Journal of Statistics & Applied Science, Issue 5
ISSN 2718-0999

Research Article

Received: date:20.04.2022
Accepted: date:26.06.2022
Published: date:30.06.2022

Medikal Sentetik Veri Üretimiyle Veri Dengelemesi

A.Fatih Deveci[1], M. Fevzi Esen [2,*]

[1] Tekhnelogos Araştırma ve Geliştirme Merkezi, İstanbul, Türkiye; fatih.deveci@tekhnelogos.com
[2] Sağlık Bilimleri Üniversitesi, Sağlık Bilişim Sistemleri Ana Bilim Dalı; fevzi.esen@sbu.edu.tr
Orcid: 0000-0002-3044-8397[1] Orcid: 0000-0001-7823-0883[2]
* Correspondence Author, e-mail: fevzi.esen@sbu.edu.tr

Öz: Sağlık hizmetleri planlaması, klinik deneyler ve araştırma geliştirme çalışmaları gibi sağlık verisi kullanımını gerektiren alanlarda, kişisel sağlık verisinin elde edilmesi ve kullanımında etik, bürokratik ve operasyonel zorluklar yaşanmaktadır. Elektronik kişisel sağlık kayıtlarının güvenliği ve kişisel veri mahremiyeti konularındaki kısıtlamalar başta olmak üzere, klinik ve saha çalışmalarından veri elde edilmesinin maliyetli ve zaman alıcı olması, gerçek veriye en yakın şekilde yapay veri üretilmesini gerekli kılmaktadır. Bu çalışmada, son dönemde sağlık alanında artan veri kullanımı ihtiyacı doğrultusunda, sentetik veri kullanımının önemi ele alınarak, sentetik veri üretiminde kullanılan SMOTE, SMOTEENN, BorderlineSMOTE, SMOTETomek ve ADASYN yöntemlerinin performanslarının karşılaştırılması amaçlanmıştır. Çalışmada, gözlem ve sınıf sayısı birbirinden farklı ve ikisi de kamuya açık, 390 hastaya ait 15 değişkenden oluşan veri seti ile 19.212 COVID-19 hastasına ilişkin 16 değişkenden oluşan veri seti kullanılmıştır. Çalışma sonucunda SMOTE tekniğinin gözlem ve sınıf sayısının fazla olduğu veri setini dengelemede daha başarılı olduğu ve sentetik veri üretiminde hibrit tekniklere göre etkin olarak kullanılabileceği sonucuna ulaşılmıştır.
Keywords: Sentetik Veri, Smote, Smoteenn, Sağlık Bilişimi

Data Balancing with Synthetic Medical Data Generation

Abstract: There are ethical, bureaucratic and operational difficulties in obtaining and using personal health data in the areas that require the use of sensitive health data such as health care planning, clinical trials and research and development studies. The cost and time consuming of obtaining data from clinical and field studies, especially the restrictions on the security of electronic personal health records and personal data privacy, necessitate the production of synthetic data as close to real data. In this study, it is aimed to compare the performances of SMOTE, SMOTEENN, BorderlineSMOTE, SMOTETomek and ADASYN methods that have been used in synthetic data production by considering the importance of synthetic data generation in line with the increasing need for data use in the health field. In the study, a dataset consisting of 15 variables belonging to 390 patients with different observation and class numbers and a dataset consisting of 16 variables related to 19,212 COVID-19 patients were used. It has been concluded that SMOTE is more successful in balancing the data sets with large number of observations and multiclass classification. This technique can be used effectively in synthetic data generation compared to hybrid techniques.
Keywords: Synthetic Data, Smote, Smoteenn, Health Informati

1. Giriş

ReportLinker [1] büyük veri ve analitiği raporuna göre son yıllarda büyük veri ve ilişkili piyasalar yıllık ortalama %4 oranında büyüyerek 76,1 milyar dolar düzeyine ulaşmış olup, bu rakamın 2025 yılında yaklaşık 117 milyar dolar seviyesine ulaşması öngörülmektedir. Söz konusu artışta, kurum ve kuruluşların uzaktan çalışma şartlarıyla birlikte araştırma, geliştirme ve pratik deneyimlerden elde edilen veriler ve diğer operasyonları sınırlayıcı önlemlerin sebep olduğu radikal değişikliklerin önemli rol oynadığı belirtilmektedir. Ayrıca, verinin elde edilmesinde kullanılan cihazların çeşitliliği ve miktarındaki artışla birlikte, cihazların nesnelerin interneti ve bulut bilişimi teknolojileriyle entegrasyonu ve gerçek zamanlı veri kullanımının karar vericiler için önemli bir ihtiyaç haline gelmesinin, büyük veri analitiği piyasasına ivme kazandırdığı ifade edilmektedir [2]. Aynı raporda, 2024 yılına kadar yapay zeka

Citation: A.F Deveci and M.F Esen , " Data Balancing with Synthetic Medical Data Generation", Journal of Statistics and Applied Sciences, no. 5, 17-27, Jul. 2022, doi:10.52693/jsas.1105599

ve veri analitiği projelerinde kullanılan verilerin yaklaşık %60'ının sentetik olarak üretileceği belirtilmektedir.

Sağlık alanında kişisel demografik verilerin yanı sıra; klinik bulgular, laboratuvar testleri ve görüntüleme yöntemleri, reçete ve sosyal sigorta kayıtları, halk sağlığının izlenmesi, bireylerin hayat tarzı ve sosyo-ekonomik düzeyi, çevre ve işyeri koşulları, sağlık ihtiyaçları ve bakım hizmetleri gibi sağlığı ilgilendiren geniş bir alanda yüksek hacimli ve karmaşık türde büyük veri birikimi yaşanmaktadır. Elektronik hasta kaydı olarak nitelendirilen bu verilerin, hastalıkların zamanında ve doğru teşhis edilmesi, tedavi planlaması, hastalık seyri hakkında yeni bilgilerin keşfedilmesi, tıbbi araştırma ve geliştirme çalışmalarının gerçekleştirilmesi ve kişiselleştirilmiş (hassas) uygulamaların geliştirilmesindeki artan önemine dikkat çekilmektedir [3]. Buna karşın, birçok sağlık kuruluşu tarafından toplanan elektronik hasta kayıtları değer yaratan bir kaynak olsa da, hasta mahremiyeti endişeleri sebebiyle araştırmacıların çoğu tarafından erişilebilir nitelikte değildir. Araştırma ve geliştirme amacı dahil olmak üzere elektronik tıbbi kayıtlara erişim, gizlilik ve güvenlik kısıtlamaları, verinin üretilmesi – işlenmesi ile ilgili düzenlemeler ve verinin değeri nedeniyle sınırlandırılmaktadır. Ayrıca, araştırmacıların sağlık verisine erişimi her ne kadar mümkün olsa bile; verinin işlenmesi, korunması ve kullanımı konusundaki yasal gerekliliklerin yerine getirilmesi uzun bir süreç gerektirmekte olup bu durum, araştırma sonucundan ve bilginin paylaşılmasından sağlanacak faydayı da ciddi şekilde geciktirmektedir [4]. Bu sebeple, sağlık verilerinin araştırmacıların kullanımına sunulabilmesi amacıyla veriden kişilerin tanımlanabilir özelliklerinin kaldırılması veya gerçek verilere dayalı olarak sentetik veri üretilmesi, yenilikçi ve adaptif uygulamalar olarak göze çarpmaktadır [5].

Avrupa Birliği Genel Veri Koruma Tüzüğü'nün (GDPR) yürürlüğe girmesinden bu yana, bireysel sağlık verisinin araştırma ve geliştirme amaçlı kullanımı dahil olmak üzere, hasta mahremiyetinin sağlanması konularında birçok tartışma yürütülmekte olup, araştırmacıların sağlıkta büyük örneklemle çalışmaları sınırlandırılmaktadır. Verilerin korunması ve veri ihlallerinin bildirilmesine ilişkin faydalarının yanı sıra; düzenleme kapsamında kişisel düzeydeki tıbbi verilere erişimin engellenmesi veya sınırlandırılması, birçok yenilikçi araştırma, geliştirme ve eğitim fırsatına da engel olmaktadır [5]. Bu durum, bilgi teknolojileri araçlarının etkin şekilde kullanılarak, kişisel sağlık verilerinin yerini tutabilecek nitelikte alternatif veri kaynaklarının oluşturulmasını gerekli kılmaktadır.

Kişisel verilerin tanımlanmasını önlemeye yönelik olarak, veri setinden kişiyi tanımlayıcı bilgilerin kaldırılmasını amaçlayan veri maskeleme, veri birleştirme, bölümleme ve veri anonimleştirme gibi birçok teknikten yararlanılmaktadır. Veri setindeki değişkenlerin içeriği ve hassasiyetini dikkate alan söz konusu tekniklerin uygulanması sonucunda orijinal veri setinden çıkartılabilecek fayda azalmakta ve belirli ölçüde bilgi kaybı yaşanmaktadır. Uygulama adımlarında karşılaşılan prosedürel güçlükler, işlemlerin maliyetli ve zaman alıcı olması, hasta verilerinin yeniden tanımlanması ve geniş veri ağlarında paylaşılması riski ve kurumların hasta mahremiyeti – bilginin mülkiyeti endişeleri, gerçek veriye dayalı sentetik veri üretim tekniklerini popüler hale getirmektedir [6,26]. Bu sebeple, verinin gizliliği ve güvenliğini sağlamak amacıyla gerçek veri setindeki özelliklere, dağılımlara ve ilişkilere sahip, gerçek veriden herhangi bir ipucu içermeyen ve verinin tanımlanması riskinin en aza indirildiği sentetik veri üretimi iyi bir alternatif olarak kabul edilmektedir [7].

Sentetik veri, güvenlik - gizlilik riski olmayan, gerçek verilerle ilişkili yapay veri olup, verinin olmadığı veya elde edilmesinin zor ve maliyetli olduğu durumlarda hacim ve çeşitlilik sorunu olmadan belirli ihtiyaca veya koşula yönelik üretilebilmektedir. Sentetik veri, gerçek veriye ulaşımın olmadığı durumlarda modelleme ve simulasyon gibi bütünleşik yöntemlerle üretilebildiği gibi, makine öğrenmesi teknikleriyle eğitilmiş modeller kullanılarak da üretilmektedir [8]. Bu durum, araştırmacıların kendi amaçları doğrultusunda gerçek verilerin yerini alabilecek, keşfedici veya doğrulayıcı nitelikte ve etik endişeler içermeyen sentetik veri kullanımını veri anonimleştirmeye iyi bir altenatif kılmaktadır[9]. Sentetik verinin üretilmesi sırasında, silme, yok etme, gürültü ekleme, genelleştirme ve baskılama gibi anonimleştirme tekniklerinden de yararlanılmakta, veri içerisinde kişiyi tarif edici nitelikler değiştirilmekte veya kaldırılmaktadır. Sentetik veri, orijinal verilerle aynı istatistiksel özellikleri ve zamana bağlı özellikleri içermekte olup, gerçek kişilerle eşleştirilmemiş, tam ve geri dönüşü olmayan bir anonimleştirme sağlamaktadır[10].

Sağlık çalışmalarında kişisel veri setlerine ulaşımdaki zorluk ve verinin işlenmesi süreçlerininin çoğunlukla kısıtlanmış olması, bilimsel tekrarlanabilirlik sorununu ortaya çıkarmaktadır. Bunun yanı sıra; özellikle sağlık bilimlerinde olgu grubundaki gözlem sayısının toplam gözlem içerisindeki oranının düşük olması ve olgulara ait elde edilen verinin yetersiz kalması, kullanılacak istatistiksel modelin tahmin doğruluğu da düşürmektedir. Dengesiz veri olarak adlandırılan bu durum, model başarısının düşmesi, aşırı veya yetersiz uyum gibi problemlere de sebep olmaktadır. Örneğin, halk sağlığı planlamasına yönelik olarak, kişilerin beslenme ve sağlık durumuna ilişkin yeterli veriye ulaşılamadığında, verinin eksik/gürültülü olduğu durumda veya nadir görülen hastalık araştırmalarında bir sınıfa ait daha fazla gözlem elde edilirken diğer sınıfa ilişkin kısıtlı sayıda gözleme ulaşıldığında, sınıf dengesizliğinden bahsetmek mümkündür. Bu sebeple, dengesizliğin azaltılması veya ortadan kaldırılmasına yönelik olarak literatürde çeşitli sentetik veri üretimi teknikleri kullanımı önerilmektedir [7].

Bu çalışmada, literatürde belirtilen [11,12] metodolojiden hareketle, veri çoğaltım ve hibrit yaklaşımlara dayalı olan SMOTE, SMOTEENN, BorderlineSMOTE, SMOTETomek ve ADASYN tekniklerinin, iki ve çok sınıflı veri setlerinde sınıf dengesizliklerinin giderilmesindeki performansları ve dolayısıyla sentetik veri üretiminin kıyaslanması amaçlanmıştır. Bu kapsamda, gözlem ve sınıf sayısı birbirinden farklı iki sağlık veri seti kullanılarak orijinal veri seti sınıf dağılımları ile söz konusu yöntemler uygulandıktan sonra oluşan sınıf dağılımları karşılaştırılmıştır.

2. Sentetik Veri Üretim Süreci

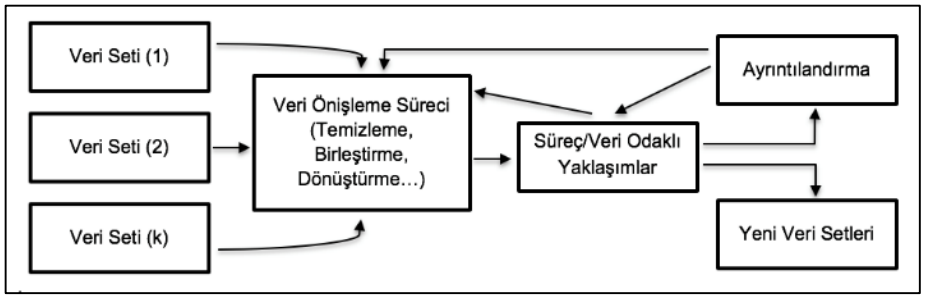
Sentetik veri üretimi, kişiyi tanımlayan ve hassas nitelikte olan verilerin prototipini sağlamaktadır. Böylelikle araştırmacıların veri gizliliği ile ilgili düzenlemeleri ihlali riskine maruz kalmasının engellenmesi ve araştırmaların zengin, geniş ve kontrol edilebilir bir veri alanı ile test edilmesine de imkan tanınmaktadır. Ayrıca sentetik veriyle, araştırmacıların gerçek tıbbi kayıt kullanımına gerek kalmadan, problemlerin çözümünde eğitilebilir modeller kullanabilmesine, çalışma sonuçlarında hassasiyetin arttırılması, istenilen tür/boyutta veri üretilmesi ve daha fazla senaryo ile çalışılması sağlanmaktadır.

Veri bilimi ve makine öğrenmesinde yeni çözümler ve hype döngüsü raporuna göre sentetik veri aşağıda belirtilen durumlarda kullanılmaktadır [14]:

- Kişisel ve hassas nitelikteki verilerin mahremiyeti yani, gizlilik gereksinimleri veri erişilebilirliğini veya nasıl kullanılabileceği sınırlandığında, veri merkezli karar destek sistemlerinin tasarlanması sürecinde,
- Mevcut veri setinin homojen olduğu ve heterojen veri setlerine ihtiyaç duyulduğunda,
- İhtiyaç duyulan verinin olmaması ya da erişilebilir veri seti bulunmadığında,
- Geçerli veri setleri olsa da, düzenlenmemiş farklı veri setlerinin birleştirilmesi veya bir araya getirilmesine ihtiyaç olduğu durumlarda,
- Gerçek verinin test edilmesi ve doğrulanmasında,
- Makine öğrenimi algoritmaları için eğitim veri setinin üretiminin çok pahalı olduğu durumlarda,
- Veri gizliliği korunurken, veriye erişimin en üst düzeye çıkarılması durumunda.
- Yazılımların test edilmesi ve doğrulanmasında,
- Akademik araştırma ve eğitim amaçlı olarak.

Sentetik verilerin üretiminde birçok farklı yaklaşım kullanılmaktadır [15]. Şekil 1'de gösterildiği üzere söz konusu yaklaşımlardaki ortak noktalar, farklı kaynaklardan elde edilen yapılandırılmış/yapılandırılmamış verilerin önişleme sürecinden geçirilerek kullanılacak sayısal yaklaşıma göre esnek ve genelleştirilebilir adımların oluşturulabilmesidir.

Eksik verilerin tamamlanmasında kullanılan klasik istatistik yaklaşımların (gürültü ekleme, döndürme, kırpma vb.) yeni kohortların ve hastalık gruplarının oluşturulmasında etkin olarak kullanılamayacağı belirtilmektedir [4]. Ayrıca, görüntü işleme, metin madenciliği ve büyük veri analitiği gibi karmaşık ve zor süreçler içeren konularda klasik istatistik yaklaşımlar yetersiz kaldığı için, benzetim teknikleri ve makine öğrenmesi yaklaşımlarından yararlanılmaktadır.



Figür 1. Sentetik Veri Üretim Süreci

Literatürde sentetik veri üretiminde kullanılan yaklaşımlar kabaca ikiye ayrılmaktadır. Süreç odaklı yaklaşımlarda, bir olayın temelini oluşturan süreçlere ilişkin hesaplamalı ve matematiksel modeller kullanılmaktadır. Kesikli olay benzetimi, etmen tabanlı modelleme ve Monte Carlo benzetimi süreç odaklı yaklaşımlara örnek olarak verilebilir. Veri odaklı yaklaşımlarda ise, makine öğrenimi kullanılmakta olup, mevcut veriler öğrenilerek çıkarımlarda bulunulmaktadır. Destek vektör makinaları, Bayes ağları ve derin öğrenme bu yöntemlere örnek olarak verilebilir.

Derin öğrenmeye dayalı yöntemler ise, bir varyasyonel otomatik kodlayıcıdan (VAE - Variational Autoencoder) ya da üretici bir rakip ağdan (GAN - Generative Adversarial Network) yararlanmaktadır. VAE'ler, kodlayıcılardan ve kod çözücülerden oluşan denetimsiz makine öğrenimi modelleridir. VAE'nin kodlayıcı kısmı, veriyi orijinal veri setini basit ve kompakt bir şekle sıkıştırmakla sorumlu olup, kod çözücü analiz işlemi ve temel verilerin temsilini oluşturmak için kullanılmaktadır. Bir VAE, hem giriş verilerinin hem de çıkış verilerinin benzer olduğu, giriş verileri ve çıkış arasında optimal bir ilişkiye sahip olmak amacıyla eğitilmektedir. GAN modellerinde, jeneratör sentetik verilerin üretilmesinden sorumlu olup, ayırıcı ise üretilen verileri gerçek bir veri setiyle karşılaştırmakta ve hangi verilerin sahte olduğunu belirlemeye çalışmaktadır. Her seferinde oluşturulan ve sahte olanların ayırt edildiği ağlar birbirlerine karşı eğitilerek gerçekçi hale getirilmektedir [17].

Tablo 1. Veri Dengeleme Yaklaşımları(Data Balancing Approachs)

Veri Azaltım Yaklaşımları	**Veri Çoğaltım Yaklaşımları (Sentetik Veri Üretimi)**	**Hibrit Yaklaşımlar (Sentetik Veri Üretimi & Veri Azaltım)**
Random Undersampling	Random Oversampling	SMOTETomek
Condensed Nearest Neighbor Rule (CNN)	Synthetic Minority Oversampling Technique (SMOTE)	SMOTEENN
Near Miss Undersampling	Borderline-SMOTE	-
Tomek Links Undersampling	Borderline Oversampling with SVM	-
Edited Nearest Neighbors Rule (ENN)	Adaptive Synthetic Sampling (ADASYN)	-
One-Sided Selection (OSS)	-	-
Neighborhood Cleaning Rule (NCR)	-	-

Sententik veri üretimi yaklaşımlarının temelinde, örneklem azaltma (under sampling), örneklem arttırma (over sampling) ve hibrit örnekleme (hybrid sampling) gibi üç temel veri dengeleme yöntemi bulunmaktadır (bkz. Tablo 1) [18]. Çoğunluk sınıfına ait veri sayısının azınlık sınıfı veri sayısına çekildiği örneklem azaltma yönteminde veri seti dengeli hale getirilmeye çalışılmaktadır. Örneklem arttırmada azınlık sınıfına ait veri sayısı çoğunluk sınıfına ait veri sayısına yaklaştırılmakta olup, yeni sentetik veriler üretilerek azınlık sınıfına eklenmektedir. Hibrit yaklaşımda ise, hem çoğunluk sınıfından örneklem azaltma yapılmakta, hem de azınlık sınıfına ait örneklerin dağılıma uygun olarak örneklem arttırılması yapılarak her iki yöntemden yararlanılmaktadır.

3. Literatür Araştırması

Sağlıkta sentetik veri üretimi konusundaki çalışmalar geniş bir uygulama alanına sahip olup, tıbbi görüntüleme başta olmak üzere kanser araştırmaları, gen dizileme, biyogözetim, ilaç geliştirme ve farmakovijilans alanında yoğunlaşmaktadır.

Gerçekleştirilen çalışmalarda sentetik veri üretiminin yanı sıra; araştırmacıların klinik kararlarını destekleyici nitelikte ve kişisel verilerin korunması ilkeleri çerçevesinde bilgi güvenliğiyle uyumlu iş akışı sağlayan sistem modelleri üstünde durulmaktadır. Örneğin, gerçek elektronik tıbbi kayıtlardan sentetik kayıtlar oluşturması amacıyla gerçekleştirilen Sentetik Elektronik Tıbbi Kayıt Oluşturucu (EMERGE) çalışmasında, salgın ve acil durumlarda kayıt kontrolu ve tarihsel veri doğrulama işlemlerinin gerçekleştirilmesi amaçlanmaktadır. Standartlaştırılmış bir test veri setinin oluşturulması farklı algoritmaların ve prosedürlerin karşılaştırılmasına izin vermenin yanı sıra, yenilikçi ve adaptif algoritmaların geliştirilmesi için de önemli bir veri kaynağı niteliği taşımaktadır [10]. Gözlemsel Tıbbi Veri Kümesi Simülatörü (OSIM) çalışmasında ise zaman, cinsiyet ve yaş gibi demografik değişkenler ele alınarak, gerçek verilerin olasılık dağılımlarına uygun hastalık sınıflarının oluşturulması ve reçete edilebilecek ilaçlar konusunda sentetik veri sentezi sunulmaktadır [19]. Bir diğer çalışmada, 823 kardiovasküler hastadan oluşan veri seti üzerinde düşük ve yüksek risk sınıflaması vakaları ele alınmış, iki sınıf arasındaki dengesizliğin farklı tekniklerle giderilmesi amacıyla kullanılabilecek teknikler üzerinde değerlendirmelerde bulunulmuştur [20]. Çalışma sonucunda SMOTE uyarlanmış veri azaltma yaklaşımının sentetik veri üretiminde etkin olarak kullanılabileceği gösterilmiştir. 13-24 yaş aralığında 132 hastaya ait sol el MR görüntüsü ve 20 farklı hastaya ait bilgisayarlı tomografi görüntülerinin kullanıldığı medikal görüntüleme çalışmasında, Evrişimli Sinir Ağı tabanlı algoritmalar ve bilgisayarlı görü teknikleri kullanılmıştır. Çalışmada, üretilen sentetik görüntülerle vücut pozu tahminleme gerçekleştirilmiştir [21].

Farklı hastalık sınıfları ve örneklem sayılarına sahip medikal veri setleri üzerine yapılan bir diğer çalışmada, Destek Vektör Makinaları, k-En Yakın Komşu yöntemi ve Çok Katmanlı Algılayıcılar algoritmaları kullanılarak toplamda 2471 hasta kaydını içeren veri setinin dengelenmesi amaçlanmıştır. Çalışmada, sınıflama algoritmalarının performansları karşılaştırılmıştır[22]. Optimizasyon, istatistik ve makine öğrenme tekniklerinin sentetik zaman serisi oluşturmadaki etkinliği üzerine gerçekleştirilen çalışmada ise 42.240 hasta kaydı içeren AreM (Activity Recognition system based on Multisensor data) veri seti ve 122 kayıttan oluşan EEG verileri ele alınmışır. Çalışma sonucunda sentetik veri setinin orijinal veri setindeki öznitelikleri koruduğu tespit edilmiştir [23]. Açık kaynak kodlu Synthea (Sentetik Sağlık) üretecinde ise, algoritma testi, araştırma sonuçları doğrulama, güvenlik ve gizlilik kontrolü, fizibilite analizi ve diğer akademik araştırmacılar için sentetik boylamsal veri oluşturulması mümkün kılınmakta olup, klinik keşif ve bilimsel çıkarım yönünden araştırmacılara karar desteği sağlanmaktadır. Karaciğer lezyonu bilgisayarlı tomografi görüntüleri kullanılarak Çekişmeli Üretici Ağlarıyla (GANs) sentetik medikal görüntü üretilmesini amaçlayan bir diğer çalışmada, Evrişimli Sinir Ağı (CNN) ile sınıflanma modeli tahmin başarısının arttırılması sağlanmıştır. Buna göre klasik sınıflandırma yaklaşımlarıyla %78,6 duyarlılık ve %88,4 özgüllük sağlanırken, sentetik veri arttırma ile %85,7 duyarlılık ve %92,4 özgüllük sağlanmıştır. Çalışma sonucunda, sentetik veri üretmeye yönelik bu yaklaşımın diğer tıbbi sınıflandırma uygulamalarına genelleştirilebileceği ve böylelikle radyologların tanıyı iyileştirme çabalarının desteklenebileceği gösterilmiştir [24]. Otomatik kodlayıcılar ve üretken çekişmeli ağların bir kombinasyonuyla yüksek boyutlu ayrışık örneklemler olşturmayı amaçlayan medGAN sisteminde ise, girdi olarak gerçek hasta kayıtları kullanılmakta olup, epidemiyoloji alanında yüksek kaliteli, çok boyutlu ve gerçeğe yakın sentetik veri üretimi sağlanmaktadır [25].

Bir başka çalışmada, 2010 ve 2015 yılları arasında teşhis edilen meme, solunum ve katı olmayan kanser vakalarından oluşan 360.000 hastaya ilişkin veri setini kullanarak olasılık modelleri, sınıflandırmaya dayalı isnat modelleri ve çekişmeli üretken ağ olmak üzere üç sınıf sentetik veri üretme yaklaşımı değerlendirilmiştir [4]. Çalışmada, üretilen sentetik veri kümelerinin kalitesini değerlendirmeye yönelik ölçütler sunulmakta olup, kategorik gizli Gauss sürecinin realistik veri üretiminde daha başarılı olduğu sonucuna ulaşılmıştır. 1042 farklı muayenehaneden elde edilen 27,5 milyon birinci basamak hastasına ilişkin veriden hareketle yapılan bir diğer çalışmada, hastaların sentetik veriyle yeniden tanımlanması

riskini ölçerek, makine öğrenmesi sınıflandırıcılarından elde edilen duyarlılık analizi sonuçlarını ele almıştır [7]. Ayrıca, grafik modelleme ile adapte edilen aykırı değer analizi de gerçekleştirilmiş olup, orijinal veriden istatistiksel olarak anlamlı derecede farklı olmayan sentetik veri kümelerinin tespit edilmesi amaçlanmıştır. Çalışma sonucunda, hastaları tanımlayıcı nitelikte veri üretme riskinin çok düşük olduğu belirtilmiştir. Acil servis hastaları üzerine gerçekleştirilen bir başka çalışmada, 27.963 acil servis taburcu verisi ele alınmış olup, değişkenlere ilişkin gizli olasılık dağılımlarının tespiti Kullback-Liebler uzaklığı ile gerçekleştirilmiştir. Bireysel gizliliği koruyan Elektronik Sağlık Kayıtları (EHR) oluşturmak için Gretel'in açık kaynaklı, sentetik veri kütüphanesinden faydalanılmış olup, gerçek veri setiyle uyumlu sentetik veri setlerinin oluşturulabileceği sonucuna ulaşılmıştır [8].

4. Veri Seti ve Yöntem

Çalışmada, sentetik veri üretiminde kullanılan SMOTE, SMOTEENN, BorderlineSMOTE, SMOTETomek ve ADASYN yöntemlerinin performanslarının karşılaştırılması amacıyla, gözlem ve sınıf sayısı birbirinden farklı iki veri seti kullanılmıştır. İlk veri setinde 390 hastaya ait 15 değişkenden oluşan diyabet hastalığı teşhisleri, ikinci veri setinde ise 19.212 COVID-19 hastasına ilişkin 16 kesikli ve sürekli değişkenden oluşan hastane yatış ve sosyodemografik veriler bulunmaktadır. Literatürde önerilen, azınlık sınıfı üzerinden yeni gözlemler üreterek azınlık sınıfı gözlem sayısını arttırmaya yönelik metodolojiden hareketle, sentetik veri üretimi için veri çoğaltım ve hibrit yaklaşımlar kullanılmıştır[38]. Söz konusu yaklaşımların ortak özelliği, çoğunluk sınıfının aşırı öğrenilmesini önlemek için azınlık sınıfından sentetik numuneler oluşturulmasıdır. Ayrıca, aşırı örneklemenin Rastgele Veri Çoğaltım veya diğer adıyla Rastgele Aşırı Örnekleme (random oversampling) metoduyla yapılmasının en basit yaklaşım olduğu belirtilen çalışmadan yola çıkarak [20], bu çalışmada ilk olarak, veri çoğaltım (oversampling) ve sonrasında örneklem azaltımı (undersampling) da içeren hibrit teknikler kullanılmıştır. Gürültülü ve eksik gözlemler veri setinden çıkartılarak ve herhangi bir normalizasyon uygulanmadan veri analize hazır hale getirilmiş, sınıflar arası heterojen yapı giderilmiştir.

5. Bulgular

Çalışmada kullanılan diyabet teşhislerine dağılım frekansları ve sınıflar arası dengesizlik oranı Tablo 3'de verilmiştir. Diyabet teşhisi iki sınıflı bir değişken olarak (0=diyabet tanısı yok, 1= diyabet tanısı var) tanımlanmıştır. Orijinal veri setinde tüm vakaların %84,61'i diyabet tanısı olmayanlardan oluşmaktadır. Sınıflar arası dengesizliği oranı ise diyabetli olmayan hasta sayısının diyabetli hasta sayısına oranlanmasıyla elde edilmiştir. Buna göre uygulama öncesi orijinal veri setinde sınıf dengesizlik oranı %81,81 olarak bulunmuştur.

Tablo 3. İki sınıflı diyabet Veri Seti Analiz Sonuçları

Yöntem	Orijinal Veri Seti-1 Sınıf Dağılımı ve Yüzde		Dengesizlik Oranı	Yöntem Uygulandıktan Sonra Sınıf Dağılımı ve Yüzde		Yöntem Uygulandıktan Sonra Dengesizlik Oranı
	Diyabet Tanısı: Yok	Diyabet Tanısı: Var		Diyabet Tanısı: Yok	Diyabet Tanısı: Var	
SMOTE	330	60	% 81,81	330	330	%0,00
SMOTEENN	330	60	% 81,81	262	312	%16,02
ADASYN	330	60	% 81,81	330	325	%1,01
BorderlineSMOTE	330	60	% 81,81	330	330	%0,00
SMOTETomek	330	60	% 81,81	325	325	%0,00

Sentetik veri üretme ve hibrit yaklaşımlar uygulandıktan sonra elde edilen sınıf dengesizliği oranına bakıldığında, SMOTE, Borderline-SMOTE ve SMOTETomek uygulamaları sonrasında dengesizliğin giderildiği sonucuna ulaşılmıştır. SMOTETomek yaklaşımında, çoğunluk sınıfı gözlem sayısından veri azaltılarak, azınlık sınıfına sentetik veri eklenerek dengeleme gerçekleştirilirken, borderlineSMOTE ve SMOTE yöntemlerinde ise azınlık sınıfı için sentetik veri üretilmiştir.

Tablo 4'de 19.212 COVID-19 vakasından oluşan dengesiz sınıf dağılımına sahip veri setinde ise hastaların çalışma statüleri 10 farklı sınıf şeklinde ifade edilmiştir. Örneğin COVID-19 hastaları içerisinde tam zamanlı çalışan sayısı 4595 kişi, engelli 4380 kişi ve asker/ordu görevi yürüten toplamda iki kişi görülmektedir. Uygulama sonrası sonuçlara bakıldığında, SMOTE veri çoğaltımı yönteminde toplamda 54,640 yeni vaka üretildiği ve her bir sınıf yüzdesinin eşitlendiği görülmektedir. Aynı şekilde, Borderline-SMOTE uygulaması sonrasında her bir sınıf orijinal veri setindeki çoğunluk sınıfındaki vaka sayısına eşitlenmiş olup, sadece son sınıfta (asker/ordu görevi olanlar) orijinal veri setindeki vaka sayısı sabit kalmıştır. SMOTETomek yönteminde ise sınıf frekanslarının birbirine yaklaşmış olduğu izlense de, eşitliğin sağlanamadığı görülmektedir.

Tablo 5'de, iki sınıflı veri setinde uygulanan yöntemlere ilişkin olarak değişkenlerin ortalama değerleri ve çarpıklık katsayıları verilmiştir. Buna göre, orijinal veri seti değerlerine en yakın ortalama ve çarpıklık katsayısı değerlerinin her bir değişken için yöntemlere göre farklılık gösterdiği tespit edilmiştir. Örneğin, orijinal veri setinde 107,33 olarak gözlenen ortalama kan şekeri değerinin ADASYN yöntemiyle üretilen sentetik veri değerine yakın iken, orijinal veri setinde 46,77 ortalama ile gözlenen yaş değişkenine en yakın değerlerin SMOTEENN yöntemiyle oluşturulduğu göze çarpmaktadır.

Tablo 4. Çok Sınıflı Veri Seti Analiz Sonuçları

Sınıf*	**Orijinal Veri Seti Frekans**	**SMOTE Uyg. Sonrası Frekans**	**SMOTEENN Uyg. Sonrası Frekans**	**ADASYN Uyg. Sonrası Frekans**	**Borderline-SMOTE Uyg. Sonrası Frekans**	**SMOTETomek Yöntemi Uyg. Sonrası Frekans**
1	5464(%28,44)	5464(%10)	345(%1,06)	5464(%22,1	5464(%11,11)	4848(%10,29)
2	4595(%23,92)	5464(%10)	3683(%11,29)	4380(%17,7	5464(%11,11)	5303(%11,25)
3	4380(%22,80)	5464(%10)	936(%2,87)	2078(%8,43)	5464(%11,11)	4883(%10,36)
4	2078(%10,82)	5464(%10)	3634(%11,14)	1522(%6,17)	5464(%11,11)	5247(%11,14)
5	1522(%7,92)	5464(%10)	3240(%9,93)	4595(%18,6	5464(%11,11)	5171(%10,98)
6	610(%3,18)	5464(%10)	5004(%15,34)	610(%2,47)	5464(%11,11)	5393(%11,44)
7	465(%2,42)	5464(%10)	4932(%15,11)	57(%0,23)	5464(%11,11)	5368(%11,39)
8	57(%0,30)	5464(%10)	5412(%16,59)	465(%1,89)	5464(%11,11)	5451(%11,57)
9	39(%0,20)	5464(%10)	5438(%16,67)	5481(%22,2	5464(%11,11)	5458(%11,58)
10	2(%0,01)	5464(%10)	-	-	2(%0,01)	-

*Hastaların çalışma statülerini ifade etmektedir. Kodlanan sınıflar: 1- İşsiz, 2- Tam zamanlı, 3- Engelli, 4- Yarı zamanlı, 5- Emekli, 6- Tam zamanlı öğrenci, 7- Bilinmiyor, 8- Serbest meslek, 9- Yarı zamanlı öğrenci, 10- Asker/ordu görevi

Tablo 5. İki Sınıflı Veri Seti Temel İstatistikleri

Yöntem	**Ort. Yaş (Çarpıklık katsayısı)**	**Ort. HDL Kolesterol (Çarpıklık katsayısı)**	**Ort. Glukoz (Çarpıklık katsayısı)**	**BMI ort. (Çarpıklık katsayısı)**	**Cinsiyet (mod)**	**Ort. Boy (Çarpıklık katsayısı)**	**Ort. Ağırlık (Çarpıklık katsayısı)**
Orijinal Veri Seti	46,77 (0,41)	50,26 (0,74)	107,33 (0,96)	28,77 (0,44)	Kadın	65,95 (-0.03)	177,40 (0,32)
SMOTE	51,69 (-0,24)	47,96 (0,55)	140,72 (1,20)	29,65 (0,36)	Kadın	65,80 (-0,16)	182,86 (0,23)

SMOTEENN	50,99 (-0,28)	47,63 (0,57)	148,30 (1,28)	29,38 (0,25)	Kadın	65,92 (-0,05)	182,42 (0,19)
ADASYN	51,22 (-0,34)	48,25 (0,62)	117,71 (0,79)	29,68 (0,70)	Kadın	65,48 (0,40)	181,25 (0,41)
Borderline SMOTE	51,21 (-0,35)	48,88 (0,56)	118,56 (0,54)	29,75 (0,69)	Kadın	65,16 (0,13)	179,83 (0,55)
SMOTE Tomek	51,68 (-0,35)	47,77 (0,51)	143,39 (1,31)	29,43 (0,35)	Kadın	65,88 (-0,09)	182,30 (0,18)

Tablo 6'da çok sınıflı veri setine ilişkin tanımsal istatistiklere bakıldığında, yaş değişkeninin orijinal veri setinde simetriğe yakın hafif sağa çarpık dağılım gösterdiği ve bu duruma en yakın SMOTE tekniğiyle veri üretildiği göze çarpmaktadır. Aynı şekilde, cinsiyet, etnisite, din, medeni durum, sağlık sigortası, hastaların polikliniğe başvuru tarihi (gün ve ay) olarak orijinal veri seti istatistiklerinin SMOTE ile üretilen veri ile uyumlu olduğu tespit edilmiştir.

Tablo 6. Çok Sınıflı Veri Seti Temel İstatistikleri

Yöntem	**Ort. Yaş (Çarpıklık katsayısı)**	**Cinsiyet (mod)**	**Etnisite (mod)**	**Din (mod)**	**Medeni durum (mod)**	**Sağlık sigortası (mod)**	**Başvuru ay (mod)**	**Başvuru gün (mod)**
Orijinal Veri Seti	50,14 (0,17)	Kadın	Beyaz	Katolik	Bekar	Medicaid	Haziran	Pazartesi
SMOTE	47,99 (0,19)	Kadın	Beyaz	Katolik	Bekar	Medicaid	Haziran	Pazartesi
SMOTEENN	41,91 (0,70)	Kadın	Beyaz	Yok	Bekar	Özel	Mayıs	Cumartesi
ADASYN	44,80 (0,69)	Kadın	Beyaz	Katolik	Bekar	Medicaid	Haziran	Pazartesi
Borderline SMOTE	41.05 (0,32)	Erkek	Beyaz	Katolik	Bekar	Medicaid	Nisan	Cumartesi
SMOTE Tomek	42,67 (0,56)	Kadın	Beyaz	Katolik	Bekar	Özel	Mayıs	Cumartesi

6. Sonuç ve Tartışma

Sağlık bilişim alt yapısı ve hasta kayıtlarının saklanma şekli, makine öğrenmesi teknikleri ve yapay zekâ tabanlı sistemlerin ihtiyaç duyduğu nicelikte veriye ulaşımı zorlaştırmaktadır. Veri paylaşımındaki teknik ve bürokratik zorluklarla birlikte, kişisel verilerin gizliliği ve güvenliği sebebiyle gerçek veriye erişimin kısıtlanması da sağlık alanında gerçekleştirilecek bilimsel araştırmaları sınırlandırmaktadır. Gerçek hasta verisine bu denli kısıtlı erişim, araştırmacıların sentetik veri ve ilişkili yaklaşımlara olan ilgisini arttırmaktadır.

Literatürde ikili sınıfları içeren veri setlerinin dengelenmesinde bir çok veri çoğaltımı, azaltımı ve hibrit yöntem önerilmiştir. Sağlık alanında incelenen problemlerin genellikle çok sınıflı dengesiz veri setlerini içermesi, çok sınıflı veri setlerinin dengelenmesinin ikili sınıflara göre güç oluşu ve bu konuda gerçekleştirilmiş çalışmaların az oluşu, çok sınıflı değişkenler içeren veri setlerinin dengelenmesi problemini literatürde sık tartışılan konulardan biri haline getirmektedir [39]. Çalışma sonucunda, iki sınıflı veri setinin dengelenmesinde sentetik veri üretilmesine dayanan SMOTE ve Borderline-SMOTE yöntemlerinin başarılı olduğu, çok sınıflı veri setininin dengelenmesinde SMOTE'un daha iyi bir performansa sahip olduğu tespit edilmiştir. Buna karşın, SMOTE yönteminde her bir sınıfta eşit frekans dağılımı (%10) görülmekteyken; borderline-SMOTE yönteminde orijinal veri setinden sentetik veri üretimi için son sınıfta (asker/ordu görevi olanlar) bulunan vaka sayısı yetersiz kalmış olup, bu sınıf için sentetik veri üretilememiştir. Bu durum borderline-SMOTE yönteminde azınlık sınıfının parçalara bölünerek veri çoğaltımının sağlanması durumuyla ilgili olup, azınlık sınıfı sayısının çok düşük olduğu durumlarda veri hacminin veri üretimi için yeterli derecede bilgi sağlayamadığına işaret etmektedir [40].

Hibrit yaklaşımlardan olan SMOTETomek algoritmasının ikili sınıf içeren veri setinin dengelenmesindeki etkinliğine karşın, çok sınıflı veri setinin dengelenmesinde aynı durumun gerçekleşmediği; veri çoğaltım yaklaşımlarının sınıf dengesizliğini gidermede daha etkin olduğu görülmektedir. Buna göre, iki ve çok sınıflı veri setleri için sentetik veri üretiminde veri çoğaltım yöntemlerinin hibrit yöntemlere göre sınıflandırma performanslarının belirgin şekilde farklılık gösterdiği tespit edilmiştir.

Hibrit yaklaşımların çok sınıflı veri setlerinin dengelenmesinde kullanımına ilişkin olarak söz konusu yaklaşımların sınıflandırma performanslarının ölçümünde doğruluk, duyarlılık, kesinlik ve seçicilik gibi metriklerin de göz önünde bulundurulmasının faydalı olacağı düşünülmektedir. Ayrıca, çalışmada elde edilen sonuçlar doğrultusunda, düşük sınıflandırma performansı ve veri setinden gözlem çıkartarak bilgi kaybına yol açması nedeniyle, hibrit yöntemlerin sağlık araştırmalarında kullanımının dezavantajlarının da göz önünde bulundurulması gerekmektedir. Ayrıca sağlık gibi çok değişkenli ve çok sınıflı veri setlerinin bulunduğu bir alanda, değişken seçimi yöntemleriyle gürültülü değişkenlerin veri setinden çıkartılması veya çarpık dağılıma sahip değişkenlerin sentetik veri üretilmesindeki etkilerinin tespit edilmesi ve düzeltilmesine yönelik yaklaşımların kullanılması önerilebilir.

Aynı anda hem azınlık sınıfı için sentetik veri üretilmesi ve çoğunluk sınıfından örnek azaltılması prensibine dayanan hibrit yöntemlerle medikal veri setlerinde daha tutarlı makine öğrenimi modellerinin test edilmesi sağlanabilir. Böylelikle, farklı veri çoğlatım teknikleri ile azınlık sınıfına sentetik veri üretilmesi sağlanıp sonrasında ise, çeşitli veri azaltım teknikleri ile sadeleştirme yapılarak her iki yöntemin avantajlı taraflarından faydalanılabilir. Büyük hacme, çeşitliliğe ve yüksek sınıf dengesizliğine sahip veri setleri için sentetik veri üretiminde kullanılan algoritmaların etkinliği konusu, gelecekte gerçekleştirilebilecek çalışmalar için önemli bir alan olarak görülmektedir.

Yazar Katkıları: Bu çalışmada A.D, literatür araştırması ve uygulama kısmının destek verilmesi, M.F.E, çalışmanın yönetilmesi verilerin hazırlanması ve analizlerin yapılması ve kontrol edilmesi; konularında katkı sağlamışlardır.

Finansman: Bu araştırma her hangi bir fon desteği almadı.

Çıkar çatışmaları: Yazarlar çıkar çatışması olmadığını beyan etmemektedir.

Not: Bu çalışma 30.09.2021-01.10.2021 tarihleri arasında düzenlenen II. Tıpta Bilişim Kongresinde özet bildiri olarak sunulan bildirinin genişletilmiş halidir.

Kaynakça

[1] ReportLinker (2021). Big Data Industry. https://www.reportlinker.com/market-report/Advanced- IT/513221/Big-Data,20.07.2021

[2] Gartner (2021). Top Strategic Technology Trends for 2021, https://www.gartner.com/en/publications/top-tech-trends-2021,13.07.2021

[3] Jacob, P.D. (2020). Management of patient healthcare information: Healthcare-related information flow, access, and availability, In Fundamentals of Telemedicine and Telehealth (ss. 35-57) (Eds. Shashi Gogia), Academic Press.

[4] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20(1), 1–40. https://doi.org/10.1186/s12874-020-00977-1

[5] 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.2019.12.136

[6] Rocher, L., Hendrickx, J.M. & de Montjoye, YA. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun, 10: 3069.

[7] Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. Npj Digital Medicine, 3(1). https://doi.org/10.1038/s41746-020-00353-9

[8] Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., Watson, A., & Eastman, J. (2020). SyntheaTM Novel coronavirus (COVID-19) model and synthetic data set. Intelligence-Based Medicine, 1–2: 100007. https://doi.org/10.1016/j.ibmed.2020.100007

[9] Dube, K. , Gallagher, T. (2014). Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use. In: Gibbons J., MacCaull W. eds. Foundations of Health Information Engineering and Systems. FHIES 2013. Lecture Notes in Computer Science, vol 8315. Berlin, Heidelberg: Springer.

[10] Buczak, A. L., Babin, S., & Moniz, L. (2010). Data-driven approach for creating synthetic electronic medical records. BMC medical informatics and decision making, 10, 59. https://doi.org/10.1186/1472-6947-10-59

[11] Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016). Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. Proceedings of 2016 IEEE International Conference of Online Analysis and Computing Science, ICOACS 2016, 2016, 225–228. https://doi.org/10.1109/ICOACS.2016.7563084

[12] Liu, N., Li, X., Qi, E., Xu, M., Li, L., & Gao, B. (2020). A novel ensemble learning paradigm for medical diagnosis with imbalanced data. IEEE Access, 8, 171263–171280. https://doi.org/10.1109/ACCESS.2020.3014362

[13] Liu, Y., Li, X., Chen, X., Wang, X., & Li, H. (2020). High-Performance Machine Learning for Large-Scale Data Classification considering Class Imbalance. Scientific Programming, 2020. https://doi.org/10.1155/2020/1953461

[14] Gartner (2020). Hype Cycle for Data Science and MachineLearning-2020, https://www.gartner.com/en/documents/3988118/hype-cycle-for-data-science-and-machine-learning-2020, 19.07.2021

[15] Ayala-Rivera, V., Portillo-Dominguez, A. O., Murphy, L., & Thorpe, C. (2016). COCOA: A synthetic data generator for testing anonymization techniques. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9867 LNCS, 163–177. https://doi.org/10.1007/978-3-319-45381-1_13

[16] Marathe, M. V. (2006). Synthetic Data for Data Mining to Support Epidemiological Modeling. Network Dynamics and Simulation Science Laboratory, Virginia Tech, 1 Ağustos 2021 tarihinde https://www.cs.dartmouth.edu/~cbk/sdm06/marathe-data.sdm.pdf adresinden alındı.

[17] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. Frontiers in artificial intelligence, 3, 4. https://doi.org/10.3389/frai.2020.00004

[18] Bekkar, M., & Alitouche, T. A. (2013). Imbalanced Data Learning Approaches Review. International Journal of Data Mining & Knowledge Management Process, 3(4). https://doi.org/10.5121/ijdkp.2013.3402

[19] Murray, R. E., Ryan, P. B., & Reisinger, S. J. (2011). Design and validation of a data simulation model for longitudinal healthcare data. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2011, 1176–1185.

[20] Rahman, M. M., & Davis, D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing, May 2014, 224–228. https://doi.org/10.7763/ijmlc.2013.v3.307

[21] Riegler, G., Urschler, M., Ruther, M., Bischof, H., & Stern, D. (2015). Anatomical Landmark Detection in Medical Applications Driven by Synthetic Data. Proceedings of the IEEE International Conference on Computer Vision, 2015-February, 85–89. https://doi.org/10.1109/ICCVW.2015.21

[22] Belarouci, S., & Chikh, M. A. (2017). Medical imbalanced data classification. Advances in Science, Technology and Engineering Systems, 2(3), 116–124. https://doi.org/10.25046/aj020316

[23] Shamsuddin, R., Maweu, B. M., Li, M., & Prabhakaran, B. (2018). Virtual patient model: An approach for generating synthetic healthcare time series data. Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, February 2019, 208–218. https://doi.org/10.1109/ICHI.2018.00031

[24] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321–331. https://doi.org/10.1016/j.neucom.2018.09.013

[25] Zhang, Z., Yan, C., Mesa, D. A., Sun, J., & Malin, B. A. (2020). Ensuring electronic medical record simulation through better training, modeling, and evaluation. Journal of the American Medical Informatics Association, 27(1). https://doi.org/10.1093/jamia/ocz161

[26] Benaim, A. R., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., Khamaisi, M., Lurie, Y., Azzam, Z. S., Khoury, J., Kurnik, D., & Beyar, R. (2020). Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies. JMIR Medical Informatics, 8(2), 1–14. https://doi.org/10.2196/16492

[27] Gherardini, M., Mazomenos, E., Menciassi, A., & Stoyanov, D. (2020). Catheter segmentation in X-ray fluoroscopy using synthetic data and transfer learning with light U-nets. Computer Methods and Programs in Biomedicine, 192, 105420. https://doi.org/10.1016/j.cmpb.2020.105420

[28] Hernandez-Matamoros, A., Fujita, H., & Perez-Meana, H. (2020). A novel approach to create synthetic biomedical signals using BiRNN. Information Sciences, 541, 218–241. https://doi.org/10.1016/j.ins.2020.06.019

[29] Shi, G., Wang, J., Qiang, Y., Yang, X., Zhao, J., Hao, R., Yang, W., Du, Q., & Kazihise, N. G. F. (2020). Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. Computer Methods and Programs in Biomedicine, 196, 105611. https://doi.org/10.1016/j.cmpb.2020.105611

[30] Stolfi, P., Valentini, I., Palumbo, M. C., Tieri, P., Grignolio, A., & Castiglione, F. (2020). Potential predictors of type-2 diabetes risk: machine learning, synthetic data and wearable health devices. BMC Bioinformatics, 21(17), 1–20. https://doi.org/10.1186/s12859-020-03763-4

[31] Vaden, K. I., Gebregziabher, M., Dyslexia Data Consortium, & Eckert, M. A. (2020). Fully synthetic neuroimaging data for replication and exploration. NeuroImage, 223. https://doi.org/10.1016/j.neuroimage.2020.117284

[32] Vilardell, M., Buxó, M., Clèries, R., Martínez, J. M., Garcia, G., Ameijide, A., Font, R., & Civit, S. (2020). Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (ModGraProDep): An application to breast cancer survival. Artificial Intelligence in Medicine, 107: 101875. https://doi.org/10.1016/j.artmed.2020.101875

[33] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020). CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. IEEE Access, 8: 91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762

[34] Dai, F., Song, Y., Si, W., Yang, G., Hu, J., & Wang, X. (2021). Improved CBSO: A distributed fuzzy-based adaptive synthetic oversampling algorithm for imbalanced judicial data. Information Sciences, 569, 70–89. https://doi.org/10.1016/j.ins.2021.04.017

[35] Karbhari, Y., Basu, A., Geem, Z. W., Han, G. T., & Sarkar, R. (2021). Generation of synthetic chest X-ray images and detection of COVID-19: A deep learning based approach. Diagnostics, 11(5), 1–19. https://doi.org/10.3390/diagnostics11050895

[36] Palmér, E., Karlsson, A., Nordström, F., Petruson, K., Siversson, C., Ljungberg, M., & Sohlin, M. (2021). Synthetic computed tomography data allows for accurate absorbed dose calculations in a magnetic resonance imaging only workflow for head and neck radiotherapy. Physics and Imaging in Radiation Oncology, 17(December 2020), 36–42. https://doi.org/10.1016/j.phro.2020.12.007

[37] Vepa, A., Saleem, A., Rakhshan, K., Daneshkhah, A., Sedighi, T., Shohaimi, S., Omar, A., Salari, N., Chatrabgoun, O., Dharmaraj, D., Sami, J., Parekh, S., Ibrahim, M., Raza, M., Kapila, P., & Chakrabarti, P. (2021). Using machine learning algorithms to develop a clinical decision-making tool for covid-19 inpatients. International Journal of Environmental Research and Public Health, 18(12), 1–22. https://doi.org/10.3390/ijerph18126228

[38] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). snopes.com: Two-Striped Telamonia Spider. Journal of Artificial Intelligence Research, 16(Sept. 28), 321–357. https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp,

[39] Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data, 7: 70.

[40] Susan, S. & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. Engineering Reports, 3: e12298. https://doi.org/10.1002/eng2.12298