T.C.
EGE ÜNİVERSİTESİ
Sağlık Bilimleri Enstitüsü

KOPYA SAYISI VARYASYONLARININ MAKİNE ÖĞRENME YÖNTEMİ İLE ANALİZ EDİLMESİ

Doktora Tezi

Uzm. Dr. Erhan PARILTAY

Sağlık Biyoinformatiği Anabilim Dalı

İzmir
2019

T.C.
EGE ÜNİVERSİTESİ
Sağlık Bilimleri Enstitüsü

KOPYA SAYISI VARYASYONLARININ MAKİNE ÖĞRENME YÖNTEMİ İLE ANALİZ EDİLMESİ

Uzm. Dr. Erhan PARILTAY

Danışman
Doç. Dr. Buket KOSOVA

Sağlık Biyoinformatiği Anabilim Dalı
Sağlık Biyoinformatiği Yüksek Lisans Dereceli Doktora

İzmir
2019

Tez Değerlendirme Kurulu Üyeleri

(Adı Soyadı)
(İmza)

Başkan	: Doç.Dr. Buket Kosova **(Danışman)**	.....
Üye	: Doç. Dr. Burak ORDİN	.....
Üye	: Prof. Dr. Muhsin Özgür ÇOĞULU	.....
Üye	: Prof. Dr. Ahmet Okay ÇAĞLAYAN	.....
Üye	: Doç. Dr. Tufan ÇANKAYA	.....
Üye	: Doç. Dr. Elçin BORA	.....

Doktora Tezinin kabul edildiği tarih: 18.12.2019

Önsöz

Tıp ve Genetik ile başlayıp Biyoinformatik ile devam eden eğitimimin üç farklı disiplini bu çalışma ile bir arada değerlendirmeye çalıştım. Geleceğin dünyasında daha da iç içe geçeceğini düşündüğüm bu üç alanın içinde bulunmaktan onur duyuyorum.

İzmir, 19.12.2019

Erhan PARILTAY

Özet

Kopya Sayısı Varyasyonlarının Makine Öğrenme Yöntemi İle Analiz Edilmesi

Kopya sayısı varyasyonları insan genomunun yaklaşık yüzde ikisinde bulunan belirli bir DNA bölgesinin kayıp veya kazançlarıdır. Yapısal varyasyonlar arasında yer alan bu grup sağlıklı popülasyonda bulunabileceği gibi ilgili bölgenin kayıp veya kazançları klinik tablolarla ilişkilendirilebilir. Tespit edilen kopya sayısı varyasyonunun klinik olarak yorumlamak aile çalışmasını da gerektiren, klinik ve genetik verinin değerlendirme sürecidir ve her zaman doğru bilgiye ulaşılamamaktadır. Kullanımı artan makine öğrenme algoritmaları tıp alanında da kullanılmakta ve özellikle büyük veri setlerinin bulunduğu genetik gibi alanlarda giderek önem kazanmaktadır. Bu çalışma ile kopya sayısı varyasyonlarının klinik açıdan değerlendirilmesinde makine öğrenme algoritmalarının kullanımı amaçlanmıştır. ClinVar veri seti ile birden fazla çok sınıflı makine öğrenme algoritmaları kullanılarak makine öğrenme modelleri oluşturulmuştur. Daha küçük veri seti ile yapılan modelde çok sınıflı karar ağacı ile ortalamada 0,96 doğruluğa ulaşılırken ana veri setinde yine çok sınıflı karar ağacı ile 0,86 doğruluğa ulaşılmıştır. Çalışmada sık karşılaşılan tanımlı varyantların daha yüksek başarı ile tanımlandığı, yine iki sınıflı benign patojenik ayrımında modelin daha başarılı olduğu gösterilmiştir. Bu çalışma kopya sayısı varyantlarının klinik değerlendirilmesinde kullanılabilecek ve tanıyı otomatikleştirebilecek öncül bir makine öğrenme modeli oluşturulabileceğini göstermiştir.

Anahtar Kelimeler; Kopya Sayısı Varyasyonları; Makine Öğrenme; Genetik; Biyoinformatik

Abstract

Analysis of Copy Number Variations with Machine Learning Technics

Copy number variations, that are losses or gains of particular DNA regions, are found about two percent in the human genome. This group which belongs to structural variations can be found in the healthy population and sometimes losses or gains specific regions can lead to clinical manifestations. Interpretation of clinical and genetic data, which also requires family work to interpret clinically the detected copy number variation, does not always provide accurate information. Increasing application of machine learning algorithms is getting widely used in the medical field and becoming increasingly important, especially in areas with large data such as genetics. The aim of this study is to use machine learning algorithms in clinical interpretation of copy number variations. Machine learning models were created using multiple multi-class machine learning algorithms with the ClinVar data set. In the model made with a smaller data set, the average accuracy of the multi-class decision forest is 0.96, while in the main data set, the accuracy of the multi-class decision forest is 0.86. In this study, it could be shown that the most common identified variants were interpreted with higher accuracy and the model was more successful when the two classes benign and pathogenic were separated. This preliminary study demonstrated that a machine learning model could be used to automate diagnosis and in the clinical interpretation of copy number variants.

Keywords; Copy Number Variations; Machine Learning; Genetics; Bioinformatics

İçindekiler

Önsöz ..... **II**
Özet ..... **III**
Abstract ..... **IV**
İçindekiler ..... **V**
Tablolar Dizini ..... **VI**
Şekiller Dizini ..... **VII**
Grafikler Dizini ..... **VIII**
Kısaltma Listesi ..... **IX**
Giriş **1**
1.1. Araştırmanın Problemi **2**
1.2. Araştırmanın Sorusu **2**
1.3. Araştırmanın Hipotezleri **3**
1.4. Araştırmanın Varsayımları **3**
1.5. Araştırmanın Sınırlılıkları **3**
1.6. Araştırmanın Amacı **3**
Genel Bilgiler **4**
2. Genom ve Varyasyonlar **4**
2.1. Kopya Sayısı Varyantları **4**
2.2. Kopya Sayısı Varyantlarının Klinik Önemi **5**
3. Makine Öğrenme ve Yapay Zeka **7**
3.1. Makine Öğrenme Çeşitleri **8**
3.1.1. Gözetimli Öğrenme **9**
3.1.2. Gözetimsiz Öğrenme **10**
Gereç ve Yöntem **12**
Bulgular **17**
Tartışma **33**
Sonuç ve Öneriler **40**
Kaynaklar **41**
Ekler **45**
Teşekkür **47**
Özgeçmiş **50**

Tablolar Dizini

Tablo 1: Sık görülen bazı mikrodelesyon sendromları ve insidansları (Devriendt and Vermeesch, 2004). 6
Tablo 2 Kopya Sayısı Varyantlarının Klinik Yorumlanmasında Kullanılan ACMG önerileri (Kearney et al., 2011) 7
Tablo 3 Veri seti etiketlerinin dağılımı 13
Tablo 4 Pilot çalışma değerlendirmesi 17
Tablo 5 Eğitim verilerinin farklı algoritmalarla analizleri 19
Tablo 6 Çok Sınıflı Karar Ağacı-Forest (32 dal) tahminlerin dağılımı 23
Tablo 7 Tahmin etiketlerine göre kayıp/kazanç durumunu dağılımı 26

Şekiller Dizini

Şekil 1 Gözetimli ve Gözetimsiz Öğrenme (Langs et al., 2018) 9
Şekil 2 Sınıflandırma ve Regresyon algoritmalarının grafiksel gösterimi 10
Şekil 3 Microsoft Azure Machine Learning Studio (classic) Örnek Çalışma Alanı . 16
Şekil 4 Pilot çalışma sonrası elde edilen çok sınıflı karar ormanı-forest 18
Şekil 5 Çok Sınıflı karar ağacı-forest (8 dal) doğru tahminlerin dağılımı 19
Şekil 6 Çok sınıflı karar ağacı-forest (16 dal) doğru tahminlerin dağılımı 20
Şekil 7 Çok sınıflı karar ağacı-forest (32 dal) doğru tahminlerin dağılımı 20
Şekil 8 Çok sınıflı karar ağacı-jungle doğru tahminlerin dağılımı 21
Şekil 9 Çok sınıflı lojistik regresyon doğru tahminlerin dağılımı 21
Şekil 10 Çok sınıflı sinir ağı doğru tahminlerin dağılımı 22
Şekil 11 İki sınıflı örneklem çoklu sınıf karar ağacı-forest(32 dal) 30
Şekil 12 İki sınıflı örneklemin iki sınıflı karar ağacı ile doğruluk analizi 31
Şekil 13 İki sınıflı örneklemin sinir ağı ile doğruluk analizi 31
Şekil 14 İki sınıflı örneklemin destek vektör makinesi (SVM) ile doğruluk analizi . 32

Grafikler Dizini

Grafik 1 Verilerin kromozomlara göre dağılımı 14
Grafik 2 CNV boyutlarının dağılımı (bç)..... 14
Grafik 3 Gerçek etiketler ile tahmin edilen etiketlerin dağılımı 24
Grafik 4 Çok Sınıflı Karar Ağacı-Forest (32 dal) Sonuçlarının kromozomlara göre dağılımı, ... 25
Grafik 5 Çok Sınıflı Karar Ağacı-Forest (32 dal) Sonuçların CNV boyutuna göre karşılaştırılması 25
Grafik 6 Tahmin etiketlerine göre kayıp/kazanç durumunu dağılım grafiği 26
Grafik 7 1. Kromozom doğru tahminlerin kromozomal lokasyonları 27
Grafik 8 1. Kromozom yanlış tahminlerin kromozomal lokasyonları 27
Grafik 9 15. Kromozom doğru tahminlerin kromozomal lokasyonları 28
Grafik 10 15. Kromozom yanlış tahminlerin kromozomal lokasyonları 28
Grafik 11 22. Kromozom doğru tahminlerin kromozomal lokasyonları 29
Grafik 12 22. Kromozom yanlış tahminlerin kromozomal lokasyonları 29

Kısaltma Listesi

SNP	:	Tek Nükleotid Polimorfizmi (Single Nucleotide Polymorphisms)
VNTR	:	Değişken Sayıda Tandem Tekrarlar (Variable number tandem repeat)
CNV	:	Kopya Sayısı Varyasyonu (Copy Number Variation)
Kb	:	Kilo Baz (Kilo Base)
Mb	:	Maga Baz (Mega Base)
VUS	:	Önemi Bilinmeyen Varyant (Variant of uncertain significance)
bç	:	Baz Çifti
SVM	:	Destek Vektör Makinesi (Support vector machine)

Giriş

İnsan genomu yaklaşık 6 milyar baz çiftinden oluşmaktadır ve yaklaşık 20000 gen 23 çift kromozom üzerinde bulunur. Kişiler arası farklılıklar genomik düzeyde %0.01'den daha azdır. Bireysel farklılıkların oluşumu DNA dizisi üzerindeki baz değişikliklerinin yanında genomik yeniden düzenlenmeleri, epigenetik ve çevresel nedenleri de içermektedir. Özellikle gelişen teknoloji genomik yapıların araştırılmasını ve genomik materyallerin değerlendirilmesini kolaylaşmıştır.

Genomik varyantlar tek nükleotid polimorfizmler, (SNP), değişken sayıdaki ardışık tekrarlar (VNTR mini ve mikro satellitler), transpozon elementlerin varlığı veya yokluğu (Alu elementleri) ve delesyon, duplikasyon inversiyon gibi yapısal değişiklikler olarak kendini gösterirler (Freeman et al., 2006). Bunun yanında SNP'lerin fenotipik çeşitlilikteki sorumlu yapılar olduğu düşünülmüş ancak sonraları kopya sayısı varyasyonlarının (CNV) daha yaygın oldukları tespit edilmiştir (Iafrate et al., 2004). İnsan genomundaki varyasyonların araştırılması sonucu genomik bölgelerini ardışık veya lineer olmayan şekillerde birden fazla kopyaya sahip olabildiği gösterilmiştir. Kompleks yeniden düzenlenmeler sonucu kopya sayısı varyasyonları ortaya çıkar.

Yaygın olarak tespit edilen kopya sayısı varyantlarının büyük çoğunluğu herhangi bir klinik öneme sahip değildir; bunun yanında yaygın gelişimsel geriliklerin ve çoklu konjenital anomalilerin eşlik ettiği ciddi klinik tablolara da yol açmaktadır. Klinik nedenselliğin araştırılması genellikle eğitimli klinisyenlerin değerlendirmesine ihtiyaç duymaktadır.

Makine öğrenme (yapay zeka) yöntemlerinin teorileri yirminci yüzyılın ilk yarılarında şekillenmeye başlamıştır fakat teknik yetersizlikler nedeniyle yaygın kullanıma kavuşması elli yıldan uzun sürmüştür. Son yıllardaki artan işlemci güçleri ve modern programlama yazılımları ile birçok alanda kullanılabilir ve yaygın uygulanabilir hale gelmiştir. Makine öğrenme başlığı altında birçok farklı algoritmalar kullanılabilmektedir.

Bu çalışmada daha önce literatürde örneğine rastlamadığımız kopya sayısı varyantlarının klinik yorumlanmasında makine öğrenme yöntemlerinin kullanımının araştırılması planlanmıştır.

1.1. Araştırmanın Problemi

Kopya sayısı varyasyonlarının değerlendirilmesi, klinik olarak yorumlanması zor bir süreçtir. Öncelikli olarak elde edilen varyasyonun daha önce tanımlanmış kopya sayısı varyasyonları ile karşılaştırılması gerekir. Bu amaçla oluşturulmuş çok uluslu DGV ve DECIPHER ("Database of Genomic Variants," n.d.; Firth et al., 2009) gibi veri tabanları yaygın olarak kullanılmaktadır. Ancak tanımlı varyantlar, tespit edilenlerle her zaman örtüşmemekte ve *de novo* varyatlar bu veri tabanlarında bulunmamaktadırlar. Ebeveyn çalışması ve kalıtım gösteren varyantların benign olarak yorumlanması da sık kullanılan yaklaşımlar arasında yer almaktadır. Ayrıca daha önceleri tanımlanan klinik durumlarda kullanılmak üzere istatistiksel yöntemlerle oluşturulan her bir gen için hesaplanan "Haployetmezlik (Haploinsufficiency) skorları ilgili kayıp kazanç bölgesindeki varyantların etkilerinin değerlendirilmesinde kullanılmaktadır.

Kopya sayısı varyasyonlarının klinik yorumlanmasında kullanılan birçok algoritmaya rağmen her zaman neden-sonuç ilişkisi kurulamamakta ve özellikle *de novo* varyantlarda klinik yorum askıda kalmakta ve alternatif bakış açısına ihtiyaç duymaktadır. Ayrıca çoğu durumda varyantların değerlendirilmesi tek tek insan gözü ve klinisyen yorumu ile yapılmaktadır. Bu sürecin otomatikleştirilmesi ve makine destekli yüksek doğrulukta klinik yorumların oluşturulması da beklenmektedir.

1.2. Araştırmanın Sorusu

Kopya sayısı varyantları tüm popülasyonun neredeyse tamamında bulunmasına rağmen bunların çok az bir kısmı klinik olarak anlamlı olarak değerlendirilirler. Başta çoklu konjenital anomali ve yaygın gelişimsel gerililikler olmak üzere birçok klinik tablo ile ilişkili olarak değerlendirilirler. Klinik etkilerinin değerlendirilmesi yaygın olmayan CNV'ler için zorlayıcı olabilmektedir. Moleküler veya sitogenetik yöntemlerle elde edilen CNV bilgilerinin değerlendirilmesi klinik uzman değerlendirmesi ile mevcut veri tabanlarındaki verilerin karşılaştırılmasına dayanmaktadır. Özellikle son yıllarda kullanım alanları artan makine öğrenme tekniklerinin bireyde tespit edilen kopya sayısı varyantının klinik etkisinin değerlendirilmesinde kullanılabilirliği sorusunu araştırmayı planladık.

1.3. Araştırmanın Hipotezleri

Makine öğrenme teknikleri patolojik varyantları tanımlamak için kullanılabilecek bir skorlama yöntemi geliştirebilir. Böylelikle genomik veriden patojenite tahminleri otomatik olarak oluşturulabilir. Makine öğrenme modelleri ile oluşturulan model yeni tespit edilecek kopya sayısı varyantlarının klinik etkileri yorumlaması değerlendirilecektir.

1.4. Araştırmanın Varsayımları

Genlerin genomik lokasyonları bilindiği için tespit edilen kopya sayısı varyantlarının hangi genleri kapsadığı bulunabilecektir. Klinik tablolarla ilişkilendirilmiş gen kayıp ve kazançları genellikle yaygın olarak tanımlanmıştır. Bu verilerin makine öğrenme ile değerlendirilmesi yeni klinik tanılar için kolaylaştırıcı olacaktır.

1.5. Araştırmanın Sınırlılıkları

Genlerin organizma üzerine olan etkileri her zaman dozaj ile ilişkili değildir. Tespit edilen CNV'nin lokasyonu, oryantasyonu ve birleşim yeri bilgileri mevcut yöntemlerle pratik olarak hesaplanamamaktadır. Ayrıca kompleks genomik regülasyon basamakları gen ekspresyonu üzerindeki diğer faktörlerin dışlanmasını da zorlaştırmaktadır. Özellikle iyi tanımlı tek gen hastalılarında etkilenen gene göre klinik tabloyu yorumlamak göreceli kolay iken birleşik gen sendromları birleşik, üst üste binmiş bir fenotip ortaya çıkarmaktadır. Ayrıca model oluşturulması tamamen veri tabanlarında tanımlı mevcut bilgiler üzerine kurulacaktır. Ancak mevcut veri tabanları hem yetersiz hem de hatalı bilgiye sahip olabilirler.

1.6. Araştırmanın Amacı

Kopya sayısı varyasyonları (CNV) genomumuzdaki gen kopyalarının sayısal varyasyonlarını tanımlar. Gen/dozaj ilişkisi ya da haploinsufficiency genlerin normal çalışma durumlarının analizi için önemli belirteçlerdir. Bu çalışma ile veri tabanlarında tanımlanmış kopya sayısı varyasyonlarını makine öğrenme teknikleri kullanarak CNV'lerin patolojik tablolarla ilişkisinin tahmin edilmesi amaçlanmıştır.

Genel Bilgiler

2. Genom ve Varyasyonlar

İnsan genomu %99.9'dan daha fazla benzerlik gösterse de varyasyonlar birçok şekilde karşımıza çıkabilir (Reich et al., 2002). Tek nükleotid değişiklikleri, küçük insersiyon ve delesyonlar, değişken sayıdaki tekrarlar, mikrosatellitler, minisatellitler ve büyük boyutlu kopya sayısı varyantları sık görülen formlar arasındadır (Feuk, Marshall, Wintle, & Scherer, 2006). Bunlar arasında en az bilinenlerden biri olan birkaç yüz kilo baz uzunluğa kadar büyüyebilen kopya sayısı varyantlarıdır (CNV) (Iafrate et al., 2004).

2.1. Kopya Sayısı Varyantları

Kopya sayısı varyantları ortalama olarak 100 Kb'den daha büyük segmentlerin yapısal olarak kayıp ve kazancu şeklinde tarif edilir (Sebat et al., 2004) . CNV tanımı genel olarak oryantasyonel bilgi içermeyen kayıp ve kazanç bölgelerini tanımlar. Yaklaşık 20 Kb'dan küçük CNV'ler hemen her bireyde bulunurlar (Redon et al., 2006). Göreceli olarak diğer genomik varyantlara göre daha geç tanımlanmışlardır. Bu gecikmenin başlıca nedeni teknik ve biyolojik kısıtlılıkların aşılması için çeşitli yeniliklere ihtiyaç duyulmuş olmasıdır. DNA seviyesindeki varyantların tespiti dizileme teknikleri ile yapılabilmektedir ancak bu yöntemlerdeki zenginleştirme işlemleri (örn. PCR) ilgili gen bölgesinin sayısal miktarının tespitini zorlaştırmaktadır. Ayrıca kromozom düzeyindeki değişikliklerin değerlendirilmesinde yaygın olarak kullanılan GTG bantlama gibi sitogenetik yöntemler ise optik kısıtlılıklara bağlı olarak 5-10 Mb çözünürlüğün altına düşememektedir. FISH (Fluorescence in situ hybridization) tekniği ise sadece spesifik 20-100 Kb uzunluğunda bölgeleri değerlendirebilmektedir. CGH (Comparative genomic hybridization) ve sonrasında geliştirilen arrayCGH yöntemleri ile tüm genomun kopya sayısı araştırmaları yaygın ve kolay uygulanabilir hale gelmiştir (Albertson & Pinkel, 2003).

Kallioniemi ve arkadaşları tarafından geliştirilen CGH yöntemi başta kanser dokusu olmak üzere hücre düzeyindeki yapısal değişiklikleri tespit etmeyi amaçlamıştır (Kallioniemi et al., 1992). SKY veya M-FISH gibi diğer floresan hibridizasyon yöntemleri ile benzeşse de bu yöntemde referans genomdan elde edilen floresan işaretli metafaz plağının hastadan elde edilen başka bir renkte floresan marker ile

işaretli genomik DNA ile hibridize edilmesi sonucu optik olarak hibridizasyon farklılıklarının araştırılması prensibiyle çalışır. Tekrarlayan dizilerin hibridizasyonu bozmaması amacıyla Cot-1 DNA ile tekrar bölgeleri bloklanır. Metafaz bağımlı CGH yönteminin en büyük kısıtlılıklarından biri floresan mikroskobunun optik çözünürlüğüdür. Bu sorunları aşmak adına bilim adamları arrayCGH yöntemini geliştirdiler. Prensipte CGH ile benzer özellikler taşısa da birkaç yönden ayrışırlar. Referans genom metafaz plağı yerine bakteriyel (BAC- bacterial artificial chromosomes) veya maya (YAC- yeast artificial chromosomes) klonlarından elde edilmiş genom bölgelerinin lam veya cam yüzeyde hibridizasyonu ile gerçekleştirilir. Böylelikle hem örnek hazırlama basamakları kolaylaştırılır hem de daha yüksek çözünürlükte kopya sayısı verilerin elde edilebilmesini sağlar(Albertson & Pinkel, 2003). Sonrasında cam yüzeyde sentezlenen oligo problar ile daha kısa segmentlerin hibridizasyonu ve hatta tek nükleotid değişikliklerinin tespiti olanaklı hale gelmiştir. Özellikle genom ilişkilendirme çalışmalarında (GWAS- Genome-wide association study) yaygın kullanımı bulunan oligo arrayler Hidden Markow modeli gibi algoritmaların yardımıyla CNV analizinde de kullanılmaya başlanmıştır. Hibridizasyon temelli bu yöntemler ile eş zamanlı yüz binlerce genom bölgesi analiz edilip 600 Kb hassasiyete kadar kopya sayısı varyasyonları değerlendirilebilir. Ayrıca yeni nesil DNA dizileme teknikleri (Masif paralel sekanslama) ile de genomik olarak kopya sayısı varyasyonları yüksek hassasiyetle tespit edilebilir.

2.2. Kopya Sayısı Varyantlarının Klinik Önemi

Kopya sayısı varyasyonları temel olarak iki hastalık grubu için önem taşır. Bunlardan ilki yaygın gelişimsel defektlerin eşlik edebildiği büyüme gelişme gerilikleridir. İyi bilinen DiGeorge(22q), Cri-du Chat(5p), Prader Willi (15p) vb. (Tablo 2.1) mikrodelesyon sendromlarının yanı sıra birçok mikrodelesyon sendromları array tabanlı yeni nesil teknolojilerle tanımlanmıştır (Slavotinek, 2008a). Kopya sayısı değişikliklerin tanımlı bölgeler içerisinde bulunması durumunda ilgili sendromdan sorumlu tutulan gen/genleri içeren segmentin varlığı genellikle tanımlı sendrom ile uyumlu olarak değerlendirilir. Ancak tanımlı herhangi bir sendromla kesişmeyen kopya sayısı değişiklikleri tanımlamak zorlaşır. Büyük oranda kalıtılabilir olan kopya sayısı değişiklikleri *de novo* olarak da karşımıza çıkmaktadır.

Tablo 1: Sık görülen bazı mikrodelesyon sendromları ve insidansları (Devriendt and Vermeesch, 2004).

Sendrom	**Kromozom Bölgesi**	**Delesyon İnsidansı**
Velocardiofacial/DiGeorge	22q11.2	1/4.000
Prader Willi	15q11.2-13 (Paternal)	1/ 20.000
Angelman	15q11.2-13 (Maternal)	1/ 20.000
Williams	7q11.23	1/ 20.000 – 1/ 50.000
Smith-Magenis	17q.11.2	1/ 25.000
Cri-du-chat	5p15.2	1/ 20.000 – 1/ 50.000
Wolf-Hirschhorn	4p16.3	1/ 50.000
Miller Dieker	17p13.3	11.7/1.000.000
1p36 delesyon sendromu	1p36	1/5.000

Germ line olmayan somatik kopya sayısı değişikliklerinin büyük çoğunluğu kanser hücrelerinde karşımıza çıkar. Kanser hücrelerinin neredeyse tamamında kopya sayısı değişiklikleri gözlemlenebilir. Çoğunlukla onkogenik hücrelerin artışına bağlı olarak hücre proliferasyonunun artışından sorumludurlar (Hieronymus et al., 2018). Kanser hücrelerindeki kopya sayısı değişiklikleri koromozomal veya ekstra kormozomal olarak (double minute) kendisini gösterebilirler.

Varyant sınıflamasında American College of Medical Genetics (ACMG) tarafından önerilen beşli sınıflandırma yaygın olarak kullanılır (Kearney, Thorland, Brown, Quintero-Rivera, & South, 2011). Bu sınıflandırmaya göre varyant sınıflandırılması Tablo 2-1'de gösterilmiştir. Ayrıca nokta mutasyonu gibi gen düzeyindeki varyantların sınıflandırılmasında kullanılan parametrelerin tek genin etkilendiği kopya sayısı varyantlarının sınıflandırılmasında kullanılması da önerilmiştir (Brandt et al., 2019; Richards et al., 2015). Çoğu durumda ebeveyn çalışması ile birlikte değerlendirilmesi önerilse de varyantların klinik sınıflandırması halen daha yorum zorlukları taşımaktadır.

Tablo 2 Kopya Sayısı Varyantlarının Klinik Yorumlanmasında Kullanılan ACMG önerileri (Kearney et al., 2011)

Tanım	**Açıklama**
Sık CNV / Benign CNV:	Toplumda sık görülen, klinik ilişkili olmadığı düşünülen kopya sayısı değişikliği.
Muhtemel Benign CNV	Klinik ilişkili olmadığına dair veriler bulunan fakat iyi huylu olduğu kesin kanıtlanamamış kopya sayısı değişikliği.
Önemi bilinmeyen varyant (VOUS)	Önemi bilinmeyen kopya sayısı değişikliği.
Muhtemel Patojenik:	Klinik ilişkili olduğuna dair veriler bulunan fakat patolojiye sebep olduğu kesin kanıtlanamamış kopya sayısı değişikliği.
Patojenik:	Klinik ilişkili olduğu düşünülen/patolojiye yol açtığı kesin kanıtlanan kopya sayısı değişikliği

3. Makine Öğrenme ve Yapay Zeka

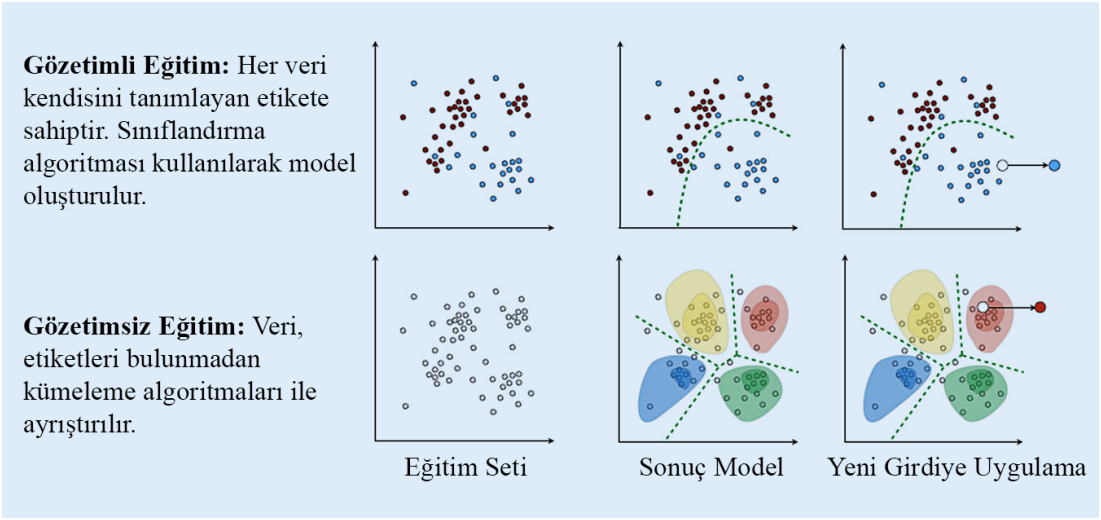
Öğrenme en basit tanımıyla deneyimlerin bilgi ve tecrübeye dönüşmesidir. Makine öğrenme veri içindeki anlamlılıkların otomatik olarak tespiti olarak tanımlanabilir. Son yıllarda büyük veri setlerinin analiz edilmesi gereken otonom araçlardan, görüntü işlemeye kadar hemen her alanda geniş kullanım alanları bulmuştur. Makine öğreneme, temel programlamadan farklı olarak programcı tarafından algılanması çok güç kompleks ilişkilerin öğrenme algoritmaları ile uygulanabilir yazılımlar haline dönüştürülmesidir (Shalev-Shwartz & Ben-David, n.d.). Çoğu durumda bu veriler insan yetenekleri ile değerlendirme yapmanın çok ötesinde aşırı kompleks ve büyük yapıdadırlar. Büyük veri setlerinin yaygın olarak kullanıldığı sağlık ve tıp alanında da birçok çalışmada makine öğrenme teknikleri kullanılmaya başlamıştır. Örneğin radyolojik görüntülerin işlenmesinde, kanser dokularının değerlendirilmesinde ve multifaktöriyel kalıtım gösteren kompleks hastalıklar için risk faktörlerinin araştırılması gibi birçok alanda kullanılmıştır (Ainscough et al., 2018a; Isakov, Dotan, & Ben-Shachar, 2017).

3.1. Makine Öğrenme Çeşitleri

Gözetimli ve Gözetimsiz (Supervised/Unsupervised) öğrenme: öğrenim işlemi veri ile etkileşimin şekline göre değişiklik gösterir. Gözetimli öğrenme önceden etiketlenmiş ve beklenilen özellikler dışarıdan eklenmiş veriler içerisinde, veri ile ilişkili fonksiyon arayan sistemleri tanımlar. Bir eğitim seti ve bir de test seti oluşturulur. Eğitim seti ile oluşturulan öğrenme sonrası test seti ile sistemin değerlendirilmesi yapılır. Gözetimsiz öğrenmede ise veri etiketleri olmadan veriler kendi içerisindeki bilinmeyen fonksiyonun tespiti için algoritmalar kullanılır (Şekil 1). Hangi yöntemin kullanılacağının seçimi ise mevcut veriye ve ulaşılmak istenilen sonuca göre karar verilmesi gereken bir sorudur. Gözetimli öğrenme önceden verilerin tasnif ve etiketlenmesinin gerektirdiği için daha zor olarak değerlendirilebilir. Ancak gözetimsiz öğrenme ise kolay uygulanabilir olmasına karşın sonuç elde etmede her zaman aynı başarıya sahip değildir.

Öğreticinin, sisteme beklenen sonucu tam olarak söyleyemediği ancak sistemin ürettiği sonuç için "doğru/yanlış" şeklinde fikir beyan ettiği öğrenme şekline de takviyeli (reinforcement) öğrenme adı verilir.

Aktif veya Pasif öğrenme: Aktif öğrenmede sistem etkileşim kuracağı bir yapıda parametreleri değerlendirirken (örneğin otonom araçlar, sensör verileri ile analiz edip araçtaki değişkenleri değiştirebilirler) pasif öğrenme etkileşim kurmadığı hazır veriler üzerinden analiz yapar (örneğin radyolojik görüntülerin değerlendirilmesi). Aynı zamanda online öğrenme ile sistem öğrenme sonuçlarını direkt olarak gözlemleyebilmektedir.



Şekil 1 Gözetimli ve Gözetimsiz Öğrenme (Langs et al., 2018)

3.1.1. Gözetimli Öğrenme

Bu sistemde eğitim için kullanılacak veriler etiketlenmiş olarak bulunurlar. Diğer bir ifadeyle eğitimde kullanılacak veri ve veriye ait sınıflar (kategoriler/etiketler) analiz öncesi önceden bilinir. Sistem öğrenir ve test verilerini bu öğrendikleriyle yorumlar. El yazısı tanımlama, görsel sınıflandırma gibi birçok alanda yaygın olarak kullanılırlar. Etiketlendirilmiş geniş veri setlerinin varlığı bu sistemlerin çalışması için elzemdir. Sınıflandırma ve regresyon algoritmalar şeklinde iki grup olarak da tanımlanabilirler (Şekil 2). Sıklıkla kullanılan algoritmalar:

En Yakın Komşuluk (k-Nearest Neighbors (KNN))

Her bir noktanın en yakın komşularının basit çoğunluk oyu (yakınlık) ile hesaplanması kullanılan sınıflandırmadır. Veriler arası mesafeler ile ayrım sağlanılır. Bu algoritma kolay uygulanır, gürültülü eğitim verisine (noisy training data) dirençli ve büyük eğitim verilerinde oldukça etkilidir.

Destek Vektör Makineleri (Support Vector Machine (SVM))

Temel olarak veri setinde en uzak noktalar arası destek fonksiyonu bulmaya ve bu destek fonksiyonunun sınırlarını belirler.

Karar Ağaçları (Decision Trees (DTs))

Veriler ağaç dallanmasına benzer şekilde gruplandırılarak sınıflandırma yapılır. Yapı düğüm ve yapraklardan oluşur. Karar ormanı-forest algoritması ise rastgele üretilen başlangıç nodu ile sonuç arasındaki ilişkiyi bulmaktır. Karar ormanı-jungle algoritması ise yönlendirilmiş akrilik grafiklerle tanımlayıcı sınıflandırma algoritması oluştururlar (Shotton et al., 2013).

Doğrusal Regresyon (Linear Regression)

Sayısal verilerin değerlendirilmesinde kullanılır, sayılar arasındaki ilişki doğrusal olarak modellenir.

Lojistik Regresyon (Logistic Regression):

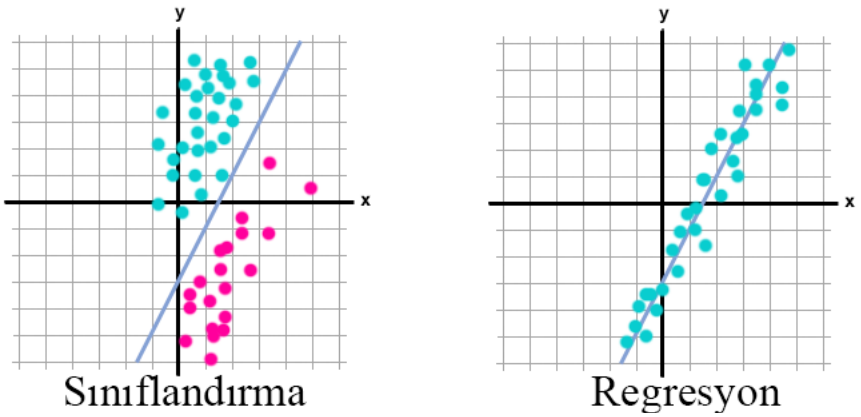
Lojistik regresyon "evet/hayır" gibi iki cevabın olduğu sistemlerin oluşturulmasında kullanılır. Sonucun hangi seçeneğe yakın olduğu tahmin edilmeye çalışılır.

Naif Bayes (Naive Bayes):

Bir seçeneğin olma ihtimalinin en yüksek olma koşulunu hesaplayan bir tür sınıflandırma algoritmasıdır. Eğitim verileri için olasılık hesaplamaları yapılır ve sisteme sunulan yeni verinin hesaplanmış olasılık değerine göre sınıflandırılması sağlanır.

Yapay Sinir Ağları (Artificial Neural Network (ANN))

Biyolojik sinir ağlarına benzer bir yapıyı matematiksel olarak modelleyen algoritmalardır. Temel sinir ağı hücresi girdiler, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktıdan oluşur. Sinir ağı sistemi ise çok sayıda sinir hücrelerinin birleşmesi ile oluşur.



Şekil 2 Sınıflandırma ve Regresyon algoritmalarının grafiksel gösterimi

3.1.2. Gözetimsiz Öğrenme

Gözetimli öğrenmeden farklı olarak verilerde herhangi bir etiketlendirme bulunmaz. Gözetimsiz öğrenme veriler arasında bağlantılar bulup birbirine yakın verilerin kümelenmesi ile oluşturulur. Sık kullanılan gözetimsiz öğrenme modelleri:

Kümeleme (Clustering)

Her bir veri benzerlik durumuna göre küme adı verilen gruplara ayrılır. Kümelerin birbirlerinden ayrılma oranları algoritmanın başarısını arttırır.

Birliktelik Kuralı (Association Rule Mining)

Kümelemeden farklı olarak bağımsız gibi görünen veriler içerisindeki ilişki bulunmaya çalışılır. Apriori, Eclat, FP-growth gibi algoritmalar örnek verilebilir.

Boyut Azaltma (Dimensionality Reduction)

Analiz için kullanılacak verideki özniteliklerin azaltılması yoluyla birbirleri arasındaki ilişkinin tanımlanması yöntemidir.

Gereç ve Yöntem

Makine öğrenme sisteminin denenmesi için açık veri setleri kullanılmıştır. Çalışma öncelikle çalışma tasarımının denenmesi amacıyla 11989 varyantın bulunduğu ISCA (International Standards for Cytogenomic Arrays) konsorsiyumunun verileri kullanılarak pilot analiz gerçekleştirilmiştir (Kaminsky et al., 2011; Miller et al., 2010). Veriler dbVar veri tabanından CSV dosyası olarak indirilmiştir ("nstd101 - ClinGen - dbVar Study - NCBI," n.d.).

Pilot çalışma sonuçları sonucu çalışmaya ClinVar veritabanına girilmiş 63156 varyantın bulunduğu nstd102 (Clinical Structural Variants) verileri kullanılmıştır ("nstd102 - Clinical Structural Variants - dbVar Study - NCBI," n.d.).

Veri setlerindeki veriler klinik özelliklerine göre sınıflandırılmış, ayrıca hangi kromozomda bulundukları, genomik lokasyonları ve kayıp kazanç durumu bilgileri kullanılmıştır.

Klinik tanımlama verilerinden

Patojenik. (Pathogenic)

Muhtemel Patojenik (Likely Pathogenic)

Etkisi bilinmeyen varyant (Unknown Significance)

Muhtemel Benign (Likeliy Benign)

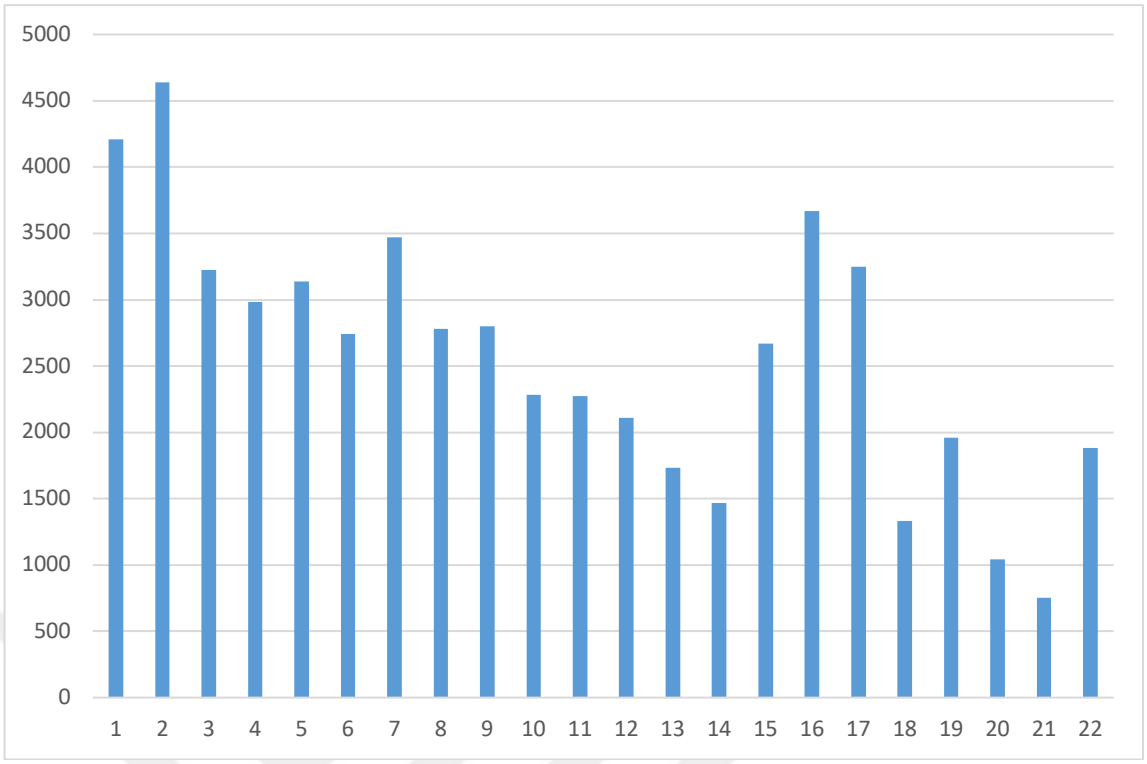
Benign

etiketleri kullanılmıştır. Veri setinde bulunan diğer etiketler ya mevcut 5 etiketten birine dönüştürülmüş veya dışlanmıştır. Ayrıca veri serisindeki sadece insan genomu GRCh38 (hg38) versiyonu veriler kullanılmıştır.

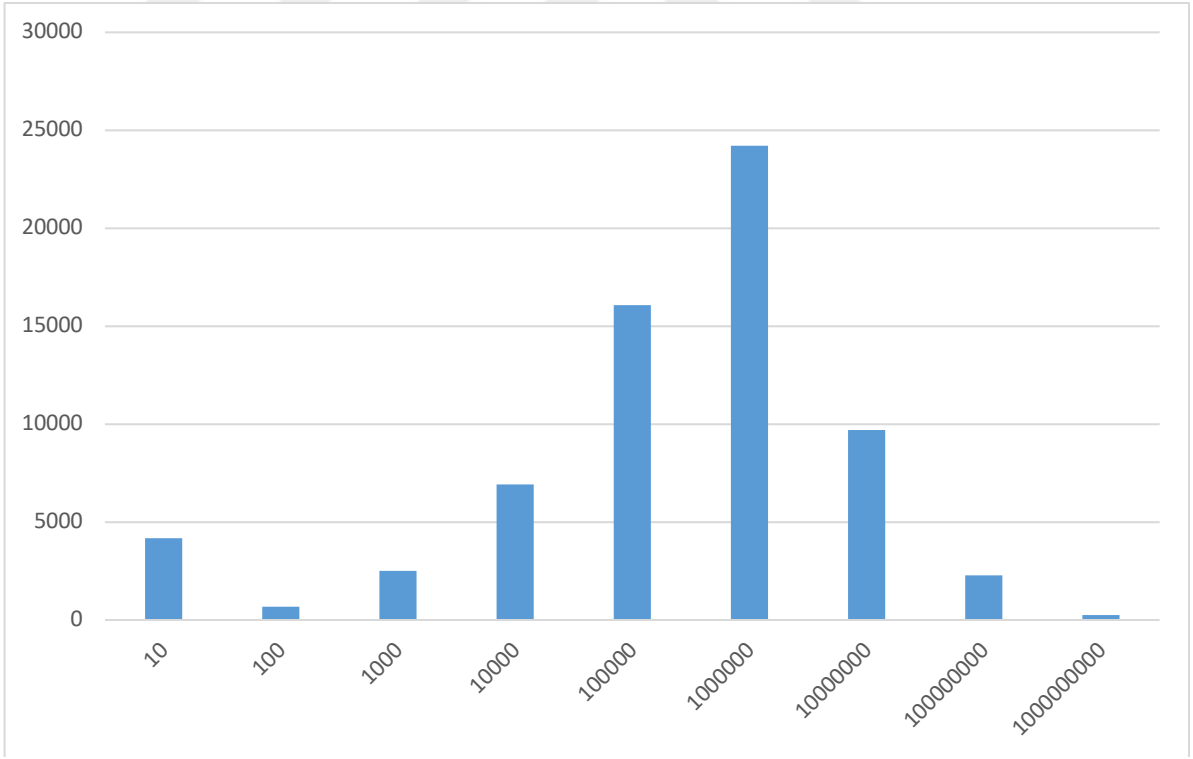
Tablo 3 Veri seti etiketlerinin dağılımı

	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

ClinVar veri setinden uygunsuz örnekler çıkarıldıktan sonra 24110 Benign, 20602 önemi bilinmeyen, 14855 patojenik, 4834 muhtemel benign ve 2402 muhtemel patojenik varyant ile analize başlanmıştır (Tablo 3). Kullanılan veri setinde cinsiyet kromozomları dışlanmış olup, kayıp ve kazançların kromozomlar arası dağılımı Grafik 1'de gösterilmiştir. Ayrıca veri setinde 35494 kopya sayısı kaybı ve 30774 kopya sayısı kazancı bulunmaktadır. CNV'ler boyutlarına göre değerlendirildiğinde ortanca değer 132406 bç iken ortalama değer 2046347 bç (Grafik 2) olarak bulunmuştur.



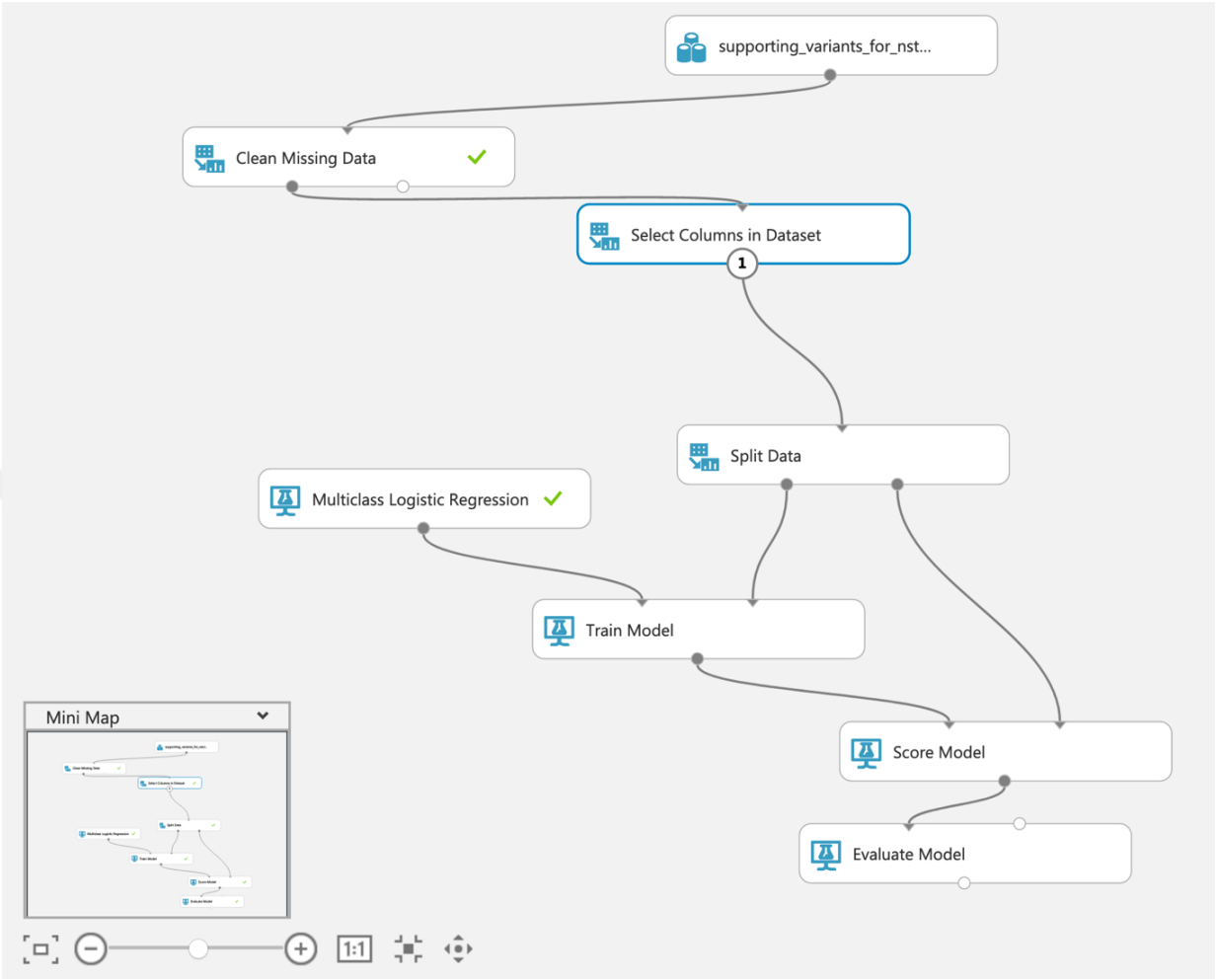
Grafik 1 Verilerin kromozomlara göre dağılımı



Grafik 2 CNV boyutlarının dağılımı (bç)

Veriler Microsoft Azure Machine Learning Studio kullanılarak analiz edilmiştir ("Microsoft Azure Machine Learning Studio (classic)," n.d.) (Şekil 3). 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.



Şekil 3 Microsoft Azure Machine Learning Studio (classic) Örnek Çalışma Alanı

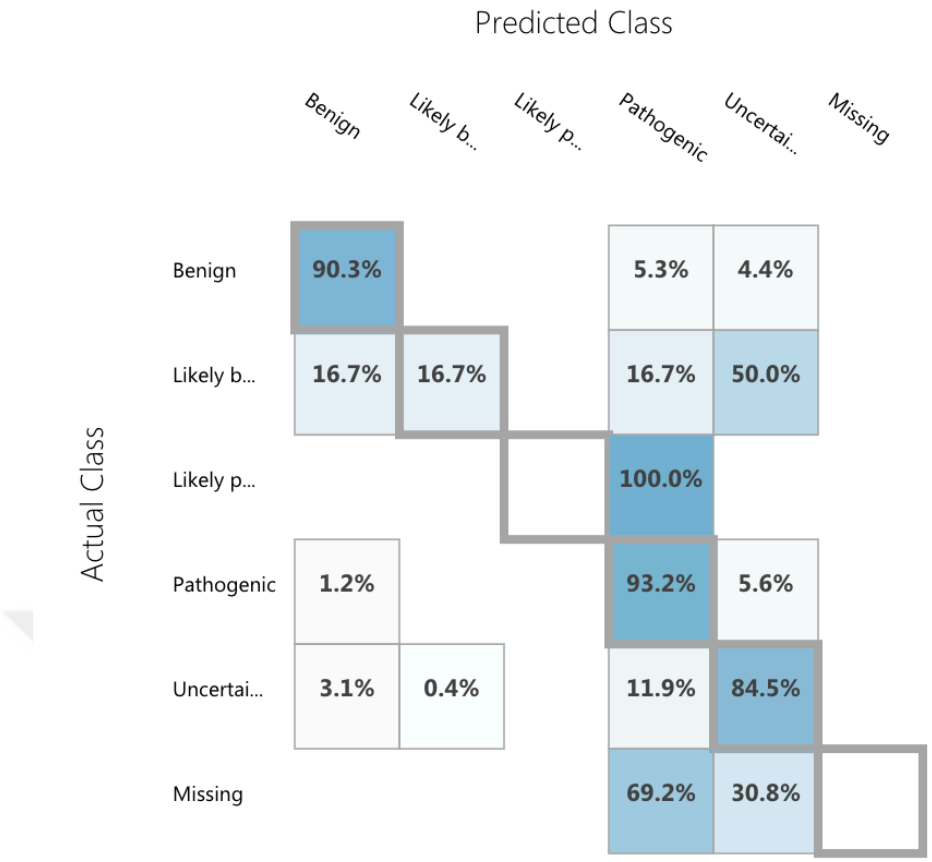
Bulgular

ISCA verileri ile yapılan çalışma sonucunda toplamda %89.241 ve ortalamada %96.4137 doğruluğa ulaşılmıştır. 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 (Şekil 4). Bu öncül pilot çalışma verileri 07-11 Kasım 2018 tarihinde Antalya'da Düzenlenen "Uluslararası Katılımlı 13. Ulusal Tıbbi Genetik Kongresi"'inde sözlü bildiri olarak sunulmuştur (Parıltay & Ece Solmaz, 2018). 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 4).

Tablo 4 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

Pilot çalışmada veri etiketleri kendi içerisinde değerlendirilmemiştir ayrıca alt grupların 5 ten fazla olmasına rağmen modelin ortalama başarısı %96'nın üzerinde olarak değerlendirilmiştir.

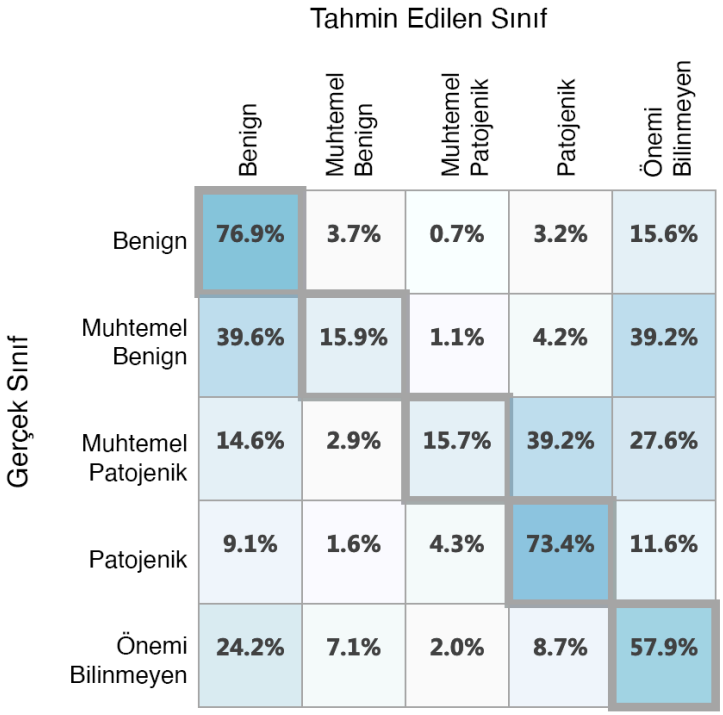


Şekil 4 Pilot çalışma sonrası elde edilen çok sınıflı karar ormanı-forest

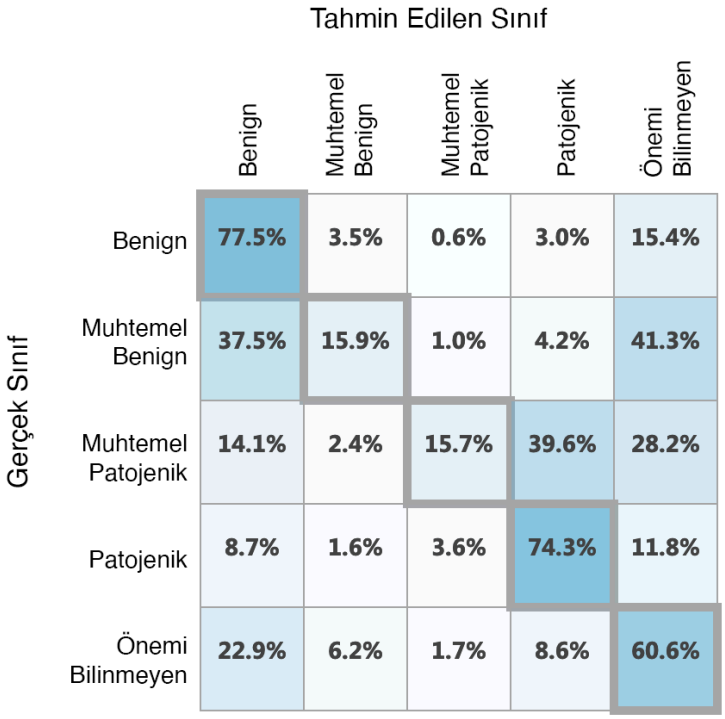
Pilot çalışma sonrası clinVar veri seti için yapılan analiz de farklı yöntemler birbirleri ile karşılaştırılmış ve bunlar arasında çoklu karar ağacı-forest (32) yaklaşık %86 ortalama doğruluk ile en başarılı yöntem olarak değerlendirilmiştir (Tablo 5). Yine bu analizler sonrası elde edilen doğru tahminlerin dağılımı Şekil 5-10'da gösterilmiştir.

Tablo 5 Eğitim verilerinin farklı algoritmalarla analizleri

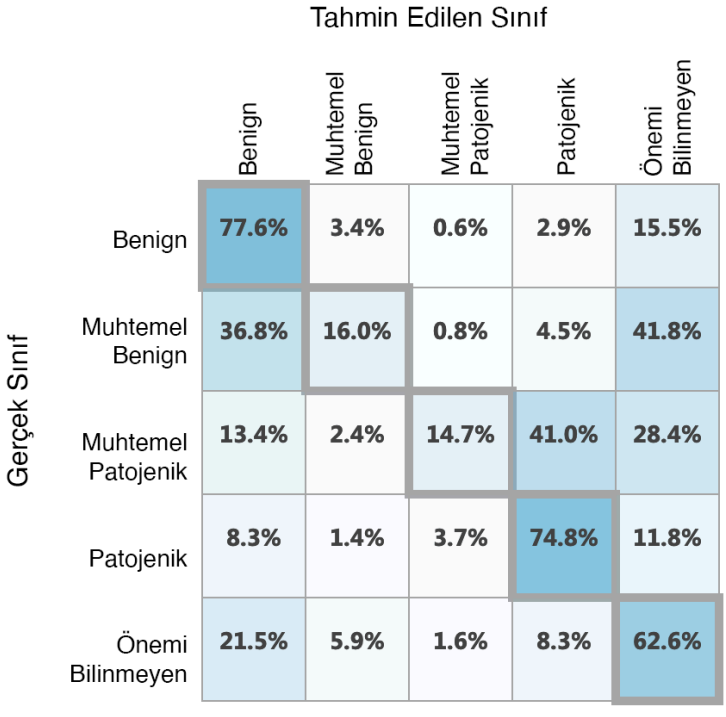
	Çok Sınıflı Karar Ağacı-Forest (8 dal)	**Çok Sınıflı Karar Ağacı-Forest (16 dal)**	**Çok Sınıflı Karar Ağacı-Forest (32 dal)**	**Çok Sınıflı Karar Ağacı-Jungle**	**Çok Sınıflı Lojistik Regresyon**	**Çok Sınıflı Sinir Ağı**
Toplam Doğruluk	0.637443	0.650167	0.657602	0.52188	0.432813	0.480415
Ortalama 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



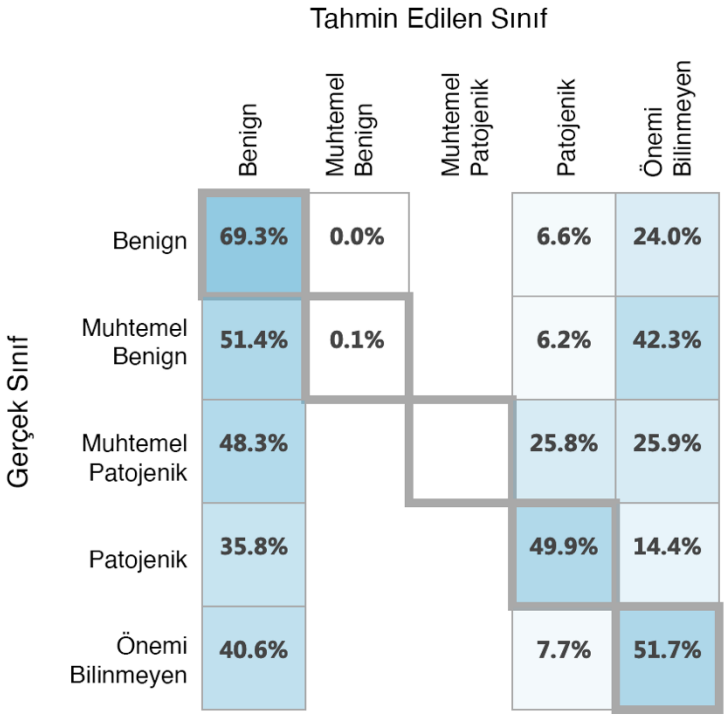
Şekil 5 Çok Sınıflı karar ağacı-forest (8 dal) doğru tahminlerin dağılımı



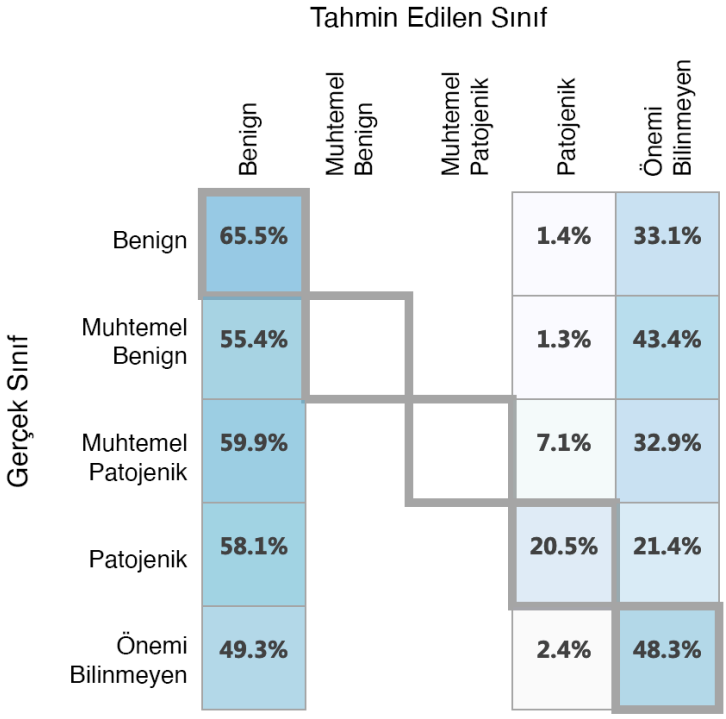
Şekil 6 Çok sınıflı karar ağacı-forest (16 dal) doğru tahminlerin dağılımı



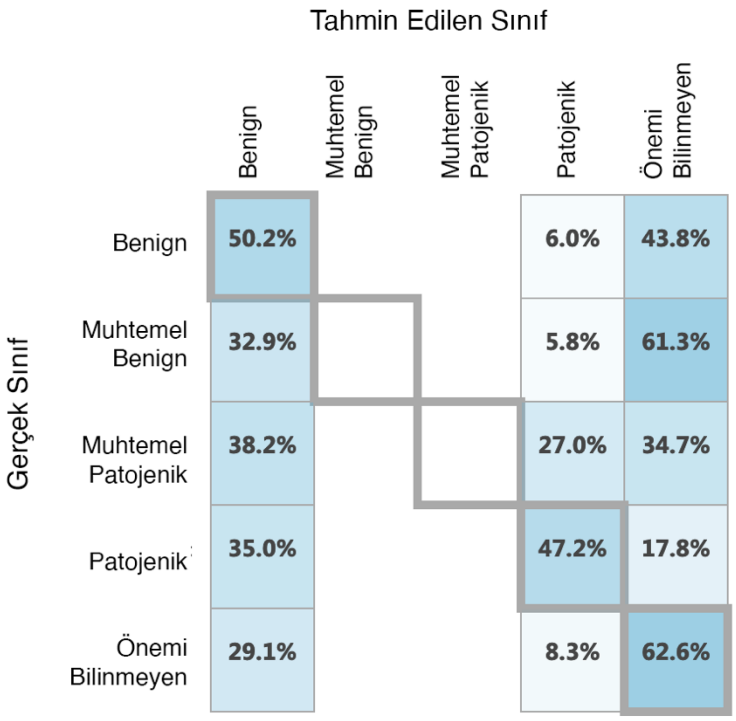
Şekil 7 Çok sınıflı karar ağacı-forest (32 dal) doğru tahminlerin dağılımı



Şekil 8 Çok sınıflı karar ağacı-jungle doğru tahminlerin dağılımı



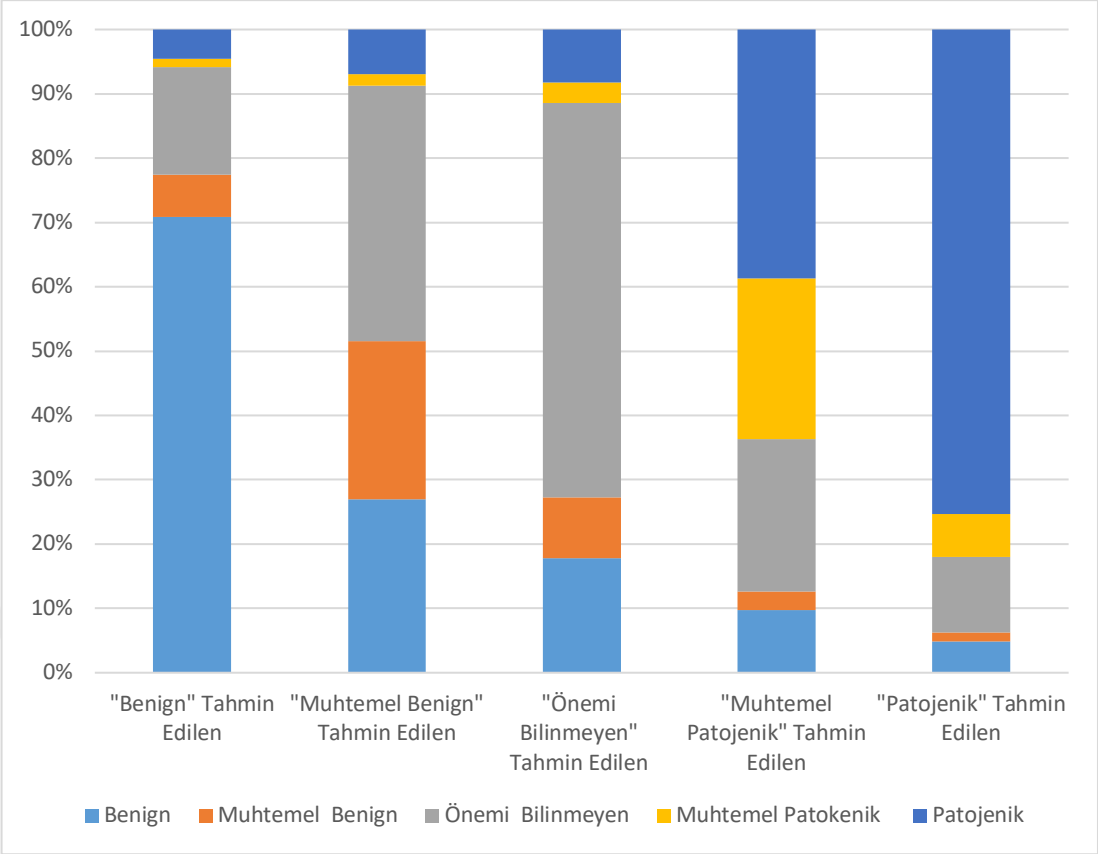
Şekil 9 Çok sınıflı lojistik regresyon doğru tahminlerin dağılımı



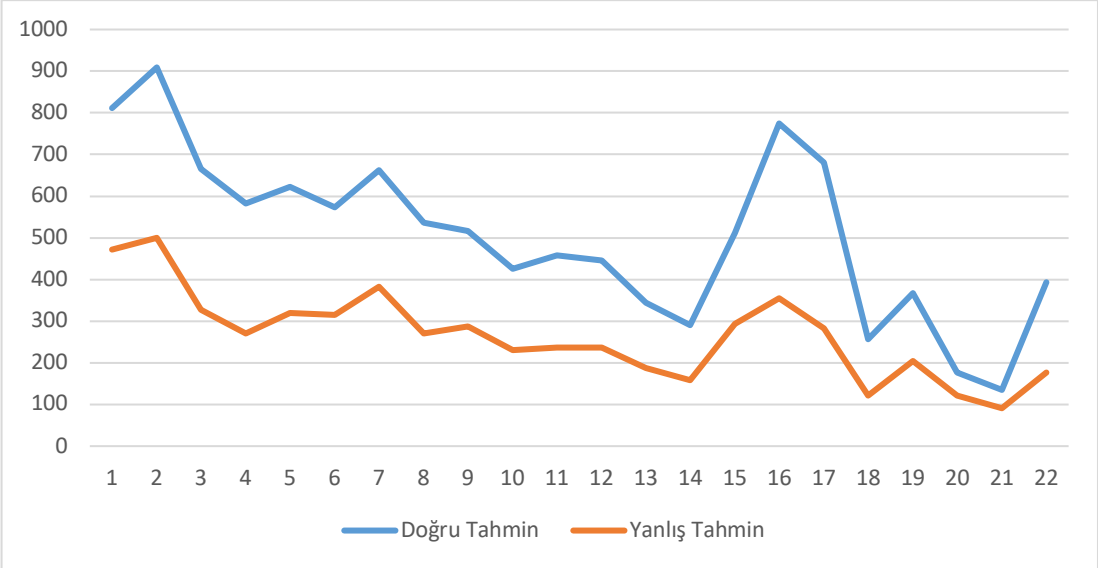
Şekil 10 Çok sınıflı sinir ağı doğru tahminlerin dağılımı

Tablo 6 Çok Sınıflı Karar Ağacı-Forest (32 dal) tahminlerin dağılımı

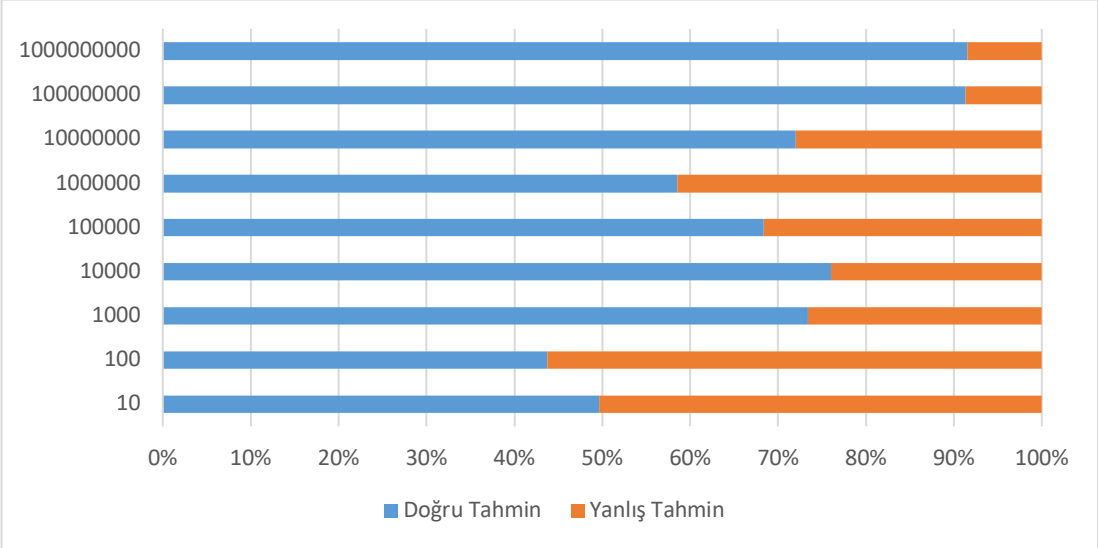
Sınıf	**"Benign" Tahmin Edilen**	**"Muhtemel Benign" Tahmin Edilen**	**"Muhtemel Patojenik" Tahmin Edilen**	**"Patojenik" Tahmin Edilen**	**"Önemi Bilinmeyen" Tahmin Edilen**	**Ortalama Logaritmik Kayıp**	**Keskinlik**	**Geri Çağırma**
Benign	5656	249	41	212	1128	0.935331263619313	0.709038485646233	0.776283283008509
Muhtemel Benign	523	227	12	64	594	828.199.462.769.923	0.245670995670996	0.159859154929577
Muhtemel Patojenik	96	17	105	293	203	115.228.531.472.405	0.249406175771971	0.147058823529412
Patojenik	367	64	163	3310	521	167.518.106.422.244	0.753642987249545	0.748022598870057
Önemi Bilinmeyen	1335	367	100	513	3881	152.805.710.223.528	0.613402876560771	0.626371852808263



Grafik 3 Gerçek etiketler ile tahmin edilen etiketlerin dağılımı



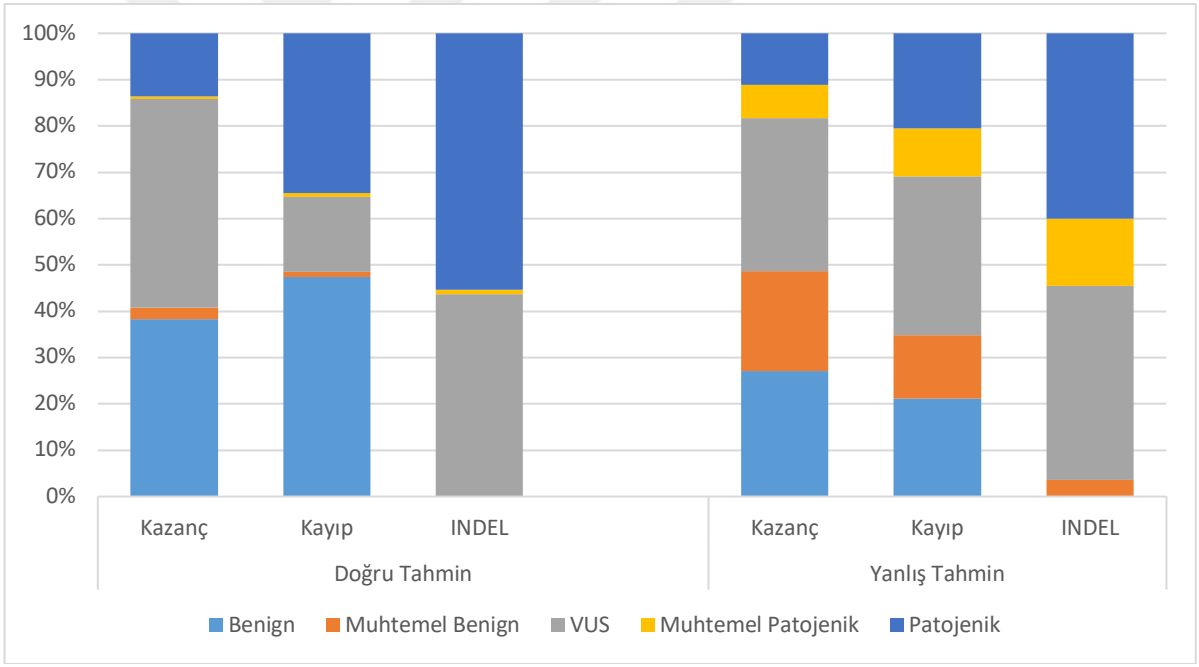
Grafik 4 Çok Sınıflı Karar Ağacı-Forest (32 dal) Sonuçlarının kromozomlara göre dağılımı,



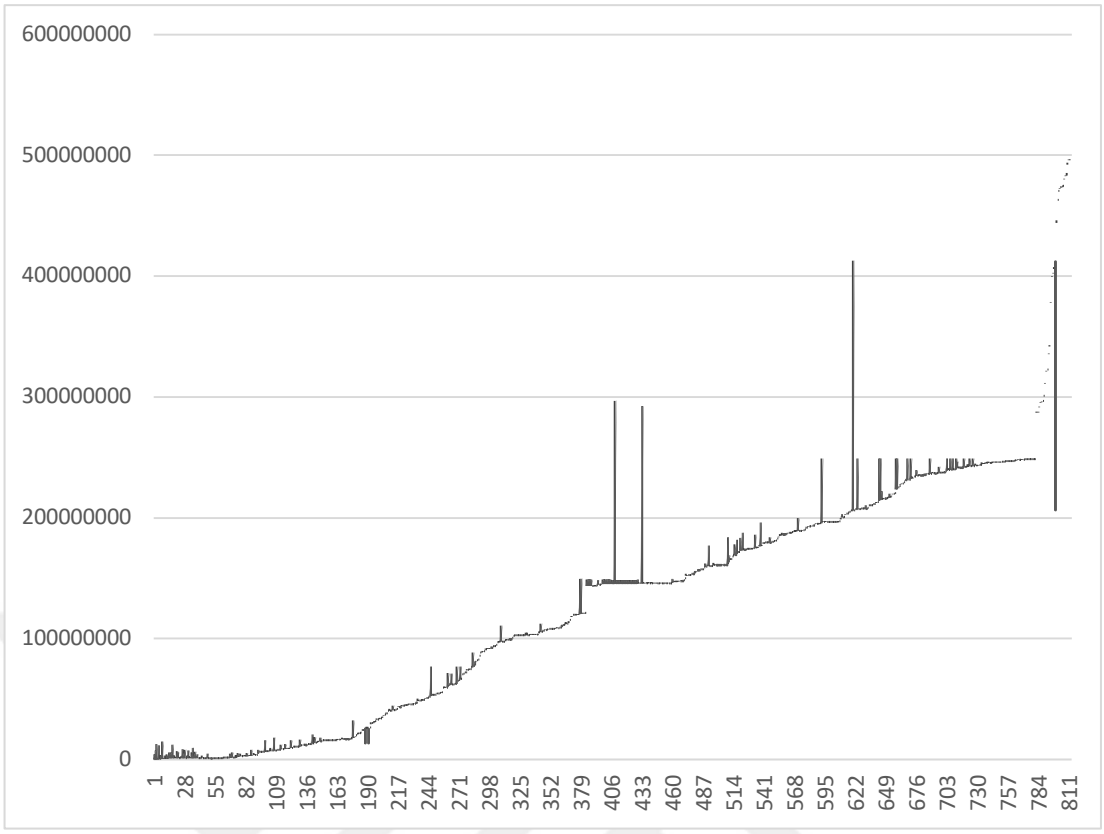
Grafik 5 Çok Sınıflı Karar Ağacı-Forest (32 dal) Sonuçların CNV boyutuna göre karşılaştırılması

Tablo 7 Tahmin etiketlerine göre kayıp/kazanç durumunu dağılımı

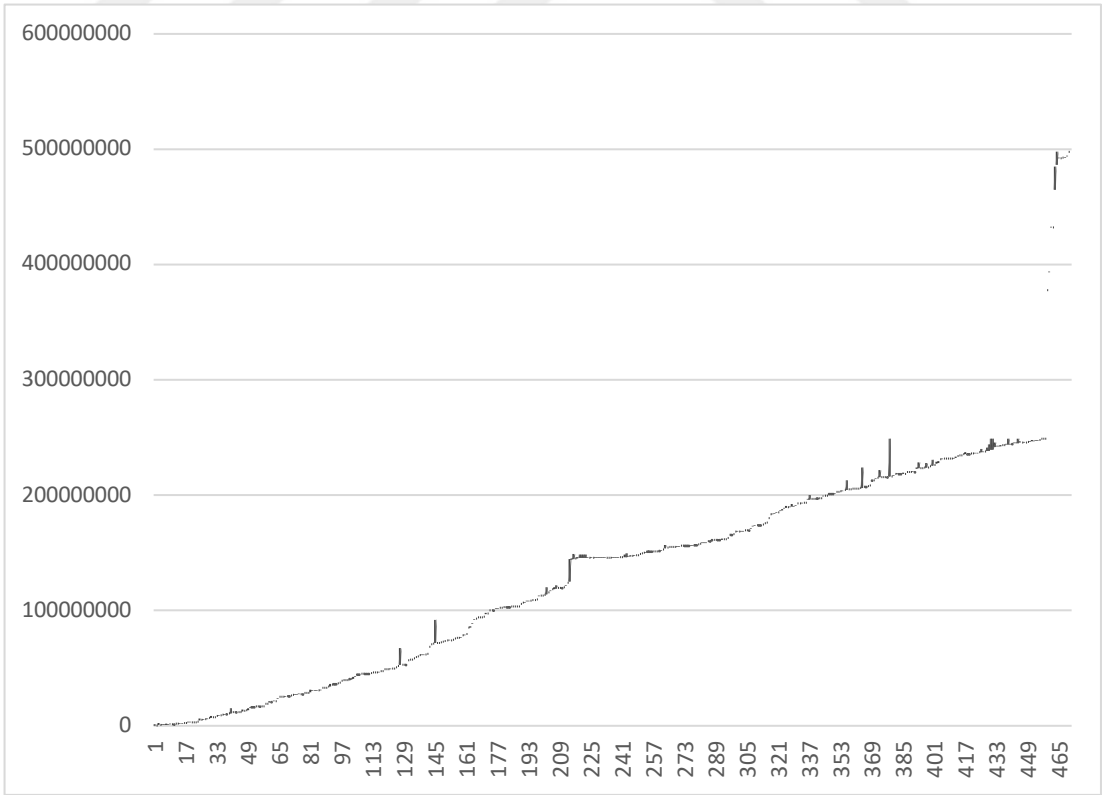
	Doğru Tahmin				**Yanlış Tahmin**		
	Kazanç	Kayıp	INDEL		Kazanç	Kayıp	INDEL
Benign	2295	3360	0		876	754	0
Muhtemel Benign	149	78	0		701	490	2
VUS	2693	1147	41		1070	1222	23
Muhtemel Patojenik	37	67	1		233	368	8
Patojenik	815	2443	52		360	733	22
Toplam	5989	7095	94		3240	3567	55



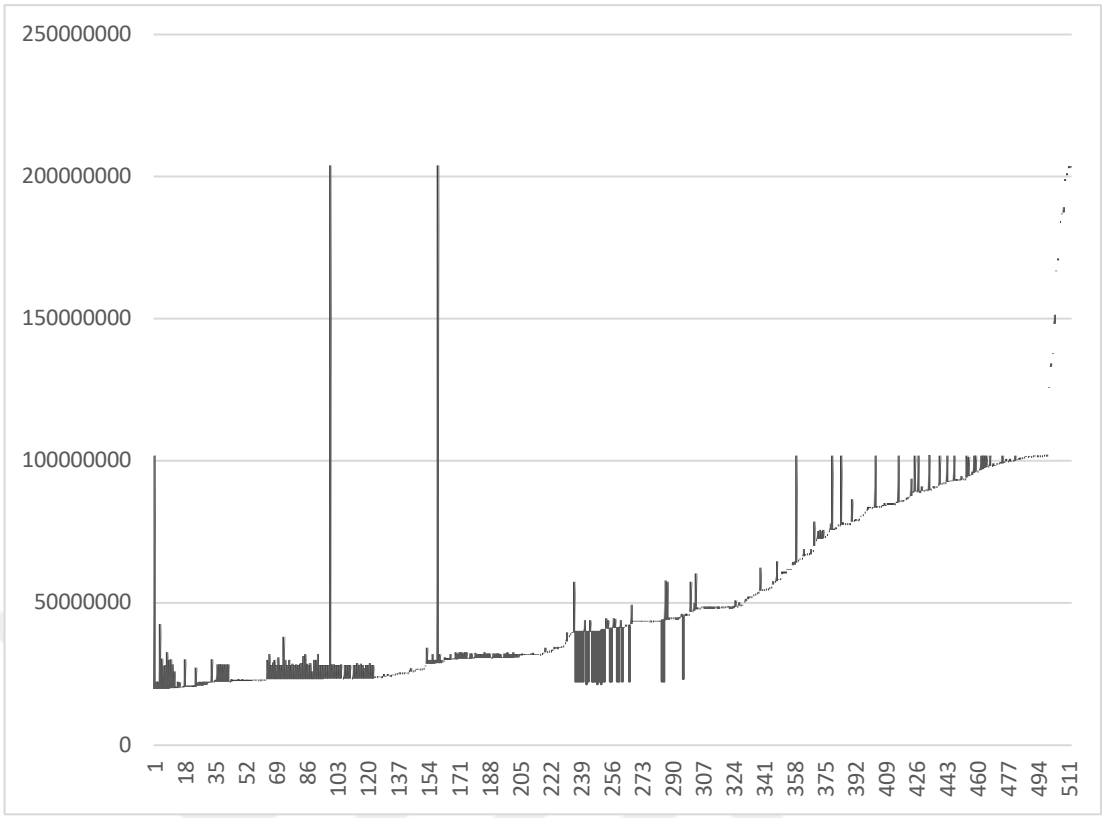
Grafik 6 Tahmin etiketlerine göre kayıp/kazanç durumunu dağılım grafiği



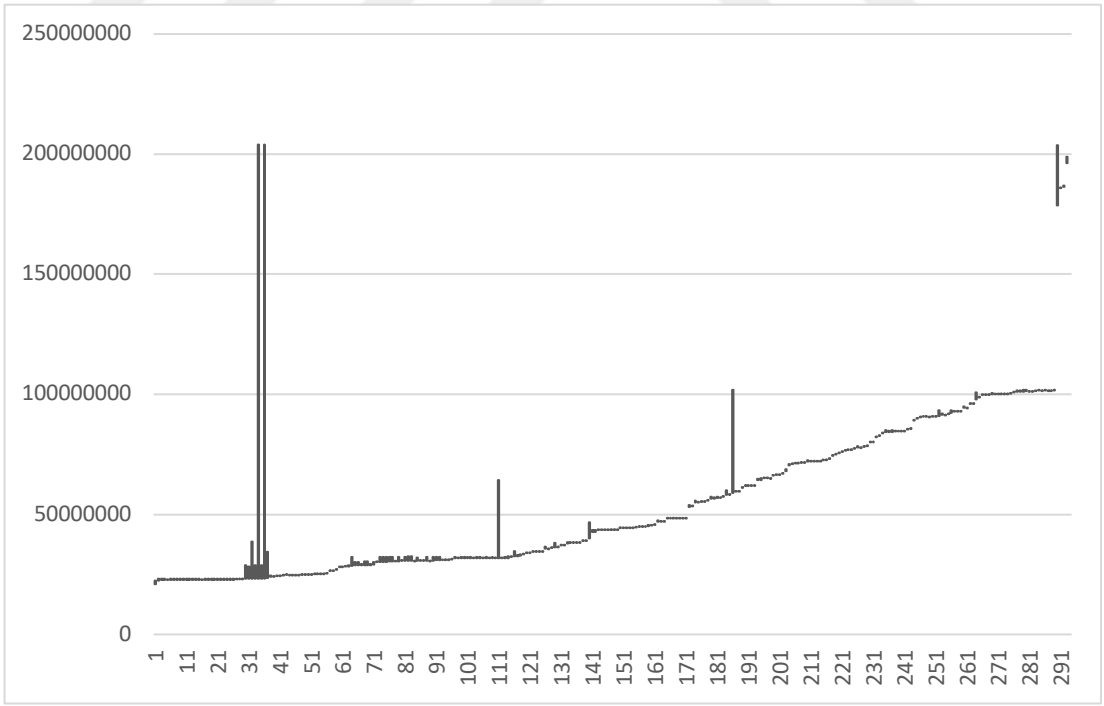
Grafik 7 1. Kromozom doğru tahminlerin kromozomal lokasyonları



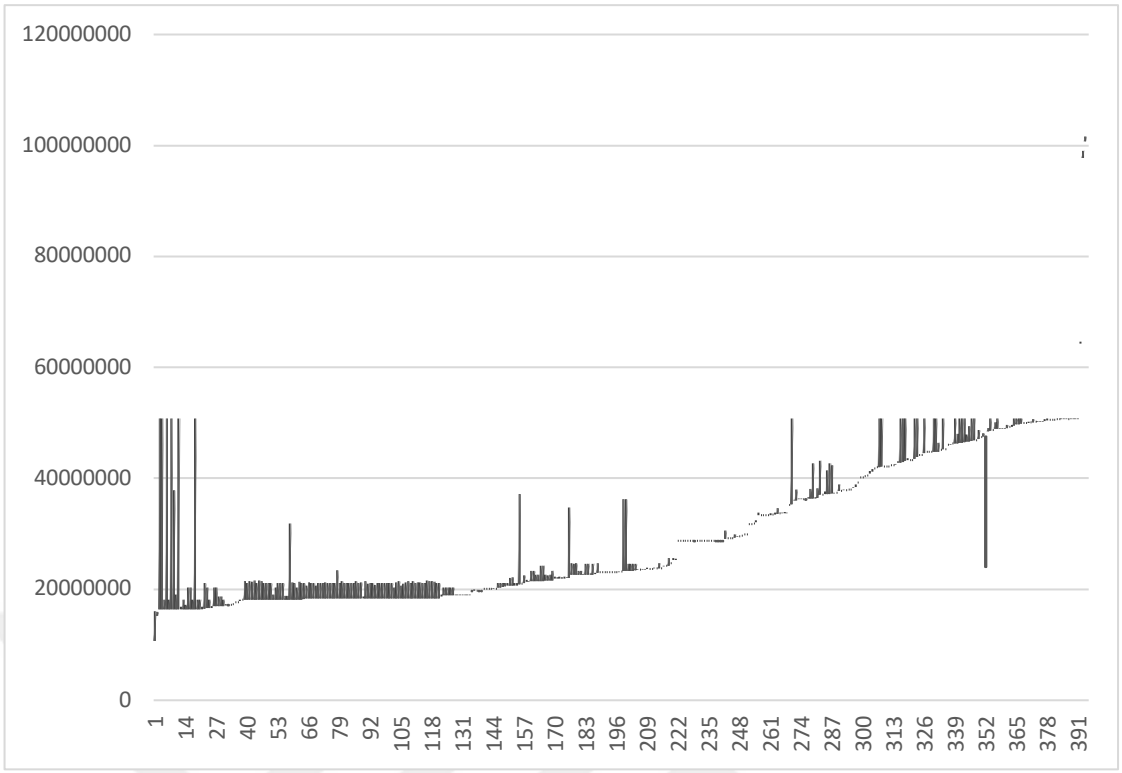
Grafik 8 1. Kromozom yanlış tahminlerin kromozomal lokasyonları



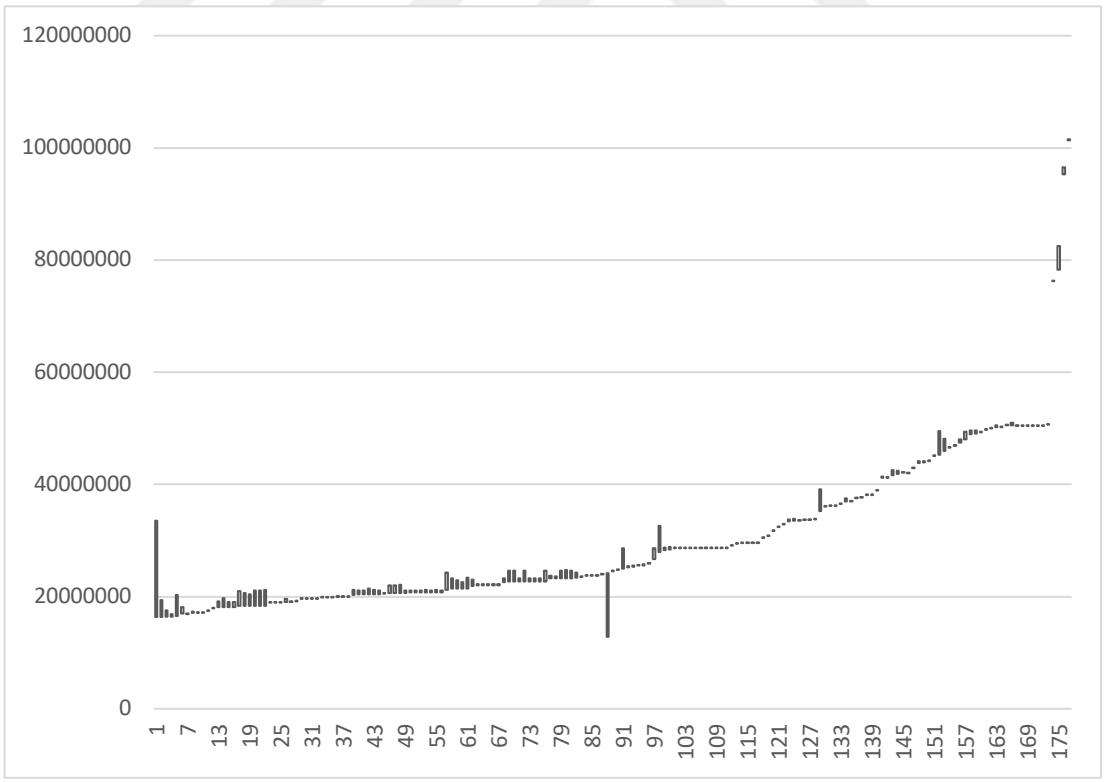
Grafik 9 15. Kromozom doğru tahminlerin kromozomal lokasyonları



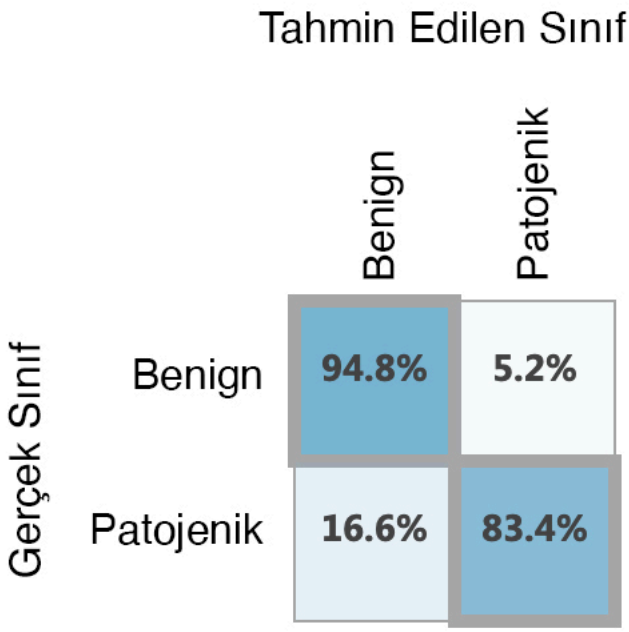
Grafik 10 15. Kromozom yanlış tahminlerin kromozomal lokasyonları



Grafik 11 22. Kromozom doğru tahminlerin kromozomal lokasyonları

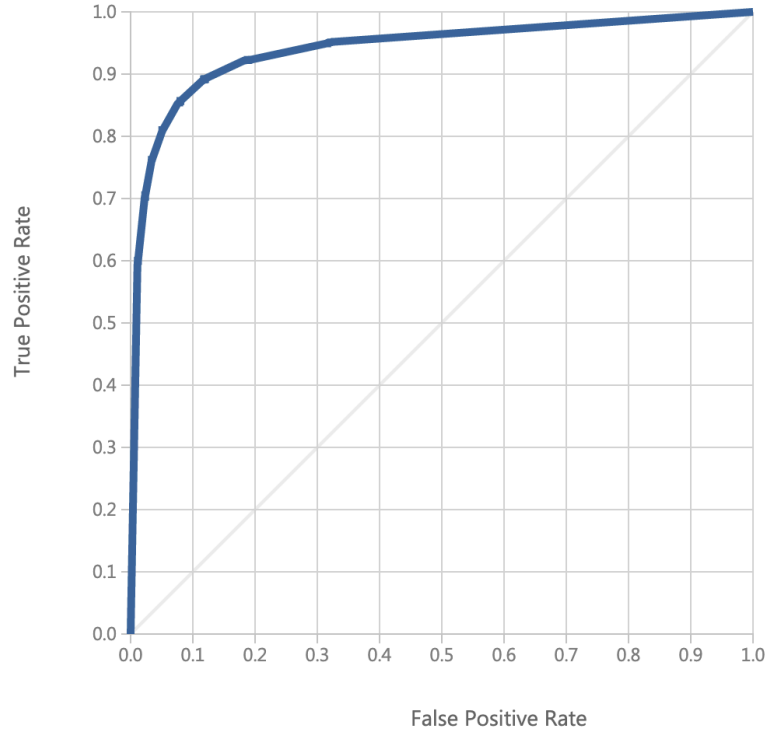


Grafik 12 22. Kromozom yanlış tahminlerin kromozomal lokasyonları

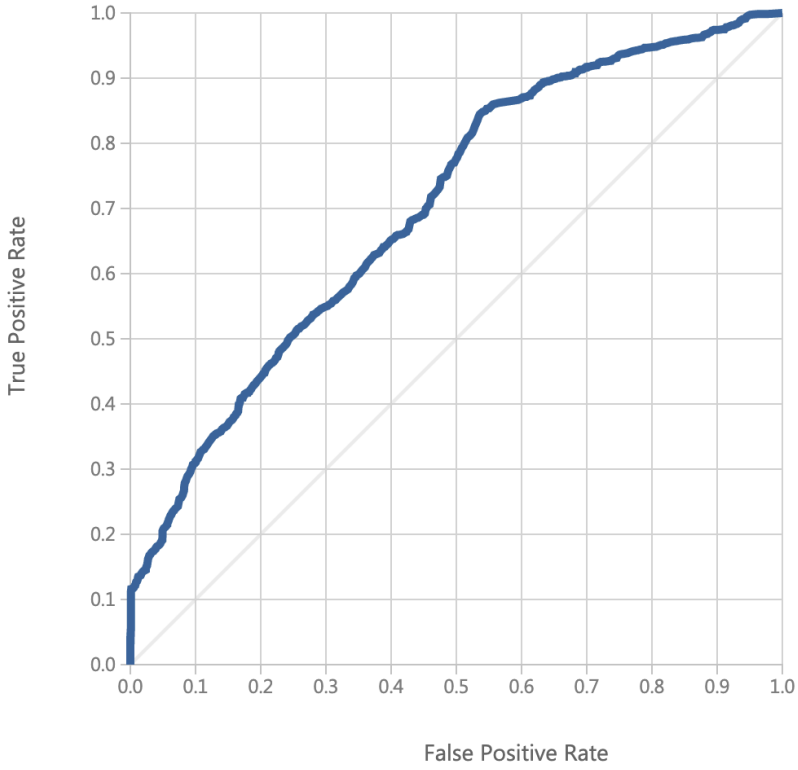


Şekil 11 İki sınıflı örneklem çoklu sınıf karar ağacı-forest(32 dal)

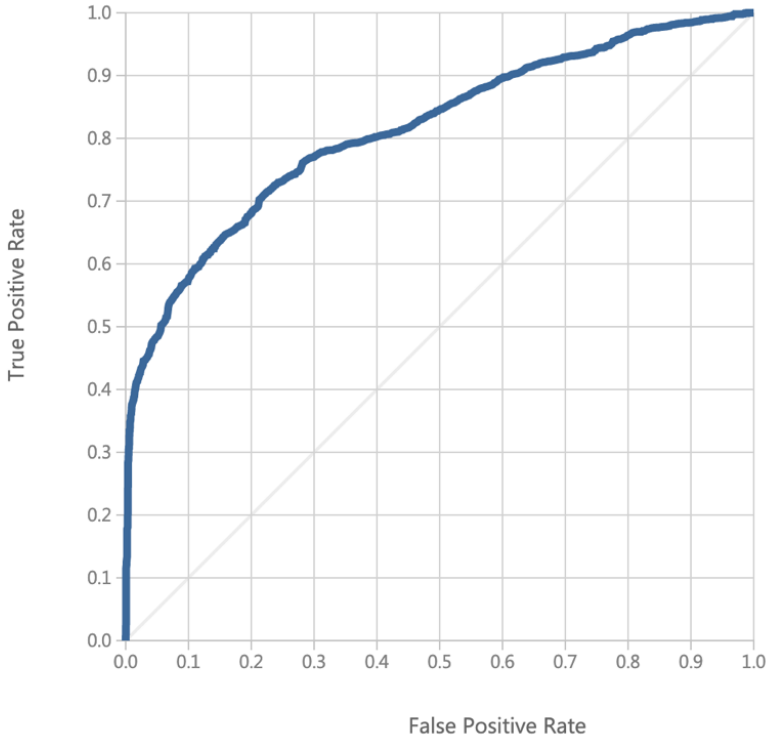
Örneklemdeki benign ve patojenik değerler dışındaki veriler dışlandıktan sonra analiz tekrarlanmış ve çok sınıflı karar ağacı-forest (32 dal) ile ortalama ve toplamda 0.905176 keskinliğe ulaşılmıştır (Şekil 11). İki sınıflı karar ağacı-forest algoritması 0.897 keskinlikte sonuç vermiştir (Şekil 12). İki sınıflı sinir ağı analizinde 0,730 keskinliğe ulaşılırken iki sınıflı destek vektör makinesi (SVM) ile 0.669 keskinliğe ulaşılmıştır (Şekil 13-14).



Şekil 12 İki sınıflı örneklemin iki sınıflı karar ağacı ile doğruluk analizi



Şekil 13 İki sınıflı örneklemin sinir ağı ile doğruluk analizi



Şekil 14 İki sınıflı örneklemin destek vektör makinesi (SVM) ile doğruluk analizi

Tartışma

Kopya sayısı varyantlarının klinik olarak değerlendirilmesi hem klinisyen hem de analitik olarak zorluklar içermektedir. Zorlukların başında kopya sayısı tespiti için kullanılacak yöntemlerin kısıtlılıkları gelmektedir. Modern array yöntemleri CNV tespitini kolaylaştırmış olsa da pozisyonel veya oryantasyonel bilgi içermemesi halen en büyük eksikliklerindendir (Albertson & Pinkel, 2003). Özellikle yeni nesil dizileme yöntemleri ile beraber yaygınlaşan exom ve genom sekanslama verileri de kopya sayısı varyasyonlarının tespiti için kullanılmaya başlanmıştır. Bilinen gen veya gen parçaları içeren bölgelerin kayıp veya kazançlarını değerlendirmek gen/dozaj ilişkisine dayalı olarak göreceli kolaylık gösterir. Haplo-yetmezlik skorları fonksiyon kayıp/kazanım (LOF/GOF) mutasyonları da dahil olmak üzere kopya sayısı varyantlarının değerlendirilmesinde kullanılırlar (Steinberg, Honti, Meader, & Webber, 2015). Ancak gen içermeyen düzenleyici genom bölgelerinin kayıp/kazançlarını değerlendirmek zorluklar taşımaktadır. Bu tarz gen dışı kopya sayısı varyantları tüm analizlerin neredeyse %95'ini oluşturmaktadır (Spielmann & Klopocki, 2013). Varyantların değerlendirilmesine yönelik olarak çeşitli öneri kararları bildirilmiştir (Kearney et al., 2011). Hatta yeni yaklaşımlarda CNV'lerin diğer nokta mutasyonları veya varyantlar gibi değerlendirilip skorlama ile klinik etkinin değerlendirilmesi önerilmiştir (Brandt et al., 2019) . Varyantların değerlendirilmesinde ilk basamak çoğunlukla ilgili varyantın veri tabanlarında taranması ile başlar (Koolen et al., 2009). *De novo* varyantların varlığı durumunda aile çalışması ve segregasyon analizleri önemli yer tutar. Tüm bunlara rağmen süreç çoğunlukla konunun uzmanları tarafından gözle değerlendirilen ve klinik tabloya göre yorumlanan basamakları içeren zaman alıcı ve maliyetli yapıdadır.

Son yıllarda gelişen teknoloji ve uygulanmasını kolaylaştırılan algoritmalar sayesinde makine öğrenme teknikleri günlük hayatın parçası olmaya başlamışlardır (Barber, 2012). Her alanda olduğu gibi sağlık alanında da yaygın kullanım alanları bulmuştur. Doğal olarak büyük verinin en çok kullanıldığı genetik de de yaygın kullanım alanı bulmuştur (Zou et al., 2019). Varyantların değerlendirilmesinde yaygın olarak kullanılmaya başlanmış (Schubach, Re, Robinson, & Valentini, 2017), kanser gibi kompleks hastalıkların aydınlatılmasında da kullanılmaya başlanmıştır (Ainscough et al., 2018b). Klinik değerlendirmenin yanı sıra yeni nesil dizileme verilerinden kopya

sayısı varyantlarının tespitinde makine öğrenme teknikleri kullanılmaktadır (Hill & Unckless, 2019).

Çalışmamız mevcut bilgilerimize göre kopya sayısı varyasyonlarının klinik yorumlanması için kullanılan ilk çalışmalardan biridir. Bu çalışma ile klinik önemi bilinmeyen varyantların makine öğrenme yöntemi ile analizine yönelik olarak makine öğrenme modellerinin çalıştırılması amaçlanmıştır. Bu amaçla dbVar veri tabanındaki nstd102 veri seti kullanılmıştır (Lappalainen et al., 2013). dbVar internette yer alan açık erişimli birçok veri tabanından biridir (Sneddon & Church, 2012). İnsan genomu için yaklaşık 6 milyon genomik bölge için 35 milyondan fazla varyant bulundurur ("NCBI Variation Summary," n.d.). Bu veriler açık erişime sunulmuş 100'den fazla çalışmadan ve birçok organizmadan köken alır. Çalışmalardan elde edilen verilenin bir kısmında klinik veri bulunmazken büyük bir kısmı popülasyon çalışmalarından gelmektedir (Mallick et al., 2016).

Bu çalışmada kullanılan veriler tüm veri tabanının sadece küçük bir kısmını oluşturmaktadır. Kullanılan ClinVar veri tabanının klinik yapısal varyantlar kısmından köken almaktadır. ClinVar verileri birçok dış laboratuvarın bildirdiği verilerden oluşmaktadır. Bu yaklaşım veri setinin büyüklüğüne katkı sunsa da homojen olmayan bir yapı ve farklı klinik yorumlar içermektedir. Mevcut etiketlerin tamamen bireyler tarafından değerlendirilmesi ve kişisel arası yorum farklılıkları bulunmasından dolayı ideal veri setinden bahsetmek güçleşmiştir. Ayrıca klinik etki etiketleri, bireydeki tabloyu oluşturan tek bir neden olsa bile tüm bulunan varyantlara uygulanabilmektedir (bireyde bir patojenik varyant olmasına rağmen tespit edilen tüm varyantların patojenik olarak veri tabanına girilmesi gibi). Yine veri tabanındaki verilerin bir kısmı bölgelerin benign/patojenik etkisinin gösterilmesine rağmen güncel olmayan "önemi bilinmeyen varyant" ve etiketlerini içermektedir.

Veri seti yaklaşık olarak 66 bin varyant içerse de insan genomu düşünüldüğünde veri seti yaklaşık 3,5 milyar bç'nin küçük bir bölümünü kapsamaktadır. Sık görülen ve genellikle iyi tanımlanmış mikrodelesyon/duplikasyon sendromları veri setinde büyük yer tutarken ender kayıp/kazanç bölgeleri için sınırlı veri bulunmaktadır. Verinin mevcut halinin zaten tanısı zor olan yapısal varyantların klinik tanısına katkısı sınırlı olacaktır.

Veri analizinin en büyük kısıtlılıklarının başında veri setinin direkt gen bilgisi içermemesi sayılabilir. Elbette ki ilgili bölgelerin kromozomal lokasyonları gen içerikleri hakkında direkt bir göstergedir ancak ilgili bölge içerisindeki genlerin analize dahil edilmesinin faydalı olacağı düşünülmektedir. Etkilenen gen/gen bölgesi verilerinin eklenmesi çalışmanın etkinliğini arttıracağı düşünülmektedir. Özellikle bilinen mikrodelesyon sendromlarında kritik gen bölgelerinin kaybı ilgili klinik tablonun oluşmasından sorumludurlar veya klinik özelliklerin ağırlıklarını belirler (Rauch et al., 2001). Genomik bölge tabanlı analizlerin gen/ekzon bilgisi içermesi analizin doğruluğunu arttıracaktır.

Varyantların patojenik olarak tanımlanması klinik tabloyu tam olarak tanımlamamaktadır. Örneğin 15q11.2 mikrodelesyon sendromlarında olduğu gibi delesyonun maternal veya paternal olması klinik tabloyu değiştirirken delesyon her iki koşulda da patojenik olarak etiketlenmektedir (Schwartz;, 1998). Makine öğrenme yöntemleri patojenite bilgisi için kullanılabileceği gibi veri seti arttıkça klinik tablonun yorumlanması için de kullanılabilecektir.

Örneklem otozomal kromozomlar için çoğunlukla dengeli dağılmış olsa da 15, 16, 17, 19 ve 22. kromozomlarda beklenilenin dışında fazlalıklar göstermektedir (Grafik 1). Bu kromozomlar diğer kromozomlara oranla daha çok varyasyon göstermedikleri gibi sık görülen bazı mikrodelesyon/duplikasyon sendromları da bu kromozomlarda yer alırlar (Slavotinek, 2008b). Bunun yanında doğru tahminlerin kromozomlara dağılımlarında 15, 16 ve 17. kromozomlar dışında farklılık göstermediği gözlemlenmiştir (Grafik 4). İlgili kromozomlardaki bu ayrımın muhtemel nedeni diğer benzer nedenlerle tanımlanış mikrodelesyon/duplikasyon sendromları bu kromozomlardaki sıklığı şeklinde gözlemlenmiştir.

Çalışmadaki örneklemde kayıp/kazançlar en çok 10Kb ila 10Mb arasında bulunmaktadır (Grafik 2). Varyantın boyutu arttıkça doğru tahminlerin arttığı ve 100 bç'den küçük varyantların yorumlanmasında yanlış tahminlerin fazlalığı gösterilmiştir (Grafik 5).

Pilot çalışma ile %96'lara varan doğruluk tespit edilmiş olsa da daha büyük veri setinde farklı yöntemler kullanılmasına rağmen doğruluk oranı daha düşük kalmıştır. Her iki çalışmada da sonuçların yüzde yüze yaklaşması istenilse de analizin aşırı uyumdan (overfitting) kaçınması önerilir. Bizim çalışmamızda da overfite

yaklaşmadığımız gözlemlenmiştir. Daha az veri olsa da pilot veri seti daha az kaynaktan köken alan iyi kürete edilmiş varyantlardan oluşmaktadır. Bunun yanında ClinVar verisi daha çok kaynaktan daha az denetlenmiş bir veri tabanıdır. ClinVar verileri doğruluğu klinik rutin kullanımda çoğu zaman sorgulanmaktadır, özellikle nokta mutasyonlarında birbirleri ile çelişkili kayıtların olması yorumu güçleştirmektedir (Peterson, Doughty, & Kann, 2013).

Eğitim için kullanılan çok sınıflı algoritmalar içerisinde karar ağacı-forest yüksek skora ulaşmıştır. Ayrıca karar ağacı içerisinde dallanma sayısı arttıkça skor artmış ancak artış sınırlı kalmıştır (8 dal: 0.854977, 16 dal: 0.860067, 32 dal:0.863041) (Tablo 5). Karar ormanları doğru tahmin oluşturma konusunda diğer yöntemlere göre daha başarılıdır ancak oluşturdukları karmaşık veri ağacı yapısı nedeniyle yorumlanması genellikle daha güçtür (Tai et al., 2019). Çok sınıflı lojistik regresyon ise kullanılan algoritmalar içerisinde en düşük skora sahip yöntem olarak değerlendirilmiştir. Lojistik regresyon örneklemi benign olarak sınıflandırma eğiliminde bulunulmuştur (Şekil 9).

Analizler değerlendirildiğinde etiketlerin büyük ölçüde önemi bilinmeyen varyant yönüne kaydığı gözlemlenmiştir. Tüm algoritmalarda gözlemlenen bu kayma bireysel değerlendirmede olduğu gibi doğru etiketin bulunamadığı veya öngörülemediği durumlarda olduğu gibi varyantın etkisinin değerlendirilememesi olarak yorumlanabilir.

Çok sınıflı karar ağacı-forest (32 dal) analizinde muhtemel benign örnekler benign yönünde değerlendirilirken (%36,8) muhtemel patojenik örnekler patojenik (%41,0) yönünde kaymıştır. Bu kayma varyant etiketlerinin klinisyenler tarafından temkinli kullanıldığı benign veya patojenik etiketlerin direk kullanımı yerine muhtemel benign/muhtemel patojenik etiketlerin kullanımının daha fazla olduğu şeklinde yorumlanabilir. Benign ve patojenik sonuçlar değerlendirildiğinde doğru tahminlerin diğer gruplara göre daha yüksek tahmin başarısına sahip olduğu gözlemlenmektedir (Grafik 3).

Varyantın etkisi klinik olarak benign veya patojenik olmasına rağmen varyantın etkisinin ispatı yapılana kadar beş basamaklı tanımlama sistemi kullanılması uluslararası kılavuzlarda önerilmiştir (Kearney et al., 2011). Ancak bu sistem varyantın doğru değerlendirilmesini zorlaştırmaktadır. Grafik 6 da görüldüğü gibi

doğru tahminler içerisinde benign ve patolojik diğer etiketlere göre daha başarılıdır, çoklu etiketlerde sınıflar arası kayma daha fazladır. Makine öğrenme algoritmaları içerisinde iki sınıflı analizler daha başarılı bir ayrım sağlayabilirler (Mayoraz & Alpaydin, 1999). Çalışmada kullanılan verilerde de muhtemel benign, muhtemel patojenik ve önemi bilinmeyen varyantlar çıkarıldıktan sonra analiz başarısı artmıştır (Şekil 11). Yine analiz yöntemleri içerisinde iki sınıflı ve çok sınıflı karar ormanları SVM veya sinir ağına göre daha yüksek performans göstermiştir.

Doğru/yanlış tahminlerin 1, 15 ve 22. kromozomal lokasyonları Grafik 7-12'de gösterilmiştir. Grafikte de görüldüğü gibi kayıp ve kazanç bölgeleri yanlış tahminler için daha dengeli olmaya eğilim gösterirken doğru tahminlerde yığılmalar olmaktadır. Beklenildiği üzere ilgili bölgeler tanımlı mikrodelesyon/duplikasyon sendromları ile komşuluk göstermektedir. Birinci kromozomda 1p36 ya denk gelen yaklaşık ilk 5 Mb, 15q11.2'ye denk gelen yaklaşık 22 Mb ile 23 Mb arası ve 22q11.2'ye denk gelen yaklaşık 21 Mb ile 24 Mb arası bölgelere yığılma göstermektedir. Mikrodelesyon sendromlarında tek bir klinik tablo gözlemlenmesi her zaman beklenilen bir tablo değildir. Özellikle kayıp bölgesindeki kritik gen/gen bölgeleri çoğunlukla primer tablonun oluşmasını tetiklerler (Carvill & Mefford, 2013). Ancak buna rağmen değişken penetrans klinik tablonun şiddetini değiştirmektedir. Genom milyarlarca baz ve on binlerce gen içeren dinamik bir yapıya sahiptir. Kompleks tabloların büyük bir çoğunluğu birbirini tetikleyen etkileşimler ile mümkündür. Genlerin fonksiyon göstermeleri transkripsiyon faktörleri, promoter düzenleyici bölgeler, metilasyon, histon asetilasyonu gibi epigenetik faktörlerin kontrolü altındadır. Genlerin fonksiyon kayıp ve kazanımları çoğu durumda düzenleyici etkenlerin durumuna da bağımlıdır. Mevcut değerlendirilme içerisinde kromozomal lokasyon, kayıp kazanç durumu ve patojenitesi kullanılmıştır ancak bu veriler moleküler düzenleyici mekanizmaların da varlığında daha anlamlı olacaktır. Bu amaçla veri tabanlarının daha detaylı klinik özellikler içeren daha homojen bir yapıya kavuşması önemlidir. Ayrıca çalışmanın zayıf yönleri arasında sağlıklı popülasyon verilerinin dahil edilmemesi gösterilebilir. Gittikçe artan sayıda açık erişimli varyant veri tabanlarının analiz keskinliğini arttıracağı düşünülmektedir. Yine gen bazlı hesaplanan haployetmezlik skorları da ilgili bölgenin değerlendirilmesine eklenilebilir.

Makine öğrenme algoritmaları ve yapay zeka uygulamaları günlük hayatımızın önemli bir parçası olmaya artan bir hızda devam edecektir. Büyük verilerin ortaya çıkması ve

çoklu parametrelerin varlığı insan yorumlama limitlerini zorlamaktadır. Özellikle tıp alanında verilerin kompleksliği özel eğitim süreçleri ve yorumlama algoritmalarıyla aşılmaya çalışılsa da 25 milyonun üzerindeki bilimsel makale günlük tıbbi uygulamayı zorlaştırmaktadır (Hanke, Gibbons, Casar Berazaluce, & Ponsky, 2019). Bu kapsamda tıp ve uygulamalarının dijitalleştirmesi kaçınılmaz olmaya başlayacaktır. Tıbbi görüntüleme sistemleri, laboratuvar araçları, EEG, EMG gibi nöronal kayıtlar ciddi veriler üretmektedir. Bu verilerin analizinde birçok yapay zeka / makine öğrenme algoritmaları kullanılmakta ve bu uygulamaların klinik verilerin doğru yorumlanması ve maliyetlerin düşürülmesinde değerli olduğu gösterilmiştir (Al-Mufti et al., 2019). Yapay zeka sistemlerinin yaygınlaşması ve iyi klinik uygulama örneklerinin oluşturulması klinik değerlendirme ve görüntüleme sistemlerini otomatikleştirebilme ve tanıya hızlı ulaşabilmeyi kolaylaştıracaktır (Kilic, 2019).

Çalışmada kullanılan Microsoft Azure Machine Learning Studio makine öğrenme basamaklarının son derece kolaylaştırmıştır. Özellikle sürükle-bırak yapısıyla kullanılan algoritmalar arası ilişkilendirme uygulamaları programlama geçmişi olmayan kullanıcılar için bile büyük kolaylıklar sağlamaktadır. Elbette ki birçok platform makine öğrenme algoritmalarının kullanımı için pratik çözümler sunmaktadır. Örneğin Weka veya Tensor Flow diğer öncül yazılımlar arasında sayılabilir (“Tensor Flow,” n.d.; Witten, Frank, & Hall, 2011). Bu tarz yazılımların kullanımındaki kısıtlılıkların başında diğer makine öğrenme kütüphanelerine göre daha az özelleştirilebilme kapasitesine sahip olmaları ve daha az parametre değişikliğine izin veriyor olmalarıdır. Bunun yanında bulut teknolojileri kullanımı yerel işlemci veya depolama gereksinimlerinden bağımsız olarak makine öğrenme analizlerinin hızlıca yapılmasına olanak sağlamaktadır. Özellikle Microsoft, Google, Amazon gibi firmaların bulutta veri bulundurma ve analiz etme sistemlerini içeren kapsamlı alt yapıları kullanıcılara sunuyor olması makine öğrenme algoritmalarının kullanımını kolaylaştırmaktadır. Öğrenciler için sunulan ücretsiz kullanım imkanları ise veri analizine ekstra bir maliyet katmamaktadır. Bu çalışmada da veri oluşturulması, depolanması ve analizinde tamamen açık kaynaklar kullanıldığı için herhangi bir maliyet oluşmamıştır.

Bu çalışma ile genetik rutini olarak uygulanan özellikle mental retardasyon/çoklu konjenital anomali hasta grubunun tanısında ilk basamaklardan kabul edilen kopya sayısı varyasyonlarını analizi için kullanılabilecek veri analiz yöntemleri için bir

basamak oluşturulmaya çalışılmıştır. Özellikle her geçen gün değişen teknoloji ve dizi analizi uygulamaları, mikroarray teknolojileri ile kopya sayısı varyantlarının tespiti kolaylaşmakta ve daha kolay veriler elde edilebilmektedir. Ancak genomik verilerin halen daha net olarak anlaşılamaması ve genomik verinin değerlendirmedeki zorluklar kopya sayısı varyantlarının klinik etkisinin yorumlanmasını kısıtlamaya devam etmektedir. Veri setlerinin gelişmesi ve makine öğrenme algoritmalarının ideale yaklaşması ile DNA analiz cihazlarının çıktılarının ara biyoinformatik basamaklar olmadan klinik veriye dönüşmesi olanaklı hale gelebilecektir.

Sonuç ve Öneriler

Bu çalışma ile kopya sayısı varyantlarının analizinde makine öğrenme algoritmalarının kullanımının varyant sınıflamasında kullanılabileceği ve klinik rutinde yer alabileceği gösterilmiştir. Elbette ki model bu hali ile uygulanabilir değildir. Sistemi idealize edebilmek için:

- Veri setlerinin büyümesi
- Standart ve iyi kürate veri setlerinin oluşturulması
- Mevcut analize daha fazla parametre eklenmesi
- Klinik verinin detaylandırılması
- Haplo-yetmezlik verilerinin hesaplamaya dahil edilmesi
- Tek nükleotid varyantları ve/veya epigenetik modifikasyon bilgilerinin dahil edilmesi önerilir.

Ayrıca makine öğrenme algoritmalarının gelişmekte ve evrilmekte olduğu da unutulmamaktadır. Her geçen gün artan genetik veri ile artan bilgi birikimimiz doğru tanı ve tedavideki dijitalleşmeyi arttıracağına inanmaktayız. Bu yönde bir basamak olan çalışmamızın sürece katkı sunacağını düşünmekteyiz.

Kaynaklar

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., … Griffith, O. L. (2018a). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature Genetics*, *50*(12), 1735–1743. https://doi.org/10.1038/s41588-018-0257-y

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., … Griffith, O. L. (2018b). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature Genetics*, *50*(12), 1735–1743. https://doi.org/10.1038/s41588-018-0257-y

Al-Mufti, F., Kim, M., Dodson, V., Sursal, T., Bowers, C., Cole, C., … Mayer, S. A. (2019). Machine Learning and Artificial Intelligence in Neurocritical Care: a Specialty-Wide Disruptive Transformation or a Strategy for Success. *Current Neurology and Neuroscience Reports*, *19*(11), 89. https://doi.org/10.1007/s11910-019-0998-8

Albertson, D. G., & Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Human Molecular Genetics*, *12*(suppl 2), R145–R152. https://doi.org/10.1093/hmg/ddg261

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Brandt, T., Sack, L. M., Arjona, D., Tan, D., Mei, H., Cui, H., … Meck, J. M. (2019). Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants. *Genetics in Medicine*, 1–9. https://doi.org/10.1038/s41436-019-0655-2

Carvill, G. L., & Mefford, H. C. (2013). Microdeletion syndromes. *Current Opinion in Genetics & Development*, *23*(3), 232–239. https://doi.org/10.1016/J.GDE.2013.03.004

Database of Genomic Variants. (n.d.). Retrieved November 3, 2019, from http://dgv.tcag.ca/dgv/app/home

Devriendt, K., & Vermeesch, J. R. (2004). Chromosomal phenotypes and submicroscopic abnormalities. *Human Genomics*, *1*(2), 126–133. https://doi.org/10.1186/1479-7364-1-2-126

Feuk, L., Marshall, C. R., Wintle, R. F., & Scherer, S. W. (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. *Human Molecular Genetics*, *15*(suppl_1), R57–R66. https://doi.org/10.1093/hmg/ddl057

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., … Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics*, *84*(4), 524–533. https://doi.org/10.1016/j.ajhg.2009.03.010

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., … Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Research*, *16*(8), 949–961. https://doi.org/10.1101/gr.3677206

Hanke, R. E., Gibbons, A. T., Casar Berazaluce, A. M., & Ponsky, T. A. (2019). Digital Transformation of Academic Medicine: Breaking Barriers, Borders, and Boredom. *Journal of Pediatric Surgery*. https://doi.org/10.1016/J.JPEDSURG.2019.10.037

Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., … Sawyers,

C. L. (2018). Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *ELife*, *7*. https://doi.org/10.7554/eLife.37294

Hill, T., & Unckless, R. L. (2019). A Deep Learning Approach for Detecting Copy Number Variation in Next-Generation Sequencing Data. *G3 (Bethesda, Md.)*, *9*(11), 3575–3582. https://doi.org/10.1534/g3.119.400596

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., … Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–951. https://doi.org/10.1038/ng1416

Isakov, O., Dotan, I., & Ben-Shachar, S. (2017). Machine Learning–Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. *Inflammatory Bowel Diseases*, *23*(9), 1516–1523. https://doi.org/10.1097/MIB.0000000000001222

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, *258*(5083), 818–821. https://doi.org/10.1126/science.1359641

Kaminsky, E. B., Kaul, V., Paschall, J., Church, D. M., Bunke, B., Kunig, D., … Martin, C. L. (2011). An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *13*(9), 777–784. https://doi.org/10.1097/GIM.0b013e31822c79f9

Kearney, H. M., Thorland, E. C., Brown, K. K., Quintero-Rivera, F., & South, S. T. (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genetics in Medicine*, *13*(7), 680–685. https://doi.org/10.1097/GIM.0b013e3182217a3a

Kilic, A. (2019). Artificial Intelligence and Machine Learning in Cardiovascular Healthcare. *The Annals of Thoracic Surgery*. https://doi.org/10.1016/J.ATHORACSUR.2019.09.042

Koolen, D. A., Pfundt, R., de Leeuw, N., Hehir-Kwa, J. Y., Nillesen, W. M., Neefs, I., … de Vries, B. B. A. (2009). Genomic microarrays in mental retardation: A practical workflow for diagnostic applications. *Human Mutation*, *30*(3), 283–292. https://doi.org/10.1002/humu.20883

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., … Church, D. M. (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Research*, *41*(Database issue), D936-41. https://doi.org/10.1093/nar/gks1213

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Mayoraz, E., & Alpaydin, E. (1999). Support vector machines for multi-class classification (pp. 833–842). Springer, Berlin, Heidelberg . https://doi.org/10.1007/BFb0100551

Microsoft Azure Machine Learning Studio (classic). (n.d.). Retrieved November 18, 2019, from https://studio.azureml.net/

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., … Ledbetter, D. H. (2010). 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*, *86*(5), 749–764. https://doi.org/10.1016/J.AJHG.2010.04.006
NCBI Variation Summary. (n.d.). Retrieved November 24, 2019, from https://www.ncbi.nlm.nih.gov/dbvar/content/org_summary/
nstd101 - ClinGen - dbVar Study - NCBI. (n.d.). Retrieved November 18, 2019, from https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd101/
nstd102 - Clinical Structural Variants - dbVar Study - NCBI. (n.d.). Retrieved November 18, 2019, from https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd102/
Parıltay, E., & Ece Solmaz, A. (2018). *Makine Öğrenme Yöntemleri İle Kopya Sayısı Varyasyonlarının Değerlendirilmesi; Kavram Ispatı*. Antalya.
Peterson, T. A., Doughty, E., & Kann, M. G. (2013). Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. *Journal of Molecular Biology*, *425*(21), 4047–4063. https://doi.org/10.1016/J.JMB.2013.08.008
Rauch, A., Schellmoser, S., Kraus, C., D□rr, H. G., Trautmann, U., Altherr, M. R., … Reis, A. (2001). First known microdeletion within the Wolf-Hirschhorn syndrome critical region refines genotype-phenotype correlation. *American Journal of Medical Genetics*, *99*(4), 338–342. https://doi.org/10.1002/ajmg.1203
Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/nature05329
Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., … Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics*, *32*(1), 135–142. https://doi.org/10.1038/ng947
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–423. https://doi.org/10.1038/gim.2015.30
Schubach, M., Re, M., Robinson, P. N., & Valentini, G. (2017). Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Scientific Reports*, *7*(1), 2959. https://doi.org/10.1038/s41598-017-03011-5
Schwartz;, S. C. (1998). Prader-Willi and Angelman Syndromes: Disorders of Genomic Imprinting. *Medicine*, *77*(2), 140–151. Retrieved from https://insights.ovid.com/crossref?an=00005792-199803000-00005
Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., … Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, *305*(5683), 525–528. https://doi.org/10.1126/science.1098918
Shalev-Shwartz, S., & Ben-David, S. (n.d.). *Understanding machine learning : from theory to algorithms*.
Shotton, J., Sharp, T., Kohli, P., Nowozin, S., Winn, J., & Criminisi, A. (2013, January 1). Decision Jungles: Compact and Rich Models for Classification. Retrieved from https://www.microsoft.com/en-us/research/publication/decision-jungles-compact-and-rich-models-for-classification/
Slavotinek, A. M. (2008a). Novel microdeletion syndromes detected by chromosome microarrays. *Human Genetics*, *124*(1), 1–17. https://doi.org/10.1007/s00439-008-0513-9

Slavotinek, A. M. (2008b). Novel microdeletion syndromes detected by chromosome microarrays. *Human Genetics*, *124*(1), 1–17. https://doi.org/10.1007/s00439-008-0513-9

Sneddon, T. P., & Church, D. M. (2012). Online resources for genomic structural variation. *Methods in Molecular Biology (Clifton, N.J.)*, *838*, 273–289. https://doi.org/10.1007/978-1-61779-507-7_13

Spielmann, M., & Klopocki, E. (2013). CNVs of noncoding cis-regulatory elements in human disease. *Current Opinion in Genetics & Development*, *23*(3), 249–256. https://doi.org/10.1016/J.GDE.2013.02.013

Steinberg, J., Honti, F., Meader, S., & Webber, C. (2015). Haploinsufficiency predictions without study bias. *Nucleic Acids Research*, *43*(15), e101. https://doi.org/10.1093/nar/gkv474

Tai, A. M. Y., Albuquerque, A., Carmona, N. E., Subramanieapillai, M., Cha, D. S., Sheko, M., … McIntyre, R. S. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artificial Intelligence in Medicine*, *99*, 101704. https://doi.org/10.1016/J.ARTMED.2019.101704

Tensor Flow. (n.d.). Retrieved November 27, 2019, from https://www.tensorflow.org/

Witten, I. H. (Ian H. ., Frank, E., & Hall, M. A. (Mark A. (2011). *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, *51*(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5

Ekler

- Uluslararası Katılımlı 13. Ulusal Tıbbi Genetik Kongresi 07-11 Kasım 2018 Bildiri Özeti

S-012 - MAKİNE ÖĞRENME YÖNTEMLERİ İLE KOPYA SAYISI VARYASYONLARININ DEĞERLENDİRİLMESİ; KAVRAM ISPATI

Erhan PARILTAY[1], Aslı ECE SOLMAZ[1],

[1]EGE ÜNİVERSİTESİ TIP FAKÜLTESİ TIBBİ GENETİK AD,

Kopya sayısı varyasyonları (CNV) genomik yapısal varyasyonlarının oldukça sık görülen formlarından biridir. CNV'ler FISH, microarray, masif paralel sekanslama gibi yöntemlerin kullanımı ile daha yaygın olarak tespit edilmeye başlanılmıştır. Başta büyüme gelişme geriliği olmak üzere birçok klinik tablo ile ilişkilendirilmiş olmalarına rağmen sağlıklı bireylerde de oldukça sık rastlanırlar. Bunun yanında bireyde tespit edilen varyasyonların klinik yorumu güçlükler içermektedir. Veri tabanlarında bulunmayan ve ilk kez tespit edilen değişikliklerin değerlendirilmesinde CNV'nin kalıtılmış veya de-novo olması önem taşımaktadır. Ancak yine de önemli bir CNV grubu etkisi bilinmeyen olarak tanımlanmaktadır. Makine öğrenme ya da yapay zeka teorileri eskiye dayansa da son yıllarda veri işleme imkanlarının artması ile tekrar önem kazanmıştır. Başta büyük veri setleri olmak üzere birçok kompleks durumda kullanım alanı bulmaktadır. Bu çalışmada makine öğrenme yöntemleri kullanılarak kopya sayısı varyasyonlarının klinik yorumuna ulaşılabilirliğin gösterilmesi amaçlanmıştır.

Bu çalışmada dbVar veri tabanında bulunan ISCA (International Standards for Cytogenomic Arrays consortium) veri seti kullanılmıştır. Bu veri setindeki klnik veri etiketi bulunan 11989 varyant bilgisi kullanılmıştır. Veriler, Microsoft Azure Machine Learning Studio ile bulut hesaplama teknolojisi kullanılarak analiz edilmiştir. Veri setinin yüzde yetmişi eğitim seti olarak kullanılırken yüzde otuzu test verisi olarak kullanılmıştır. Eğitim modeli olarak farklı algoritmalar denenmiş ve çok sınıflı karar ormanı en yüksek veri keskinliğine ulaşmıştır.

Makine öğrenme analizi sonucunda toplamda %89.241 ve ortalamada %96.4137 doğruluğa ulaşılmıştır. 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.

Bu çalışma daha geniş ölçekli yapılması planlanan kopya sayısı varyasyonlarının makine öğrenme yöntemi kullanılarak analiz ve klinik yorumlanmasına dair çalışmanın ilk basamağını oluşturmaktadır. Temel olarak iyi düzenlenmiş ve seçilmiş bir veri grubunda analizin çalışma başarısının değerlendirilmesi amaçlanmıştır. Mevcut veri grubu klinik kullanım ve genomun tamamı için analiz yapmaya yeterli değildir. Daha büyük veri setleri ve sağlıklı popülasyon verilerinin eklenmesi ile veri keskinliğinin arttırılması planlanmaktadır.

ANAHTAR KELİMELER: KOPYA SAYISI VARYASYONLARI, MAKİNE ÖĞRENME, VERİ MADENCİLİĞİ

TEZ ONAY SAYFASI

Kurum Adı : Ege Üniversitesi Sağlık Bilimleri Enstitüsü

Anabilim Dalı : Sağlık Biyoinformatiği

Program : Sağlık Biyoinformatiği

Tez Konusu : Kopya Sayısı Varyasyonlarının Makine Öğrenme Yöntemi İle Analiz Edilmesi

Danışman : Doç. Dr. Buket Kosova

Tezi Hazırlayan : Erhan Parıltay

Değerlendirme Kurulu Üyeleri :

Adı Soyadı	:
Başkan(Danışman)	: Doç. Dr. Buket Kosova
Üye / İmza	: Doç. Dr. Burak ORDİN
Üye / İmza	: Prof. Dr. Muhsin Özgür ÇOĞULU
Üye / İmza	: Prof. Dr. Ahmet Okay ÇAĞLAYAN
Üye / İmza	: Doç. Dr. Tufan ÇANKAYA
Üye / İmza	: Doç. Dr. Elçin BORA

Tezin Kabul Edildiği Tarih : 18.12.2019

T. C.
EGE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZ SAVUNMA JÜRİSİ ÜYELERİNE CİLTLİ TEZİN TESLİMİ TUTANAĞI

Öğrencinin;
Adı, Soyadı : Erhan Parıltay
Numarası :o93140000333
Anabilim Dalı : Sağlık Biyoinformatiği
Programı : ☐ Yüksek Lisans ■ Doktora

Danışman Onayı :
Danışman Adı Soyadı : Doç. Dr. Buket Kosova
İmza :

Tez Savunma Sınav Jürisi Onayı

Enstitü Yönetim Kurulu'nun 04 /12/2019 tarih ve 52/7 sayılı toplantısında "Tez Savunma Sınavı Jüri Üyesi" olarak görevlendirildiğim yukarıda adı, soyadı programı yazılı öğrencinin tezi tarafıma teslim edilmiştir.

Tez Savunma Sınav Jürisi
Asıl Üyeler

Ünvanı, Adı Soyadı	Anabilim Dalı/ Kurumu	Tarih	İmza
1. Doç. Dr. Buket KOSOVA (Danışman)	Tıbbi Biyoloji/Ege Üniversitesi	09/12/2019	
2. Doç. Dr. Burak ORDİN	Bilgisayar Bilimleri/Ege Üniversitesi	11/12/2019	
3. Prof. Dr. Muhsin Özgür ÇOĞULU	Tıbbi Genetik/Ege Üniversitesi	11/12/2019	
4. Prof. Dr. Ahmet Okay ÇAĞLAYAN	Tıbbi Genetik/Dokuz Eylül Üni.	13/12/2019	
5. Doç. Dr. Tufan ÇANKAYA	Tıbbi Genetik/Dokuz Eylül Üni.	13/12/2019	
6. Doç. Dr. Elçin BORA	Tıbbi Genetik/Dokuz Eylül Üni.	13/12/2019	

Yedek Üyeler

Ünvanı, Adı Soyadı	Anabilim Dalı/ Kurumu	Tarih	İmza
1. Dr. Öğr. Üyesi Altuğ KOÇ	Tıbbi Genetik/Dokuz Eylül Üni.	/ /	.....
2. Prof. Dr. Cenk SELÇUKİ	Biyokimya/Ege Üniversitesi	/ /	.....
3.	/	/ /	.....

*Bu form Ege Üniversitesi Eğitim-Öğretim Yönetmeliği'nin ..... maddelerine göre düzenlenmiştir.Ayrıntılar ve yapılacak işlemlerle ilgili bilgi için Lisansüstü Eğitim - Öğretim Yönetmeliği'ne web sayfasından (http://sbe.ege.edu.tr/) ulaşılabilir.

T.C
EGE ÜNİVERSİTESİ
SAĞLIK BİLİMLER ENSTİTÜSÜ

DOKTORA TEZ SAVUNMASI
SINAV TUTANAĞI

Öğrencinin Adı-Soyadı :Erhan PARILTAY
Anabilim Dalı :Sağlık Biyoinformatiği
Programın Adı :Doktora
Sınavın Tarihi ve Saati :18.12.2019
Sınavın Yapıldığı Yer :Sağlık Bilimleri Enstitüsü

Jüri tarafından; ☐ Oy Çokluğu / ☒ Oy Birliği ile

☒ tezi kabul edilmiş ve başarılı bulunmuştur.

☐ tez başlığı değiştirilerek tezi kabul edilmiş ve başarılı bulunmuştur.

- Yeni tez başlığı:.....

.....

☐ tezinde düzeltme gerektiğine karar verilmiştir.

(Bu durumda öğrenci en geç altı ay içinde gerekli düzeltmeyi yaparak tezini aynı jüri önünde yeniden savunur.)

☐ tezi reddedilmiştir.

Doç.Dr.Buket KOSOVA
(Danışman)

Prof.Dr.A.Okay ÇAĞLAYAN

Prof.Dr.M.Özgür ÇOĞULU

Doç.Dr.Burak ORDİN

Doç.Dr.Tufan ÇANKAYA

Doç.Dr.Elçin BORA

- *Bu tutanak ve jüri üyelerinin hazırladığı "Tez İnceleme ve Değerlendirme Formları" ilgili Anabilim Dalı Başkanlığının ön yazısı ile Sağlık Bilimleri Enstitüsü Müdürlüğü'ne 3 iş günü içerisinde ulaştırılacaktır.*
- *Bu belgenin elektronik kopyasına http://sbe.ege.edu.tr/formlar adresinden ulaşabilirsiniz.*

Teşekkür

Tez yazımı süresince desteklerini esirgemeyen başta tez danışmanım Doç. Dr. Buket Kosova olmak üzere tez izleme jürim Doç Dr. Burak Ordin ve Doç. Dr. Elçin Boraya, her zaman yol gösterici olan ve desteğini esirgemeyen Prof. Dr. Fazıl Apaydın'a ve eğitimimde katkıları olan tüm hocalarıma teşekkür ederim.

İzmir, 19.12.2019

Erhan PARILTAY

Özgeçmiş

Uzm. Dr. Erhan PARILTAY

Elazığ / 21.04.1979

pariltay@gmail.com

Mesleki Deneyim:

Ege Üniversitesi Tıp Fakültesi Tıbbi Genetik AD **İzmir**	Uzm. Dr.	Kasım 2013 - Halen
Şevket Yılmaz Eğitim ve Araştırma Hastanesi **Bursa**	Uzm. Dr. (Birim İdari Sorumlusu)	Ekim 2011- Kasım 2013
Gülhane Askeri Tıp Akademisi Tıbbi Genetik BD (Askerlik) **Ankara**	Uzm. Dr.	Kasım 2010- Eylül 2011
Ali Osman Sönmez Onkoloji Hastanesi **Bursa**	Uzm. Dr.	Mayıs 2010- Eylül 2010
Ege Üniversitesi Tıp Fakültesi Tıbbi Genetik AD **İzmir**	Asist. Dr.	Haziran 2005- Mayıs 2010
70. yıl Sağlık Ocağı Pertek/**Tunceli**	Dr.	Kasım 2004- Aralık 2004

Eğitim Bilgileri:

- Ege Üniversitesi Tıp Fakültesi Tıbbi Genetik AD Tıpta Uzmanlık 2005-2010
- Ege Üniversitesi Tıp Fakütesi 1997-2003
- Elazığ Anadolu Lisesi 1990-1997
- Elazığ Atatürk İlkokulu 1986-1990

Sınav Bilgileri:

- 2011 Ales Sonbahar Dönemi (27.11.2011)
  - Sayısal 80.776
- 2012 Üds ilkbahar Dönemi (18.03.2012)
  - Sağlık Blimleri İngilizce 70.00
- 2017 Yökdil
  - İngilizce 91.25

Yabancı Diller:

İngilizce / Almanca

Uzmanlık Tezi:

- İdiyopatik Mental Retardasyonlu Olguların Array Çipi ile Tüm Genom Kopya Sayısı Analizi Subat 2010

Makale:

1. Seckel Syndrome With Morgagni Hernia, Onder A, Cogulu O, Ekmekci A, Pariltay E, Kirbiyik O, Ozkinay F., Clin Dysmorphol. 2007 Jul16(3):209-10.
2. The Evaluation Of The Referral Reasons Of Patients At A Tertiary Pediatric Genetic Center İn Izmir, Turkey. Durmaz B, Alpman A, Pariltay E, Akgul M, Ataman E, Kirbiyik O, Cogulu O, Ozkinay F. Genet Test Mol Biomarkers. 2009 Apr13(2):163-6.
3. Reasons For Adult Referrals For Genetic Counseling At A Genetics Center İn Izmir, Turkey: Analysis Of 8965 Cases Over An Eleven-Year Period. Cogulu O, Ozkinay F, Akin H, Onay H, Karaca E, Durmaz Aa, Durmaz B, Aykut A, Pariltay E, Kirbiyik O, Gunduz C, Ozkinay C. J Genet Couns. 2011 Jun20(3):287-93. Epub 2011 Jan 8.
4. Demonstration of uniparental-isodisomy on chromosome 22q11.2 in a patient with childhood schizophrenia and facial dysmorphology by whole-genome analysis. Cogulu O, Pariltay E, Durmaz AA, Aykut A, Gunduz C, Ozbaran B, Aydin HH, Erermis S, Aydin C, Ozkinay F. J Neuropsychiatry Clin Neurosci. 2012 Dec 1;24(1):E13-4.
5. Genome wide analysis in a discordant monozygotic twin with caudal appendage and multiple congenital anomalies. Cogulu O, Pariltay E, Koroglu OA, Aykut A, Ozyurek R, Levent E, Kultursay N, Ozkinay F. Genetic Counseling (Geneva, Switzerland) [2013, 24(1):85-91]
6. Low Zip 4 gene expression levels in RPMI-8226 and ARH-77 cell lines support the possible role of zip 4 transporter protein in plasma cell tumorogenesis. Çoban ZD, Torun D, Avcu F, Ural AU, Parıltay E, Kozan S, Güran S Cumhuriyet Med J 2013; 35: 9-13

7. Genome-wide copy number variation analysis in idiopathic intellectual disability/multiple congenital anomalies. Pariltay E, Durmaz A, Durmaz B, Aykut A, Onay H, Ak H, Aydin HH, Ozkinay F, Cogulu O. Genet Couns. 2014;25(2):221-9.
8. The phenotypic and molecular genetic spectrum of Alström syndrome in 44 Turkish kindreds and a literature review of Alström syndrome in Turkey. Ozantürk A, Marshall JD, Collin GB, Düzenli S, Marshall RP, Candan S, Tos T, Esen I, Taşkesen M, Cayır A, Oztürk S, Ustün I, Ataman E, Karaca E, Ozdemir TR, Erol I, Eroğlu FK, Torun D, Parıltay E, Yılmaz-Güleç E, Karaca E, Atabek ME, Elçioğlu N, Satman I, Möller C, Müller J, Naggert JK, Ozgül RK. J Hum Genet. 2014 Oct 9. doi: 10.1038/jhg.2014
9. A novel splice site mutation of FGD1 gene in an Aarskog-Scott syndrome patient with a large anterior fontanel. Parıltay E, Hazan F, Ataman E, Demir K, Etlik Ö, Özbek E, Özkan B. J Pediatr Endocrinol Metab. 2016 Sep 1;29(9):1111-4. doi: 10.1515/jpem-2015-0482.
10. An X-Linked Hyper-IgM Patient Followed Successfully for 23 Years without Hematopoietic Stem Cell Transplantation. Kutukculer N, Karaca NE, Aksu G, Aykut A, Pariltay E, Cogulu O. Case Reports Immunol. 2018 Oct 14;2018:6897935. doi: 10.1155/2018/6897935. eCollection 2018.
11. Chronic granulamatous disease: Two decades of experience from a paediatric immunology unit in a country with high rate of consangineous marriages. Kutukculer N, Aykut A, Karaca NE, Durmaz A, Aksu G, Genel F, Pariltay E, Cogulu Ö, Azarsız E. Scand J Immunol. 2019 Feb;89(2):e12737. doi: 10.1111/sji.12737. Epub 2019 Jan 23.
12. Profile Templates to Use during Surgery in Precision Rhinoplasty. Apaydin F, Garcia RFF, Pariltay E. Facial Plast Surg. 2019 Feb;35(1):111-112. doi: 10.1055/s-0039-1677830. Epub 2019 Feb 13.
13. A Novel TTC37 Mutation Causing Clinical Symptoms of Trichohepatoenteric Syndrome Such as Pyoderma Gangrenosum and Immunodeficiency Without Severe Diarrhea. Karaca Edeer N, Aykut A, Pariltay E, Aksu G, Cogulu O, Kutukculer N. J Investig Allergol Clin Immunol. 2019;29(5):396-398. doi: 10.18176/jiaci.0418. Epub 2019 May 27.

Sözel Bildiri

1. Prenatal Dıagnosıs Pallster-Kıllıan Syndrome (Mosaıc Tetrosomy 12p) Hilmi Bolat, Erhan Parıltay, Aslı Ece Solmaz, Çağrı Güven, Burak Durmaz, Emin Karaca, Özgür Çoğulu, Haluk Akın Gevher Nesibe Tıp Günleri 2016 & Tıbbi Genetik ve Klinik Uygulamaları Kongresi 11-13 Şubat 2016 Kayseri
2. A Recurrent Pregnancy Loss Case Wıth T(1;7)(P26;Q36) Translocatıon Tuba Sözen Türk, Erhan Parıltay, Aslı Ece Solmaz, Burak Durmaz, Emin Karaca, Haluk Akın, Özgür Çoğulu Gevher Nesibe Tıp Günleri 2016 & Tıbbi Genetik ve Klinik Uygulamaları Kongresi 11-13 Şubat 2016 Kayseri
3. Prematür Over Yetmezliği Olgusunda Mikroarray ile Saptanan kompleks X, Y Translokasyonu E.Parıltay 12. Ulusal Tıbbi Genetik Kongresi 5-9 Ekim 2016 Çeşme

Poster

1. The evaluation of referral reasons for genetic counseling and prenatal diagnosis at a tertiary genetic center: a Turkish experience O. Cogulu1, A. Alpman1, B. Durmaz1, E. Pariltay2, M. Akgul2, O. Kirbiyik2, E. Ataman2, H. Akin2, F. Ozkinay1, C. Ozkinay2; 39th European Human Genetics Conference (EHGC) in Nice, France, in June 2007
2. Retrospective evaluation of the referral reasons of patients at a tertiary paediatric genetic center in Izmir, Turkey B. Durmaz1, E. Pariltay2, M. Akgul2, A. Alpman1, E. Ataman2, O. Kirbiyik2, O.Cogulu1, F. Ozkinay139th European Human Genetics Conference (EHGC) in Nice, France, in June 2007
3. Cytogenetic analysis of 135 myelodysplastic syndrome patients E. Pariltay[1], A. Alpman[1], E. Karaca[2], B. Durmaz[2], O. Cogulu[2], F. Ozkinay[2] 57th Annual Meeting, The American Society of Human Genetics, in San Diego, USA September 2007
4. Novel Findings By Genome-wide copy Number Analysis on chromosome 22 in a case with mild Facial Dysmorphology and Autistic/schizophrenic Behaviours E. Pariltay, O. Cogulu, A. Aykut, A. Alpman, B. Ozbaran, S. Erermis, C. Aydin, F. Ozkinay; European Human Genetics Conference May 2009 Vienna, Austria

5. Identification of a Deletion on Chromosome 3p(12.3) by Whole Genome Analysis in a Discordant monozygotic twins with a tail and multiple congenital Anomalies O. Cogulu, E. Pariltay, A. Alpman, O. Altun, N. Kultursay, R. Ozyurek, F. Ozkinay; European Human Genetics Conference May 2009 Vienna, Austria
6. Prenatal Detection of Pericentric inversion of chromosome 9 in 5358 Referrals at a Reference Genetic center E. Karaca, E. Pariltay, O. Cogulu, H. Akin, F. Ozkinay; European Human Genetics Conference May 2009 Vienna, Austria
7. The cytotoxic effect of docetaxel and gemcitabine on plasma cell leukemia cell line and downregulation of zip4 expression due to the treatment. ZD Coban, S Guran, E Pariltay, S Kozan, AU Ural, F Avcı. Rna&Oligonucleotide therapeutics 4-7 December 2011
8. Farklı Yöntemlerle Trombofili Yatkınlık Paneli Çalışılan Olguların Karşılaştırılması ve Sonuçlarının Değerlendirilmesi. E Parıltay, ŞÖ Sağ, YH Özön. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
9. Ailesel Akdeniz Ateşi Öntanısıyla 12 Sık Mutasyon Analizi Yapılan 360 Olgunun Mutasyon Profillerinin Değerlendirilmesi. E Parıltay, ŞÖ Sağ, O Görükmez. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
10. Yeni Açılmış Bir Tıbbi Genetik PolikliniğIne Başvuran Hasta Profili O Görükmez, ŞÖ Sağ, E Parıltay, S Barış. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
11. Rpmı-8226 ve arH-77 hücre hatlarında düşük zip 4 gen ekspresyon düzeyi bulgusu plazma hücrelerinden köken alan tümörlerin gelişiminde zip 4 taşıyıcı proteininin olası rolünü desteklemektedir. Çoban ZD, Torun D, Avcu F, Ural AU, Parıltay E, Kozan S, Güran S. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
12. Wolf–Hirschhorn sendromunda, İlk defa tariflenen ektopik böbrek anomalisi. F Hazan, E Parıltay, A Öztürk, K Yararbaş, A Tükün. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
13. Duplikasyon 3q sendromlu bir olgu sunumu. ŞÖ Sağ, E Parıltay, O Görükmez, Ö Etlik. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa

14. Frontonazal displazi tanılı bir olgu sunumu. ŞÖ Sağ, O Görükmez, E Parıltay, Ö Etlik. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
15. X Kromozomu Mozaiklikleri; Sitogenetik ve FISH Sonuçlarının Karşılaştırılması Erhan Parıltay, Haluk Akın, Emin Karaca, Asude Durmaz, Burak Durmaz, Ayça Aykut, Ayşe Nur Güleçoğlu, Hilmi Bolat, Özgür Çoğulu 11. Ulusal Tıbbi Genetik Kongresi 24-27 Eylül 2014 İstanbul
16. der(19)t(1;19) in Childhood Acute Lymphoblastic Leukemia Patient E. Pariltay, E. Karaca, A. Aykut1, B. Durmaz ve ark. ESHG 2015 6-9 Haziran 2015 Glasgow / İngiltere
17. Azoospermia and Varicocele in Noonan Syndrome F. Hazan , E. Pariltay , B. S. Ada , E. Ataman , A. Tukun ESHG 2016, 21-24 Mayıs 2016 Barselona / İspanya
18. TACI, ICOS and BAFFR mutation analysis in 449 Common Variable Immunodefi ciency (CVID) patients. E. Pariltay ,A. Aykut , A. Durmaz, F.Hazan, N. Gulez, N. Karaca, H. Onay , O. Ardeniz, G. Aksu, F. Genel, N.Kutukculer, F. Ozkinay ASHG 2016 Annual Meeting 16-22 Ekim 2016 Vancouver/Kanada
19. A review of a 20-year experience with prenatal diagnosis records, 9,297 cases from Turkey. H. Bolat, B. Durmaz, Z. Cengisiz, E. Karaca, A. Durmaz, A. Aykut, E. Pariltay, T. ve ark. ASHG 2016 Annual Meeting 16-22 Ekim 2016 Vancouver/Kanada
20. E. Pariltay, A. Aykut, O. Cogulu, F. Ozkinay. A case of Nager Syndrome, IVth Dysmorhology Days, Craniorare Educational Contribution konferansı dahilinde , "IVth Dysmorhology Days, Craniorare Educational Contribution", bildiri kitapçığındaki "A case of Nager Syndrome", 2-2 pp.,İstanbul, Türkiye, Nisan, 2009 (Sözlü sunu)
21. Ö. Çoğulu, E. Parıltay, A. Aykut, A. Alpman, C. Gündüz, B. Özbaran, H. Aydın, S. Erermiş, C. Aydın, F. Özkınay, 20. Ulusal Çocuk ve Ergen Ruh Sağlığı ve Hastalıkları Kongresi konferansı dahilinde, "20. Ulusal Çocuk ve Ergen Ruh Sağlığı ve Hastalıkları Kongresi", bildiri kitapçığındaki "Minör Anomalilerle Birlikte Mental Retardasyon ve Erken Başlangıçlı Şizofreni Tanılarıyla İzlenen Bir Olguda 22q11.2 de Tüm Genom Analizi ile Segmental Uniparental İsodizominin Gösterilmesi", 140-140 pp.,Bodrum, Muğla, Nisan, 2010 (Poster sunum)

22. E. Parıltay, H. Akın, E. Karaca, A. Durmaz, B. Durmaz, A. Aykut, A.N. Güleçoğlu, H. Bolat, Ö. Çoğulu, " 11. Ulusal Tıbbi Genetik Kongresi" bildiri kitapçığındaki "X Kromozomu Mozaiklikleri; Sitogenetik ve FISH Sonuçlarının Karşılaştırılması", 110-110 pp., İstanbul, Türkiye, Eylül, 2014 (Poster sunu)

Çalıştığı Projeler:

1. Emziren anne ve bebeklerindeki Vitamin D reseptör polimorfizimlerinin anne saç kökü hücrelerinde ekspresyonları, anne sütünde ve bebek serumundaki vitamin D düzeyleri ilişkisi
2. İdiyopatik Mental Retardasyonlu Olguların Array-CGH ile Değerlendirilmesi
3. Peptik ülserli ohgularda HRH2 gen polimorfizmlerinin araştırılması
4. Düşük Materyalinde Matriks Metalo Proteinaz 2 ve 9 Gen Polimorfizmlerinin Araştırılması
5. İnsanda kuyruk oluşumuna yol açabilecek gen bölgelerinin array-CGH yöntemiyle araştırılması

Kurs-Eğitim-Seminer

1. Meet Öğrenci Değişim Programı Heinrich Heine Universitat Düsseldorf Germany 1-30 Ağustos 2003
2. 7th Bioinformatics Course for Molecular Biologist (Bologna University Bertinoro İtalya 18-22 Mart 2007)
3. Kanser Çalıştayı 8 Mart 2010 İzmir
4. II. Ulusal Fetal Tanı ve Post Mortem Kursu 26-27 Mayıs 2011 Ankara
5. Olympus Mikroskopi Semineri 22 Mayıs 2012 Bursa

Kongre ve Sempozyumlar

1. Ulusal Sinirbilim Ogrenci Kongresi (14-16 Eylül 2001, İzmir Atatürk Kültür Merkezi) (Orgnizasyon Komitesi Üyeliği)
2. Ulusal Moleküler Biyoloji ve Genetik Öğrenci Kongresi (2-4 Mayıs 2003, Ege Üniversitesi Tıp Fakültesi Muhiddin Erel Anfisi) (Orgnizasyon Komitesi Üyeliği)

3. Ulusal Endokrinoloji Öğrenci Kongresi (30 Nisan – 2 Mayıs 2004, Ege Üniversitesi Tıp Fakültesi Muhiddin Erel Amfisi) (Orgnizasyon Komitesi Üyeliği)
4. II. Ege Dahili Tıp Günleri 26-29 Mart 2003 İzmir
5. 1. Ege genetik sempozyumu İzmir Kasım 2005
6. 2. Ege genetik sempozyumu Afyon 24 Kasım 2006
7. 3. Ege genetik sempozyumu Denizli 1 Aralık 2007 (Olgu Sunumu)
8. 57th Annual Meeting, The American Society of Human Genetics, in San Diego, USA September 2007
9. IV. Dismorfoloji Günleri 24-25 Nisan 2009(Olgu Sunumu)
10. European Human Genetics Conference May 2009 Vienna, Austria
11. 5. Ege genetik sempozyumu İzmir 19 Şubat 2010
12. 62nd Annual Meeting, The American Society of Human Genetics, in San Francisco USA September 6-10 November 2012
13. 10. Ulusal Tıbbi Genetik Kongresi 19-23 Aralık 2012 Bursa
14. 11. Ulusal Tıbbi Genetik Kongresi 24-27 Eylül 2014 İstanbul
15. ESHG 2015 6-9 Haziran 2015 Glasgow / İngiltere
16. 12. Ulusal Tıbbi Genetik Kongresi 5-9 Ekim 2016 Çeşme
17. ASHG 2016 Annual Meeting 16-22 Ekim 2016 Vancouver/Kanada
18. ESHG 2018 Milano İtalya
19. Balkan Genetik Kongresi 2019 Edirne

Bilimsel Kuruluşlara Üyelikler :

1. Tıbbi Genetik Derneği
2. American Society of Human Genetics
3. İzmir Tabip Odası
4. Ege Perinatoloji Derneği

İlgi alanı ve tecrubeler

Bioinformaitik, Bilgi teknolojileri ve medikal uygulamalari

Ege Universitesi Tip Fakultesi Ogrenci Bilgisayar Klubu kuruculugu ve uyeligi (1997-2003)

Ege Üniversitesi BITAM"da (Bilgi Iletisim Teknolojileri Arastirma ve Uygulama Merkezi) 2000- 2010 Uzaktan eğitim uygulamaları, Video codec, stream, video on demand uygulamaları, Windows server yönetimi, Linux/Unix işletim sistemleri uygulama ve yonetimi, İmage capture / non-linear video kurgulama, canli video yayini, radyo yayin ve otomasyon sistemleri uygulamalari

Bursa Şevket Yılmaz Eğitim ve Araştırma Hastanesi Sitogentik ve Moleküler Laboratuvarı Kurulumu ve Genetik Hastalıklar Tanı Merkezi Başvurusu