Araştırma Makalesi / *Research Article*

Ege Tıp Dergisi / *Ege Journal of Medicine* 2025; 64 (1): 143-152

Kopya sayısı varyasyonlarının makine öğrenmesi algoritmaları kullanılarak biyoinformatik analizi

Bioinformatic analysis of copy number variations using machine learning algorithms

Erhan Parıltay[1] Buket Kosova[2]

[1] Ege Üniversitesi Tıp Fakültesi Tıbbi Genetik Anabilim Dalı, İzmir, Türkiye

[2] Ege Üniversitesi Tıp Fakültesi Tıbbi Biyoloji Anabilim Dalı, İzmir, Türkiye

ÖZ

Amaç: Kopya sayısı varyasyonları, insan genomunun yaklaşık yüzde ikisinde bulunan belirli DNA bölgelerinin kayıp veya kazançlarıdır. Yapısal varyasyonlar arasında yer alan bu grup, sağlıklı popülasyonda bulunabileceği gibi ilgili bölgenin kayıp veya kazançları klinik tablolarla da ilişkilendirilebilir. Tespit edilen kopya sayısı varyasyonunun klinik olarak yorumlanması, aile çalışmasını da gerektiren karmaşık bir süreçtir. Klinik ve genetik verilerin yorumlanması sürecinde her zaman doğru bilgiye ulaşılamamaktadır. Kullanımı artan makine öğrenme algoritmaları giderek tıp alanında da kullanılmakta ve özellikle büyük veri setlerinin bulunduğu genetik gibi alanlarda giderek önem kazanmaktadır. Bu çalışma ile kopya sayısı varyasyonlarının klinik değerlendirilmesinde makine öğrenme algoritmalarının kullanımı amaçlanmıştır.

Gereç ve Yöntem: Araştırmada öncelikle 11989 varyant bulunan ISCA konsorsiyumu verileri ile pilot analiz gerçekleştirilmiş, sonrasında ClinVar veri tabanından elde edilen 63156 varyantlı veri seti kullanılmıştır. Beş ana sınıfta (Benign, Muhtemel Benign, VUS, Muhtemel Patojenik ve Patojenik) bulunan varyantlar, Microsoft Azure Machine Learning Studio platformunda, %70 eğitim ve %30 test verisi olarak ayrılmış ve çeşitli makine öğrenmesi algoritmaları (Çok Sınıflı Karar Ağaçları, Lojistik Regresyon ve Sinir Ağı) kullanılarak analiz gerçekleştirilmiştir.

Bulgular: ISCA veri seti ile yapılan modelde çok sınıflı karar ağacı ile ortalamada 0,96 doğruluğa ulaşılırken, ClinVar veri setinde yine çok sınıflı karar ağacı ile 0,86 doğruluğa ulaşılmıştır. Bu modelde patojenikler %74.8, benignler %77.6 ve önemi bilinmeyen varyantlar %62.6 oranında doğru tahmin edilmiştir. Çalışmada sık karşılaşılan varyantlar daha yüksek başarı ile tanımlanmış ve örneklemin benign ve patojenik olarak iki sınıflı haline getirilmesi durumunda ise ortalama ve toplamda 0.90 doğruluğa ulaşılmıştır.

Sonuç: Bu çalışma, kopya sayısı varyantlarının klinik değerlendirilmesinde kullanılabilecek ve tanıyı otomatikleştirebilecek öncül bir makine öğrenme modeli oluşturulabileceğini göstermiştir.

Anahtar Sözcükler: Kopya sayısı varyasyonları, makine öğrenme, genetik, biyoinformatik.

ABSTRACT

Aim: *Copy number variations (CNVs), comprising approximately two percent of the human genome, represent specific DNA segment deletions or duplications. While these structural variations may be present in healthy populations, they can also manifest clinically significant implications. The clinical interpretation of identified CNVs constitutes a complex process necessitating comprehensive family studies.*

Sorumlu yazar: Erhan Parıltay
Ege Üniversitesi Tıp Fakültesi Tıbbi Genetik Anabilim Dalı, İzmir, Türkiye
E-posta: *erhan.pariltay@ege.edu.tr*
Başvuru tarihi: 17.04.2024 Kabul tarihi: 04.02.2025

The interpretation of clinical and genetic data often presents challenges in achieving definitive conclusions. Machine learning algorithms have emerged as increasingly valuable tools in medical applications, particularly in genetics where large-scale datasets predominate. This investigation aimed to evaluate the implementation of machine learning algorithms for the clinical assessment of copy number variations.

Materials and Methods: *The study methodology comprised an initial pilot analysis utilizing ISCA consortium data (n=11,989 variants), followed by a comprehensive analysis of ClinVar database variants (n= 66803). The variants were stratified into five clinical classification categories (Benign, Likely Benign, VUS, Likely Pathogenic, and Pathogenic). Analyses were conducted using the Microsoft Azure Machine Learning Studio platform, implementing various machine learning algorithms (Multiclass Decision Trees, Logistic Regression, and Neural Network) with a 70:30 training-testing data partition.*

Results: *The ISCA dataset analysis demonstrated an average accuracy of 0.96 utilizing multiclass decision trees, while the ClinVar dataset achieved 0.86 accuracy with the same algorithmic approach. The model exhibited predictive accuracies of 74.8%, 77.6%, and 62.6% for pathogenic, benign, and variants of unknown significance, respectively. Frequently occurring variants demonstrated superior predictive accuracy, and binary classification (benign/pathogenic) yielded an enhanced average accuracy of 0.90.*

Conclusion: *This investigation demonstrates the feasibility of developing a preliminary machine learning model for the clinical evaluation and potential automated classification of copy number variants.*

Keywords*: Copy number variations, machine learning, genetics, bioinformatics .*

GİRİŞ

İnsan genomu 23 çift kromozom üzerinde yaklaşık 20.000 gen ve 6 milyar baz çifti içerir. Bireyler arasındaki genomik farklılıklar %0.01'den azdır. Bu farklılıklar DNA baz değişiklikleri (tek nükleotid polimorfizmleri (SNP), mikrosatellitler, minisatellitler, değişken sayıda ardışık tekrarlar (VNTR)), transpozon yapılar (Alu elementleri gibi), yapısal değişiklikler (delesyon, duplikasyon, inversiyon, oryantasyon değişikliği) ve epigenetik değişimler gibi farklı şekillerde ortaya çıkar. Kromozomlardaki ortalama 100 Kb'den büyük DNA bölgelerinin beklenilen kopya sayısından farklı olarak (kayıp ya da kazanç) halinde bulunduğu durumlar kopya sayısı varyasyonları (CNV) olarak adlandırılır (1). CNV tanımı genel olarak oryantasyonel bilgi içermeyen kayıp ve kazanç bölgelerini tanımlar. Yaklaşık 20 Kb'dan küçük CNV'ler hemen her bireyde bulunurlar ve karmaşık genomik yeniden düzenlemeler sonucu oluşur (2). GTG bantlama, DNA dizileme teknikleri ve PCR gibi zenginleştirme işlemleri gen bölgelerinin sayısal ölçümünü zorlaştırmaktadır. Floresan In Situ Hibridizasyon (FISH) tekniği ise yalnızca 20-100 Kb uzunluğundaki spesifik bölgeleri inceleyebilmektedir. CGH ve sonrasında geliştirilen arrayCGH/Oligo Array teknolojileri sayesinde, tüm genomun kopya sayısı analizleri daha erişilebilir ve yaygın hale gelmiştir (3). Ayrıca yeni nesil DNA dizileme teknikleri (Masif paralel sekanslama) ile de genomik olarak kopya sayısı varyasyonları yüksek hassasiyetle tespit edilebilir. Çoğu CNV klinik açıdan önemsiz olmakla birlikte, bazıları gelişimsel gerilik ve doğumsal anomaliler gibi ciddi klinik durumlara neden olabilir. İyi bilinen DiGeorge(22q), Cri-du Chat(5p), Prader Willi (15p) vb. mikrodelesyon sendromlarının yanı sıra birçok mikrodelesyon sendromları array tabanlı yeni nesil teknolojilerle tanımlanmıştır (4). Bu varyasyonların klinik öneminin belirlenmesi uzman hekimlerin değerlendirmesini gerektirir (5). Kopya sayısı varyasyonlarının klinik yorumlanması karmaşık bir süreçtir. Bu süreçte, tespit edilen varyasyonlar DGV ve DECIPHER gibi uluslararası veri tabanlarındaki varyasyonlarla karşılaştırılır (6,7). Ancak yeni ortaya çıkan (*de novo*) varyantlar bu veri tabanlarında bulunmayabilir veya tanımlı varyantlarla tam eşleşme göstermeyebilir. Bu durumda, ebeveyn analizleri yapılır ve kalıtsal varyantlar genellikle benign olarak değerlendirilir. Ayrıca, her gen için hesaplanan haployetmezlik skorları, kopya sayısı değişikliklerinin klinik etkilerini değerlendirmede önemli bir araç(4) olarak kullanılmaktadır. Varyant sınıflamasında American College of Medical Genetics (ACMG) tarafından önerilen beşli sınıflandırma Benign, Muhtemel Benign, Önemi bilinmeyen varyant (VUS), Muhtemel Patojenik ve Patojenik yaygın olarak kullanılır (8). Kopya sayısı varyasyonlarının klinik yorumlanmasında mevcut algoritmalar her zaman kesin sonuçlar verememektedir.

Makine öğrenmesi teorileri yirminci yüzyılın ilk yarılarında ortaya çıkmış, ancak teknolojik sınırlamalar nedeniyle pratik uygulamaları

günümüze kadar gecikmiştir. Özellikle gelişmiş işlemci kapasiteleri ve modern yazılımlar sayesinde çeşitli alanlarda yaygın kullanım imkanı bulmuş ve farklı algoritmalarla uygulanabilir hale gelmiştir. Öğrenme, deneyimlerin bilgi ve tecrübeye dönüşümü olarak tanımlanırken, makine öğrenmesi verilerdeki anlamlı ilişkilerin otomatik olarak tespit edilmesi sürecidir. Geleneksel programlamadan farklı olarak, karmaşık ilişkileri öğrenme algoritmaları aracılığıyla yazılıma dönüştüren bu teknoloji, son yıllarda otonom araçlardan görüntü işlemeye kadar geniş bir kullanım alanı bulmuştur (9). İnsan kapasitesini aşan büyük ve karmaşık veri setlerinin analizi için özellikle sağlık ve tıp alanında yaygın olarak kullanılmaya başlanmış, radyolojik görüntüleme, kanser dokusu değerlendirmesi ve kompleks hastalıklarda risk faktörlerinin belirlenmesi gibi çeşitli alanlarda önemli uygulamalar geliştirilmiştir (10, 11).

Makine öğrenmesi yaklaşımları, veri etkileşim biçimlerine göre farklı kategorilerde incelenir. Gözetimli öğrenmede, önceden etiketlenmiş veriler üzerinden eğitim ve test setleri oluşturularak sistem değerlendirilir. Gözetimsiz öğrenmede ise etiketlenmemiş veriler üzerinden algoritmaların kendi içinde fonksiyon tespiti yapması sağlanır. Her iki yöntemin de kendine özgü avantaj ve dezavantajları vardır. Takviyeli (reinforcement) öğrenme, sistemin ürettiği sonuçlara doğru/yanlış şeklinde geri bildirim verilen bir ara form olarak tanımlanır. Öğrenme süreçleri ayrıca aktif ve pasif olarak da sınıflandırılır; aktif öğrenmede sistem parametrelerle doğrudan etkileşim kurabilirken (örneğin otonom araçlar), pasif öğrenmede hazır veriler üzerinden analiz yapılır (örneğin radyolojik görüntü değerlendirmesi). Online öğrenme sayesinde sistem, öğrenme sonuçlarını anlık olarak gözlemleyebilmektedir (12, 13).

Gözetimli öğrenme, önceden etiketlenmiş veriler üzerinden eğitim yapan bir makine öğrenmesi yaklaşımıdır. Bu sistemde, eğitim verileri ve bunlara ait sınıflar/kategoriler önceden belirlenmiş durumdadır. Sistem, bu etiketli verilerden öğrendiklerini kullanarak test verilerini yorumlar ve el yazısı tanıma veya görsel sınıflandırma gibi çeşitli alanlarda yaygın olarak uygulanır. Başarılı bir gözetimli öğrenme için geniş, etiketlenmiş veri setlerinin varlığı şarttır ve bu sistemler temel olarak sınıflandırma ve regresyon algoritmaları olmak üzere iki kategoride incelenir (12). Sıklıkla kullanılan algoritmalar: En Yakın Komşuluk (k-Nearest Neighbors (KNN)), Destek Vektör Makineleri (Support Vector Machine (SVM)), Karar Ağaçları (Decision Trees (DTs)), Doğrusal Regresyon (Linear Regression), Lojistik Regresyon (Logistic Regression), Naif Bayes (Naive Bayes), Yapay Sinir Ağları (Artificial Neural Network (ANN))'dır (14). Makine öğrenmesi yöntemleri arasında, KNN mesafe bazlı sınıflandırma yaparken, SVM destek fonksiyonları kullanır, karar ağaçları veriyi dallanmalar halinde gruplar, regresyon analizleri sayısal/kategorik ilişkileri modeller ve yapay sinir ağları biyolojik sinir sistemini taklit eden matematiksel modellerdir. Bu algoritmaların her biri farklı veri türleri ve problemler için optimize edilmiş olup, spesifik avantaj ve kısıtlamalara sahiptir (15). Gözetimsiz Öğrenme ise gözetimli öğrenmeden farklı olarak verilerde herhangi bir etiketlendirme bulunmaz ve sistem veriler arasında bağlantılar bulup birbirine yakın verilerin anlamlılıklarının arar. Bu amaçla Kümeleme (Clustering), Birliktelik Kuralı (Association Rule Mining), Boyut Azaltma (Dimensionality Reduction) gibi yöntemler kullanılır (16, 17).

Özellikle *de novo* varyantların değerlendirilmesinde klinik yorumlar belirsiz kalabilmekte ve farklı yaklaşımlar gerekebilmektedir. Günümüzde varyant değerlendirmeleri çoğunlukla manuel olarak, klinisyenlerin bireysel değerlendirmeleriyle yapılmaktadır. Bu nedenle, sürecin otomatikleştirilmesi ve yapay zeka destekli, yüksek doğrulukta klinik yorumlama sistemlerinin geliştirilmesi önemli bir ihtiyaç olarak görülmektedir. Bu çalışma ile makine öğrenme tekniklerinin bireyde tespit edilen kopya sayısı varyantının klinik etkisinin değerlendirilmesinde kullanılabilirliğini sorusunu araştırmayı planladık. Bu çalışma ile açık veri tabanlarında bildirilmiş kopya sayısı varyasyonları ile makine öğrenme teknikleri kullanarak ilgili CNV klinik sınıfının tahmin edilmesi amaçlanmıştır.

GEREÇ ve YÖNTEM

Öncelikle çalışma tasarımının denenmesi amacıyla 11989 varyantın bulunduğu ISCA (International Standards for Cytogenomic Arrays) konsorsiyumunun verileri kullanılarak pilot analiz gerçekleştirilmiştir (18, 19). Veriler dbVar veri tabanından CSV dosyası olarak alınmıştır. (20). Pilot çalışma sonuçları sonucu çalışmaya ClinVar veritabanına girilmiş 63156 varyantın bulunduğu nstd102 (Clinical Structural Variants), GRCh38 (hg38) versiyonu cinsiyet kromozomları hariç

tutularak çalışmada kullanılmıştır (21). Veri setlerindeki veriler klinik özelliklerine göre sınıflandırılmış ve hangi kromozomda bulundukları, genomik lokasyonları ve kayıp kazanç durumu bilgileri kullanılmıştır. Bu veriler içerisinden klinik özelikleri belirten Benign, Muhtemel Benign, Önemi bilinmeyen varyant (VUS), Muhtemel Patojenik ve Patojenik sınıfları kullanılmıştır. Veri setinde bulunan farklı tanımlanmış varyantlar uyumlu sınıf etiketine dönüştürülmüş, dönüştürülemeyen 535 varyant dışlanarak 66268 varyant ile çalışmaya devam edilmiştir. Oluşturulan veri setinde 35494 kopya sayısı kaybı ve 30774 kopya sayısı kazancı bulunmaktadır (Tablo-1). CNV'ler boyutlarına göre değerlendirildiğinde ortanca değer 132406 bç iken ortalama değer 2046347 bç olarak bulunmuştur. Veriler Microsoft Azure Machine Learning Studio kullanılarak analiz edilmiştir (22). Verilerden öncelikle gerekli sütunlar ayrılmış, sonrasında boş veriler temizlenmiştir (Clean Missing Data fonksiyonu ile). Verilerin analizi için ilgili varyantın bulunduğu kromozom bilgisi, kayıp kazanç bilgisi, kromozomal pozisyon başlangıç ve bitiş noktası ve klinik durum bilgisi kullanılmıştır. Örneklem %70 eğitim ve %30 test verisi olacak şekilde randomizasyon fonksiyonu ile iki gruba bölünmüştür. Eğitim modeli klinik etikete göre oluşturulmuştur. Analiz sırasında farklı makine öğrenme algoritmaları denenmiş ve sonuçlar birbirleriyle karşılaştırılmıştır. Eğitim modeli olarak: Çok Sınıflı Karar Ağacı-Forest (8 dal), Çok Sınıflı Karar Ağacı-Forest (16 dal), Çok Sınıflı Karar Ağacı-Forest (32 dal), Çok Sınıflı Karar Ağacı-Jungle, Çok Sınıflı Lojistik Regresyon ve Çok Sınıflı Sinir Ağı kullanılmıştır. Eğitim verisi test verisi ile karşılaştırılmış (Score Model) ve modelin başarısı değerlendirilmiştir (Evaluate Model). Ayrıca Sonuçlar Microsoft Excel kullanılarak değerlendirilmiştir ve grafikler oluşturulmuştur.

Tablo-1 Veri seti etiketlerinin dağılımı.

	Kazanç	**Kayıp**	**INDEL**	**Toplam**
Benign	10678	13427	5	24110
Muhteml Benign	2927	1889	18	4834
VUS	12454	7922	226	20602
Muhtemel Patojenik	854	1517	31	2402
Patolojenik	3861	10739	255	14855

BULGULAR

ISCA verileri ile yapılan çalışma sonucunda toplamda %89.241 ve ortalamada %96.4137 doğruluğa ulaşılmıştır. Plot veri setinde patojenik olarak tarif edilen örnekler %93,2 oranında patolojik %5,6 belirsiz etki, %1,2 benign olarak işaretlenirken, benign örnekler ise %90,3 oranında benign, %5,3 patojenik ve %4,4 belirsiz etki olarak işaretlenmişlerdir. Çok sınıflı lojistik regresyon (Multiclass Logistic Regression), çok sınıflı karar ormanı-Jungle (Multiclass Decision Jungle) ve çok sınıflı karar ormanı-forest (Multiclass Decision Forest) algoritmaları denenmiş ve bunlar içerisinde çok sınıflı karar ağacı en yüksek doğruluğa ulaşmıştır (Tablo-2). Pilot çalışmada veri etiketleri kendi içerisinde değerlendirilmemiştir ayrıca alt grupların 5 ten fazla olmasına rağmen modelin ortalama başarısı %96'nın üzerinde olarak değerlendirilmiştir.

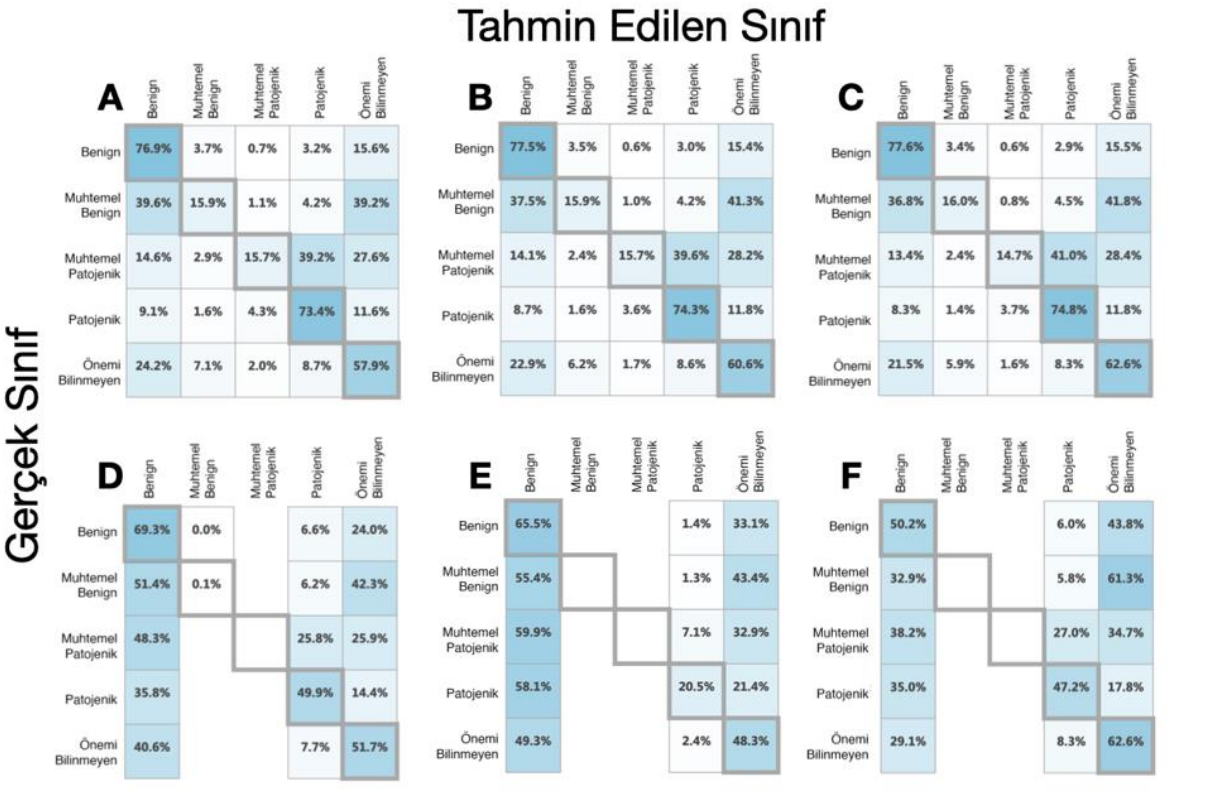
Pilot çalışma sonrası ClinVar veri seti için yapılan analiz de farklı yöntemler birbirleri ile karşılaştırılmış ve bunlar arasında çoklu karar ağacı-forest (32) yaklaşık %86 ortalama doğruluk ile en başarılı yöntem olarak değerlendirilmiştir (Tablo-3). Yine bu analizler sonrası elde edilen doğru tahminlerin dağılımı Şekil-1'de gösterilmiştir.

Tablo-2 Pilot çalışma değerlendirmesi.

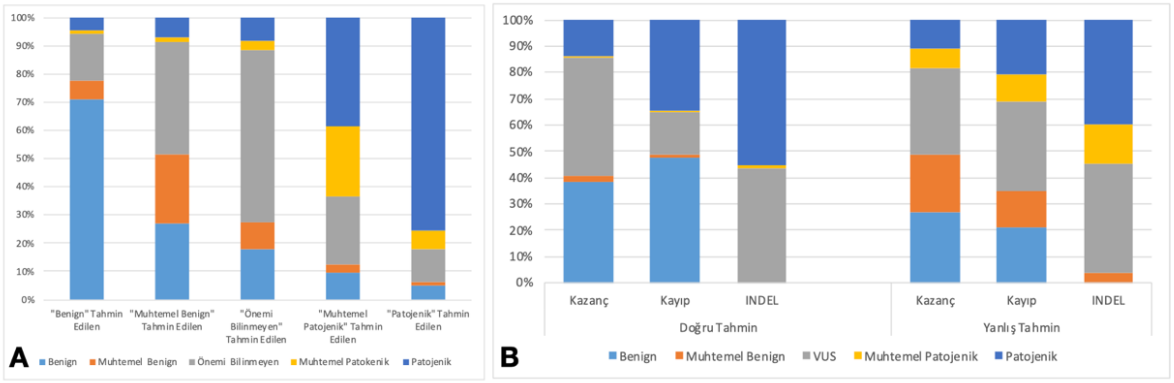
	Çok Sınıflı Lojistik Regresyon	**Çok Sınıflı Karar Ormanı-Jungle**	**Çok Sınıflı Karar Ormanı-Forest**
Toplam Doğruluk	0.46094	0.727829	0.89241
Ortalama Doğruluk	0.820313	0.909276	0.964137

Tablo-3. Eğitim verilerinin farklı algoritmalarla analizleri (NaN:verilerde numerik değer bulunmadığı için analiz edilememiştir.)

	Çok Sınıflı Karar Ağacı-Forest (8 dal)	**Çok Sınıflı Karar Ağacı-Forest (16 dal)**	**Çok Sınıflı Karar Ağacı-Forest (32 dal)**	**Çok Sınıflı Karar Ağacı-Jungle**	**Çok Sınıflı Lojistik Regresyon**	**Çok Sınıflı Sinir Ağı**
Toplam Doğruluk	0.637443	0.650167	0.657602	0.52188	0.432813	0.480415
Ortalama Doğruluk	0.854977	0.860067	0.863041	0.808752	0.773125	0.792166
Mikro-ortalama hassasiyet	0.637443	0.650167	0.657602	0.52188	0.432813	0.480415
Makro-ortalama hassasiyet	0.49486	0.510174	0.514232	NaN	NaN	NaN
Mikro-ortalama hatırlama	0.637443	0.650167	0.657602	0.52188	0.432813	0.480415
Mikro-ortalama hatırlama	0.479561	0.488108	0.491519	0.341952	0.268661	0.320082

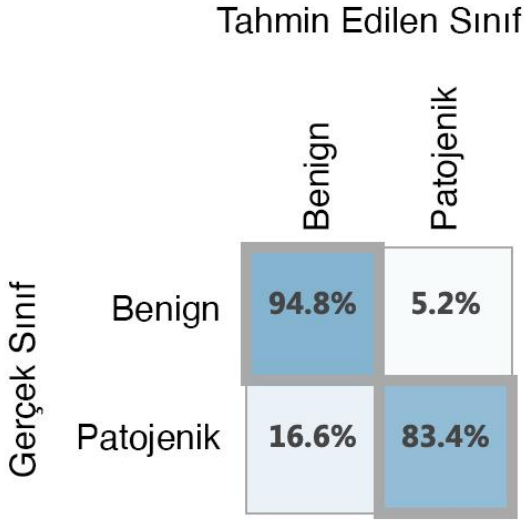


Şekil-1. A) Çok Sınıflı karar ağacı-forest (8 dal), B) Çok sınıflı karar ağacı-forest (16 dal), C) Çok sınıflı karar ağacı-forest (32 dal), D) Çok sınıflı karar ağacı-jungle, E) Çok sınıflı lojistik regresyon, F) Çok sınıflı sinir ağı doğru tahminlerin dağılımı



Şekil-2. Tahminlerin etiketlere göre dağılımı A) Gerçek etiketlerin tahmin edilen etiketler içerisindeki oranları B) Doğru ve Yanlış tahminlerin CNV türüne göre oranları

Çok sınıflı karar ağacı-forest (32 dal) analizinde muhtemel benign örnekler benign yönünde değerlendirilirken (%36,8) muhtemel patojenik örnekler patojenik (%41,0) yönünde kaymıştır. Benign ve patojenik kategorilerdeki tahmin başarısı diğer gruplara kıyasla daha yüksek bulunmuştur. Çalışmada incelenen varyantların çoğunluğu 10Kb-10Mb aralığında yer almakta olup, varyant boyutu arttıkça tahmin doğruluğunun yükseldiği, özellikle 100 baz çiftinden küçük varyantlarda yanlış tahmin oranının daha fazla olduğu tespit edilmiştir. (Şekil-2).



Şekil-3 İki sınıflı örneklem çoklu sınıf karar ağacı-forest(32 dal).

Örneklemdeki benign ve patojenik değerler dışındaki veriler dışlandıktan sonra analiz tekrarlanmış ve çok sınıflı karar ağacı-forest (32 dal) ile ortalama ve toplamda 0.905176 keskinliğe ulaşılmıştır (Şekil-3). İki sınıflı karar ağacı-forest algoritması 0.897 keskinlikte sonuç vermiştir. İki sınıflı sinir ağı analizinde 0,730 keskinliğe ulaşılırken iki sınıflı destek vektör makinesi (SVM) ile 0.669 keskinliğe ulaşılmıştır.

TARTIŞMA

Kopya sayısı varyantlarının (CNV) klinik değerlendirmesi önemli zorluklar içermektedir. Modern array teknolojileri ve yeni nesil dizileme yöntemleri CNV tespitini kolaylaştırsa da, özellikle pozisyonel ve oryantasyonel bilgi eksikliği gibi kısıtlılıklar devam etmektedir (3). Gen içeren bölgelerdeki CNV'lerin değerlendirilmesi, haplo-yetmezlik skorları ve gen/dozaj ilişkisi sayesinde nispeten kolay olsa da, düzenleyici genom bölgelerindeki CNV'lerin (tüm analizlerin %95'i) değerlendirilmesi oldukça zordur (23). Bu zorlukları aşmak için çeşitli değerlendirme kriterleri ve skorlama sistemleri önerilmiştir (24) . Değerlendirme süreci genellikle veri tabanı taraması ile başlar, *de novo* varyantlarda aile çalışmaları ve segregasyon analizleri önem kazanır (25). Ancak süreç halen uzman değerlendirmesi gerektiren, zaman alıcı ve maliyetli bir yapıdadır.

Makine öğrenme teknikleri, teknolojik gelişmeler ve algoritmaların iyileştirilmesi sayesinde günlük yaşamın önemli bir parçası haline gelmiştir (26). Özellikle sağlık sektöründe geniş uygulama alanı bulan bu teknolojiler, büyük veri analizinin yoğun olduğu genetik alanında da kendine önemli bir yer edinmiştir (27). Makine öğrenme teknikleri, genetik varyantların değerlendirilmesinden kanser gibi karmaşık hastalıkların araştırılmasına kadar geniş bir yelpazede kullanılmaktadır (28–30). Ayrıca, yeni nesil dizileme verilerinden kopya

sayısı varyantlarının belirlenmesinde de bu tekniklerden yararlanılmaktadır (31).

Bu araştırma, klinik önemi belirsiz olan genetik varyantların değerlendirilmesinde makine öğrenme modellerinin kullanımını hedeflemiştir. Çalışmada, açık erişimli bir veri tabanı olan dbVar'ın nstd102 veri seti kullanılmıştır (32). dbVar, insan genomunda yaklaşık 6 milyon genomik bölgeyi kapsayan 35 milyondan fazla varyant içermekte olup, 100'ün üzerinde bilimsel çalışmadan elde edilen verileri barındırmaktadır. Bu verilerin önemli bir kısmı popülasyon çalışmalarından elde edilmiş olup, bazı verilerde klinik bilgiler mevcut değildir (33–35). İnsan genomunun büyüklüğü göz önüne alındığında, 66 bin varyantlık veri seti oldukça sınırlı bir kapsamı temsil etmektedir. Veri setinde yaygın mikrodelesyon/duplikasyon sendromları iyi temsil edilirken, nadir görülen yapısal değişiklikler için yeterli veri bulunmamaktadır. Ayrıca, veri setinin en önemli eksikliği doğrudan gen bilgisi içermemesidir. Her ne kadar kromozomal lokasyonlar gen içeriği hakkında fikir verse de, etkilenen gen ve gen bölgelerinin analize dahil edilmesi çalışmanın etkinliğini önemli ölçüde artıracaktır. Özellikle mikrodelesyon sendromlarında, kritik gen bölgelerinin kaybı klinik tabloyu belirlediğinden, genomik bölge analizlerinin gen/ekzon bilgisini içermesi sonuçların doğruluğunu güçlendirecektir (36). Ayrıca ilgili bölgenin pozisyonel durumu, epigenetik derişikliklerin olabilmesi, cinsiyetler arası farklılıklar gibi birçok etken bu çalışmanın başarısını kısıtlamaktadır.

Kromozomal dağılımlar incelendiğinde, otozomal kromozomların çoğunlukla dengeli bir dağılım gösterdiği, ancak 15, 16, 17, 19 ve 22. kromozomlarda beklenenden daha fazla varyasyon bulunduğu gözlemlenmiştir. Bu kromozomlardaki farklılıkların, buralarda sık görülen mikrodelesyon/duplikasyon sendromlarıyla ilişkili olduğu düşünülmektedir (37). Doğru tahminlerin kromozomal dağılımında ise sadece 15, 16 ve 17. kromozomlarda belirgin farklılıklar tespit edilmiş olup, bu durumun da yine bu bölgelerdeki mikrodelesyon/duplikasyon sendromlarının sıklığıyla bağlantılı olabileceği öne sürülmüştür.

Pilot çalışmada yüksek doğruluk oranları (%96) elde edilmesine rağmen, daha geniş veri setinde farklı yöntemler denenmesine karşın doğruluk oranları daha düşük seyretmiştir. Pilot çalışmadaki veri seti daha küçük olmasına rağmen, iyi kürate edilmiş ve sınırlı kaynaktan gelen varyantlardan oluşurken, ClinVar veritabanı daha çok kaynaktan gelen ancak daha az denetlenmiş veriler içermektedir. ClinVar'ın klinik kullanımda güvenilirliği sıklıkla sorgulanmakta, özellikle nokta mutasyonlarındaki çelişkili kayıtlar nedeniyle yorumlama zorlukları yaşanmaktadır (38).

Yapılan analizlerde, varyant sınıflandırmalarının çoğunlukla "önemi bilinmeyen varyant" kategorisine kayma eğilimi gösterdiği tespit edilmiştir. Uluslararası kılavuzlar, varyantların klinik etkisi kesin olarak kanıtlanana kadar beş basamaklı bir sınıflandırma sisteminin kullanılmasını önermektedir (39, 40). Ancak bu çok kategorili sistem, doğru değerlendirmeyi zorlaştırmaktadır. Araştırma sonuçları, benign ve patojenik kategorilerdeki tahminlerin daha başarılı olduğunu, ancak ara kategorilerde sınıflar arası kaymaların daha fazla olduğunu göstermiştir. Özellikle, muhtemel benign, muhtemel patojenik ve önemi bilinmeyen varyantlar çıkarıldığında analiz başarısının arttığı gözlemlenmiştir. Makine öğrenme algoritmaları arasında, iki sınıflı ve çok sınıflı karar ormanları, SVM ve sinir ağı yöntemlerine kıyasla daha yüksek performans sergilemiştir.

Yapay zeka ve makine öğrenme teknolojileri, özellikle tıp alanında giderek daha fazla önem kazanmaktadır. Artan bilimsel veri hacmi ve karmaşık parametreler, insan yorumlama kapasitesini aşmakta ve tıbbi uygulamaların dijitalleşmesini zorunlu kılmaktadır (41). Tıbbi görüntüleme, laboratuvar ve nörolojik kayıtlar gibi sistemlerden elde edilen büyük veriler, yapay zeka algoritmaları ile analiz edilmekte ve bu uygulamalar klinik verilerin doğru yorumlanması ve maliyet etkinliği açısından değerli sonuçlar vermektedir (42,43). Microsoft Azure Machine Learning Studio gibi platformlar, kullanıcı dostu arayüzleriyle programlama deneyimi olmayan kişiler için bile makine öğrenme süreçlerini kolaylaştırmaktadır.

Bu çalışma, özellikle mental retardasyon ve çoklu konjenital anomali hastalarının tanısında kullanılan kopya sayısı varyasyonlarının analizi için veri analiz yöntemlerinin geliştirilmesini amaçlamıştır. Gelişen teknoloji ve dizi analizi uygulamaları, mikroarray teknolojileri sayesinde kopya sayısı varyantlarının tespiti kolaylaşmakla birlikte, genomik verilerin tam olarak anlaşılamaması ve değerlendirme zorlukları halen devam etmektedir. Çalışma, makine öğrenme algoritmalarının klinik uygulamalarda

kullanılabilirliğini göstermiş olup, modelin optimize edilmesi için veri setlerinin genişletilmesi, standart verilerin oluşturulması ve çeşitli parametrelerin iyileştirilmesi gerekmektedir. Gelişen teknoloji ve artan genetik veri birikimiyle birlikte, bu çalışmanın genetik tanı ve tedavide dijitalleşme sürecine önemli katkılar sağlaması beklenmektedir.

Çıkar çatışması: Bu çalışmada yazarlar arasında çıkar çatışması bulunmamaktadır.

Kaynaklar

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004 Jul 23;305(5683):525–8.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006 Nov;444(7118):444–54.
3. Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. Hum Mol Genet. 2003 Oct 15;12(suppl 2):R145–52.
4. Slavotinek AM. Novel microdeletion syndromes detected by chromosome microarrays. Hum Genet [Internet]. 2008 Aug 30 [cited 2019 Nov 3];124(1):1–17. Available from: http://link.springer.com/10.1007/s00439-008-0513-9
5. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res [Internet]. 2006 Aug 1 [cited 2019 Jul 9];16(8):949–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16809666
6. Database of Genomic Variants [Internet]. [cited 2019 Nov 3]. Available from: http://dgv.tcag.ca/dgv/app/home
7. Firth H V., Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics [Internet]. 2009 Apr [cited 2019 Nov 3];84(4):524–33. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929709001074
8. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genetics in Medicine. 2011 Jul 15;13(7):680–5.
9. Shalev-Shwartz Shai, Ben-David Shai. Understanding machine learning : from theory to algorithms. 397 p.
10. Isakov O, Dotan I, Ben-Shachar S. Machine Learning–Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. Inflamm Bowel Dis. 2017 Sep 1;23(9):1516–23.
11. Ainscough BJ, Barnell EK, Ronning P, Campbell KM, Wagner AH, Fehniger TA, et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nat Genet. 2018;50(12):1735–43.
12. Barber D. Bayesian reasoning and machine learning. Cambridge University Press; 2012. 697 p.
13. Alpaydin E. Machine Learning - Ethem Alpaydin. 2016 [cited 2025 Jan 10];112–8. Available from: https://mitpress.mit.edu/9780262529518/machine-learning/
14. Beam AL, Drazen JM, Kohane IS, Leong TY, Manrai AK, Rubin EJ. Artificial Intelligence in Medicine. New England Journal of Medicine [Internet]. 2023 Mar 30 [cited 2024 Aug 9];388(13):1220–1. Available from: https://www.nejm.org/doi/full/10.1056/NEJMe2206291
15. Shotton J, Sharp T, Kohli P, Nowozin S, Winn J, Criminisi A. Decision Jungles: Compact and Rich Models for Classification [Internet]. 2013 [cited 2019 Nov 27]. Available from: https://www.microsoft.com/en-us/research/publication/decision-jungles-compact-and-rich-models-for-classification/
16. Mayoraz E, Alpaydin E. Support vector machines for multi-class classification. In Springer, Berlin, Heidelberg ; 1999 [cited 2019 Nov 25]. p. 833–42. Available from: http://link.springer.com/10.1007/BFb0100551
17. Ainscough BJ, Barnell EK, Ronning P, Campbell KM, Wagner AH, Fehniger TA, et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nat Genet [Internet]. 2018 [cited 2019 Nov 10];50(12):1735–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30397337
18. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. The American Journal of Human Genetics. 2010 May 14;86(5):749–64.

19. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med. 2011 Sep;13(9):777–84.
20. nstd101 - ClinGen - dbVar Study - NCBI [Internet]. [cited 2019 Nov 18]. Available from: https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd101/
21. nstd102 - Clinical Structural Variants - dbVar Study - NCBI [Internet]. [cited 2019 Nov 18]. Available from: https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd102/
22. Microsoft Azure Machine Learning Studio (classic) [Internet]. [cited 2019 Nov 18]. Available from: https://studio.azureml.net/
23. Spielmann M, Klopocki E. CNVs of noncoding cis-regulatory elements in human disease. Curr Opin Genet Dev. 2013 Jun 1;23(3):249–56.
24. Brandt T, Sack LM, Arjona D, Tan D, Mei H, Cui H, et al. Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. Genetics in Medicine. 2019 Sep 19;1–9.
25. Koolen DA, Pfundt R, de Leeuw N, Hehir-Kwa JY, Nillesen WM, Neefs I, et al. Genomic microarrays in mental retardation: A practical workflow for diagnostic applications. Hum Mutat. 2009 Mar 1;30(3):283–92.
26. Barber D. Bayesian reasoning and machine learning. Cambridge University Press; 2012. 697 p.
27. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019 Jan 26;51(1):12–8.
28. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet [Internet]. 2019 Jan 26 [cited 2019 Nov 24];51(1):12–8. Available from: http://www.nature.com/articles/s41588-018-0295-5
29. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature [Internet]. 2021 Aug 26 [cited 2024 Jul 26];596(7873):583–9. Available from: https://pubmed.ncbi.nlm.nih.gov/34265844/
30. de Sainte Agathe JM, Filser M, Isidor B, Besnard T, Gueguen P, Perrin A, et al. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. Hum Genomics [Internet]. 2023 Dec 1 [cited 2024 Jul 26];17(1). Available from: https://pubmed.ncbi.nlm.nih.gov/36765386/
31. Hill T, Unckless RL. A Deep Learning Approach for Detecting Copy Number Variation in Next-Generation Sequencing Data. G3 (Bethesda) [Internet]. 2019 Nov 5 [cited 2019 Nov 17];9(11):3575–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31455677
32. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013 Jan;41(Database issue):D936-41.
33. Sneddon TP, Church DM. Online resources for genomic structural variation. Methods Mol Biol [Internet]. 2012 [cited 2019 Nov 24];838:273–89. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22228017
34. NCBI Variation Summary [Internet]. [cited 2019 Nov 24]. Available from: https://www.ncbi.nlm.nih.gov/dbvar/content/org_summary/
35. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature [Internet]. 2016 Oct 13 [cited 2019 Nov 24];538(7624):201–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27654912
36. Rauch A, Schellmoser S, Kraus C, D&#xFFFD;rr HG, Trautmann U, Altherr MR, et al. First known microdeletion within the Wolf-Hirschhorn syndrome critical region refines genotype-phenotype correlation. Am J Med Genet. 2001 Apr 1;99(4):338–42.
37. Slavotinek AM. Novel microdeletion syndromes detected by chromosome microarrays. Hum Genet. 2008 Aug 30;124(1):1–17.
38. Peterson TA, Doughty E, Kann MG. Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. J Mol Biol. 2013 Nov 1;425(21):4047–63.
39. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genetics in Medicine [Internet]. 2011 Jul 15 [cited 2019 Nov 3];13(7):680–5. Available from: http://www.nature.com/doifinder/10.1097/GIM.0b013e3182217a3a

40. Brandt T, Sack LM, Arjona D, Tan D, Mei H, Cui H, et al. Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. Genetics in Medicine [Internet]. 2019 Sep 19 [cited 2019 Nov 3];1–9. Available from: http://www.nature.com/articles/s41436-019-0655-2
41. Hanke RE, Gibbons AT, Casar Berazaluce AM, Ponsky TA. Digital Transformation of Academic Medicine: Breaking Barriers, Borders, and Boredom. J Pediatr Surg [Internet]. 2019 Nov 9 [cited 2019 Nov 27]; Available from: https://www.sciencedirect.com/science/article/pii/S0022346819307729?via%3Dihub
42. Al-Mufti F, Kim M, Dodson V, Sursal T, Bowers C, Cole C, et al. Machine Learning and Artificial Intelligence in Neurocritical Care: a Specialty-Wide Disruptive Transformation or a Strategy for Success. Curr Neurol Neurosci Rep [Internet]. 2019 Nov 13 [cited 2019 Nov 27];19(11):89. Available from: http://link.springer.com/10.1007/s11910-019-0998-8
43. Kilic A. Artificial Intelligence and Machine Learning in Cardiovascular Healthcare. Ann Thorac Surg [Internet]. 2019 Nov 7 [cited 2019 Nov 27]; Available from: https://www.sciencedirect.com/science/article/pii/S0003497519316121?via%3Dihub