sürecine aktif müdahalesine imkân tanır. Bu yaklaşım, kullanıcıların her iterasyonda üretilen sentetik verilerin kalitesini gerçek zamanlı olarak değerlendirmesine, hiperparametreleri iteratif olarak ayarlamasına ve mod çöküşü veya yakınsama başarısızlığı gibi sorunları erken tespit etmesine olanak sağlar. Çalışmamız, GAN eğitiminde kullanılan veri miktarını aşamalı olarak artıran bir yaklaşım sunarak, GAN'ların temel zorluklarını ele alırken yüksek kaliteli sentetik veri üretimini daha erişilebilir ve verimli hale getirmektedir.

979-8-3315-6655-5/25/$31.00 ©2025 IEEE

II. LİTERATÜR ÖZETİ

Sentetik veri üretimindeki özellikle son gelişmeler, kapsamlı araçlar üretimi, sofistike değerlendirme metodolojileri ve yinelemeli rafinasyon süreçlerinin gerekliliğini vurgulamaktadır.

Sentetik veri üretimi için geliştirilen araçlar, son yıllarda önemli gelişmeler göstermiştir. Patki ve diğer. (2016) tarafından geliştirilen Synthetic Data Vault (SDV), tabular verilerin sentetik versiyonlarını otomatik olarak üretmek için tasarlanmış açık kaynaklı bir araçtır [4]. Bu araç, veri bilimcilere farklı veri tipleri için sentetik veri üretme imkanı sunsa da, kullanıcı etkileşimi ve gerçek zamanlı değerlendirme konusunda sınırlı kalmaktadır.

Sentetik veri kalitesinin değerlendirilmesi sıklıkla veri dağılımlarının orijinal veriye uyumu etrafında şekillenmektedir. Arnold ve Neunhoeffer (2020), farklılaştırılmış gizli veriler için dağılımların benzerliğine odaklanmış, Campbell ve diğer. (2020) ise bunu elektromiyografi (EMG) verileri alanına genişleterek, özellik uzayı projeksiyonları aracılığıyla nitel ve nicel bir değerlendirme önermişlerdir [5][6].

Yinelemeli üretim ve değerlendirme yöntemleri, Razghandi ve diğer. (2022) akıllı şebeke verileri için VAE-GAN kullanarak sapma metrikleri ile uyumsuzluk ölçümleri kullanarak üretilen veri kalitesini iteratif olarak değerlendirmesiyle ivme kazanmıştır [7]. Zhang ve diğer. (2022) ve Cheung ve diğer. (2022), sırasıyla elektronik sağlık kayıtlarını simüle ederken ve akış sitometrisi için sentetik veri setleri üretirken yinelemeli metodolojiler kullanmış ve progresif veri rafinasyonunu sağlamışlardır [8][9].

Literatürdeki çalışmalar sentetik veri üretiminde kalite değerlendirmesinin önemini vurgulamış, yineleme ve görselleştirmenin veri kalitesini artırmadaki rolünü belirtmiştir. Ancak sentetik veri kalitesini değerlendirirken sadece dağılım benzerliği ve özellik korelasyonu gibi dar metrik setlerine dayanmak, verinin aslına uygunluğu konusunda yanıltıcı sonuçlar doğurabilir. Özellikle karmaşık GAN tabanlı derin öğrenme algoritmalarının kullanıldığı üretim süreçleri, yüksek hesaplama maliyetleri ve uzun işlem süreleriyle zorlaşmakta, büyük veri kümeleriyle çalışırken ciddi darboğazlar oluşmaktadır. Hiperparametre seçimi sentetik veri kalitesini büyük ölçüde etkilemesine rağmen, mevcut yöntemler bunların ayarlanmasında yetersiz kalmaktadır.

Çözümümüz, kullanıcıların sentetik verinin orijinale yakınsamasını yinelemeli olarak görselleştirmelerine olanak tanır. Sistem, KL sapmasını optimum aralıkta tutmak ve kayıpları stabilize etmek üzere tasarlanmıştır, böylece tüm eğitim iterasyonlarının tamamlanması beklenmeden yakınsama sinyali alınabilir. Bu yaklaşım sadece değerlendirmeyi hızlandırmakla kalmaz, aynı zamanda hiperparametre değişikliklerinin etkilerini de gösterir. Kullanıcılar bu sayede parametreleri iteratif şekilde ayarlayabilir ve sentetik veri üretimini gerçek zamanlı gözlemleyerek daha verimli ve kaliteli sonuçlar elde edebilir.

III. METOT

Sentetik Veri Üretici Değerlendirme Görselleştirme Aracı, büyük veri kümeleri için GAN tabanlı modellerin hesaplama yoğunluğu ve hiperparametre ayarlama zorluklarına odaklanarak sentetik veri kalitesini yinelemeli değerlendirmek üzere tasarlanmıştır. Araç genel olarak tüm GAN tabanlı modellerle uyumlu olsa da, bu çalışmada tabular verilerde etkinliği kanıtlanmış olan CTGAN modeli tercih edilmiştir [10]. Bu model, kategorik ve sürekli değişkenleri etkili işleyebilmesi ve veriye özgü koşullu dağılımları öğrenebilmesi nedeniyle seçilmiştir.

Kullanıcılar Excel dosyasında veri kümesini yükledikten sonra kolon adı ve tipi gibi parametreleri seçerek sistemi yapılandırır ve görselleştirme modunu belirler. Görselleştirme modları çizgi, renk veya her ikisi olabilir: Çizgi modu, orijinal ve sentetik veriler arasındaki mesafeyi çizgi kalınlığıyla temsil eder (kalın çizgi daha yakın uyum); renk modu ise yakınlığı yeşilden (yakın) kırmızıya (uzak) değişen gradyanla gösterir. Her iki mod birlikte kullanıldığında, farklı dağılım segmentleri boyunca sentetik verilerin orijinale göre durumunu gösteren çok boyutlu bir görsel elde edilir.

Algoritma 1 aracın iteratif bir şekilde sentetik veri üretme sürecini izah etmektedir. İlk iterasyonda araç daha önce kullanılmamış orijinal veri kümesinden kullanıcı tarafından seçilen örneklem sayısınca örneklem alır. Bu yeni veriler eğitim setine eklenir ve GAN modeli eğitilir. Eğitilen model kullanılarak mevcut örneklem sayısınca sentetik veri üretilir. Üretilen verilerin dağılımı ve metrikleri ekrana yansıtılır ve bir sonraki iterasyona geçilir. İkinci iterasyonda orijinal verinin kullanılmamış kısmından seçilen örneklem sayısınca örneklem alınır, ve bu veriler eğitim setine eklenir böylece örnek boyutu iki katına çıkar. Seçilen GAN modeli elde edilen örneklemler ile tekrardan eğitilir, ve süreç böylece devam eder. Her bir adımda, dağılım grafikleri güncellenir. Modeli eğitmek için, kullanımı kolay özellikleri nedeniyle ydata-sentetik kütüphanesinin RegularSynthesizer'ı kullanılmıştır [11].

ALGORİTMA 1: İTERATİF SENTETİK VERİ ÜRETİMİ ALGORİTMASI

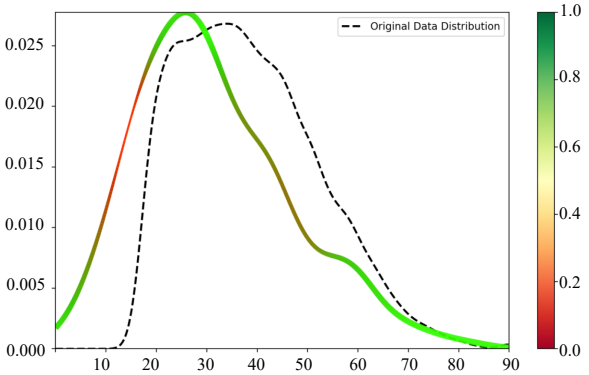
```
df ← orijinal veri seti
params ← seçilen eğitim ve model parametreleri
sentezör ← RegularSynthesizer
toplananVeri ← {}
kalanVeri ← df
for i ← 1 to [len(df) // params.ÖrneklemMiktarı] do
    yeniÖrneklem ← kalanVeri'den boyutu
    params.ÖrneklemMiktarı olan örneklem al
    toplananVeri ← toplananVeri ∪ yeniÖrneklem
    kalanVeri ← kalanVeri - yeniÖrneklem
    sentezör.FIT(toplananVeri, params.eğitimparametreleri)
    sentetikData ← sentezör.SAMPLE(len(toplananVeri))
    DAĞILIMGÖRSELLEŞTİR(sentetikData,
    params.GörselleştirmeModu)
end for
```

Aynı zamanda, araç her iterasyon için orijinal ve sentetik dağılımlar arasındaki KL sapmasını hesaplar ve hem generatör hem de diskriminatör için en son 'epoch' kayıplarını ekranda kullanıcıya verir. Bu metrikler, sentetik verilerin gerçek verilere olan yakınsamasının nicel göstergeleri olarak hizmet eder. Yakınsama metriklerini gözlemleyen veya tatmin edici bir görsel eşleşmeye ulaşan kullanıcılar, süreci durdurmayı veya farklı yapılandırmalarla deney yapmayı seçebilirler. Bu özellik, hesaplama kaynaklarının ve zamanın hat safhada önemli olduğu

büyük veri kümelerini yönetmede aracı özellikle etkili kılan bir esneklik ve kontrol seviyesi sunar. Bu yinelemeli süreç, gerçek zamanlı görsel ve nicel geri bildirim ile donatılmış olup, kullanıcılara GAN model parametrelerini yinelemeli olarak iyileştirmek için gerekli içgörüleri sağlar.

IV. BULGULAR

Çalışmanın test ve sonuç aşamasında, UCI Makine Öğrenmesi Repo'sundan alınan "Adult" veri seti kullanılmıştır [12]. Veri seti, yaş, çalışma sınıfı, eğitim, medeni durum gibi kategorik ve sayısal değişkenlerden oluşmaktadır. Bu çalışmada test edilen özellik "age" (yaş) olup, sürekli sayısal veri niteliğindedir, ve özellik dağılımı **Şekil 1**'de kesikli çizgi ile gösterilmiştir. Sistemin seçilen parametreleri **TABLO I**'de verilmiştir. Ayrıca, CTGAN mimarisinin ortaya atıldığı çalışmada belirtildiği gibi Üretici (Generator) ve Ayırıcı (Discriminator) ağlar için kayıp fonksiyonu olarak "Wasserstein GAN with Gradient Penalty (WGAN-GP)" ve Adam optimizasyon algoritması kullanılmıştır. İlk iterasyonda, **Şekil 1**'de tasvir edildiği üzere, kesikli çizgi orijinal seçilen yaş kolonu için orijinal veri dağılımını temsil ederken, yeşil-kırmızı çizgi üretilen sentetik veri dağılımını göstermektedir. Sentetik veriler orijinal dağılımdan önemli bir sapma göstermiş ve çoğu noktada kırmızı renk ve ince çizgi gözlemlenmiştir. Ayrıca sentetik veri dağılım grafiği orijinal verinin dağılım grafiğindeki genel trendini yakalamada oldukça başarısızdır.

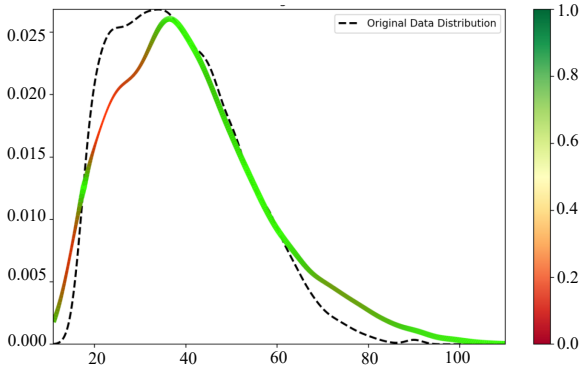


Şekil. 1.İlk İterasyon için Orijinal ve Sentetik Veri Dağılımları

TABLO I. EĞİTİM İÇİN SEÇİLEN PARAMETRELER

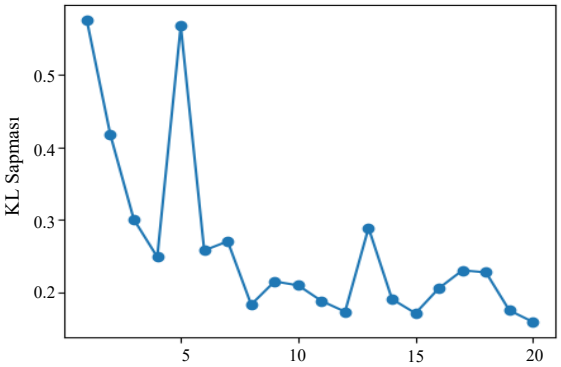
Parametre İsimleri	**Değerler**
Kolon İsmi	Yaş
Kolon Tipi	Sayısal
Görselleştirme Modu	Her ikisi
Epoch Sayısı	10
Batch Boyutu	500
Öğrenme Oranı	0.0001
Beta 1	0.6
Beta 1	0.9
Model İsmi	CTGAN
Örneklem Miktarı	500

Şekil 2'de ise 20.000 satırdan oluşan veri setinden 500 örneklem miktarı ile yapılan bu çalışmada toplam 40 iterasyonun orta noktası olan 20. iterasyondan önce sentetik verilerin gerçek dağılıma yakınsadığı gözlenmiştir. Sentetik veri dağılımı orijinal veri dağılımının sağ tarafında neredeyse birebir eşleşme yakalerken genel trendi yakalamada da oldukça başarılıdır. Sadece 20 iterasyon içinde hızlı bir yakınsama, özellikle işlenen büyük veri hacmi göz önüne alındığında, aracın verimliliğine dair güçlü bir kanıttır.



Şekil. 2.20. İterasyon için Orijinal ve Sentetik Veri Dağılımları

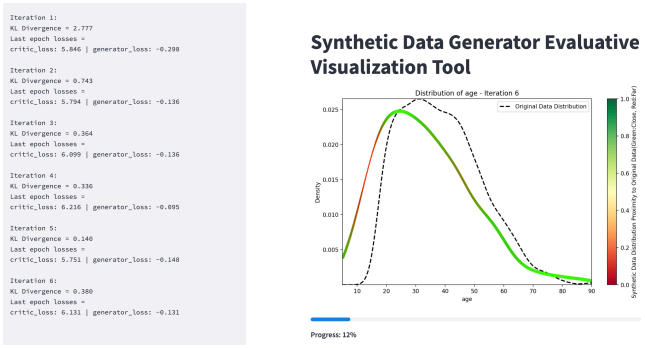
Şekil 3, bu iddiayı desteklemektedir, burada KL sapması iterasyonlar boyunca çizilmiştir. Bu değerin, özellikle ilk 20. İterasyon içerisindeki hareketi, araştırma tezini destekleyen güçlü kanıtlar sunmaktadır. Azalan KL sapma değerleri, sentetik veri kalitesindeki görsel iyileşmelerle uyumlu iken, GAN modelinin bir dengeye yaklaştığını göstermektedir. Özellikle 7. iterasyondan sonra, KL sapması değerinin neredeyse doyuma ulaştığı görülebilir. Bu denge, sentetik verilerin sadece rastgele bir yaklaşım olmaktan çıkıp orijinal veri setinin istatistiksel olarak oldukça güçlü bir temsili haline geldiğine işaret eder. Bu metriklerin iteratif sürecin orta noktasından çok daha önce yakınsaması, sistemin orjinal veri dağılımını etkili bir şekilde öğrendiğini gösterir.



Şekil. 3.İlk 20 İterasyon için KL Sapması Değerleri

Sentetik Veri Üretici Değerlendirme Görselleştirme Aracı, çeşitli veri setlerinde yüksek kaliteli sentetik veri üretmede etkinliğini kanıtlamıştır. İterasyonlar ilerledikçe KL sapması önemli yakınsama göstermiş ve değerler kabul edilebilir aralıklarda stabilize olmuştur. Görselleştirme modları veri

kalitesi hakkında anlık ve sezgisel geri bildirim sağlamış, kullanıcıların dağılım farklılıklarını kolayca izleyebilmesine olanak tanımıştır. **TABLO I**'de verilen parametreler ile çalıştırılan uygulamanın 6. iterasyon sonundaki arayüz görüntüsü **Şekil 4**'de verilmiştir. Sol panelde iterasyonlara göre KL sapması ve kayıp değerleri, sağ tarafta ise değişen sentetik ve orijinal veri dağılımları görüntülenmektedir.



Şekil. 4.6. İterasyon için Uygulama Arayüzü Ekran Görüntüsü

V. TARTIŞMA VE GELECEK ÇALIŞMALAR

Sentetik Veri Üretici Değerlendirme Görselleştirme Aracımız, büyük veri kümeleri için geleneksel yaklaşımların gerektirdiği yüksek hesaplama maliyetlerini aşan ikna edici sonuçlar ortaya koymuştur. Daha küçük örneklemlerle başlayıp aşamalı genişleme stratejisi uygulayan aracımız, KL sapması metriğinin 7. iterasyondan sonra doyuma ulaşmasıyla, tüm veri kümesini kullanmadan bile kaliteli sentetik veri üretilebileceğini kanıtlamıştır.

Bununla birlikte yaklaşımımızın bazı sınırlamaları vardır. Normal dağılım göstermeyen veya çok modlu dağılımlara sahip veri kümeleri için, azaltılmış örneklemler orijinal verinin istatistiksel özelliklerini tam olarak yakalayamayabilir. Bu durum, üretilen sentetik verilerin gerçek dağılımdan sapmasına neden olabilir. Çözüm olarak, dağılım karakteristiklerinin karmaşıklığına göre örneklem büyüklüğünü dinamik olarak belirleyen adaptif bir örnekleme stratejisi geliştirilebilir.

Küçük veri kümeleri için her iterasyonda modeli sıfırdan eğitme yaklaşımı hesaplama açısından verimli olmayabilir. Bu durumlarda "çevrimiçi öğrenme" yaklaşımı daha uygundur; bu yöntemle model parametreleri korunarak yeni gelen verilerle güncellenir ve her iterasyonda sıfırdan başlamak yerine bilgi birikimi artırılır, böylece özellikle küçük veri kümeleri için araç verimliliği önemli ölçüde yükseltilebilir.

Gelecek çalışmalarda, öğrenilen parametreleri iterasyonlar boyunca koruyan gelişmiş bir çevrimiçi öğrenme mekanizması entegre edilebilir, eğitim sürecinin daha detaylı analizi için KL sapması dışında yeni metrikler eklenebilir ve kullanıcıların belirli veri aralıklarını hedefleyebildiği daha interaktif bir süreç tasarlanabilir. Ayrıca, farklı GAN mimarilerinin otomatik karşılaştırılmasını sağlayan bir modül, kullanıcılara en uygun modeli seçme konusunda rehberlik ederek hiperparametre ayarlamasını kolaylaştırabilir.

VI. SONUÇ

Bu çalışmada sunulan Sentetik Veri Üretici Değerlendirme Görselleştirme Aracı, sentetik veri üretimi ve değerlendirmesi alanında önemli bir ilerleme sağlamıştır. Geliştirilen araç, GAN modelleriyle ilişkili hesaplama zorluklarını etkili bir şekilde ele almakta ve sezgisel görselleştirmeler aracılığıyla sentetik veri kalitesinin gerçek zamanlı değerlendirilmesine olanak tanımaktadır. KL sapması metriğinin erken iterasyonlarda doyuma ulaşması, tüm veri kümesini işlemeye gerek kalmadan yüksek kaliteli sentetik veri üretilebileceğini kanıtlamaktadır. Renk ve çizgi kalınlığı gibi görsel ipuçları sayesinde kullanıcılar, sentetik verilerin orijinal dağılıma yakınlığını kolayca değerlendirebilmektedir. Bu çalışma, makine öğrenimi ve veri biliminde giderek önem kazanan sentetik veri üretimi literatürüne değerli bir katkı sağlamakta ve veri gizliliği, dengesizlik ve erişilebilirlik sorunlarının çözümünde önemli potansiyel taşımaktadır.

VII. BİLGİLENDİRME

Çalışmamızda kullanılan kodların Github linkine aşağıdaki adresten erişim sağlanabilir:

https://github.com/anonymsynthetic/Synthetic_Data_SIU/tree/main

KAYNAKLAR

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y., "Generative adversarial nets", Advances in Neural Information Processing Systems, 2014.

[2] Borji, A., "Pros and cons of GAN evaluation measures", Computer Vision and Image Understanding, 2019.

[3] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., "TensorFlow: A system for large-scale machine learning", 12th USENIX Symposium on Operating Systems Design and Implementation, 2016.

[4] Patki, N., Wedge, R., & Veeramachaneni, K., "The Synthetic Data Vault", IEEE International Conference on Data Science and Advanced Analytics, 2016.

[5] Arnold, C., & Neunhoeffer, M., "Really useful synthetic data - A framework to evaluate the quality of differentially private synthetic data", ArXiv, 2020.

[6] Campbell, E., Cameron, J. A. D., & Scheme, E., "Feasibility of data-driven EMG signal generation using a deep generative model", 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020.

[7] Razghandi, M., Zhou, H., Erol-Kantarci, M., & Turgut, D., "Variational autoencoder generative adversarial network for synthetic data generation in smart home", ICC 2022 - IEEE International Conference on Communications, 2022.

[8] Cheung, M., Campbell, J. J., Thomas, R., Braybrook, J., & Petzing, J., "Systematic design, generation, and application of synthetic datasets for flow cytometry", PDA Journal of Pharmaceutical Science and Technology, 2022.

[9] Zhang, Z., Yan, C., & Malin, B. A., "Keeping synthetic patients on track: Feedback mechanisms to mitigate performance drift in longitudinal health data simulation", Journal of the American Medical Informatics Association: JAMIA, 2022.

[10] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K., "Modeling Tabular Data using Conditional GAN", Advances in Neural Information Processing Systems (NeurIPS), 2019.

[11] YData Synthetic. "Synthetic data generation with YData," GitHub Repository, 2024. [Online]. Available: https://github.com/ydataai/ydata-synthetic.

[12] Kohavi, R., & Becker, B., "Adult Data Set," UCI Machine Learning Repository, 1996. [Online]. Available: https://archive.ics.uci.edu/dataset/2/adult.