Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics
Turkiye Klinikleri J Biostat. 2020;12(2):204-10

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-73403

Genom-Boyu İlişki Çalışmalarında, Makine Öğrenimi ve Derin Öğrenme Yöntemlerinin Farklı Örnek Genişliklerinde Performanslarının Değerlendirilmesi

Evaluation of Machine Learning Methods and Deep Learning Method Performance in Different Sample Size in Genome-Wide Association Studies

Ragıp Onur ÖZTORNACI[a], Erdal COŞGUN[b], Bahar TAŞDELEN[a]

[a]Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Mersin, TÜRKİYE
[b]Microsoft Research, 14820 NE 36th Street, Building 99, Redmond, WA, 98052, USA

Bu çalışma, "XXI. Ulusal ve IV. Uluslararası Biyoistatistik Kongresi (26-29 Ekim 2019, Antalya)'nde sözlü olarak sunulmuştur.

ÖZET Amaç: Makine öğrenimi (ML) ve derin öğrenme (DL) yöntemlerinin genetik ilişki çalışmalarında (GWAS) kullanımı giderek yaygınlaşmaktadır. Bu yöntemler, klasik GWAS metodolojisine ek olarak elde edilen sonuçların doğrulanmasına yardımcı olmaktadır. Bu nedenle; bu çalışmanın amacı, en yaygın kullanılan ML yöntemleri ve DL yönteminin farklı örnek genişliklerinde kıyaslanarak en doğru hasta-kontrol ayırımını yapan sınıflama yöntemini bulmaktır. **Gereç ve Yöntemler:** ML ve DL yöntemleri için farklı örnek genişliklerinde aynı prevelansa ve minör allel frekansına (MAF) sahip ve eşit sayıda hasta-kontrol içeren veriler üretilmiştir. Tüm veri setlerinde, tüm yöntemler için 10-katlı çapraz geçerlilik yöntemi uygulandı ve tüm yöntemler için, %70 eğitim veri seti %30 test veri seti olarak belirlenerek, modellerin geçerlilikleri ve tahmin güçleri sınandı. Simülasyonlar için PLINK yazılımı, ML için R programlama dili ve DL için de Python programlama dili Microsoft Azure GPU makinalar ile kullanmıştır. **Bulgular:** Örnek genişliği; N=200 için; en iyi sınıflama performanslarını ML yöntemleri içinde, Random Forest (RF) 0.73 Doğruluk (Acc.) ve CART 0.73 Acc. ile gösterirken, DL 0.60 Acc. değeri elde etmiştir. N=600 için; ML yöntemlerinden DVM 0.78 Acc. ile en iyi sınıflama başarısını elde ederken, DL 0.64 Acc. değeri elde etmiştir. Maksimum örnek genişliğinde, hem ML hem de DL için doğru sınıflama başarısının değişiminin etkilenmediği söylenebilir. **Sonuç:** Son yıllarda özellikle bağımsız değişkenler arasında ilişkilerin çok yüksek olduğu veri setlerinde yüksek doğru sınıflama oranı DVM ile edilmektedir. Bu çalışmada örneklem genişliğinden bağımsız olarak en yüksek sonuç DVM ile elde edilmiştir. Bu da genetik verilerde sık karşılaşılan doğrusal olmayan ML yöntemlerinin başarısını kanıtlamaktadır.

Anahtar Kelimeler: Tek nükleotid polimorfizmi; makine öğrenimi; derin öğrenme

ABSTRACT Objective: The use of Machine Learning (ML) and deep learning (DL) methods in genetic association studies (GWAS) is becoming increasingly common. These methods help confirm the results obtained in addition to the classical GWAS methodology. Therefore; The aim of this study was to compare the most commonly used ML methods and DL methods in different sample widths and find the most accurate patient-control classification method. **Material and Methods:** For ML and DL methods, data with the same prevalence and minor allele frequency with different sample size and equal number of patient-controls were generated. In all data sets, 10-fold cross validity method was applied for all methods and 70% training data set was determined as 30% test data set for all methods and validity and predictive power of the models were tested. PLINK software for simulations, R programming language for ML and Python programming language for DL have been used with Microsoft Azure GPU machines. **Results:** Sample size; n=200; The best classification performances in ML methods, Random Forest 0.73 Accuracy (Acc.) and CART™ 0.73 Acc. while DL 0.60 Acc. n=600; MLM methods of DVM 0.78 Acc. while DL 0.64 Acc. value. **Conclusion:** In recent years, DVM method is used for high estimation ratio, especially in data sets where the relationships between independent variables are very high. In this study, the highest result was obtained with DVM regardless of sample size. This proves the success of nonlinear ML methods that are frequently encountered in genetic data.

Keywords: Single nucleotide polymorphism; machine learning; deep learning

Correspondence: Ragıp Onur ÖZTORNACI
Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Mersin, TÜRKİYE/TURKEY
E-mail: onur.oztornaci@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 07 Jan 2020 ***Received in revised form:*** 26 Feb 2020 ***Accepted:*** 26 Feb 2020 ***Available online:*** 28 May 2020

Genom-boyu ilişki analizi çalışmaları [genome-wide association studies (GWAS)], belirli bir hastalığa neden olduğu düşünülen genleri ve genom bölgelerini belirlemek için hastalık durumu ve genetik varyasyon arasındaki ilişkiyi inceleyen çalışmalardır. GWAS adımları takip edilirken, başlangıçta belirli bir hipotez doğrultusunda yola çıkmak yerine tüm genom incelenerek hastalıkla ilişkili olduğu düşünülen bölgeler tespit edilir. Aynı anda tüm kromozomlara dolayısıyla da birçok gene bakılarak gen-gen ve gen-çevre etkileşimleri incelenir.[1,2] GWAS, hastalığın genetik yapısını araştırmak için en yaygın kullanılan stratejidir. GWAS metodolojisi, genomdaki tek nükleotid polimorfizmlerine, [single nucleotide polymorphisms (SNPs)] göre hastalar ve sağlıklı bireyler arasındaki farkları inceler. GWAS ile belirli bir hastalıkla ilgili SNP'leri tespit etmeye çalışırken istatistiksel önem eşiğinin tanımlanması esastır. Bu nedenle, büyük eşik değerleri kullanmak, hastalıkla ilişkili yanlış-pozitif SNP sayısını potansiyel olarak artırabilir, alternatif olarak, küçük eşik değerleri kullanmak, hastalıkla ilişkili gerçek pozitif SNP miktarını azaltabilir. Her iki durumda da yanıltıcı sonuçlar üretilebilir. GWAS için 5 x 10e-8 p-değeri eşiğini kullanmak standart bir uygulama hâline gelmiştir.[3] Makine Öğrenimi [Machine Learning (ML)] algoritmalarının uygulanması bu durumda bazı avantajlar sağlar; i) Çok sayıda muhtemel korelasyona sahip değişken birlikte modellenebilir; ii) Herhangi bir önem eşiği gerekli değildir; iii) Gen-gen etkileşimleri değerlendirilebilir ve iv) Normallik ve varyansların homojenliği gibi klasik istatistiksel varsayımlar gerekli değildir.[4] Dolayısıyla günümüzde gelişen teknoloji ile birlikte, ML ve Derin Öğrenme [Deep Learning (DL)] yöntemlerinin hasta ve sağlıklı ayrımı için kullanılması giderek yaygınlaşmaktadır. Bu nedenle de doğrusal olmayan sınıflama problemleri için hangi yöntemin daha iyi sonuç vereceği önemli hâle gelmiştir.

Bu çalışmada, PLINK programı yardımıyla, popülasyona dayalı vaka-kontrol incelemeleri için büyük miktarda SNP veri kümeleri oluşturmak üzere hastalık prevalansı %1, minör allel frekansı [minor allele frequency (MAF)] %5 olan farklı örnek genişliklerinde veriler üretilmiştir.[5]

GEREÇ VE YÖNTEMLER

MAKİNE ÖĞRENİMİ

ML 1950'li yıllardan bu yana matematiğe ve istatistiğe yeni bir bakış açısı getiren, günümüzde de gelişimini sürdüren çalışma alanıdır.[6,7] Bir ML algoritmasının performansı değerlendirilirken en önemli ölçüt doğru tahmin edebilme yeteneğidir.[8] Dolayısıyla birçok farklı algoritma, doğrusal olmayan sınıflama problemleri için doğru tahmin yeteneğini artırmak amacıyla geliştirilmiştir ve de her geliştirilen algoritmanın kendisine ait ayırt edici özellikleri vardır.

Breiman ve ark. tarafından bulunan Sınıflama ve Regresyon Ağaçları [Classification and Regression Trees (CART™)], heterojen yapıdaki veriden homojen alt sınıflar elde edebilmek amacıyla kullanılan klasik karar ağacı algoritmasıdır. Kategorik yapıdaki değişkenler için Gini ve Twoing kurallarına uygun olarak sınıflama gerçekleştirilirken, sürekli yapıdaki değişkenlerde ise en küçük kareler sapması kriteri kullanılır.[9,10] CART™ algoritması, kayıp verilerden etkilenmeyen ve hem kategorik hem de sürekli değişkenlerle kullanılabilen bir sınıflama yöntemidir.[11]

2001 yılında Breiman tarafından bulunan Rastgele Orman [Random Forest (RF)], oluşturulan birçok rastgele karar ağacının kombinasyonudur. Ağaçlar inşa edilirken "bootstrap" örneklem seçimi kullanılır. Sınıflamada Gini katsayısı kullanılır ve oluşturulan ağaçlar için budama yapılmaz, amaç, daha yüksek doğru sınıflama oranı elde edebilmektir. Değişken sayısının fazla olduğu veri setleri için de kullanılabilir olması avantaj iken, iyi donanıma sahip yüksek geçici belleği olan bilgisayarlarda kullanılması gerekliliği bir dezavantaj olarak görülebilir.[12,13]

Sık tercih edilen bir diğer sınıflandırma yöntemi olan Destek Vektör Makineleri (DVM), ilk olarak 1992 yılında Vladmir Vapnik ve ark. tarafından önerilmiştir. Yüksek güvenirliliği, ezberlemeye karşı olan direnci ile DVM, iki boyutlu uzayda doğrusal, üç boyutlu uzayda düzlemsel, çok boyutlu uzayda hiper düzlem şeklindeki ayırma mekanizmaları ile veriyi iki ya da daha çok sınıfa ayırma yeteneğine sahiptir.[14-16]

Kümeleme ve sınıflandırma için kullanılan doğadan, sinir hücrelerinden ve bilhassa insan beyninin öğrenme yeteneğinden esinlenerek bilgisayar teknolojilerine uyarlanan diğer bir ML yöntemi de yapay sinir ağlarıdır.[17] Bir sinir hücresindeki elemanların hepsi yapay sinir ağında da mevcuttur. Yapay sinir ağları iki ya da üç katmandan meydana gelirler. Bunlar girdi, gizli ve çıktı katmanlarıdırlar. Bazı durumlarda sadece girdi ve çıktı katmanları oluşarak iki katmanlı yapay sinir ağını meydana getirirler.[18] Geri besleme tekrarlarının (feed-back loops) olmadığı yapay sinir ağlarına ileri beslemeli sinir ağı (feed-forward nueral network) adı verilir.[19] Girdi katmanı, gizli katmanlar ve çıktı katmanı olmak üzere üç temel ögeden meydana gelen çok katmanlı algılayıcılar [multilayer perceptrons (MLP)], temelde ileri beslemeli birçok yapay sinir ağından oluşmaktadırlar. Genelleştirilmiş delta öğrenme kuralı tabanlıdırlar ve doğrusal yapıda olmayan sınıflama modelleri için kullanılırlar.[20,21]

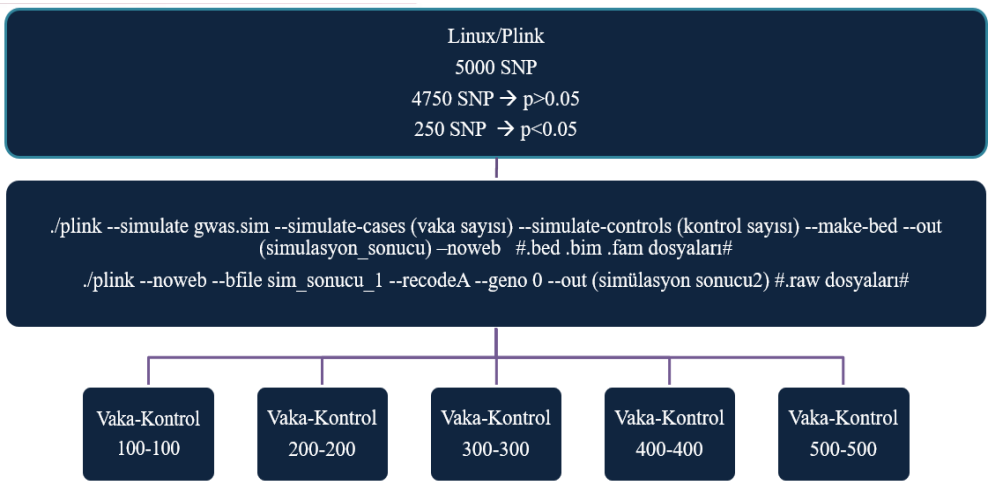
DERİN ÖĞRENME

ML'nin alt alanı olan DL, temelde birçok yapay sinir ağından meydana gelmektedir. Dolayısıyla çok fazla katmandan oluşmaktadır ve bu katmanlar, eğitim aşamasında öğrenme işlemlerini gerçekleştirirler.[22] DL yöntemleri, verilen çok sayıda girdiye göre ayırt edici özellikleri kendisi öğrenir. Geleneksel ML algoritmalarında eğitim aşamasından önce manuel olarak belirlenen özelliklerin hesaplanması gerekirken; DL, bu işlemi katmanlar arasında kendisi yapar.[23] Bir DL modelini eğitmek genellikle çok zaman alır, bu nedenle birçok DL süreci, eğitim süresini kısaltmak için çok çekirdekli işlemcilerden ve çok çekirdekli Grafik İşleme Ünitesi [Graphics Processing Unit (GPU)]nden yararlanır. Sık kullanılan bazı DL kütüphaneleri şunlardır: "Microsoft Cognitive Tool Kit (CNTK)", Tensorflow, Caffe, Torch ve MXNet.[24-29] CNTK, Python ve C ++ gibi dillerin desteğinin yanı sıra verimli kaynak tüketimi ile de son derece optimize edilmiştir.

SİMÜLASYON ADIMLARI

Popülasyona dayalı vaka-kontrol sınıflamalarında, doğrusal olmayan sınıflama problemlerinin performanslarını karşılaştırmalı olarak görmek amacıyla, PLINK programı yardımıyla, büyük miktarda SNP veri kümeleri oluşturulmuştur. Hastalık prevalansı %1, MAF %5 olan farklı örnek genişliklerinde veriler üretilmiştir.

Veri simülasyonları, Linux Ubuntu 17.04 işletim sisteminde PLINK yazılımı kullanılarak, %5 MAF gözetilerek, 4.750 hastalıkla ilişkisiz SNP ve 250 hastalıkla ilişkili SNP, %1 hastalık prevalansı, 1:1 vaka-kontrol oranı ve heterozigotluk odds oranları da dikkate alınarak, toplam 200 örnek genişliğinden başlanarak, 1.000 örnek genişliğine kadar, kalıtım modeli: AA, AB, BB olacak şekilde GWAS çalışmalarındaki güç dikkate alınarak üretilmiştir (Şekil 1).[30,31] Ayrıca DL modelinin eğitim aşamasında CNTK GPU kütüphanesinde "Convolutional neural networks (CNNs)" algoritması 1.000 katman, 10 mini batch, 100 epoch ve 0,0001 öğrenme hızı (learning rate) ile kullanılmıştır.



ŞEKİL 1: Simülasyon adımları ve örnek kodlar.

BULGULAR

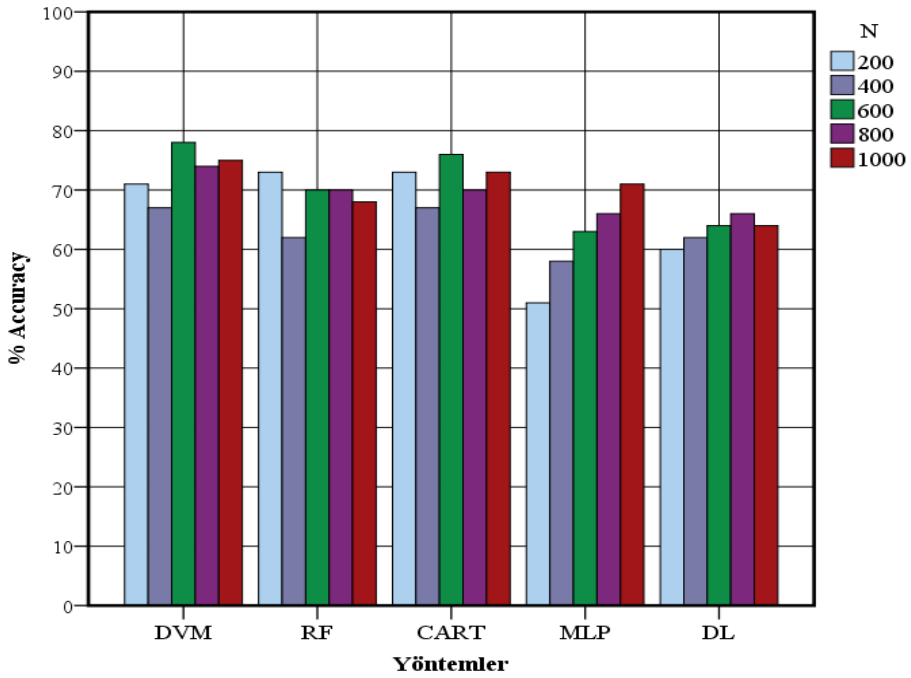
En küçük örnek genişliği için ML yöntemlerinden, RF 0,73 doğruluk sınıflama oranı [accuracy (Acc.)] ve CART™ 0,73 Acc. elde ederken, MLP 0,51 Acc. değeri ile en düşük sınıflama performansını göstermiştir. DVM ise 0,71 Acc. elde etmesine rağmen Acc. için güven aralığı oldukça geniştir [0,58;0,82]. Öte yandan DL, 0,60 Acc. değeri elde etmiştir. n=400 için; ML yöntemlerinde, DVM ve CART™ 0,67 Acc ile en iyi performanslara sahiplerdir RF 0,62 Acc. değerini [0,61;0,63] kendi performansı içerisinde (örnek genişliği değişimine rağmen) en dar güven aralığı ile elde etmiştir. MLP yöntemi 0,62 Acc. ile örnek genişliğinin artmasından olumlu yönde etkilenmiştir. DL'de ise, 0,62 Acc. ile stabil Acc. elde edilerek 200 örnek genişliğine göre güven aralığının daha dar olduğu görülmüştür. n=600 için; ML yöntemlerinden DVM 0,78 Acc. ile en iyi sınıflama başarısını elde ederken, diğer tüm yöntemlerin de Acc. değerlerinin arttığı gözlemlenmiştir. CART™ yöntemi, 0,94 gibi oldukça yüksek spesifisite değeri elde ederken, 0,58 sensitivite değeri elde etmiştir. RF yönteminde ise tam tersi olduğu; 0,54 sensitivite, 0,85 spesifisite elde edildiği görülmektedir. MLP yönteminin, örnek genişliği arttıkça başarısındaki artış da devam etmektedir. DL 0,64 Acc. değeri elde etmiştir. n=800 için; ML yöntemlerinde, en iyi sınıflama performansını DVM 0,74 Acc. değeri ile göstermiştir. Sensitivite ve spesifisite bakımından da incelenecek olursa, CART™ yönteminin 0,89 ile en iyi spesifisiteye sahip olduğu görülürken; 0,50 ile en düşük sensitivite değerini elde ettiği dikkat çekmektedir. MLP'nin performansındaki artış da 0,66 Acc. değeri ile devam etmektedir. DL ise 0,66 Acc. ile stabil sınıflama performansını sürdürmektedir. n=1.000'e çıkartıldığında, hem ML hem de DL için doğru sınıflama başarısının değişiminin etkilenmediği söylenebilir, ancak MLP'deki artışın 0,71 Acc. seviyesine ulaştığı da gözden kaçmamaktadır. ML için DVM 0,75 Acc. değeri elde edilirken, DL'de bu değerin 0,64 olduğu görülmüştür (Tablo 1).

TABLO 1: Makine Öğrenimi ve Derin Öğrenme yöntemlerinin sınıflama performansları.

Yöntemler	n (Vaka-Kontrol)	Acc. %95 GA	Sensitivite	Spesifisite
Destek Vektör Makineleri	100-100	0,71 [0,58;0,82]	0,70	0,73
	200-200	0,67 [0,58;0,75]	0,70	0,65
	300-300	0,78 [0,72;0,84]	0,84	0,73
	400-400	0,74 [0,68;0,79]	0,80	0,68
	500-500	0,75 [0,69;0,79]	0,74	0,75
Random Forest	100-100	0,73 [0,60;0,83]	0,73	0,73
	200-200	0,62 [0,61;0,63]	0,63	0,61
	300-300	0,70 [0,62;0,76]	0,85	0,54
	400-400	0,70 [0,75;0,66]	0,75	0,66
	500-500	0,68 [0,62;0,73]	0,68	0,68
CART™	100-100	0,73 [0,60;0,83]	0,70	0,76
	200-200	0,67 [0,58;0,75]	0,40	0,95
	300-300	0,76 [0,69;0,82]	0,58	0,94
	400-400	0,70 [0,63;0,75]	0,50	0,89
	500-500	0,73 [0,67;0,78]	0,53	0,93
MLP	100-100	0,51 [0,38;0,64]	0,46	0,56
	200-200	0,58 [0,48;0,67]	0,58	0,58
	300-300	0,63 [0,62;0,70]	0,67	0,60
	400-400	0,66 [0,59;0,72]	0,72	0,60
	500-500	0,71 [0,65;0,76]	0,70	0,72
DL	100-100	0,60 [0,52;0,64]	0,65	0,57
	200-200	0,62 [0,58;0,63]	0,64	0,61
	300-300	0,64 [0,62;0,65]	0,60	0,57
	400-400	0,66 [0,58;0,68]	0,62	0,59
	500-500	0,64 [0,62;0,65]	0,58	0,60

Acc.: Accuracy, CART™: Classification and Regression Trees, MLP: Multilayer perceptrons, DL: Deep learning.

Bu çalışmada, örnek genişliğindeki değişimden en çok etkilenen MLP yöntemi olmuştur, örnek genişliği arttıkça MLP'nin sınıflama performansının da açık bir şekilde arttığı görülmüştür (Şekil 2).



ŞEKİL 2: Makine Öğrenimi ve Derin Öğrenme yöntemlerinin sınıflama performansları.

DVM: Destek Vektör Makineleri, RF: Random Forest, CART™: Classification and Regression Trees, MLP: Multilayer perceptrons, DL: Deep learning.

TARTIŞMA

Bu çalışmada, günümüzde her alanda kullanımı giderek yaygınlaşan ML ve DL yöntemlerinin GWAS sınıflama performansları değerlendirilmiştir. GWAS'ta maliyetin yüksek olması sebebiyle simülasyonlara başvurmak kaçınılmazdır. PLINK programı yardımıyla, gerçek veri yapısına en yakın olacak şekilde SNP'leri üretmek mümkündür. Bu sayede zaman ve maliyet tasarrufu yapmak adına, amaca yönelik en uygun yöntemi belirlemek önem arz etmektedir. Doğruluk oranı bakımından DVM, genel anlamda iyi performans gösterirken, sensitivite ve spesifisite değerleri de mutlaka göz önünde bulundurulmalıdır. RF ve CART™ yöntemleri, benzer şekilde performans gösterseler de sensitivite ve spesifisite değerleri arasında farklılık görülmektedir ve CART™ yönteminin genel olarak RF'ye göre spesifisite başarısının daha yüksek olduğu söylenebilir. Benzer çalışmalarda olduğu gibi, bu çalışmada da doğruluk oranının tek başına sınıflama başarısını belirlemek için yeterli olmadığı, diğer parametrelerin de göz önüne alınması gerektiği görülmektedir.[9,10,15] DL'nin, daha dar güven aralıkları ve farklı örnek genişliklerinde dahi istikrarlı sonuçlar vermesi nedeni ile kanser araştırmaları gibi özel veri setlerinde güvenle kullanılabilir olduğu görülmektedir.

SONUÇ

ML ve DL yöntemleri, GWAS'ta hasta-sağlıklı ayrımı için klasik istatistiksel yöntemlere iyi birer alternatif olarak kullanılabilir. Maliyet düşünülünce, düşük örnek genişliklerinde çalışmak GWAS için kaçınılmazdır. Klasik istatistiksel yöntemler (lineer ve lojistik regresyon) örnek genişliklerinden etkilenirken, ML ve DL'nin nasıl etkilendiği kıyaslandığında; düşük örnek genişliklerinde DVM ve CART™'ın en uygun yöntemler oldukları, ancak DL'nin, uygulanması bakımından diğer yöntemler ile kıyaslandığında daha kolay

ve hızlı olduğu söylenebilir. Hem ML hem de DL, GWAS metodolojisi için vaka-kontrol ayrımında iyi birer geçerlilik (validation) yöntemi olarak düşünülebilir.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Fikir/Kavram: *Ragıp Onur Öztornacı, Erdal Coşgun;* ***Tasarım:*** *Ragıp Onur Öztornacı, Erdal Coşgun, Bahar Taşdelen;* ***Denetleme/Danışmanlık:*** *Ragıp Onur Öztornacı, Erdal Coşgun, Bahar Taşdelen;* ***Veri Toplama ve/veya İşleme:*** *Ragıp Onur Öztornacı;* ***Analiz ve/veya Yorum:*** *Ragıp Onur Öztornacı, Erdal Coşgun, Bahar Taşdelen;* ***Kaynak Taraması:*** *Ragıp Onur Öztornacı;* ***Makalenin Yazımı:*** *Ragıp Onur Öztornacı, Erdal Coşgun, Bahar Taşdelen;* ***Eleştirel İnceleme:*** *Ragıp Onur Öztornacı, Erdal Coşgun, Bahar Taşdelen;* ***Kaynaklar ve Fon Sağlama:*** *Ragıp Onur Öztornacı, Erdal Coşgun, Bahar Taşdelen.*

KAYNAKLAR

1. Zhang W, Zhang P, Gao F, Zhu Y, Liu R. Simulation comparison of statistical approaches and procedures in building SNP based prediction models for drug response. Acta Scientific Pharmaceutical Sciences. 2020;4(1):38-43.
2. Battaloğlu E, Başak AN. Kompleks hastalık genetiği: güncel kavramlar ve nörolojik hastalıkların tanısında kullanılan genomik yöntemler. Klinik Gelişim Dergisi. 2010;23:128-33.
3. Fadista J, Manning AK, Florez JC, Groop L. The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet. 2016;24(8):1202-5. [Crossref] [PubMed] [PMC]
4. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, et al. Machine learning in genome-wide association studies. Genet Epidemiol. 2009;33(Suppl 1):S51-7. [Crossref] [PubMed]
5. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform. 2012;10(2):117-22. [Crossref] [PubMed] [PMC]
6. Bell J. Machine Learning: Hands-On for Developers and Technical Professionals. 2nd ed. USA: John Wiley & Sons; 2020. p.432. [Crossref]
7. Michie D, Spiegelhalter DJ, Taylor CC. Machine Learning, Neural and Statistical Classification. 1st ed. New York: Horwood; 1994. p.289.
8. Kononenko I, Bratko I. Information-based evaluation criterion for classifier's performance. Machine Learning. 1991;6(1):67-80. [Crossref]
9. Akpınar H. Data Veri Madenciliği Veri Analizi. 1. Baskı. İstanbul: Papatya Yayıncılık; 2013. p.212.
10. Tan PN, Steinbach M, Kumar V. Chapter 8 Cluster analysis: basic concepts and algorithms. Introduction to data mining, 6th ed. Boston: Peason Addison Wesley; 2006. p.486-568.
11. Temel GO, Çamdeviren H, Akkuş Z. Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma. 2005.
12. Breiman L. Random forests. Machine Learning. 2001;45:5-32. [Crossref]
13. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99(6):323-9. [Crossref] [PubMed] [PMC]
14. Jiawei H, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. San Francisco: Morgan Kaufmann, Elsevier; 2012. p.673.
15. Alpaydın E. Introduction to Machine Learning. 1st ed. Cambridge, Mass: The MIT Press; 2004. p.415.
16. Rathasamuth W, Pasupa K, Tongsima S. Selection of a Minimal Number of Significant Porcine SNPs by an Information Gain and Genetic Algorithm Hybrid Model. 2019. arXiv preprint arXiv:1905.09059.
17. Koç ML, Balas CE, Arslan A. [Preliminary design of ruble mound breakwaters by using artificial neural networks]. İMO Teknik Dergi. 2004;3351-75.
18. Silahdaroğlu G. Veri Madenciliği Kavram ve Algoritmaları. 2. Baskı. İstanbul: Papatya Yayıncılık; 2013. p.121.
19. Cai YD, Liu XJ, Xu XB, Zhou GP. Support Vector Machines for predicting protein structural class. BMC Bioinformatics. 2001;2:3. http://www.biomedcentral.com/1471-2105/2/3 [Crossref] [PubMed] [PMC]
20. Elmas Ç. Yapay Zeka Uygulamaları: (Yapay Sinir Ağı, Bulanık Mantık, Genetik Algoritma). 3. Baskı. Ankara: Seçkin Yayıncılık; 2007. p.425.
21. Heidari AA, Faris H, Mirjalili S, Aljarah I, Mafarja M. Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. In: Mirjalili S, Dong JS, Lewis A, eds. Nature-Inspired Optimizers: Theories, Literature Reviews and Applications. 1st ed. Cham, Switzerland: Springer; 2020. p.23-46. [Crossref]
22. Korkmaz S. Small drug molecule classification using deep neural networks. Turkiye Klinikleri J Biostat. 2019;11(2):93-101. [Crossref]
23. Gürsakal N. Makine Öğrenmesi Derin Öğrenme. 1. Baskı. Bursa: Dora Yayıncılık; 2017. p.309.
24. Yu D, Eversole A, Seltzer ML, Yao K, Huang Z, Guenter B, et al. An introduction to computational networks and the computational network toolkit. Microsoft Technical Report, MSR-TR-2014-112. 2014. p.168.
25. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016. [Link]
26. Jia Y, Shelhamer E , Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional architecture for fast feature embedding. arXiv: 1408.5093v1, 2014. [Crossref]
27. Collobert R, Kavukcuoglu K, Farabet C. Torch7: a matlab-like environment for machine learning. BigLearn. NIPS Workshop. 2011, no. EPFL-CONF-192376.

28. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, et al. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274, 2015. [Link]
29. Seide F, Agarwal A. CNTK: Microsoft's open-source deep-learning toolkit. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, August. p.2135. [Crossref]
30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. [Crossref] [PubMed] [PMC]
31. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform. 2012;10(2):117-22. [Crossref] [PubMed] [PMC]