T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

MEFV GEN VARYANTLARININ SINIFLANDIRILMASINDA ORTAYA ÇIKAN BELİRSİZLİKLERİN YAPAY ZEKA ALGORİTMALARIYLA ÇÖZÜMLENMESİ

Mustafa Tarık ALAY

DANIŞMAN
Prof. Dr. Murat KİRİŞÇİ

Biyoistatistik Anabilim Dalı

Biyoistatistik ve Tıbbi Bilişimi Programı

Kasım, 2022

Aileme ithaf ediyorum…

BÜTÇE DESTEKLERİ

MEFV GEN VARYANTLARININ SINIFLANDIRILMASINDA ORTAYA ÇIKAN BELİRSİZLİKLERİN YAPAY ZEKA ALGORİTMALARIYLA ÇÖZÜMLENMESİ

Bu tez çalışması için herhangi bir kurumdan bütçe desteği alınmamıştır.

TEŞEKKÜR

Yüksek lisans eğitimim boyunca bana her zaman değerli vaktini ayırarak benimle tecrübe ve deneyimlerini paylaşan, şefkatli ve sabırlı yaklaşımı ile rehber olan, değerli ve çok kıymetli hocam Sayın Prof. Dr. Murat Kirişçi'ye, Lisansüstü eğitimim süresince bilgilerini ve desteğini esirgemeyen tüm değerli hocalarıma,

Hayatım boyunca değerli desteklerini hissettiğim benim için sayısız fedekarlıklarda bulunan annem, çalışmayı ve hesaplamayı bize sevdiren fedekar babam, azmiyle bana örnek olan, yol gösteren ilköğretmenim abim ve ablama, değerli desteğini her zaman yanımda hissettiğim kıymetli amcama ve biyometrinin babası Galton'a

En içten ve kalbi teşekkürlerimi sunarım

Kasım 2022 Mustafa Tarık ALAY

İÇİNDEKİLER

Sayfa No

TEZ KABUL VE ONAYI ..... **ii**
BEYAN ..... **iii**
BÜTÇE DESTEKLERİ ..... **v**
TEŞEKKÜR ..... **vi**
İÇİNDEKİLER ..... **vii**
ŞEKİL LİSTESİ ..... **xi**
TABLO LİSTESİ ..... **xii**
SİMGE VE KISALTMA LİSTESİ ..... **xiii**
ÖZET ..... **xv**
ABSTRACT ..... **xvii**
1. GİRİŞ **1**
2. KAVRAMSAL ÇERÇEVE **3**
2.1. Yapay Zeka 3
2.1.1. Tarihçe 3
2.1.2. Yapay Zeka ve Kapsamı 3
2.1.3. Yapay Zeka ve İnsan Aklı 4
2.2. Makine Öğrenmesi'ne Giriş 5
2.2.1. Veri 5
2.2.2. Tanım ve Amaç 5
2.2.3. Kullanım Alanları ve Biyolojik Bilimlerdeki Yeri 5
2.2.4. Makine Öğrenmesi Aşamaları 5
2.2.4.1. Veri önişleme *6*
2.2.4.2. DAA *6*
2.2.4.3. TBA *6*
2.2.4.4. DAA ile MANOVA arasındaki benzerlikler ve farklılıklar *6*
2.2.4.5. DAA ve Logistik regresyon arasındaki benzerlik ve farklılıklar *7*
2.2.4.6. DAA ve TBA arasındaki benzerlikler ve farklılıklar *7*
2.3. Makine Öğrenmesi Tipleri 7

2.3.1. Denetimli Öğrenme 7
2.3.1.1. Destek Vektör Makineleri *8*
2.3.1.2. K-en Yakın Komşuluk *8*
2.3.1.3. Logistik Regresyon *8*
2.3.1.4. Naive Bayes *9*
2.3.1.5. Karar Ağaçları *9*
2.3.1.6. Rasgele Ormanlar *10*
2.3.2. Denetimsiz öğrenme 10
2.3.3. Yarı denetimli Öğrenme 11
2.3.4. Pekiştirmeli Öğrenme 11
2.4. Makine öğrenmesi ile ilişkili temel kavramlar 11
2.4.1. Model genellenebilirliği, overfitting ve underfitting problemleri 11
2.4.2. Boyut indirgeme 12
2.4.3. Randomizasyon ve Cross-Validation 12
2.4.3.1. Randomizasyon *12*
2.4.3.2. K-kat Cross Validasyon *12*
2.4.4. Performans Ölçütleri 12
2.4.5. Kısıtlılıkları 13
2.4.5.1. Makine Öğrenmesi yaklaşımlarında yaşanan zorluklara bir örnek: Multi-omik Yaklaşımlarda karşılaşılan zorluklar *14*
2.5. Ailevi Akdeniz Ateşi 14
2.5.1. Genel Bilgiler 14
2.5.2. Epidemiyoloji 14
2.5.3. Genetik ve Kalıtım 15
2.5.4. Genetik Testler 15
2.6. Yeni Nesil Dizileme Teknikleri 15
2.6.1. İnsan Genomunun Yapısı 15
2.6.2. Yeni Nesil Dizileme Teknikleri 15
2.7. Biyoinformatik veritabanları 16
2.8. Varyantların Sınıflandırılması ve Tahmin Edilmesinde kullanılan Kılavuzlar 16
2.9. *İn silico* Tahmin Araçları 20
2.9.1. KORUNMUŞLUK SKORLARI 21
2.9.1.1. GERP++ SKORU *21*
2.9.2. Aminoasit Değişimi Tahmin Skorları 22
2.9.2.1. Sorting Intolerant From Tolerant (SIFT): *22*

2.9.2.2. PROVEAN *22*
2.9.2.3. Polyphen-2 ..... *22*
2.9.2.4. Mutation Assessor *22*
2.9.2.5. MutationTaster: *23*
2.9.3. METAPREDİKTÖRLER 23
2.9.3.1. CADD SKORU *23*
2.9.3.2. DANN SKORU *24*
2.9.3.3. ClinPred *24*
2.9.3.4. MetaLR ve MetaSVM *24*
2.9.3.5. IMHOTEP *24*
2.9.3.6. REVEL SKORU *25*
2.9.3.7. MISTIC SKORU *25*
2.9.3.8. CAPICE SKORU *25*
2.9.3.9. RENOVO *26*
2.10. Makine öğrenmesinin varyantların sınıflandırılmasında kullanılması 27
2.10.1. MEFV 27
2.10.2. BRCA 27
2.11. Araştırmanın Amacı ve Türü 27
3. YÖNTEM 28
3.1. Veri kümelerine erişim 28
3.2. Skorların Elde Edilmesi 28
3.2.1. Franklin Genoox 28
3.2.2. dbNSFP 4.0 28
3.3. İstatistiki Karşılaştırmalar 29
3.4. Makine Öğrenmesi Yöntemleri 29
4. BULGULAR 30
4.1. Tanımlayıcı İstatistikler 30
4.1.1. Revel Tanımlayıcı İstatistikler 30
4.1.2. SIFT Tanımlayıcı İstatistikler 31
4.1.3. FATHMM Tanımlayıcı İstatistikler 31
4.1.4. MetaLR Tanımlayıcı İstatistikler 32
4.2. ANOVA test sonuçlarının Değerlendirilmesi 33
4.3. ANOVA sonrası Posthoc test sonuçlarının değerlendirilmesi 34
4.4. Makine Öğrenmesi Yöntemleri 36

4.4.1. Eğitim Setinin Değerlendirilmesi 36
4.4.1.1. K en yakın komşuluk *36*
4.4.1.2. Rasgele Ormanlar ve Karar Ağaçları *37*
4.4.1.3. Destek Vektör Makineleri *38*
4.4.1.4. DAA *40*
4.4.1.5. Skorların Sınıflandırmaya Etkileri *41*
4.4.2. Prediksiyon Sonuçlarının Değerlendirilmesi 43
5. TARTIŞMA **46**
6. SONUÇ VE ÖNERİLER **50**
KAYNAKLAR **51**
EK-1 ..... **58**
İNTİHAL RAPORU İLK SAYFASI **64**
ETİK KURUL İZİN YAZISI **65**
KURUM İZNİ YAZILARI **66**
ÖZGEÇMİŞ **67**

ŞEKİL LİSTESİ

Sayfa No

Şekil 1. Yapay zeka, Makine öğrenmesi ve Derin öğrenmenin arasındaki ilişkinin görselleştirilmesi....4

Şekil 2. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (Revel) 30

Şekil 3. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (SIFT) 31

Şekil 4. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (FATHMM)....32

Şekil 5. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (MetaLR) 33

Şekil 6. Kruskall Wallis-H testinde istatistiki açıdan anlamlı farklılık bulunan skorların post-hoc değerlendirilmesi 35

Şekil 7. KNN en uygun komşuluk sayısının gösterimi 36

Şekil 8. Konfüzyon matrisi (KNN) 36

Şekil 9. RO ve KA konfüzyon matrisinin gösterilmesi....37

Şekil 10.Destek Vektör Makinesi yöntemlerine ait Konfüzyon matrisleri 38

Şekil 11. DAA Konfüzyon Matrisi....40

Şekil 12. Skorların sınıflandırmaya olan etkileri.....41

Şekil 13. Skorların katkısının ısı haritası şeklinde gösterilmesi....42

Şekil 14. Skorlar arası korelasyon grafiği 42

TABLO LİSTESİ

Sayfa No

Tablo 1. Karar ağaçları ile ilgili bazı kavramların gösterimi 9

Tablo 2. Makine öğrenmesi değerlendirme ölçütleri..... 13

Tablo 3. Varyantların veri tipi ve gücüne göre sınıflandırılması 17

Tablo 4. Tahmin skorlarına göre ANOVA test sonuçları..... 33

Tablo 5. Tukey testi kullanılarak gruplar arası farklılığın değerlendirilmesi (Revel)..... 34

Tablo 6. Tukey testi kullanılarak gruplar arası farklılığın değerlendirilmesi (SIFT)..... 34

Tablo 7. Tukey testi kullanılarak gruplar arası farklılığın değerlendirilmesi (MetaLR)..... 35

Tablo 8. KNN eğitim verisetine ait performans metriklerinin gösterilmesi..... 37

Tablo 9. Eğitim verisetinde kullanılan RO ve KA'nın performans metriklerine ait sonuçlar.. 38

Tablo 10. Polinomiyal Destek Vektör Makinesine ait Performans Metrikleri..... 39

Tablo 11.Doğrusal DVM'ye ait Performans Metrikleri 39

Tablo 12. RBF'e ait Performans Metrikleri 40

Tablo 13. DAA'ya ait performans metrikleri 41

Tablo 14. Prediksiyon sonuçları 43

Tablo 15. Makine öğrenmesi yöntemi prediksiyon sonuçlarının skorlara göre ortalamalarının dağılımı 44

Tablo 16.Makine öğrenmesi yöntemi prediksiyon sonuçlarının skorlara göre standart sapmalarının dağılımı 45

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar	Açıklama
ACMG	: American College of Medical Genetics
B	: Benign
DAA	: Doğrusal Ayırma Analizi
DDBJ	: Japonya DNA Bilgi Bankası
DSÖ	: Dünya Sağlık Örgütü
DVM	: Destek Vektör Makinesi
EMBL	: Avrupa Moleküler Biyoloji Laboratuarları
FMF	: Ailevi Akdeniz Ateşi
Genbank	: Amerikan Ulusal Biyoteknoloji Merkezi
GBM	: Gradient Boosting Machine
HMM	: Hidden Markov Modeli
KA:	: Karar Ağaçları
LB	: Olası Benign
LP	: Olası Patojenik
LR	: Logistik Regresyon
MANOVA	: Çok Değişkenli Varyans Analizi
MEFV	: Ailevi Akdeniz Ateşi
ML	: Makine Öğrenmesi
NC	: Kategorize Edilmemiş

P : Patojenik

RO : Rasgele Orman

TBA : Temel Bileşen Analizi

VIMs : Değişken Önem Ölçütleri

VUS : Klinik Önemi Bilinmeyen

YSA : Yapay Sinir Ağları

ÖZET

YÜKSEK LİSANS TEZİ

MEFV GEN VARYANTLARININ SINIFLANDIRILMASINDA ORTAYA ÇIKAN BELİRSİZLİKLERİN YAPAY ZEKA ALGORİTMALARIYLA ÇÖZÜMLENMESİ

Mustafa Tarık ALAY

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Biyoistatistik Anabilim Dalı

Biyoistatistik ve Tıbbi Bilişimi Programı

Danışman : Prof. Dr. Murat KİRİŞÇİ

Giriş: Ailevi Akdeniş Ateşi(FMF), ataklar halinde seyreden tekrarlayan ateş ve seröz dokularda enflamasyonla karakterize MEFV genindeki fonksiyon kazandırıcı mutasyonlar kaynaklı olarak meydana gelen otoinflamatuar bir hastalıktır. Bildirilen MEFV gen varyantlarının yarıdan fazlasının klinik önemi bilinmemektedir. Makine öğrenmesi klinik önemi belirlenemeyen varyantların sınıflandırılmasında önemli faydalar sağlayabilir.

Materyal-Method: Infevers veri tabanında yer alan *MEFV* missense ve sessiz 266 varyant dört kategoriye ayrıldı: olası patojenik (LP), olası benign (LB), unresolved, önemi bilinmeyen varyant. Eğitim verisi olarak LP ve LB varyantlar seçildi. Diğer varyantlar belirsiz olarak kabul edildi ve test verisi olarak tercih edildi. Revel, SIFT, MetaLR ve FATHMM skorları 7 makine öğrenmesi yöntemine tabi tutuldu. Sonrasında bu skorlar basamaklı bir bütün makine öğrenmesi yöntemlerinin etkisinin değerlendirildiği bir skorlamaya göre değerlendirildi. Buna göre 4 ve 5 puan alanlar belirsiz geri kalanlar ise doğru sınıflandırılmış olarak kabul edildi.

Sonuç: Regresyon Ağacı ve Rasgele ormanlar eğitim verisini % 100 doğrulukla sınıflandırdı. Yedi makine öğrenmesi metodunun da kullanıldığı basamaklı sınıflandırma metodumuz 168 varyantın 139'unu (%83) yakınını LP ve yada LB olarak sınıflandırdı.

Tartışma: Kullandığımız sınıflandırma metodu *MEFV* varyantlarının sınıflandırılmasında mevcut algoritmalar arasında en iyi sonucu vermektedir. *MEFV* varyantlarının %60'ının önemi bilinmeyen varyant olarak değerlendirildiği göz önüne alındığında mevcut değerlendirme yöntemi klinik bilgilerin eşliğinde klinisyenlere varyant yorumu için daha kesin bir yorumlama imkanı ve kolaylık sağlayabilir. Ancak, kullandığımız sınıflandırma algoritmasının diğer genler üzerindeki etkinliğinin belirlenebilmesi için daha fazla sayıda çalışmaya ihtiyaç vardır.

Kasım 2022 , 87 sayfa.

Anahtar kelimeler: FMF, MEFV, makine öğrenmesi, VUS

ABSTRACT

M.Sc. THESIS

ANALYSIS OF UNCERTANTIES ARISING DURING THE CLASSIFICATION OF MEFV GENE VARIANTS WITH ARTIFICIAL INTELLIGENCE ALGORITHMS

Mustafa Tarık ALAY

İstanbul University-Cerrahpaşa

Institute of Graduate Studies

Department of Biostatistics

Biostatistics and Medical Informatics

Supervisor : Prof. Dr. Murat KİRİŞÇİ

Introduction: Familial Mediterranean Fever (FMF) is an autoinflammatory disease caused by functional mutations in the MEFV gene, characterized by recurrent fever and inflammation in serous tissues. The clinical significance of more than half of the reported MEFV gene variants is unknown. Machine learning algorithms can provide significant benefits in classifying clinical significance unknown variants.

Material-Method: MEFV missense and silent overall 266 variants in the Infevers database were divided into four categories: likely pathogenic (LP), likely benign (LB), unresolved, and variation of unknown significance. While LP and LB variants were selected as training data, others were considered indeterminate and preferred as test data. Revel, SIFT, MetaLR, and FATHMM scores were subjected to 7 machine-learning methods. These scores were then evaluated according to a scoring system in which the effect of all cascading machine-learning methods was evaluated. Accordingly, those who scored 4 and 5 were considered indeterminate, and the rest were correctly classified.

Result: Regression Trees and Random forests classified the training data with 100% accuracy. Our classification method classified 139 (83%) of 168 variants as LP or LB.

Discussion: Our classification method provides the best results among existing algorithms for classifying MEFV variants. Considering that 60% of MEFV variants are evaluated as variants of unknown significance, the current evaluation method may provide clinicians with a more precise interpretation opportunity and convenience for variant interpretation in the light of clinical information. However, more studies are needed to determine the effectiveness of our classification algorithm on other genes.

November 2022, 87 pages.

Keywords: FMF, MEFV, Machine Learning, VOUS

1. GİRİŞ

Makine öğrenmesi çok boyutlu veri kümelerini kullanarak tahmine dayalı ürettiği veri kümelerini test ederken mümkün olan en doğru tahmini yapmayı amaçlar. Makine öğrenmesi sayesinde çok uzun süren ve yüksek maliyet gerektiren karmaşık deneyler içeren çalışmaların ortaya koydukları sonuçlar daha kısa sürede ve düşük maliyetle gerçek değerlerine çok yakın bir şekilde, kabul edilebilir hata sınırları içinde tahmin edilebilmiştir. Makine öğrenmesi, biyolojik bilimlerde yeni ilaçların keşfinde, protein yapılarının 3 boyutlu modellemelerinin yapılmasında, türler arası dizi benzerliklerinin saptanmasında, yeni genlerin bulunmasında, hastalıkların teşhis, prognozu ve tedavisinin tahmin edilmesinde yaygın olarak kullanılmaktadır.

Genetik varyantların sınıflandırılmasıyla ilgili ilk kılavuz American College of Medical Genetics(ACMG) tarafından 2015 yılında yayınlanmış ve içerisinde yer alan kriterler ACMG kriterleri olarak adlandırılmıştır. Bu tarihten sonra bu kılavuzun klinik kullanımda varyantları sınıflandırmada yeterli olmadığı görülmüş ve kılavuzda birtakım güncellemeler yapılmıştır. Bu yetersizliklerden birisi bu kriterlerin gene özgü değerlendirme yapmaktan henüz çok uzak olmalarıdır . Her ne kadar varyantların değerlendirilmesi ile ilgili Clingen gibi gruplar bazı gen ve hastalıklara özgü değerlendirme tavsiyelerinde bulunuyor olsalar da insanda hastalık yapıcı varyant sayısı oldukça fazladır ve halen tespit edilen varyantların önemli bir kısmının klinik önemi bilinmemektedir. Klinik önemi bilinmeyen varyant sayısının oldukça fazla saptandığı bu genlerden birisi *MEFV* genidir. Infevers veri tabanına göre içerdiği varyantların yarısından fazlası sınıflandırılamamıştır. Makine öğrenmesi klinik önemi belirlenemeyen *MEFV* gen varyantların sınıflandırılmasında önemli faydalar sağlayabilir.

2020 yılında yayınlanan bir makalede, REVEL skoru kullanılarak makine öğrenmesi metotları kullanılarak sınıflandırılamayan ya da önemi bilinmeyen varyantlar sınıflandırılmaya çalışılmıştır. Patojenik ve olası patojenik varyantlar olası patojen ve olası benign ve olası benign varyantlar olası benign olarak kabul edilmiştir. Olası benign ve olası patojenik varyantlar eğitim seti olarak kullanılarak en doğru sınıflandırmayı yapan makine öğrenmesi algoritması tahmin edilmeye çalışılmıştır. Buna göre en yüksek doğruluk oranı Doğrusal Ayırma Analizinde (DAA) saptanmış olup bundan sonra DAA kullanılarak yeni bir eşik değeri belirlenmiştir. Buna

göre varyantlar üç grup altında yeniden sınıflandırılmıştır: olası patojen, önemi bilinmeyen ve olası beningn. Yeniden sınıflandırılan varyantlar tekrar değerlendirildiğinde ACMG kriterlerine göre klinik önemi belirlenemeyen varyant oranı %61,6'dan %17,6'ya düşmüştür. Ancak, bu çalışmada kullanılan modelin sensitivitesi yeterli olmayıp spesifitesi düşüktür. Sabit bir eşik değer içerdiği için klinik önemi belirsiz yeni bir varyant eklendikçe sınıflandırma başarısının artışı beklenmemektedir. Bu durum eklenen patojenik ya da benign olduğu kanıtlanmış her bir yeni varyant için Alıcı işlem karakteristik (ROC) eğrisi hesaplanması ve eşik değer belirlenmesi ile çözüme kavuşturulabilir. Ancak, bu durum pratik kullanım için uygun değildir. Ayrıca, eklenen varyantların patojenik ya da benign olarak değerlendirilmesi göre tekrar bir eşik değer hesaplansa dahi bu durumun eklenen varyantın patojenik ya da benign olmasına göre yanlılık oluşması ihtimali de bulunmaktadır. Bir diğer kısıtlılık ise çalışmada sadece tek bir skor (REVEL) en başarılı makine öğrenmesi metodu (DAA) kullanılarak değerlendirilmektedir. Ancak, REVEL skoru klinikte kullanılan tek skor değildir. *MEFV* gen varyantlarını sınıflandırmasında optimum sayıda skor kullanılarak daha yüksek doğruluk oranı elde edilebilir. Ayrıca, tek bir makine öğrenmesi algoritması üzerinden sınıflama yapmak her zaman için doğru sonucu vermeyebilir. Bu nedenle klinik önemi bilinmeyen *MEFV* varyantlarının sınıflandırılması konusunda yeni çalışmalara ihtiyaç vardır.

Bu tezde infevers veri tabanında bulunan tek nükleotid değişimlerinin bulunduğu *MEFV* gen varyantlarının bir çok *in silico* tahmin araçları kullanılarak yeniden yüksek bir doğruluk oranıyla tahmin edilmesi, sınıflandırılamayan varyantların sınıflandırılarak hastalığın tanı,izlem ve tedavisine yardımcı olması amaçlanmaktadır. *MEFV* gen varyantlarının sınıflandırılmasında birden fazla skorun hem birbirleriyle, hem de kombine bir halde çeşitli makine öğrenmesi metotları kullanılarak karşılaştırıldığı bir çalışma bulunmamaktadır. Bu tezde optimum sayıda veritabanı ve en ideal makine öğrenmesi yöntemi kullanılması amaçlanarak *MEFV* geninde yer alan klinik önemi belirsiz varyantların sınıflandırılması amaçlanmaktadır.

2. KAVRAMSAL ÇERÇEVE

2.1. Yapay Zeka

Yapay zeka terimi ilk defa John McCarthy tarafından zeki makineler yapma bilimi ve mühendisliği olarak tanımlanmıştır (1). Yapay zeka kavramının tanımlanması yeni olsa da insansı rabotlar fikri oldukça eski dönemlere antik Çin'e dayanmaktadır. İnsansı robotlar terimi İslam dünyasında ilk defa 12. yüzyılda Cezeri ve Avrupa'da 14.yüzyılda Leonardo da Vinci tarafından kullanılmıştır (2).

2.1.1. Tarihçe

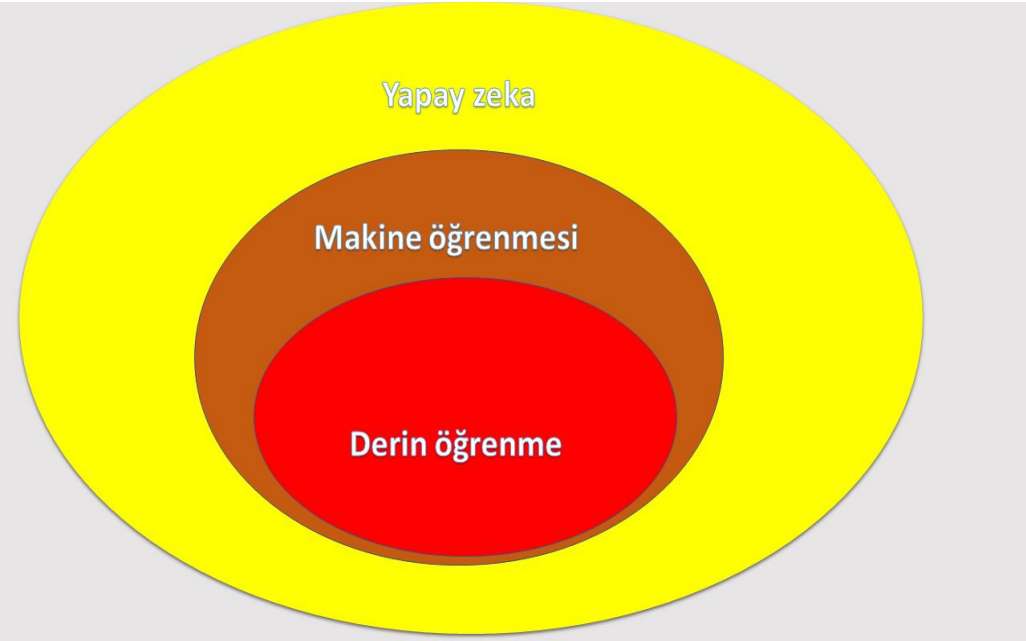
Alan Turing 1940'larda geliştirdiği algoritmalar yardımıyla II.Dünya savaşı sırasında Alman ordusunun şifrelerini çözen makineyi oluşturmayı başarmıştır (3). 1948'de ise John von Neumann'ın kendi kendini kopyalayabilen makine fikrini ortaya atmıştır (4). Alan Turing, John von Neumann'ın kendi kendini kopyalayabilen makine fikrini ortaya atmasından iki yıl sonra yayınladığı makalesinde akıllı makinelerin özelliklerini tanımlayan bir test oluşturmuştur (3). Bu test sonraki yıllarda Turing Testi olarak tanımlanmıştır. 6 yıl sonra John McCarthy tarafından "Yapay Zeka" kavramı ilk defa akıllı makineler yapma bilimi ve mühendisliği olarak tanımlanmıştır (5). Makine öğrenmesinin yaygınlaşması deterministik ve stokastik yaklaşımların formülize edilmesiyle 1960'lı yıllara rastlamaktadır (6). Bununla birlikte, yapay zekanın bir dalı olarak kabul edilen yapay sinir ağları fikrinin ortaya çıkışı ise daha öncesine McCulloch ve Pits 1943 yılında yayınladıkları bir makalede insan nöronlaryla karşılaştırılabilen bir algoritmaya sahip bilgisayarları tanımlamasına dayanmaktadır (7).

2.1.2. Yapay Zeka ve Kapsamı

Yapay zeka, günümüzde tıptan bilgisayar mühendisliğine, eğitimden uçak yapımına, güvenlikten web tasarımcılığına kadar insan aklının alabileceği neredeyse bütün disiplinlerde kullanılmaya başlanmıştır. Yapay zekanın alt disiplinleri arasında robotik, doğal dil işleme, makine öğrenmesi, uzman sistemler ve "kognitif" modelleme gibi bir çok alan yer almakta olup matematik, felsefe, bilgisayar ve psikoloji gibi birçok disiplinle etkileşim halindedir (8).

Günlük hayatta makine öğrenmesi, yapay zeka ve derin öğrenme gibi kavramlar sıklıkla birbiri yerine kullanılabilmektedir. Ancak, bu kavramlar her ne kadar birbiriyle ilişkili olsalar

da kapsamları ve sınırlılıkları ile birbirlerinden ayrılmaktadırlar. Bu bağlamda, yapay zeka oldukça kapsamlı olup, makine öğrenmesi ve derin öğrenme yapay zekanın sadece bir bölümünü oluşturumaktadır. Derin öğrenme aynı zamanda makine öğrenmesinin de bir alttipidir (9). Yapay zeka, makine öğrenmesi ve derin öğrenme arasındaki ilişki Şekil 1.'de görselleştirilmiştir.



Şekil 1. Yapay zeka, Makine öğrenmesi ve Derin öğrenmenin arasındaki ilişkinin görselleştirilmesi

2.1.3. Yapay Zeka ve İnsan Aklı

Bilgisayarlar insan zekasının altından kolaylıkla kalkamayacağı kompleks hesaplamaları çok kısa sürede yapabilir. Ancak idrak yeteneğinden yoksundurlar. Bu sebeple hem insanın idrak yeteneğini taklit edebilecek hem de hızlı işlem yapacak makinelere ihtiyaç vardır. Yapay zeka teknolojisi bu ihtiyacı karşılamak için geliştirilmiştir. Özellikle de teknolojinin gelişmesi, bilgisayarların çok hızlı işlemler yapabilmesi nedeniyle yapay zeka günümüzde oldukça popüler hale gelmiştir. Ancak, yapay zekadaki gelişmeler ne kadar devasa olursa olsun yapay zeka insan zekasının bir ürünüdür ve henüz insanın idrak yeteneğini taklit etmekten oldukça uzaktadır(10).

2.2. Makine Öğrenmesi'ne Giriş

2.2.1. Veri

Veri bilginin en ilkel halidir. İşlenmiş veriye ise informatik denilir. Bilginin en üst düzey hali ise farkındalık yani bilgelik olarak kabul edilmektedir. Bu duruma ulaşan insanın veriyi özümsediği kabul edilmektedir(11).

2.2.2. Tanım ve Amaç

Makine öğrenmesi çok boyutlu veri kümelerini kullanarak tahmine dayalı modeller üretir. Sonraki aşamada, üretilen bu modeller test edilir. Bütün bu işlemler yapılırken amaç mümkün olan en doğru tahmini yapmaktır.

2.2.3. Kullanım Alanları ve Biyolojik Bilimlerdeki Yeri

Teknolojide yaşanan çığır açıcı değişimlerden sonra makine öğrenmesine duyulan ihtiyaç daha da artmıştır. Bu durum, makine öğrenmesini günlük yaşamın neredeyse her alanında kullanılır hale getirmiştir. Biyolojik bilimler de bu gelişmelerin etkisinde kalmış ve makine öğrenmesi yeni ilaçların keşfi, protein yapılarının 3 boyutlu modellerinin, türler arası dizi benzerliklerinin saptanması, yeni genlerin bulunması, hastalıkların teşhisi, prognozu ve tedavisinin tahmin edilmesi gibi birçok konuda kullanılmaya başlanmıştır. Çok uzun süren ve yüksek maliyet gerektiren bir çok karmaşık deney gerektiren çalışmaların ortaya koydukları sonuçlar makine öğrenmesi algoritmaları ile daha kısa sürede ve düşük maliyetle gerçek değerlerine çok yakın bir şekilde, kabul edilebilir hata sınırları içinde tahmin edilebilmiştir. Buna örnek olarak, makine öğrenmesi yardımıyla deneysel çalışmaların *in silico* olarak modellenebilmesi verilebilir. (Bu durum, *in vivo* ya da *in vitro* olarak tasarlanan deneysel çalışmaların yüksek maliyetleri ve uzun süreleri göz önüne alındığında muazzam bir gelişmedir)(12).

2.2.4. Makine Öğrenmesi Aşamaları

Farklı makine öğrenmesi metotları olsa da bütün makine öğrenmesi metotlarında izlenen aşamalar 5 ana başlık altında toplanabilir(13–15).

(1) Veri Temini

(2) Veri Önişleme

(3) Model Seçimi

(4) Model Eğitimi

(5) Model Değerlendirilmesi

2.2.4.1. Veri önişleme

Veri ile ilgili yapılacak olan işlemlerin optimum sonucu verebilmesi için verinin işlem yapılabilecek en uygun hale getirilmesi gerekmektedir. Bu nedenle verinin doğru kürete edilmesi, kürete edilen verilerin ise doğru bir şekilde etiketlenmesi gerekmektedir. Ancak, veri iyi bir şekilde kürete edilse ve doğru bir şekilde etiketlense dahi bu durum verinin ideal bir şekilde etiketlendiği manasına gelmez. Kullanılacak veriyle ilgili grupların sayısının arttırılması ya da azaltılması gerekebilir. Bu nedenle, veri ön işlemede her bir değişken için gereken optimum sayıda sınıfı bulmak ya da kümeleme yapmak için bazı metotlar kullanılır. Bu metotlara örnek olarak Temel Bileşen Analizi (TBA) ve Doğrusal Ayırma Analizi (DAA) verilebilir**.**

2.2.4.2. DAA

DAA boyut indirgemede sıklıkla kullanılan gözetimli bir makine öğrenmesi metodudur. Grupları sınıflandırmak amacıyla kullanılır. Çok boyutlu veriyi daha az sayıda boyuta indirger. Bu sayede varyanslar arası farklılıkları minimum ancak ortalamalar arası farkları maksimum yapar. DAA ayırma analizi birden fazla sürekli değişken ve iki grup için kullanılırsa iki gruplu ayırma analizi olarak isimlendirilirken, birden fazla sürekli değişken ve ikiden fazla grup için kullanıldığında ise çoklu ayırma analizi olarak isimlendirilir.

Ayırma analizi iki aşamalı bir işlem olarak ele alınabilir. İlk aşamada ayırma fonksiyonlarının önemlilik testlerine bakılır, ikinci aşamada ise sınıflama yapılır. DAA'nın çalışma prensipleri 3 aşamada incelenebilir. Öncelikle gruplar arası ayrılabilirlik ölçülmelidir. Sonrasında her bir sınıfın ortalamaları arası farklar hesaplanmalıdır. En düşük boyutlu uzayı oluşturmak için gruplar arası varyans maksimum grup içi varyans ise minimum tutulmalıdır.

2.2.4.3. TBA

TBA boyut indirgemede sıklıkla kullanılan gözetimsiz bir makine öğrenmesi yöntemi olup grupları kümelemek amacıyla kullanılır. Gruplar arası varyansı maksimum yapmayı hedefler.

2.2.4.4. DAA ile MANOVA arasındaki benzerlikler ve farklılıklar

MANOVA'da bağımsız değişkenler gruplar, bağımlı değişkenler ise her grupta yer alan değişkenlerdir. MANOVA istatistiksel anlamlılık verir. Ancak, MANOVA'da elde edilen sınıflandırma yüksek oranda anlamlı bir sınıflandırma anlamına gelmez. MANOVA yalnızca

gruplara yapılacak olan doğru sınıflandırmayı şans eseri meydana gelip gelmemesine göre değerlendirir(16).

2.2.4.5. DAA ve Logistik regresyon arasındaki benzerlik ve farklılıklar

Ayırma analizinde bağımsız değişkenlerin sayısal veri türünde olması, varyans-kovaryans matrislerinin homojen olması ve gruplar arasındaki değişkenlerin çok değişkenli normal dağılıma uyması şartlarını sağlaması beklenir. Bu şartlar logistik regeresyonda aranmaz(17).

2.2.4.6. DAA ve TBA arasındaki benzerlikler ve farklılıklar

Her iki metot da boyut indirgeme için kullanılsa da metotlar arası birtakım farklılıklar mevcuttur. TBA'da varyanslar arası farklılıklar maksimum yapılmaya çalışılırken, DAA'da gruplar arası ortalamaların farkları maksimum ancak grup içi varyans farklılıkları minimum tutulmaya çalışılır. DAA bir gözetimli öğrenme metodu iken TBA gözetimli öğrenme metodu değildir.

2.3. Makine Öğrenmesi Tipleri

Makine öğrenmesi yöntemleri çok çeşitli metotlardan oluşmaktadır. Bu yöntemler arasında en yaygın kullanılanlar denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeli öğrenmedir. Etiketli verilerin kullanıldığı denetimli öğrenme tiplerinde modelin başarısı test edilebilirken, etiketlerin kullanılmadığı denetimsiz öğrenme tiplerinde ise modelin başarısı test edilememekte, ancak tespit edilelen örüntüler ve bu örüntülerin bağlı olduğu örnekler üzerinden bir takım yorumlar yapılabilmektedir(18,19).

2.3.1. Denetimli Öğrenme

Denetimli öğrenmede makineye öğretilmek istenen model etiketli örnekler ile sunulur. Bu öğrenme tipinde her bir veri kümesinde yer alan her örnek için bir etiket bulunmaktadır. Bu etiketler kullanılarak makineye bir model öğretilir. Daha sonra bu öğretilen model kullanılarak yeni ve etiketi bilinmeyen örnekler tahmin edilir. Modelin öğretilmesi için kullanılan veriler eğitim kümesini oluştururken, öğretilen bu modelin başarısının değerlendirildiği veriler ise test kümesini oluşturmaktadır. Veri kümesinde kullanılan değişkenler sırasız nitel (nominal) ya da sıralı nitel (ordinal) değişkenler gibi kategorik verilerden oluşuyorsa sınıflandırıcı olurken, sürekli değişkenlerden oluşuyorsa regresyon analiz modeli olur(20). Denetimli öğrenmeye örnek olarak k- en yakın komşuluk (KNN) , destek vektör makineleri (DVM), yapay sinir ağları

(YSA), logistik regresyon (LR), naive bayes (NB), karar ağaçları (KA) , rasgele ormanlar (RO) örnek olarak verilebilir.

2.3.1.1. Destek Vektör Makineleri

DVM düzlem üzerinde yer alan verileri sınıflandırmada kullanılır. Bu modelde veriler doğrusal olarak sınıflandırılabiliyorsa doğrusal çizgi veya hiperdüzlem yardımıyla veriler ayrılır. Ancak, veriler doğrusal olarak sınıflandırılamıyorsa o zaman doğrusal olmayan destek vektör makine yöntemleri kullanılır. Doğrusal olarak sınıflandırılamayan verileri sınıflandırabilmek için kernel yöntemleri kullanılır (21). Birçok kernel metodu olduğu gibi bu metotlara ait birçok uygulama alanı da bulunmaktadır(22).

DVM kompleks verisetlerinde gözardı edilen paternleri tanıma konusunda oldukça başarılıdır. Kullanım alanı oldukça geniştir. El yazılarının algılanması, sahte kredi kartlarının belirlenmesi, konuşmaların tanınması ve yüzlerin ayırt edilmesi gibi bir çok durumda kullanılmaktadır(23). Bunlar arasında, tıp ve biyoloji de yer almaktadır. Örneğin, periferik kan ya da dokudan alınan DNA ya da RNA örnekleri kullanılarak tümör tanısında ya da klinik seyrin değerlendirilmesinde de kullanılabilmektedir(24).

2.3.1.2. K-en Yakın Komşuluk

Bir örnek nokta seçilir. Bu örneğe en yakın k adet komşu belirlenir. Bilinmeyen örnek olarak seçilen nokta ile diğer tüm noktalar arasındaki uzaklık hesaplanır. Belirlenen örnek noktaya en yakın k adet gözlem seçilir. Veri türüne göre analiz yapılır. Sınıflandırma ya da regresyon hangisi yapılacaksa tahmin değeri için farklı yöntemler kullanılır. Sınıflandırmada, bu örnek k adet noktadan en fazla hangi etiketli veri bulunuyorsa o değerle, regresyonda ise ortalama değeri ile etiketlenir, bu değer tahmin değeri olarak seçilir**.** Bu durum tüm verisetine uygulanır. En büyük kısıtlılığı büyük veri kümeleri için oldukça yavaş çalışmasıdır(13).

2.3.1.3. Logistik Regresyon

Logistik regresyon (LR) sağlık başta olmak üzere bir çok bilim dalında dikotom verilerin analizinde kullanılmaktadır. Overfittinge oldukça yatkındır(25). Yapilan bir sistematik derleme ve meta-analizde klasik logistik regresyon modellerinin diğer makine öğrenmesi metotları ile benzer sonuçlar ortaya koyduğu gösterilmiştir. Ancak, önemli bir dezavantajı gradyan boosting gibi bazı metotlar kullanılarak diğer makine öğrenmesi yöntemlerinin performasında iyileşmeye gidilebilmekte iken bu durum LR için olası gözükmemektedir(26). Çok merkezli olarak tasarlanan bir çalışmada, siğillerde klinik seyrin bozulmasıyla ilgili olarak logistik regresyon ve makine öğrenmesi yöntemleri karşılaştırılmıştır. Buna göre, RO, gradient

boosted machine (GBM), bagged trees (BT) ,DVM, LR, KNN, KA,ve MEWS karşılaştırılmış. Performans açısından en başarılı algoritma RO olarak tespit edilmiştir. LR , KNN le beraber diğer makine öğrenmesi metotlarından daha kötü tahmin başarısına sahip olarak bulunmuştur(27).

2.3.1.4. Naive Bayes

Bayes teoremi bir olasılık teorisidir, ve sıklıkla matematikçiler tarafından kullanılmaktadır. Bu teorem, 18.yüzyılda matematikçi Thomas Bayes tarafından geliştirilmiştir(28). 1960'lı yıllarda, Hall bayes teoreminin tıpta da yaygın olarak kullanılması gerektiğini önermiştir. Yayınladığı makalede doktorların hastalarına bilgi verirken muallak ifadelerden kaçınmasının Bayes teoremi yardımıyla mümkün olabileceğini ifade etmiştir. Ön olasılık, koşullu olasılık ve sonsal olasılık gibi bayes teoremine ait terimleri açıklamış ve tıpta kullanıma dair öneriler sunmuştur(29). Günümüzde, Bayes teoremi kullanılarak hastalığa neden olan faktörler ve bu faktörlerin birbirleri ile olan ilişkisi açıklanabilmekte, hastalık prediksiyonu yapılabilmektedir(30).

Naïve Bayes algoritması Bayes teoremi baz alınarak 1960'lı yıllarda gelişitirilmiştir. Bu yaklaşımın kullanılabilmesi için Bayes varsayımlarına ek olarak tüm öngörücü değişkenlerin birbirinden koşullu olarak bağımsız olduğu şartını sağlaması gerekmektedir. Yani bütün özellikler kendi yüklerini kendileri taşırlar ya da diğer bir deyişle " Her koyun kendi bacağından asılır" şartını sağlamaları gerekmektedir(31). Aslında, bu durum gerçeklikten uzaktır; çünkü bütün özniteliklerin birbirinden bağımsız olması imkansıza yakın bir durumdur. Ancak, neyseki bu metot özniteliklerin birbirine bağımlı olduğu durumlarda dahi iyi sonuçlar ortaya koyabilmektedir(32).

2.3.1.5. Karar Ağaçları

KA nesne ile ilişkili soruların sorarak sınıflandırma yapar. Her bir soru bir düğümün içindedir. İçeride yer alan her bir düğüm noktasından yavru düğüm noktasına kadar sorular yanıtlanır. Bu anlamda sorular hiyeraşik bir düzen içerisinde ilerlemektedir. Bu hiyeraşi karar ağaçları olarak adlandırılmaktadır. Hiyeraşi içerisinde yer alan her bir ağaca "evet","hayır"" şeklinde sorular sorulur ve yanıt beklenir(33). Bir KA üç tip düğümden oluşur. Karar düğümleri, şans düğümleri ve uç düğümler. Karar düğümleri genellikle karelerle, şans düğümleri dairelerle ve uç düğümler ise üçgenlerle temsil edilir.(34)

Tablo 1. Karar ağaçları ile ilgili bazı kavramların gösterimi

Kök düğüm	Bütün örneği temsil eder. Başlangıç noktasıdır. Bu düğüm sonrasında iki ya da daha fazla kümeye ayrılır.
Parçalama	Bir düğümün iki ya da daha fazla düğüme bölünmesidir.
Budama	Parçalama işleminin tam tersidir. Karar düğümünün alt düğümleri kaldırılır.
Karar Düğümü	Bir alt düğüm başka alt düğümlere bölünürse, karar düğümü olarak adlandırılır.
Yaprak Düğümü	Daha fazla bölünmeyen düğümlere yaprak düğümü olarak adlandırılır. Terminal düğüm olarak da ifade edilebilir.
Alt ağaç	Tüm ağacın bir alt kısmı. Şube ağaç olarak da adlandırılabilir.
Ana ağaç	Alt düğümlere ayrılmış olan ağacın kendisi
Yavru ağaç	Ana ağaç tarafından meydana getirilen ağaç, alt düğümlere ayrılmış olan ağacın parçalarından herbirisi

2.3.1.6. Rasgele Ormanlar

RO algoritmalarında bir tahmin kuralı koyulur ve bu tahmin kuralına verilen cevaba göre verilere değerlendirilir ve sıralanır. Her bir prediksiyon için değişken önem ölçütleri(VIMs) otomatik olarak hesaplanır. RF VIMs ler etkileşimde bulunan tahmin edicileri başarılı bir şekilde tahminlediğine inanılmaktadır(35).

Büyük miktarda girdi ve çok sayıda değişken kullanıldığında sıklıkla RO kullanılır. Eksik veri kullanılabilmesi, uç değerlerlerle çalışılabilmesi ve sınıflandırmanın özelliklerini tahmin etmesi gibi bir çok avantajı bulunmaktadır. Her bir KA kökler, düğümler ve yapraklardan oluşur.

2.3.2. Denetimsiz öğrenme

Makineye öğretilmek istenen model etiketler olmadan sunulur. Örnekler arası ilişkiler ve benzerlik kullanılarak bir takım örüntüler tespit edilir ve bu örüntülere göre anlamlı

çıkarımlar yapılır. Amaç, veri kümeleri arasında yer alan gizli örüntülerin tespit edilmesidir. Etiketli veriler bulunmadığı için modelin başarısı test edilememektedir. Yorumlamalar tespit edilen örüntüler ve bu örüntülerin bağlı olduğu örnekler kullanılarak yapılmaktadır. Denetimsiz öğrenmede kümeleme, öznitelik çıkarımı ve boyut indirgeme kullanılır(36). Denetimsiz öğrenme yöntemlerine örnek olarak k-ortalamalar ve hiyeraşik kümeleme örnek olarak verilebilir. Biyolojide sıklıkta gen ifadesi ya da mikrodizin analizlerinin çözümlemelerinde kullanılır(20).

2.3.3. Yarı denetimli Öğrenme

Etiketsiz veriyi toplamak etiketsiz veriye göre daha kolaydır. Bu nedenle bazı verisetlerinde hem denetimli hem de denetimsiz öğrenmenin birlikte kullanıldığı yöntemler bulunur. Bu hibrit öğrenme yöntemine yarı denetimli öğrenme adı verilir(20).

2.3.4. Pekiştirmeli Öğrenme

Bir davranışsal öğrenme yöntemidir. Kullanılan ajan öğretilmek istenen veriyi etki-tepki ile öğrenmektedir. Ajan öğretilmek istenen sisteme bir tepki verir. Bu tepki karşısında olumlu ya da olumsuz bir takım yanıtlar alır. Buna göre çevreyi öğrenir. Karar süresi çevreyi algılama, eylem yapma ve hedefe ulaşma gibi aşamalardan oluşmaktadır.

2.4. Makine öğrenmesi ile ilişkili temel kavramlar

2.4.1. Model genellenebilirliği, overfitting ve underfitting problemleri

Makine öğrenmesi algoritmalarının amacı en doğru sonuca ulaşabilecek veriler üretmektir. Ancak bu algoritmalar aynı zamanda başka verisetleri için de genellenebilir sonuçlar üretebilmelidir. Yani, başka veri kümelerinde de kullanıldığında doğru sonuçlar üretebilmelidir. Bu amaçla model ideal bir kompleksliğe sahip olmaldır. Eğer model ideal bir kompleksliğe sahip değilse underfitting, eğer aşırı kompleks ve bu aşırı kompleks verisetiyle ilişkili yeterli sayıda veri bulunmuyorsa overfitting durumundan söz edilir. Yani yüksek miktarda yanlılık içeren modellerde underfitting yüksek varyans içeren modellerde ise overfittingden söz ederiz(20). Bunu bir örnekle ifade edecek olursak hastanın sabah ve akşam glukoz değerlerinden insülin seviyelerini tahmin etmek istediğimizde bir regresyon eğrisi elde ederiz. Elde edilen bu regresyon eğrisinin grafiği için ideal eğri iki boyutlu polinominal eğri olsun. Eğer bu eğri yedi boyutlu ise o zaman overfitting durumundan ya da tek boyutlu ise o zaman underfitting durumundan söz ederiz. Her ikisi de makine öğrenmesi yöntemlerinin doğruluğunu olumsuz yönde etkiler. Makine öğrenmesinde yaşanan overfitting ve underfitting

problemlerinin çözümünde örnek sayısının arttırılması ya da cross-validation gibi bir çok yöntem kullanılır(19).

2.4.2. Boyut indirgeme

Özellikle multivariate yapıya sahip verikümelerinin hem hesaplamalarında hem de görselleştirmelerinde büyük problemler yaşanır. Bu durum "çoklu boyut laneti" olarak tariflenir. Boyut indirgeme bu problemlerin çözümünde büyük kolaylıklar sağlar. Az sayıda veri ile yapılan hesaplamalar sıklıkla daha hızlıdır. Birbirine yakın veriler bir grup altında toplanır.Böylece zaman ve efordan tasarruf edilir. Doğruluk tahmini artar. İki ya da üç boyutlu bir verinin görselleştirmesi daha kolaydır. Uç veriler daha belirgin olarak gözlenebilir ve denetimsiz öğrenme kullanıldığında kümelenme eğilimi gözlenebilir(20). Boyut indirgemede iki metot kullanılır: feature selection ve feature extraction

2.4.3. Randomizasyon ve Cross-Validation

2.4.3.1. Randomizasyon

Randomizasyon öğrenme sürecinde elde edilen sonuçların veri seçiminden bağımsız olarak elde edildiğinden emin olmak için gereklidir. Bu gerçek dünya problemlerinde sıklıkla yaşanan bir problemdir. Eğer veriseti yeterince büyükse veriler rasgele olarak K parçaya ayrılır. Sonrasında ise ayrılan K adet parça rasgele olarak eğitim ve test setlerine ayrılır. Ancak, verisetleri nadiren yeterli sayıda genişliğe sahiptir. Bu durumda randomizasyon aynı verisetinin farklı şekillerde bölünmesi ile defalarca kullanılarak yapılabilir. Bu yönteme *cross-validation* adı verilir(20).

2.4.3.2. K-kat Cross Validasyon

Veriseti k-kat parçaya ayrılır. Sonrasinda k-1 parça eğitim verisi olarak kalan parça ise eğitim verisi olarak kullanılır. Böylece modelin validasyonu sağlanır. K sayısı arttıkça eğitim serisi daha kararlı hale gelmeye başlar. Ancak, validasyon için kullanılan veri sayısı azalır. Bu nedenle büyük veri setlerinde k-kat cross validasyondan kaçınılmalıdır.

2.4.4. Performans Ölçütleri

Kullanılan modelin performansının belirlenmesinde bir çok metot kullanılmaktadır. Bunlar arasında yaygın olarak kullanılanları Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1 metrikleri ve ROC eğrileri yer almaktadır. Doğruluk bir modelin başarısını ölçmek için en sık başvurulan metrik olup her zaman için tek başına yeterli olmayabilir. Kesinlik (Precision) ise pozitif olarak tahminlediğimiz değerlerin gerçekten kaç

adedinin pozitif olduğunu göstermektedir. Duyarlılık (Recall) ise pozitif olarak tahmin etmemiz gereken işlemlerin ne kadarını pozitif olarak tahmin ettiğimizi gösteren bir metriktir. F1 skoru ise kesinlik ve duyarlılık metriklerinin harmonik ortalamasıdan oluşur.

Özellikle eşit dağılmayan unbiased veri kümelerinde model doğruluğu her zaman için yeterli olmamaktadır. Örneğin nadir hastalıklarda yanlış negatif sonuç istenilmez; çünkü bu durum teşhis edilemeyen bir hastalık anlamına gelir. Daha belirgin bir ifadeyle, zaten nadir görülen bir hastalığa sahip bir bireyi yanlışlıkla sağlıklı olarak sınıflandırmak anlamına gelir. Kesinlik ise yanlış pozitif sonuçların istenmediği durumlarda daha sık olarak karşımıza çıkar. Yanlış negatif verilerin değerlendirmede oluşturduğu yanlılığı engellemek için ise duyarlılık bir metrik olarak karşımıza çıkar. F1 skorunun kullanılmasının en temel sebebi eşit dağılmayan veri kümelerinde hatalı bir model seçimini engellemektedir.

Tablo 2. Makine öğrenmesi değerlendirme ölçütleri

$$\text{Doğruluk(Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Kesinlik(Precision)} = \frac{TP}{TP + FP}$$

$$Duyarlılık(Recall) = \frac{TP}{TP + FN}$$

$$\text{Özgünlük} = \frac{TN}{TN + FP}$$

$$F - skor = \frac{2 * TP}{2 * TP + FN + FP}$$

2.4.5. Kısıtlılıkları

Makine öğrenmesi sınıflandırması sıklıkla ikili sınıflandırma problemlerini çözmeye çalışır. Ancak genellikle, ikiden fazla sınıf varsa bu durumda çoklu sınıflandırma problemi açığa çıkar. Karar ağaçları (KA) gibi veri setleri doğası gereği bu problemleri çözülebilirken SVM gibi metotlarda bu sorun farklı yöntemler kullanılarak çözülmektedir. Bu yöntemler genellikle çoklu sınıflandırma problemini çözmek için veriyi çoklu ikili sınıflamaya dönüştürmektir. Bu durum bir sınıfın karşısında diğer sınıfların yer alması şeklinde de açıklanabilir. Böylece bütün ikili sınıflamalar değerlendirilir ve en yüksek güven aralığı olan sınıflama tercih edilir. Tek sınıf dağılım kestirimi ise sınıflar arasında uyumsuzluk

bulunduğunda kullanılır. Örnğin, nadir hastalıkların tahmininde bir uyumsuzluk bulunur; çünkü araştırılan hastalık sağlıklı insanlara göre oldukça nadir olarak gözlenir(20). Bu veri bu haliyle eğitime tabi tutulduğunda doğru sonuç elde etmek olası değildir. Bu nedenle doğru ağırlıklandırma yapılmalıdır. Standart makine öğrenmesi resim tanıma ve konuşmayı algılama gibi bazı problemlerin çözümünde zorluklar yaşamaktadır. Bu sebeple alternatif yaklaşımlar geliştirilmiştir. Bu amaçla sırasıyla tek katmanlı yapay sinir ağları, çok katmanlı yapay sinir ağları ve derin öğrenme algoritmaları oluşturulmuştur(37).

2.4.5.1. Makine Öğrenmesi yaklaşımlarında yaşanan zorluklara bir örnek: Multi-omik Yaklaşımlarda karşılaşılan zorluklar

Verilerin heterojen olması, az olması, uç veriler, eskik veriler, sınıflar arası uyumsuzluklar, overfitting problem, veriden daha fazla özellik olması, hesaplamaların ve hesaplanan verileri depolamanın maliyetlerinin yüksek olması multi-omik yaklaşımlarda karşılaşılan zorluklar arasında yer alır. Bir diğer zorluk ise hangi algoritmanın hangi durum için daha iyi çalışacağını belirlemeye dayanmaktadır. Yukarıda zorluklar aşılsa dahi multi-omik bir çalışma için genomik, transkriptomik, proteomik ve metabolomik verilerin birleştirilmesi gerekmektedir. Bu birleştirme metotları kabaca üç gruba ayrılır: birleştirmeye dayalı, modele dayalı ve dönüşüme dayalı. Bütün bu problemlerin çözümünde multidisipliner bir yaklaşım faydalı olacaktır (38).

2.5. Ailevi Akdeniz Ateşi

2.5.1. Genel Bilgiler

Ailevi Akdeniş Ateşi(FMF), ataklar halinde seyreden tekrarlayan ateş, abdominal ağrı, artralji, artrit, serozit, eritem benzeri erizipelle karakterize sıklıkla doğu akdeniz bölgesinde yaşayan toplulukları etkileyen dünyada en yaygın görülen kalıtsal otoinflamatuar hastalık olup *MEFV* genindeki fonksiyon kazandırıcı mutasyonların etkisi ile genin kodladığı pyrin proteininin doğuştan gelen bağışıklık sistemi üzerindeki düzenleyici etkisini bozması kaynaklı ortaya çıkmaktadır(39).

2.5.2. Epidemiyoloji

FMF, otoinflamatuar hastalıklar arasında en sık olarak rastlanılanıdır. Sıklıkla Türkler, Seferad yahudileri, Ermeniler ve Arap toplumunda gözlenir. Türkiye'de saptanan FMF vakalarının çoğu (%70) İç Anadolu ve Karadeniz kökenlidir(40).

2.5.3. Genetik ve Kalıtım

FMF, 16p13 bölgesinde yer alan *MEFV* genindeki mutasyonlar sonucu açığa çıkar. *MEFV* geni 10 ekzondan oluşmaktadır. Gende meydana gelen mutasyonların önemli kısmı 10.ekzonda yer almaktadır. Bugüne kadar hastalıkla ilişkilendirilmiş 7 fenotip bulunmaktadır. 15 farklı transkripti bulunmakta olup ana transkriptinin NM erişim numarası **"**NM_000243.3" ve ENST numarası "ENST00000219596.6" olarak belirtilmiştir. Bu transkriptte 3506 baz çifti bulunmakta olup gen ürünü pyrin proteini 781 aminoasitten meydana gelmektedir (Bu bilgiler 30/07/2022 tarihinde ensembl.org ve uniprot.org'dan alınmıştır).

2.5.4. Genetik Testler

FMF hastalalarında *MEFV* geninde yapılan dizileme analizi sonucunda bu hastaların %80'ninde her iki allelde de varyant tespit edilmiştir(41). Yeni nesil dizileme tekniklerinin kullanıldığı 3230 otoinflamatuar hastalık ön tanısı ile ilgili merkeze refere edilen hastaları içeren bir çalışmada hastaların 1839'unda(%56,9)'unda *MEFV* geninde yanlış anlamlı mutasyon saptanmıştır. Bu çalışmaya göre 56 farklı mutasyon ve 141 genotip tespit edilmiştir. Bu mutasyonlar ACMG kriterlerine göre 7'sinin patojenik, 3'ünün olası patojenik, 6'sının bening ve 40'ının ise önemi bilinmeyen varyant olarak değerlendirilmiştir(42). Clinvar'da yer alan 802 tek nükleotit değişiminden 416 (%51,23) tanesi missense mutasyondur. Bu 416 mutasyondan 320'si (%76,92) klinik önemi bilinmeyen varyant iken, 41'i (%9,86) hakkında çelişkili yorumlar mevcuttur. 5'i benign, 13'ü olası benign, 5'i olası patojenik ve 9'u ise patojeniktir (Bu bilgiler Clinvar veritabanından 27.09.2022 tarihinde alınmıştır).

2.6. Yeni Nesil Dizileme Teknikleri

2.6.1. İnsan Genomunun Yapısı

İnsan genomu 3,2 milyar baz çiftinden meydana gelir. Bu genomun yalnızca %1-1,5'i protein kodlayan DNA dizilerinden meydana gelir. Protein kodlayan DNA dizilerine ekzon adı verilir. Ancak insanda hastalık oluşturan kalıtsal hastalıkların %80-85'inin ekzonik bölgelerede meydana geldiği tespit edilmiştir.

2.6.2. Yeni Nesil Dizileme Teknikleri

Yeni nesil dizileme teknikleri insan genomunun tamamı ya da belirli bir bölgesinde yer alan her bir baz çifti için DNA ya da RNA dizilerini tespit etmeye olanak sağlar.Yeni nesil dizileme teknolojileri ile tek bir örnek çoğaltılarak tüm ekzonik ya da genomik bölgelerin tespit edilmesine olanak sağlanmıştır. Tüm ekzonik bölgelerin tespit edildiği yönteme tüm ekzom

dizileme, tüm genomik bölgelerin tespit edildiği yönteme ise tüm genom dizileme adı verilir. Bunun yanı sıra protein kodlayan ya da kodlayaman transkript bölgelerinin ya da kromatin bölgelerinin analizi de olası hale gelmiştir. Transkript bölgelerin inceleyen tekniğe RNA-Seq, kromatin yapısını inceleyen tekniğe ise CHIP-seq adı verilmektedir. Kullanılan teknikler farklı da olsa bir yeni nesil dizileme tekniğinde kullanılan yöntemler temelde aynıdır.

2.7. Biyoinformatik veritabanları

Biyoinformatik içerisinde matematik, biyoloji, istatistik ve makine öğrenmesi gibi bir çok disiplini barındıran bir bilimdalıdır. Biyoinformatik kelimesi "biyoloji" ve "enformasyon" terimlerinin birleşimden meydana gelir. Biyoloji canlılarla ilgilenen bir bilim dalıdır. Enformasyon ise verinin işlenmiş haline verilen isimdir. Biyoinformatikle ilişkili veriler biyoinformatik veritabanlarında bulunur.

2.8. Varyantların Sınıflandırılması ve Tahmin Edilmesinde kullanılan Kılavuzlar

Herediter varyantların sınıflandırılmasında, American College of Medical Genetics (ACMG) kriterleri yaygın olarak kullanılırken; somatik, farmakogenetik ve multilokus herediter hastalıkların sınıflandırılmasında kullanılan başka kriterler de bulunmaktadır.

ACMG kriterleri 16 adet patojenik ve 12 adet benign olmak üzere toplam 28 adet kriterden oluşmaktadır. Buna göre, patojenik varyantları 4 başlık altında derecelendirilmektedir: çok güçlü(PVS1), güçlü(PS1-4), orta (PM1-PM6) ve destekleyici(PP1-5). Benign varyantlar ise 3 başlık altında derecelendirilir: tek başına(BA1), güçlü(BS1-4) ve destekleyici(BP1-6)(43). Varyantların veri tipi (benign/malign) ve gücüne göre sınıflandırılmaları Tablo 1.'de gösterilmiştir.

Tablo 3. Varyantların veri tipi ve gücüne göre sınıflandırılması

Varyant tipi	**Benign**		**Pathogenic**			
Varyant derecelendirmesi	S (-18.7)	P (-2.08)	P (2.08)	M (4.33)	S (18.7)	VS (350.0)
Toplum frekansı	BA1,BS1, BS2			PM2	PS4	
Bilişimsel ve Öngörüsel		BP1, BP3, BP4, BP7	PP3	PM4 PM5	PS1	PVS1
Fonksiyonel	BS3		PP2	PM1	PS3	
Segregasyon	BS4		PP1			
De novo				PM6	PS2	
Allelik		BP2		PM3	PS4	
Diğer veritabanları		BP6	PP5		PS3	
Diğer		BP5	PP4			

Bu varyantların içeriği şu şekildedir(43):

Tek başına

BA1: Allel sıklığı > %5 (1000 Genome Project, Exome Agregation Consortium (ExAC), The Genome Aggregation Database (gnomAD) gibi toplumsal veri tabanları)

Güçlü

BS1: Allel sıklığı hastalık için beklenen sıklıktan daha yüksektir.

BS2: Sağlıklı bireylerde saptanmıştır (X'e bağlı, resesif, dominant). Ayrıca hastalık tam penetrans göstermekte olup erken yaşlarda ortaya çıkmaktadır.

BS3: Fonsiyonel çalışmalarda protein fonksiyonu ya da kesim bölgesi üzerine hasar verici etkisinin olmadığı gösterilmiştir.

BS4: Aile bireyleri arasında segregasyon bulunmamaktadır. (Hastanın yakınlarında penetrans ortaya çıkmamıştır.)

Destekleyici

BP1: Genellikle trunkasyona uğrayan varyantların hastalık nedeni olduğu bir gende yanlış anlamlı bir mutasyon saptanması

BP2: Tam penetrans gösteren otozomal dominant bir hastalıkta trans pozisyonda patojenik varyant saptanması ya da cis pozisyonda herhangi bir kalıtım paterni için patojenik mutasyon saptanması

BP3: Herhangi bir fonksiyonu bilinmeyen tekrarlayan dizide çerçeve kayması oluşturmayan insersiyon/delesyonlar

BP4: Bir çok tahmin programında gen veya gen ürünü üzerine bir etkisi olmadığının gösterilmesi

BP5: Varyantın vakada hastalığın alternatif bir moleküler kökeninden gelişmesi (Hastalığın muhtemel başka bir olası mekanizma kaynaklı olması)

BP6: Varyant yakın zamanda benign olarak bildirilmiştir fakat kanıtlar laboratuvar denemeleri için uygun değildir.

BP7: Sinonim bir varyant ve uç kesim tahmin algoritmalarının splice consensus site üzerine bir etkisinin olmadığını göstermesi ya da yeni bir splice bölge oluşturmaması ve nükleotidin yüksek oranda korunmuş olmaması

Patojenik Etkinin Çok Güçlü Kanıtları,

PVS1: null varyant (kanonik ± 1, 2 kesim bölgesi, nonsense, başlangıç kodonu, hastalık nedeni olan fonksiyon kaybı yaptığı bilinen tek ya da birden fazla ekzonun delesyonu

Patojenik Etkinin Güçlü Kanıtları

PS1: Nükleotit değişiminden bağımsız olarak aynı aminoasit değişimine yol açan başka bir mutasyonun saptanması

PS2: Aile öyküsünün bulunmadığı *de novo* (anne ve baba doğrulanmış) mutasyonun saptanması

PS3: *İn vivo* ya da *in vitro* fonksiyonel çalışmalarda gen ya da gen ürünü üzerine hasar verici etkisi gösterilmiş

PS4: Varyantın prevalansının kontrollerle kıyaslandığında etkilenmiş bireylere göre önemli ölçüde artması

Patojenik Etkinin Orta Dereceli Kanıtları

PM1: Varyantın hot-spot bölgede ve benign varyasyonun olmadığı iyi tanımlanmış fonksiyonel domainde (enzimin aktif bölgesi gibi) bulunması

PM2: Varyantın kontrollerde bulunmaması (ya da hastalık resesif ise oldukça düşük bir düzeyde saptanması)

PM3: Resesif hastalıklarda *trans* pozisyonununda patojenik varyantın tespit edilmesi

PM4: Stop-Loss varyant ya da tekrar dizisi bulunmayan bölgede in-frame delesyon/insersiyon nedenli protein boyutunda değişimler

PM5: Aynı bölgede daha önce patojenik olarak tespit edilen ancak başka bir aminoasit rezidüsünü etkileyen, yeni bir yanlış anlamlı mutasyonun saptanması

PM6: *de novo* olduğu kabul edilen ancak anne-baba konfirmasyonunun gerçekleşmediği mutasyon

Patojenik Etkiyi Destekleyen Kanıtlar

PP1: Hastalık yaptığı bilinen bir genden etkilenmiş olan aile bireylerinden köken alması (Segregasyon verisi arttıkça kanıt düzeyi artar.)

PP2: Gende missense varyantların nadir gözlenmesi ancak hastalık yapmada missense varyantların yaygın bir rol üstlenmesi

PP3: *İn silico* veritabanlarına göre patojenik olarak saptanması

PP4: Hastanın fenotipik bulguları ve aile öyküsünün bulunması

PP5 : Varyant yakın zamanda patojenik olarak bildirilmiş ancak kanıtlar laboratuvar denemeleri için uygun değildir.

Daha sonra bu kriterler Sherloc kılavuzu ile revize edilmiştir(44)**.**

2.9. *İn silico* Tahmin Araçları

Günümüzde teknolojinin gelişmesi ve maliyetlerin azalmasıyla yeni nesil dizileme teknolojileri tıbbın neredeyse bütün disiplinlerinde kullanılır hale gelmiştir. Ancak bu durum yorumlanması gereken veri sayısında da bir artış meydana getirmektedir. Yorumlanması gereken veri sayısındaki artışla birlikte klinik olarak önemi tespit edilemeyen (VUS) varyantların sayısı da artmaktadır. Genom düzeyinde saptanan VUS varyantların fonksiyonel çalışmalarının yapılması sonucunda bu varyantların hastalık yapıcı rolleri hakkında kıymetli bilgiler elde edilebilmektedir. Ancak, fonskiyonel çalışmaların VUS olarak tespit edilen varyant sayısının fazla olması, bu çalışmaların yüksek maliyet, uzun uğraş ve zaman gerektirmesi nedeniyle pratikte uygulanması o kadar kolay değildir. Bu nedenle bu varyantların protein üzerine olan etkisini ya da korunmuşluk düzeyini tahmin edebilecek *in silico* araçlar geliştirilmiştir. Bu araçlar fonksiyonlar çalışmalar gibi kesin bir sonuç vermezler. Ancak, klinisyenlere varyant yorumlamasında önemli kolaylıklar sağlamaktadır.

İn silico araçlar iki durumu tahmin etmeye yarar: 1) İlgili varyantta meydana gelen değişimin protein yapısı üzerine etkisi 2) Evrimsel korunmuşluk düzeyi. İlgili varyantın protein üzerine etkisini tahmin eden *in silico* araçlara örnek olarak SIFT, Polyphen-2, Mutationtaster-2 ve korunmuşluk düzeyini tahmin eden araçlara ise PhyloP, GERP+ örnek olarak verilebilir. Bu araçlardan korunmuşluk skorlarını protein etkisini tahmin eden skorlara göre daha etkili olduğunu kabul eden görüşler bulunmaktadır. Bu tahmin araçları belirli algoritmaları kullanarak kendilerin bir tahmin yapmakta ve bu tahmine göre bir skor üretmektedir.

Kendi skorlarını oluştururken birden fazla aracı kullanan algoritmalar metaprediktörler olarak adlandırılmaktadır. En yaygın metaprediktörlere Revel, CADD, DANN örnek olarak verilebilir. Metaprediktörler arasında yer alan Revel skoru tek nükleotit değişimi bulunan nadir varyantları bulmada en iyi sonucu verdiğini iddia etmektedir. Ancak daha sonraki yıllarda geliştirilen CAPICE skoru sadece tek nükleotit değişimlerini değil aynı zamanda indelleri de hesaba katarak en optimal sonucu ortaya koyduğunu iddia etmektedir. Revel skoru AF < %3 olan varyantlarda hastalık yapıcı olanları nötral varyantlardan ayırmada en başarılı sonucu elde ederken AF %3-5 olan varyantlarda MetaLR ve MetaSVM daha başarılı olarak saptanmıştır. CADD ve DANN skorları missense varyantları sınıflandırma konusunda REVEL, MetaLR ve MetaSVM kadar başarılı olmasa da bu skorlar kodlamayan bölgelere ve düzenleyici bölgelere de ilişkin skorlar sundukları için oldukça değerlidir(12).

Herhangi bir varyantın saptanabilmesi için öncelikle ilgili DNA dizisi belirlenir ve burada meydana gelen değişim tanımlanır. Sonrasında bu değişimin tanımlanmasında 4 metot kullanılır: (1) dizi homolojisi, (2) dizi yapısallığı, (3) dizi benzerliği ve (4) korunmuşluk düzeyi.

2.9.1. KORUNMUŞLUK SKORLARI

Dizi korunmuşluğu negatif seçilim nedeniyle biyololojik fonksiyonu göstermek adına güçlü bir işarettir. Dizi korunmuşluğunu göstermek için bir çok araç kullanılır. Bunlardan en bilinenleri GERP, PhastCons ve phyloP'dir.

PhastCons evrimsel korunmuşluk düzeyleri arasında ayrım yapmak için phylo-Hidden Markov Modeli(HMM)'ni kullanır. HMM bioyolojik veritabanlarında sıklıkla kullanılmasına rağmen değişim olasılıklarının tahminleri genomik segment uzunluğuna bağlıdır. GERP ise "Expectation Maximization" algoritmasını kullandığı için yavaş çalışır. Ancak GERP++ getirdiği dinamik programlama yöntemiyle korunmuşluk düzeylerini, p değerleri ve yanlış pozitiflikleri üzerinden tespit ederken GERP'den çok daha hızlı çalışır. Oldukça büyük bir insan genom parçasında bulunan nadir ancak uzun korunmuş bölgeleri çok az bir yanlış pozitiflikle tespit edebilir(45). Her iki skoru da kullanarak evrimsel korunmuşluk düzeyini tespit eden internet temelli bir arayüz programı ise "phastWEB" ismini almıştır(46).

2.9.1.1. GERP++ SKORU

İnsan genomunda yaşamsal fonksiyonları kontrol eden protein-kodlayan ekzonlar, kodlamayan RNA'lar ve genlerin transkripsiyonunu düzenleyen milyonlarca dizi mevcuttur. Ancak bu fonksiyonel diziler DNA yapısında gömülü bir halde bulunmaktadır. GERP++ karşılaştırmalı genomik yönteminlerinden çoklu dizi hizalama yöntemini kullanarak varsayılan işlevsel dizileri bulmaya çalışır. Buna göre, az sayıda değişiklik içeren dizilerin daha korunmuş olması ve nötral bir etki göstermiş olması beklenir. GERP++ yöntemi fonksiyonel dizilerin çözümlenmesinde gelişmeye yol açmış ve evrimsel olarak olarak korunmuş olan insan genom dizilerinin tahmin edilenden daha fazla olduğunu ortaya koymuştur GERP++, kendinden önceki araçlardan pozisyon spesifik skorlama yaparken maksimum olabilirlik evrimsel oran tahminini kullanan dinamik bir programlama yaklaşımıdır. Böylece, kendisinden sonraki korunmuş bölgeler tanımlanabilir. GERP++ bir çok aday kırılma noktasını belirler ve bunları istatistiki önemine göre sıralar. Böylece, bu metot kullanılarak insan genomunda yayılmış halde bulunan ve yaklaşık %7'sini oluşturan korunmuş bölgeler saptanabilir. GERP++ hem nükleotid hem de element bazlı korunmuşluk skorları çoklu dizi hizalamaları ile kullanan evrimsel korunmuşluğun gösterilmesinde etkili bir araç olarak kabul edilmektedir.(45).

2.9.2. Aminoasit Değişimi Tahmin Skorları

2.9.2.1. Sorting Intolerant From Tolerant (SIFT):

SIFT nötr ve zararlı aminoasit değişimleri arasındaki farklılıkları ortaya çıkarmak için dizayn edilmiş bir programdır. Varyantları hasar verici düzeyden tolere edilebilir düzeye doğru sıralar. Tolareted, aminoasit değişiminin etkisi tolere edilebilir düzeyde olduğunu ve damaging aminoasit değişiminin etkisinin hasar verici düzeyde olduğunu göstermektedir. SIFT hizalama yapılan diziye bağımlıdır. Hizalama yapılan dizi ve hedef dizi oldukça benzer olduğu zaman buradaki değişimler nötral oldukları halde hastalık yapıcı olarak değerlendirilebilir. Ancak benzer olmayan dizilerdeki performansı daha iyidir(47).

2.9.2.2. PROVEAN

PROVEAN hem tek nükleotit değişimlerinin hem de indellerin proteinlerin biyolojik fonksiyonu üzerine bir etkisi olup olmadığının incelemeye yönelik olarak oluşturulmuş bir *in silico* yazılım aracıdır. Provean aminoasit değişimlerini iki sınıfa ayırır: Neutral ve Deleterious. Neural aminoasit değişiminin protein üzerine olan etkisinin nötr, deleterious ise aminoasit değişiminin protein üzerine olan etkisinin hasar verici olduğunu göstermek için kullanılır(48,49).

2.9.2.3. Polyphen-2

PolyPhen-2 hesaplamalarını naif Bayes algoritmasının posteriori olasılığını kullanarak mutasyonları sınıflandırmaya çalışır. Ayrıca, mutasyonları nitel olarak benign (iyi huylu), probably damaging (muhtemel hasar verici) ve daha muhtemel zarar verici (possibly damaging) olmak üzere üç kategoride inceler. Polyphen-2 tarafından eğitilmiş HumVar'ın varyantlar arası ayrımlarda, nadir allelerin ayrımında ise eğitilmiş HumDiv'in kullanılması önerilmektedir(50).

2.9.2.4. Mutation Assessor

Kombinasyonel entropi algoritması kullanarak protein rezidülerini ve proteinlerin evrimsel süreçte yer alan fonksiyonel alt ailelerini tanımlar. Bu metot kullanılarak tahmin edilen spesifik rezidüler ile bilinen fonksiyonel rezidüler arasında uyumlu bir sonuç ortaya çıkmıştır. Bu fonksiyonel bileşenler kullanılarak evrimsel süreçte ve hastalık oluşumunda yer alan mutasyonlar ortaya çıkabilir. Kombinasyonel kompleks problemelerin spesifik rezidüleri çözmedeki problemlerine karşı geliştriliş bir metottur. Hiyerarşik kümeleme optimum sonuca ulaşmak için çeşitli kümelenme yörüngeleri üzerinde alternatif kümelemelerin alanını keşfetmek için kullanıılmaktadır(51). Metodun önceki metotlara göre sunduğu yenilik homolog

dizilerin çoklu dizi hizalaması ile belirlenen korunmuşlukla beraber protein alt aileleri arasındaki evrimsel korunmuşluktan yararlanmasıdır.

2.9.2.5. MutationTaster:

Evrimsel krounmuşluk skorları, protein değişimleri, alternatif kırpılma bölgeleri ve mRNA'yı etkileyebilecek değişimleri kullanarak tahminleme yapar(52,53).

2.9.3. METAPREDİKTÖRLER

2.9.3.1. CADD SKORU

"Combined Annotation Dependent Depletion' (CADD) skoru Tek nükleotit varayantı (SNV) ve kısa insersiyon ve delesyonlar InDels'ler de dahil olmak üzere genetik varyantları insan referans genomu assemblysi kullanarak skorlamaktadır. CADD skorları sekans bağlamı (sequence context), gen anotasyonu (gene model annotations), evrimsel krounmuşluk (evolutionary constraint), epigenetik ölçümler (epigenetic measurements) ve fonksiyon tahmini (functional predictions) gibi değişik genomik özelliklerden teşekkül etmektedir. Bu skorda, yorumlanabilirliğin arttırılması için 9 milyar olası SNV'nin genom çapındaki dağılımından yararlanılmakta olup az sayıda varyant içeren patojenik veya benign varyantlar için eğitim setine dahil edilmemektedir. Bunu yerine, yanlılığı azaltmak için çok daha büyük eğitim setlerini kullanılmaktadır.

CADD yapısı(frameworkleri) eğitim verileri için sistematik ve objektif bir etiketleme sağlamaktadır. Böylece neredeyse herbir özellik reference assembly coordinatelerine bağlanabilir. Ayrıca hem kodlayan hem de kodlamayan varyantlar için bir skor üretir. Her CADD modelinde her bir iterasyon 30.000.000 varyant ve anotasyondan köken alan 100'den fazla özellik için eğitilmiştir. Bu durum overfitting olasılığını riskini azaltmaktadır. Ancak, her ne kadar CADD skoru model eğitilirken manuel bir varyant küratasyonuna dayanmazken, model tuninginde kullanılacak parametreler seçilirken manuel olarak kürete edilen varyantlar kullanılmaktadır.

Anotasyon yapılırken kullanılan değişkenler arasında türler arası korunmuşluk, epigenetik modifikasyon, fonksiyonel tahminleme ve genetik yapı önemli bir yer tutar. Varyantlar anote edilirken proxy-neutral ve proyx-deleterous varyantlar kullanılır. Bu Anote edilen bu varyantlar logistik regresyon metodu ile eğitilir. Aynı zamanda anotasyon yapılırken kullanılan değişkenler kullanıcıların sağladığı varyantlarla da anotasyona tabi tutulur. Varyantların LR ile eğitilmesi ve kullanıcılar tarafından sağlanılan varyantların anote

edilmesinden sonra yapılan varyant skorlarması PHRED skalasına göre değerlendirildekten sonra CADD skoru hesaplanır(54).

2.9.3.2. DANN SKORU

CADD skoru hem kodlayan hem de kodlamayan varyantları anote etmektedir. Ancak "eğitim" aşamasında gözlenen varyantları simule edilen varyantlardan ayırmak içn lineer kernel destek vektör makinelerini (SVM) kullanmaktadır. Gözlenen genetik varyantların insan genomu ve insan-şempanze atası genomundaki farklılıklardan kaynaklanığı düşünülmektedir. Doğal seleksiyonun etkisi ile gözlenen varyantlarda hasar verici varyantlar tükenirken, simule edilen genetik varyantlar hasar verici varyantlar yönünden zenginleşmektedir.

CADD skorunun kullandığı DVM veriyi doğrusal bir şekilde kullandığı için bu testin performasında kısıtlılıklara yol açar. Bu durumun üstesinden gelmek için derin sinir ağları (DSA)'nın kullanıldığı yöntem deleterious annotation of genetic variants using neural networks (DANN) olarak bildirilmektedir. DNN bir çok gizli katmanı içinde barındıran yapay sinir ağıdır. DNN aynı zamanda hesaplama yapmak için olan zamanı da kısaltır. Ayrıca bir çok özellik ve örneğin kullanıldığı yöntemlerde SVM ve LR gibi metotlara göre daha iyi sonuç verdiğini iddia etmektedir(55).

2.9.3.3. ClinPred

2.9.3.4. MetaLR ve MetaSVM

Sahra altı Afrika'da yapılan bir çalışmaya göre, yeni nesil dizileme analizi sonucunda elde edilen varyantlar MetaSVM ve MetaLR skorları kullanılarak değerlendirilmiştir(56).

2.9.3.5. IMHOTEP

ANNOVAR, dbNFSP gibi bir çok aracı bir araya getiren paket programlar bulunmaktadır. Bu durum bir çok varyantın tahmin edilmesi için bir çok kolaylık sağlamaktadır. Ancak, buradan elde edilen sonuçların yorumlanması oldukça zordur. Bunun nedenleri arasında bu araçların farklı skorlama sistemleri kullanması, bazı varyantlar için sonuç elde edilebilirken bazıları için elde edilememesi gibi nedenler yer almaktadır. Bununla birlikte, tahmin programları varyantın protein etkisini öngörmek üzere tasarlanmıştır. Bu nedenle tahmin etkilerinin bir araya getirildiği ve bu etkilerin standardize edildiği ve kantitatif değerlerin bir araya getirilerek tek bir skorun elde edildiği yeni tahmin programlarına ihtiyaç bulunmaktadır. Condel, CAROL, metaSVM ve MetaLR gibi araçlar tekli tahmin programlarından daha iyi çözümler sunmuştur. Ancak bu araçların çoğunda toplu sorguları çalıştırma konusunda bir takım sorunlar yaşanmıştır. IMHOTEP ('Integrating Molecular

Heuristics and Other Tools for Effect Prediction") dokuz adet skoru (PolyPhen-2, SNPs&GO, MutPred , SIFT, MutationTaster2, Mutation Assessor, FATHMM, PhyloP, Grantham score) standardize etmeyi başarıp tek bir skor haline getirmiş ve bu esnada bir çok farklı istatistiki metottan yararlanmıştır. Bunlar basit toplama, çoğunluk oyu, LR, KA, RO'dur . Ayrıca, bu entegrasyon bir araya getirilirken 20.000'in üzerinde varyant kullanılmıştır. (57).

2.9.3.6. REVEL SKORU

Yeni nesil dizileme metotları ile beraber çoğunlukla nadir varyantlar saptanmaktadır. Güncel ekzom ve genom analizlerine göre sinonim olmayan varyantların %85'inden fazlasının alternatif allel sıklığı %0,5'in altındadır ve kabaca kişi başına 100-400 arasında nadir varyant tespit edilmektedir. Yeni nesil dizileme yöntemleri yardımıyla saptanan sinonim olmayan varyantların birçoğunun klinik önemi bilinmemektedir. Bu nadir varyantlara yönelik olarak tasarlanan assosiasyon çalışmaları çok sayıda hasta gerektirdiği ve çok sayıda varyantın deneysel olarak doğrulanması mümkün olmadığı için bu bu varyantların patojenik etki gösterip göstermeyeceğini tahmin edecek metaprediktörlere ihityaç vardır. Mevcut araçlar arasında aminoasit ya da nükleotid korunmuşluk skoru ve aminoasit değişimlerinin biyokimyasal etkileri üzerinden tahminleme yaparlar. Nadir varyantları tahmin etmeye çalışan tahmin programları az sayıda olup genellikle yaygın nötral varyantlar açısından eğitilmiştir. Oysaki, hastalık yapıcı varyantların çoğu nadir varyantlardır. Revel skoru ensembl bir yöntem olup nadir varyantlar için en başarılı sonucu verdiği düşünülmektedir(12).

2.9.3.7. MISTIC SKORU

Yanlış anlamı mutasyonların zarar verici özelliklerini tahmin eden bir metaprediltördür. Makine öğrenmesi yöntemi olarak RO ve LR kullanır. Yöntemler arası integrasyon yumuşak oylama sistemi aracılığıyla yapılır. 113 adet özeliğin seçimine dayanır. MISTIC diğer tahmin araçları ile karşılaştırıldığında (Eigen, FATHMM-XF, REVEL, M-CAP, ClinPred ve PrimateAI) kodlama bölgelerinde yer alan zararlı değişkenleri tahmin etmede sıralamada en iyi performansa sahiptir. Bu araç aynı zamanda VUSları azaltmakta ve hastalık yapıcı varyantların sıralamasında iyileşmeye yol açmaktadır. Bu araçta yanlış anlamlı mutasyonların etkisini tanımlamak için 714 özellik 4 ana başlık altında toplanmıştır(58).

2.9.3.8. CAPICE SKORU

Computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations (CAPICE) skoru moleküler düzeydeki değişim ve allel frekansından bağımsız olarak önceki metotlara göre patojenik varyantları nötral varyantlardan en yüksek

doğrulukta ayırmaktadır. Aynı zamanda bu metot bütün Tek nükleotit varyantlarını ve inDels'leri önceliklendiren ilk metottur. Bu skorla ilgili ilk makale 2020 yılında yayınlanmıştır. CADD skorlarına göre kıyaslandığında bu metot nadir varyantları, çerçeve kayması, uç birleştirme ve durdurma kodonu oluşturan varyantları daha iyi tahmin yapmıştır. Nadir varyantları saptamada ortaya koyduğu başarı Revel skoru ile kıyaslanabilir düzeydedir.

Clinvar gibi bazı toplumsal bazlı veritabanlarındaki bilgiler çelişkili ya da güncel olmayabilir. nedenle bazı araştırmacılar bu veritabanından yararlanırken yüksek kaliteli verilerden yararlanır. Aynı zamanda burada yer alan bazı yaygın varyantlar kullanılan filtereler nedeniyle atlanılabilir. CAPICE diğer programlardan farklı olarak düşük kaliteli veriyi de çalışmaya dahil eder, ancak ağırlıklandırmasını daha düşük kabul eder. Ayrıca, araştırmayı tasarlayanlar modelin eğitimi sırasında kullanılan yüksek kaliteli verinin düşük kaliteli veriye göre bir üstünlük sağlamadığını saptamıştır. Sonrasında, modelin test edilmesinde patojenik varyantlar için yüksek kaliteli veri ve nötral varyantlar için toplumsal çalışmalarda kullanılan bütün nadir ve ultra-nadir varyantlar bilinen patojenik varyantlar ve kalıtım kalıpları filtre edildikten sonra dahil edilerek kullanılmıştır. Bu sayede veri seçimindeki yanlılık ve önemli bazı bilgilerin kaybolmasının önüne geçilmiştir(59).

2.9.3.9. RENOVO

Renovo skoru ile ilgili ilk yayın 2021 yılında yayınlanmıştır. Buna göre, klinik önemi bilinmeyen germline varyantları makine öğrenmesi metotları kullanılarak yeniden sınıflandırılmıştır. Varsome ya da Intervar gibi "computational tools" varyantları kılavuzlara göre sınıflandırmaktadır. Yeni varyantların bir çoğu genotip-fenotip ilişkisi kurulamadığı için VUS olarak sınıflnadırılmaktadır. Ayrıca, sınıflrandırılmasında zorluk yaşanan diğer varyantlar conflicting interpretation of pathogenicity (CIP) olarak da adlandırılmaktadır. Bu varyantlar, bazı laboratuarlar tarafından patojenik olarak bazıları tarafından ise benign olarak sınıflandırılmaktadır.

VUSların klinik yorumu oldukça büyük bir problem teşkil eder, bir çok laboraturar VUSların raporlanmaması gerektiğini savunmaktadır. Ancak, bunun yerine bu varyantların belirli aralıklarla takip edilmesi önerilmektedir. Eskiden VUS olarak bildirilen bir çok varyant benign ya da patojenik olarak sınıflandırılabilmektedir. Ancak varyantların ikili sınıflandırılmasında (patojenik veya benign) bir problem ortaya çıkmaktadır. RENOVO Clinvar veritabanınından elde ettiği bilgiler yardımıyla bu problemin çözülmesini ve sürekli değişkenlerden oluşan patojenik-olabilirlilik oranını sağlayan random forest algoritmalarını

kullanan bir tahmin programıdır. Bu sayede VUS olarak nitelendirilen varyantların %67'sini büyük bir güvenilirlilik oranıyla sınıflandırmayı başarmıştır(60).

2.10. Makine öğrenmesinin varyantların sınıflandırılmasında kullanılması

2.10.1. MEFV

INFEVERS veritabanından alınan 216 yanlış anlamlı *MEFV* gen mutasyonu REVEL skorlarına göre yeniden sınıflandırıldığı bir çalışmada 5 farklı algoritma kullanılmıştır: DAA, Karar ağaçları (KA), K-en yakın komşu (KNN), Destek Vektör Makineleri (Lineer kernel) ve rassal orman (RO). Revel skorlarının tahmin etkisini saptayabilmek için olası benign ve olası patojenik varyantlar bu analize dahil edilmiştir. En iyi sonuç DAA ile edilmiştir (doğruluk %75, duyarlılık %62,5 ve özgüllük %87,5) Yaklaşık cut-off değeri 0,29 olarak kabul edilmiştir. Böylece REVEL skorlarını da kullanarak yeni bir sınıflama yapmayı başarmışlardır. Bu sınıflandırma özellikle benign varyantları saptamada oldukça başarılı sonuçlar ortaya çıkarmıştır. Ayrıca, araştırmacılar pyrin fonksiyonel domaini üzerinde yer alan patojenik varyantların rasgele dağılmadığını ortaya koymuştur. Yazarlar, uzman bilgileri ve duyarlı tahmin programlarının birlikte kullanıldığı çalışmalar bir araya *MEFV* varyantlarının sınıflandırılmasında daha doğru sonuçlar elde edileceğini vurgulamıştır(61).

2.10.2. BRCA

BRCA 1-2 genlerinde yer alan yanlış anlamlı mutasyonların fonksiyonel etkilerini tahmin etmek için yapılan bir çalışmada yüzlerce model farklı algoritmalarla test edilmiştir. Bu model BRCA-ML olarak isimlendirilmiştir. Bu çalışmaya göre, BRCT ve RING domainleri dışında kalan BRCA1 bölgeleri test edilmiş ve BRCA-ML %74 sensitivite ve %98 spesifite göstermiştir(62)

2.11. Araştırmanın Amacı ve Türü

Bu çalışmada infevers veri tabanında yer alan *MEFV* genine ait missense ve sinonim varyantları sınıflaması için optimum sayıda tahmin skoru ve 7 adet makine öğrenmesi kullanılarak en ideal saptamayı yapan kombinasyonu bulmak hedeflenmektedir.

3. YÖNTEM

3.1. Veri kümelerine erişim

İnfevers veritabanında yer alan 389 adet varyant kürete edildi (son erişim tarihi:04.05.2022) Bunlar arasından kodlayan bölgede yer alan missense ve sessiz varyantlar çalışma kapsamına alındı. Kodlamayan bölgeler, terminasyon kodonu oluşturan bölgeler,insersiyonlar, çerçeve kayması olan ya da olmayan delesyonlar, indeller çalışma kapsamından çıkarıldı. Bu koşulları sağlalayan 266 varyant çalışmaya dahil edildi. Çalışma kapsamında, LP ve P varyantlar LP varyant , LB ve B varyantlar LB varyant ve bu varyantın dışındaki sınıflandırılmayan ya da klinik önemi bilinmeyen varyantlar VUS ve çözülemeyen varyantlar ise Unresolved olarak toplam 4 kategoride değerlendirilmiştir.

3.2. Skorların Elde Edilmesi

3.2.1. Franklin Genoox

Franklin, varyantları ACMG kriterleri, Genom bazlı SNV ve kopya sayısı değişimi (CNV) değerlendirmeleri ve birçok farklı aracı bir araya getirerek kendi oluşturduğu algoritmaları kullanarak yorumlar. Franklin aday varyantları gelişmiş bir yapay zeka yorumlama motoru yardımıyla bulmayı hedeflemektedir (63). Franklin genoox içeriğinde birçok skoru ihtiva eder.

3.2.2. dbNSFP 4.0

Varyantların önceliklendirilmesi varyant yorumlanması için önemli basamaklardan biridir. bNSFP 4.0 içeriğinde 30'dan fazla skoru barındırır. 84 milyondan fazla sinonim olmayan ve kesim bölgesi SNV'sini ihtiva eder. Bu amaçla kullanıcılar için bir çok skora erişim kolaylığı sağlar(64).

Çalışmada skorların elde edilmesinde dbNSFP (38 prediksiyon ve 9 korunmuşluk skoru) ve franklin.genoox (23 skor) veritabanları değerlendirilmiş olup her iki veritabanında da yer alan skorlar arasından eksik veri bulundurmayan skorlar seçilmiştir.

3.3. İstatistiki Karşılaştırmalar

Çalışmadaki verilerin istatistiksel analizinde python (sürüm: 3.7.3) ve SPSS (Windows için sürüm 25.0, IBM, Chicago, IL) kullanılmıştır. Nicel değişkenler (Kesikli ya da sürekli sayısal değişkenler) normal dağılım koşullarını sağladığında ortalama ve standart sapma ile sağlamadığında ise medyan ve çeyrekler arası aralık (lAR) olarak ifade edildi. Nitel (Nominal ve ordinal değişkenler) değişkenler sayı ve yüzde kullanılarak açıklanmıştır. Ordinal değişkenler tabloda hiyeraşik düzendeki yerlerine göre dizilmiştir.

Normallik varsayımının test edilmesinde grafiksel ve analitik yöntemlerden yararlanıldı. Grafiksel yöntemlerden Q-Q plot, Detrended plot, boxplot, histogram ve steam and leaf yöntemleri, analitik yöntemlerden Kolmogorov-Simirnov testinden yararlanıldı. Skewness (çarpıklık) ve basıklık değerlerinin ±1 sınırları içerisinde, çarpıklık ve basıklık indekslerinin (Skewness (çarpıklık) ve kurtosis (basıklık) değerlerinin kendi standart hatalarına (Std. Error) bölünmesi) ±2 sınırları içinde 0'a yakın olması, normal dağılımın varlığına kanıt olarak değerlendirildi.

Sayısal değişkenler normallik varsayımı yönünden test edilmiş ve normal dağılım gösterdiğinde Student t ve ANOVA testi uygulanmış, normallik varsayımı göstermediğinde ise Mann Whitney U ve Kruskal Wallis-H testlerinden yararlanılmıştır. ANOVA ve Kruskall Wallis H testinde istatistiki açıdan anlamlılık saptandığında alt gruplara post-hoc test (ANOVA için Tukey, LSD, Kruskall Wallis H için Dunn-Bonferroni) uygulanmıştır. Bütün istatistiki hesaplamalar değerlendirilirken güven aralığı (Confidence interval, CI) % 95 alınmış , (α) değeri 0.05, 1-β değeri 0.80 olarak belirlenmiş ve P değeri 0.05'ten küçük saptanan değerler istatistiki olarak anlamlı kabul edilmiştir.

3.4. Makine Öğrenmesi Yöntemleri

Test veriseti olarak LP ve LB varyantlar kullanıldı. Test veriseti kullanılarak VUS ve Unresolved varyantlar 7 farklı makine öğrenmesi yöntemi 10 katlı bir cross-validasyon yöntemi ile test edildi: DAA, CART, KNN, DVM(lineer), DVM(polynomial), DVM(radial basis function) ve RO. Test verisetinin Makine öğrenmesi metotlarının değerlendirilmesinde Sensitivite, spesifite, F1 skoru ve Accuracy kullanılmıştır.

4. BULGULAR

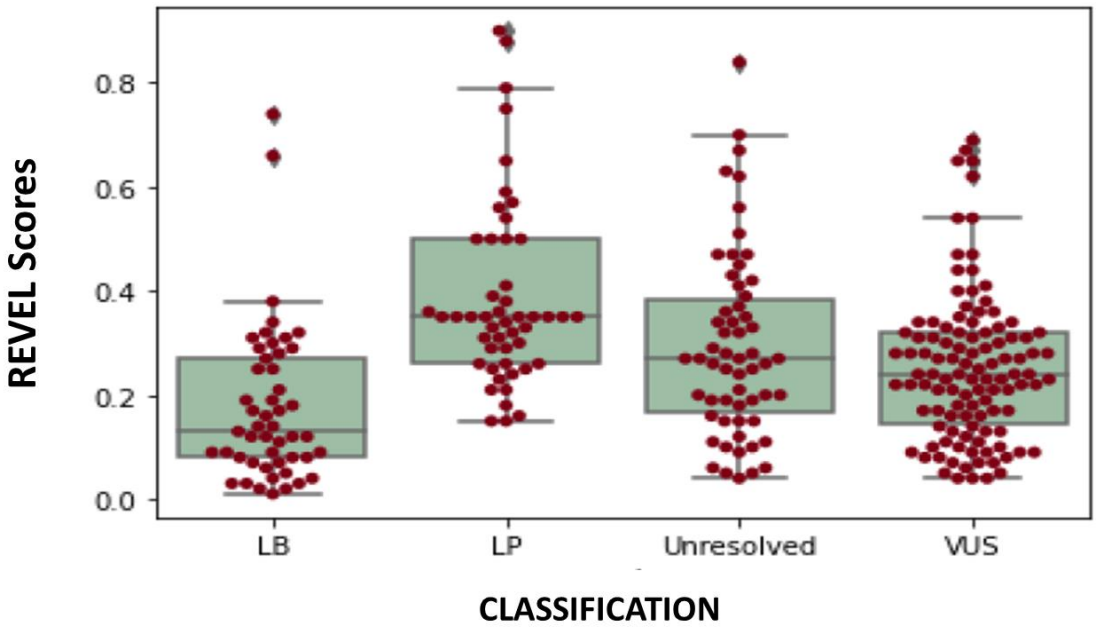
Bu bölümde infevers varyantları gözetimli öğrenme yöntemleri kullanılarak farklı *in silico* araçları kullanılarak karşılaştırılacaktır. Araştırmanın bulguları;

4.1. Tanımlayıcı İstatistikler

Toplam 266 varyant değerlendirildi. 49 varyant LB, 49 varyant LP, 110 VUS ve 58 unresolved varyant saptandı. Bu varyantlar arasında VUS ve unresolved olarak belirtilen varyant sayısı 168 (%63,15) olarak belirlendi.

4.1.1. Revel Tanımlayıcı İstatistikler

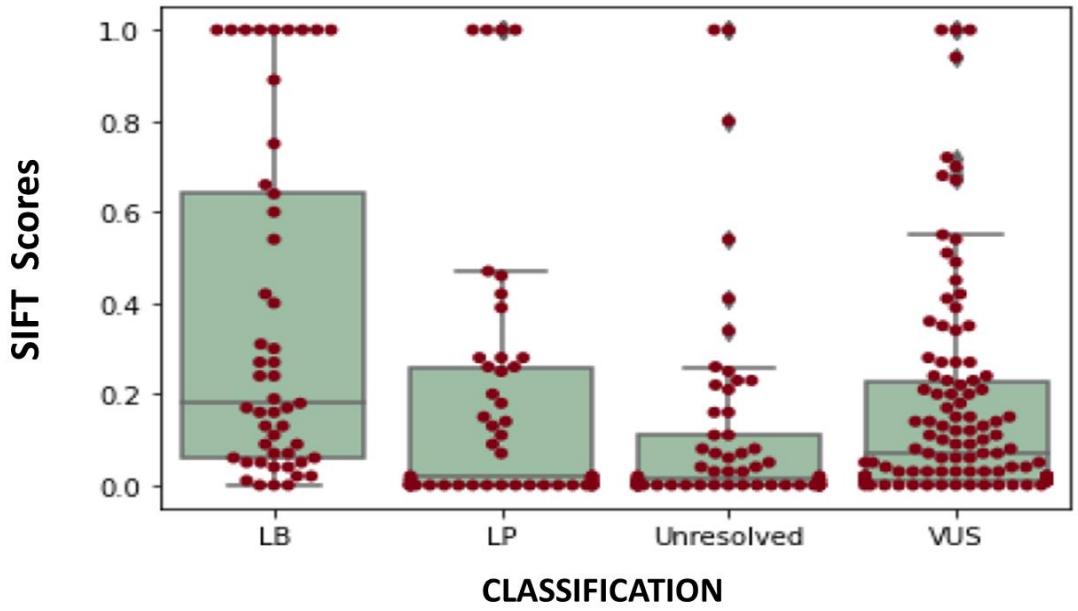
Infevers varyantlarının Revel skorlarının sınıflandırıldıkları gruplara göre ortalama ve standart sapmaları kutu grafiği olarak Şekil 2.'de gösterilmiştir. Buna göre LB varyantların ortalama ve standart sapması 0.176531 ± 0.149994 , LP varyantların ortalama ve standart sapması 0.385918 ± 0.177716 , VUS varyantların ortalama ve standart sapması 0.251545 ± 0.143042 çözülememiş (Unresolved) olan varyantların ortalama ve standart sapması 0.293448 ± 0.178410 olarak tespit edildi.



Şekil 2. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (Revel)

4.1.2. SIFT Tanımlayıcı İstatistikler

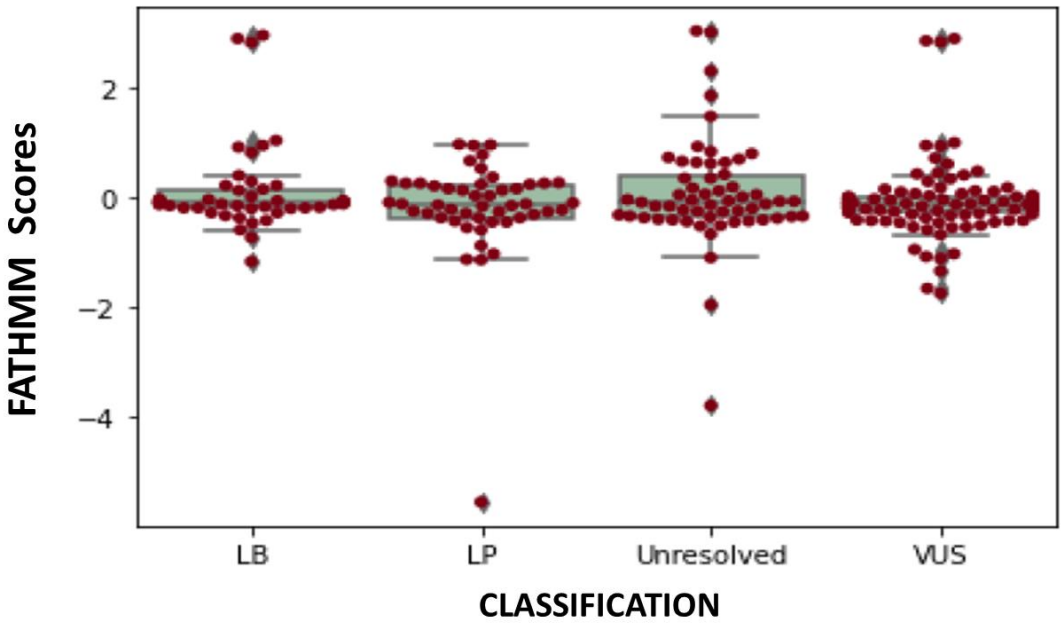
Infevers varyantlarının SIFT skorlarının sınıflandırıldıkları gruplara göre ortalama ve standart sapmaları kutu grafiği olarak Şekil 3.'de gösterilmiştir. Buna göre LB varyantların ortalama ve standart sapması 0.360204 ± 0.369276 , LP varyantların ortalama ve standart sapması 0.173878 ± 0.283966 , VUS varyantların ortalama ve standart sapması 0.170909 ± 0.235027 çözülememiş (Unresolved) olan varyantların ortalama ve standart sapması 0.115690 ± 0.224735 olarak tespit edildi.



Şekil 3. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (SIFT)

4.1.3. FATHMM Tanımlayıcı İstatistikler

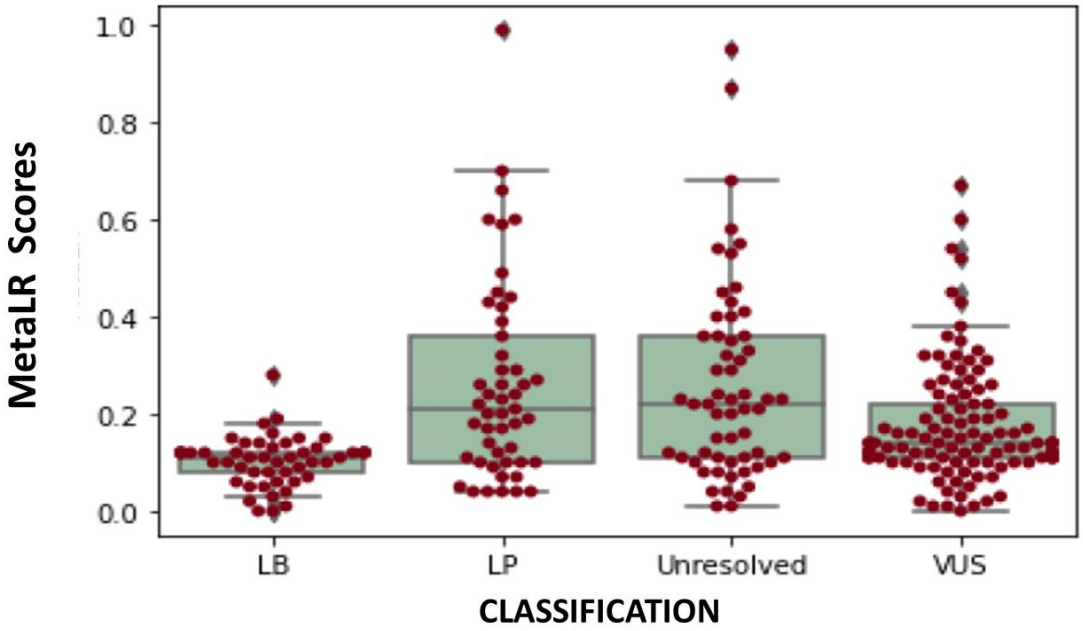
Infevers varyantlarının FATHMM skorlarının sınıflandırıldıkları gruplara göre ortalama ve standart sapmaları kutu grafiği olarak Şekil 4.'te gösterilmiştir. Buna göre LB varyantların ortalama ve standart sapması 0.135102 ± 0.813696 , LP varyantların ortalama ve standart sapması -0.161429 ± 0.924959, VUS varyantların ortalama ve standart sapması -0.065182 ± 0.649711 çözülmemiş (Unresolved) olan varyantların ortalama ve standart sapması 0.084828 ± 1.001858 olarak tespit edildi.



Şekil 4. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (FATHMM)

4.1.4. MetaLR Tanımlayıcı İstatistikler

Infevers varyantlarının MetaLR skorlarının sınıflandırıldıkları gruplara göre ortalama ve standart sapmaları kutu grafiği olarak Şekil 5.'te gösterilmiştir. Buna göre LB varyantların ortalama ve standart sapması 0.10349 ± 0.050808 , LP varyantların ortalama ve standart sapması 0.262857± 0.206377, VUS varyantların ortalama ve standart sapması 0.174727 ± 0.121932 çözülememiş (Unresolved) olan varyantların ortalama ve standart sapması 0.258966 ± 0.202638 olarak tespit edildi.



Şekil 5. Infevers varyantlarının her bir kategoriye göre ortalama ve standart sapmaları (MetaLR)

4.2. ANOVA test sonuçlarının Değerlendirilmesi

ANOVA testine göre Revel, SIFT ve MetaLR skorları arasında varyant sınıflandırm aları arasında istatistiki açıdan anlamlı bir farklılık tespit edilirken FATHMM skorunda anlam lı bir farklılık tespit edilmedi. ANOVA test sonuçları Tablo 5'te gösterilmiştir.

Tablo 4. Tahmin skorlarına göre ANOVA test sonuçları

	Sum of square	Df	F	P değeri
Revel				
grup	1.155163	3.0	15.192404	3.822690e-09
Rezidüel	6.640441	262.0		
SIFT				
grup	1.787722	3.0	8.082903	0.000036
Rezidüe	19.315793	262.0		
FATHMM				
grup	3.008870	3.0	1.492437	0.216982
Rezidüel	176.070819	262.0		
MetaLR				
grup	0.918275	3.0	13.083851	5.471664e-08
Rezidüel	6.129390	262.0		

4.3. ANOVA sonrası Posthoc test sonuçlarının değerlendirilmesi

ANOVA sonucunda anlamlı bulunan skorlar Tukey HSD testi ile değerlendirilmiştir. Revel skoruna göre Tukey HSD testi sonuçları Tablo 6'da gösterilmiştir.

Tablo 5. Tukey testi kullanılarak gruplar arası farklılığın değerlendirilmesi (Revel)

Grup 1	Grup 2	meandiff	p-adj	Alt sınır	Üst sınır	Hipotez reddi
LB	LP	0.2094	0.001	0.1262	0.2926	Evet
LB	Unresolved	0.1169	0.0011	0.037	0.1968	Evet
LB	VUS	0.075	0.327	0.0043	0.1457	Evet
LP	Unresolved	-0.0925	0.0159	-0.1723	-0.0126	Evet
LP	VUS	-01344	0.001	-0.2051	-0.0637	Evet
Unresolved	VUS	-0.0419	0.3688	-0.1087	0.0249	Hayır

SIFT skoruna göre Tukey testi sonuçları Tablo 7'de gösterilmiştir.

Tablo 6. Tukey testi kullanılarak gruplar arası farklılığın değerlendirilmesi (SIFT)

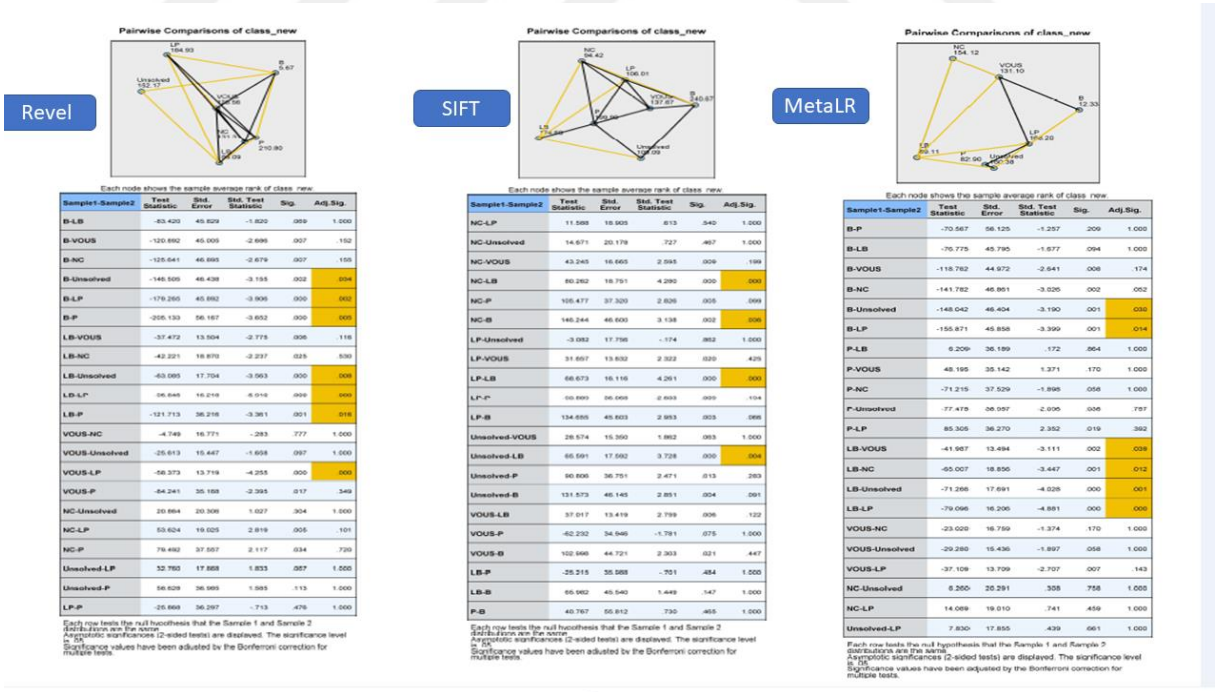
Grup 1	Grup 2	meandiff	p-adj	Alt sınır	Üst sınır	Hipotez reddi
LB	LP	-0.1863	0,0044	-0,3282	-0,0445	Evet
LB	Unresolved	-0.2445	0,0*	-0,3807	-0,1083	Evet
LB	VUS	-0.1893	0,0004	-0,3099	-0,0687	Evet
LP	Unresolved	-0.0582	0,6871	-0,1944	0,078	Hayır
LP	VUS	-0,003	0,9999	-0,1235	0,1176	Hayır
Unresolved	VUS	0,0552	0,5935	-0,0587	0,1691	Hayır

MetaLR skoruna göre Tukey testi sonuçları Tablo 8'de gösterilmiştir.

Tablo 7. Tukey testi kullanılarak gruplar arası farklılığın değerlendirilmesi (MetaLR)

Grup 1	Grup 2	meandiff	p-adj	Alt sınır	Üst sınır	Hipotez reddi
LB	LP	0,1594	0,0*	0,0795	0,2393	Evet
LB	Unresolved	0,1555	0,0*	0,0788	0,2322	Evet
LB	VUS	0,0713	0,0356	0,0033	0,1392	Evet
LP	Unresolved	-0,0039	0,9992	-0,0806	0,0728	Hayır
LP	VUS	-0,0881	0,005	-0,1561	-0,0202	Evet
Unresolved	VUS	-0,0842	0,0044	-0,1484	-0,0201	Evet

Kruskal Wallis-H testinin post-hoc test sonuçları aşağıda verilmiştir.



Şekil 6. Kruskall Wallis-H testinde istatistiki açıdan anlamlı farklılık bulunan skorların post- hoc değerlendirilmesi

(İkili karşılaştırmalarda Dunn-bonferroni testi kullanılmıştır.)

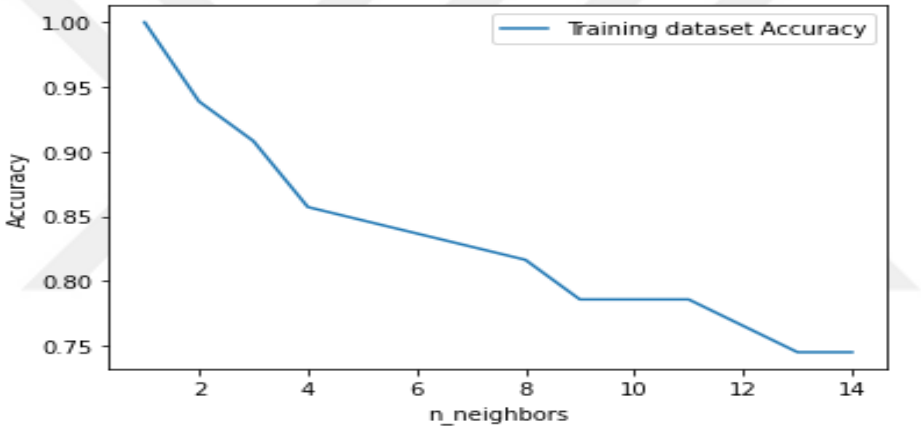
4.4. Makine Öğrenmesi Yöntemleri

4.4.1. Eğitim Setinin Değerlendirilmesi

Verisetinde yer alan LP ve LB varyantlar (n=98) eğitim kümesine dahil edildi. Bu kümeye dahil edilen varyantlar DAA, KNN, DVM, RO ve KA kullanılarak tahmin edilmeye çalışıldı. Eğitim kümesine dahil edilen varyantlar üzerinden Unresolved ve VUS varyantlar (n=168) predikte edilmeye çalışıldı. En başarılı sınıflandırmayı RO,KA ve KNN algoritmalarının yaptığı belirlendi.

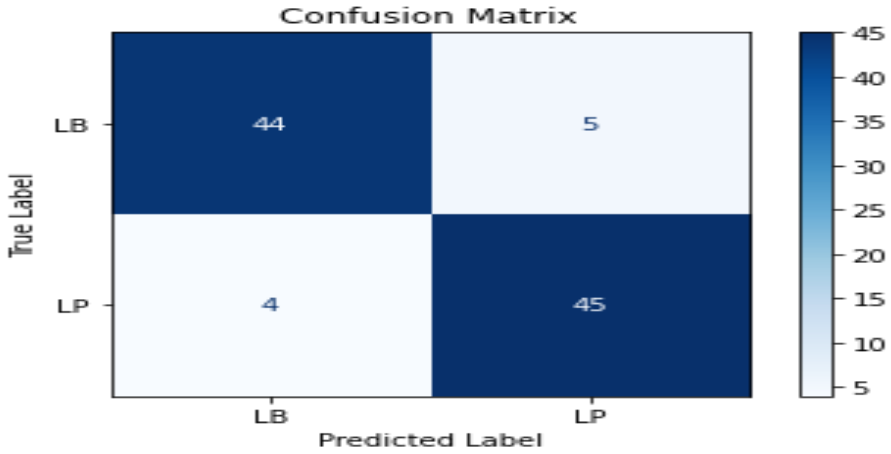
4.4.1.1. K en yakın komşuluk

KNN algoritmasında k komşuluk için saptanan doğruluk oranları Şekil 7.'de gösterildi. Buna göre, KNN için en uygun komşuluk sayısı üç (3) olarak belirlendi.



Şekil 7. KNN en uygun komşuluk sayısının gösterimi

Konfüzyon matrisi Şekil 8'de gösterilmiştir.



Şekil 8. Konfüzyon matrisi (KNN)

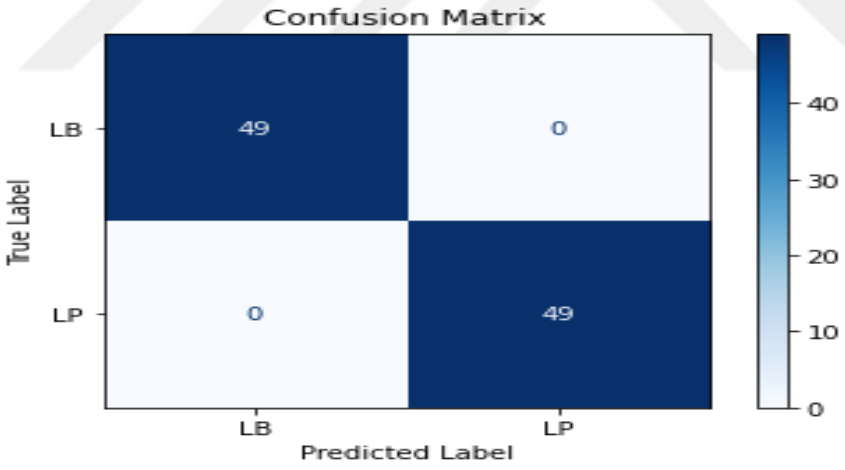
KNN'de kullanılan performans metriklerine ait sonuçlar Tablo 9'da gösterilmiştir.

Tablo 8. KNN eğitim verisetine ait performans metriklerinin gösterilmesi

	Kesinlik	**Duyarlılık**	**F1-skoru**	**Destek**
LB	0,92	0,90	0,91	49
LP	0,90	0,92	0,91	49
Doğruluk			0,91	98
Macro AVG	0,91	0,91	0,91	98
Weighted AVG	0,91	0,91	0,91	98

4.4.1.2. Rasgele Ormanlar ve Karar Ağaçları

RO'lar ve KA'lar verileri %100 doğrulukla sınıflandırmıştır. Konfüzyon matrisi Şekil 9'da gösterilmiştir.



Şekil 9. RO ve KA konfüzyon matrisinin gösterilmesi

RO ve KA'ya ait performans ölçütleri Tablo 10'da gösterilmiştir.

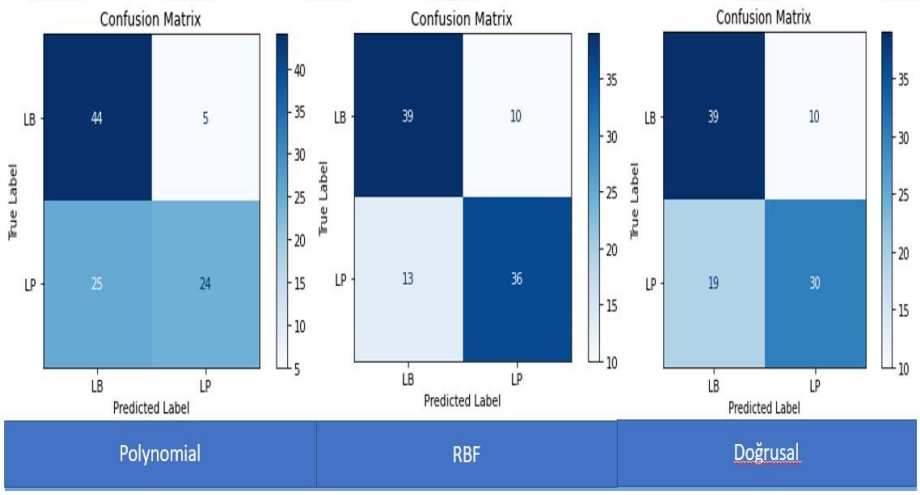
Tablo 9. Eğitim verisetinde kullanılan RO ve KA'nın performans metriklerine ait sonuçlar

	Kesinlik	**Duyarlılık**	**F1-skoru**	**Destek**
LB	1,00	1,00	1,00	49
LP	1,00	1,00	1,00	49
Doğruluk			1,00	98
Macro AVG	1,00	1,00	1,00	98
Weighted AVG	1,00	1,00	1,00	98

4.4.1.3. Destek Vektör Makineleri

Üç tip destek vektör makinesi metodu kullanıldı: Doğrusal, polinomiyal ve dairesel baz fonksiyonları(RBF). Konfüzyon matrisleri Şekil 10'da gösterilmiştir.

Şekil 10. Destek Vektör Makinesi yöntemlerine ait Konfüzyon matrisleri



Polinomiyal destek vektör makinesine ait performans metrikleri Tablo 11'de gösterilmiştir.

Tablo 10. Polinomiyal Destek Vektör Makinesine ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1-skoru**	**Destek**
LB	0,64	0,90	075	49
LP	0,83	0,49	0,62	49
Doğruluk			0,91	98
Macro AVG	0,73	0,69	0,68	98
Weighted AVG	0,73	0,69	0,68	98

Doğrusal DVM'ye ait performans metrikleri Tablo 12'de gösterilmiştir

Tablo 11. Doğrusal DVM'ye ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1-skoru**	**Destek**
LB	0,67	0,80	0,73	49
LP	0,75	0,61	0,67	49
Doğruluk			0,70	98
Macro AVG	0,71	0,70	0,70	98
Weighted AVG	0,71	0,70	0,70	98

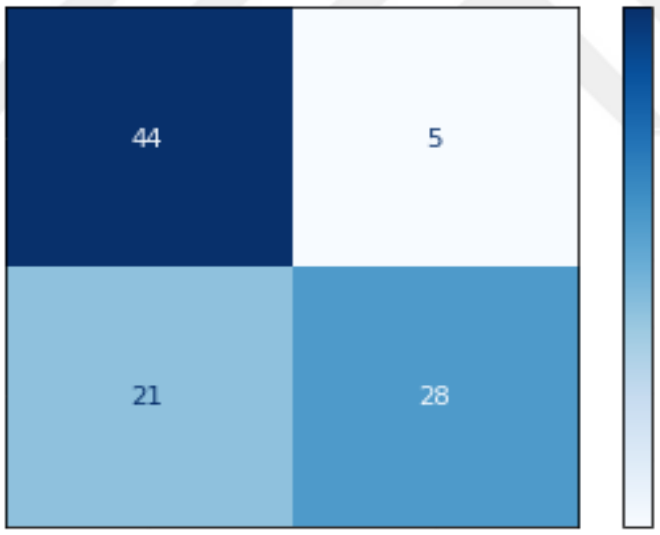
DVM (RBF)'e ait performans metrikleri Tablo 13'te gösterilmiştir.

Tablo 12. RBF'e ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1-skoru**	**Destek**
LB	0,75	0,80	0,77	49
LP	0,78	0,73	0,76	49
Doğruluk			0,77	98
Macro AVG	0,77	0,77	0,77	98
Weighted AVG	0,77	0,77	0,77	98

4.4.1.4. DAA

DAA'ya ait konfüzyon matrisi Şekil 11'de gösterilmiştir.



Şekil 11. DAA Konfüzyon Matrisi

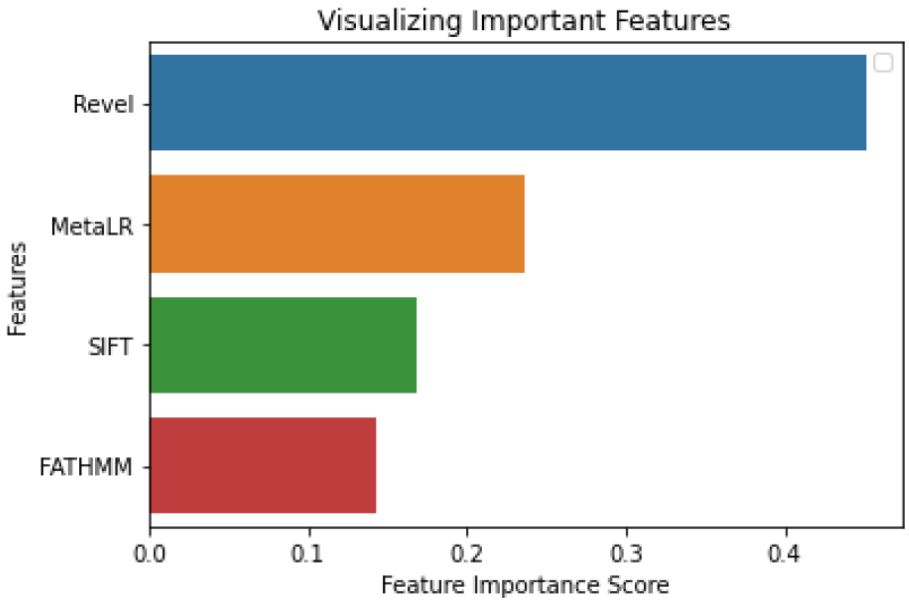
DAA'ya ait performans metrikleri Tablo 14'te gösterilmiştir.

Tablo 13. DAA'ya ait performans metrikleri

	Kesinlik	**Duyarlılık**	**F1-skoru**	**Destek**
LB	0,68	0,90	0,77	49
LP	0,85	0,57	0,68	49
Doğruluk			0,73	98
Macro AVG	0,76	0,73	0,73	98
Weighted AVG	0,76	0,73	0,73	98

4.4.1.5. Skorların Sınıflandırmaya Etkileri

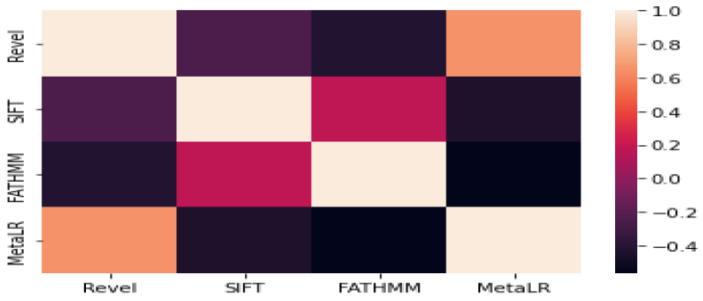
Skorların sınıflandırmaya olan etkileri Şekil 12.'de gösterilmiştir. Buna göre sonuçlar üzerinde en etkili olan skor Revel'dir.



Şekil 12. Skorların sınıflandırmaya olan etkileri.

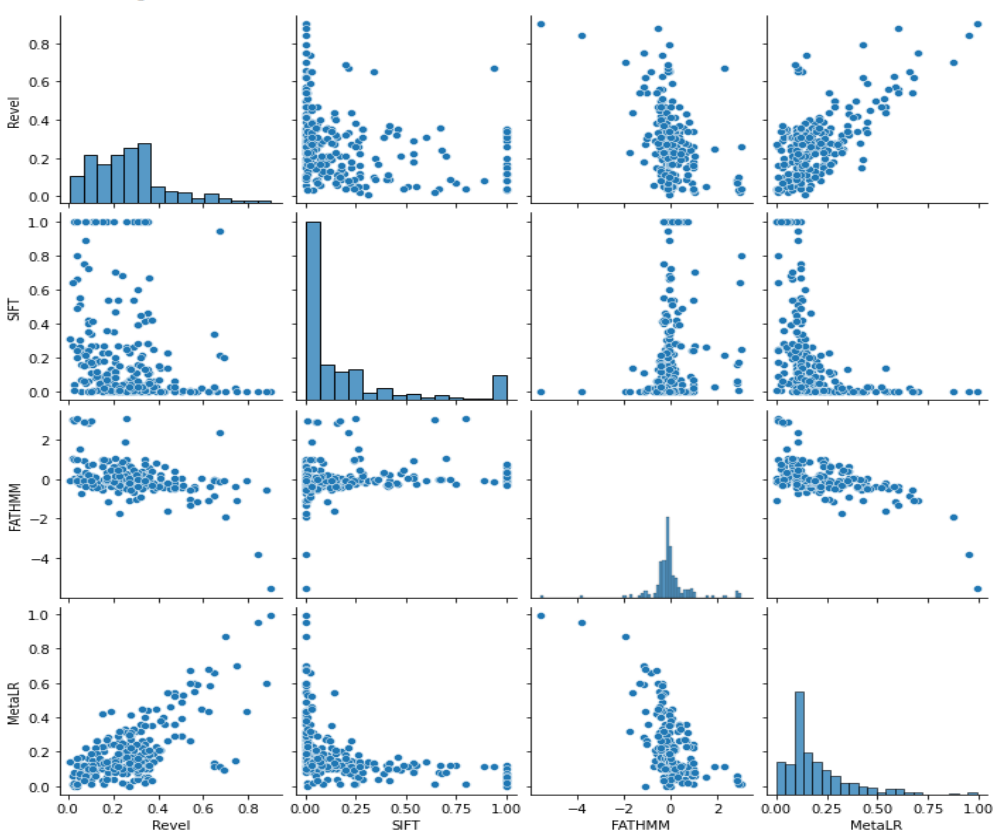
Revel: 0,451244, MetaLR: 0.236516, SIFT: 0.169073, FATHMM: 0.143167

Skorlara ait ısı haritaları Şekil 13.'te gösterilmiştir.



Şekil 13. Skorların katkısının ısı haritası şeklinde gösterilmesi

Skorlar arası korelasyon grafikleri Şekil 14.te gösterilmiştir. Buna göre MetaLR ve Revel skorları arasında orta düzeyde bir korelasyon bulunmaktadır ($r=0.649$, $p<0.001$). Ancak bu korelasyon düzeyi veriler arasında çoklu bağlantı sorunu ortaya çıkaracak kadar yüksek kabul edilmemektedir. Bu nedenle bu iki değişkenin de modelde yer almasına karar verilmiştir.



Şekil 14. Skorlar arası korelasyon grafiği

4.4.2. Prediksiyon Sonuçlarının Değerlendirilmesi

Prediksiyon sonuçları Tablo 15.'te verilmiştir. Prediksiyon sonuçlarına ait ham veriler ise Ek 1.'de yer almaktadır. Buna göre 7 skorun tamamının tahmin ettiği varyant sayısı 76'dır(%45,23). 6 skorun LP ya da LB tahmin ettiği varyant sayısı 38(%22.67)'dir. 5 skorun LP ya da LB tahmin ettiği varyant sayısı 25'tir(%14,88). Basamaklı tahmin algoritmamıza göre beş, altı ve yedi makine öğrenmesi metodunun aynı sınıflandırmayı yaptıkları varyantlar doğru olarak kabul edildi (%82.78). Buna göre 3 ve 4 makine öğrenmesi yönteminin aynı kabul ettiği 29 varyant ise belirsiz olarak değerlendirildi. Bu 29 varyanttan 14'ünde dört makine öğrenmesi metodu LP, 15'inde dört makine öğrenmesi metodu LB tespit edilmişti. Dört makine öğrenmesi metodunun LP dedikleri veriler belirsiz-patojene yakın, dört makinesi yönteminin LB dedikleri yöntemler ise belirsiz-benigne yakın olarak kabul edildi. 7 makine öğrenmesinin tamamının LP dedikleri patojen (n=33, %19,64) , 5 ve 6'sının LP dedikleri LP (n=30,%17,86) ; 7 makine öğrenmesinin tamamının LB dedikleri Benign(n=43,%25,6) , 5 ve 6'sının LP dedikleri LB (n=33,%19,64) olarak kabul edildi.

Tablo 14. Prediksiyon sonuçları

LB					**LP**				
					Sütun				
Satır	0	1	2	3	4	5	6	7	**Toplam**
0								33	33
1							19		19
2						11			11
3					**14**				14
4				**15**					15
5			14						14
6		19							19
7	43								43
Toplam	43	19	14	15	14	11	19	33	168

ML metodlarının prediksiyon sonuç ortalamaları Tablo 12'de standart sapmaları Tablo 13'te gösterilmiştir.

Tablo 15. Makine öğrenmesi yöntemi prediksiyon sonuçlarının skorlara göre ortalamalarının dağılımı

ML metodu	**Tahmin skoru ortalamaları**			
	SIFT	**FATHMM**	**MetalR**	**Revel**
RO				
LB	0.216117647	0.174	0.111294118	0.179764706
LP	0.086024096	-0.205301205	0.298554217	0.354337349
Genel Toplam	**0.151845238**	**-0.013392857**	**0.203809524**	**0.266011905**
KNN				
LB	0.224880952	0.159880952	0.11797619	0.1925
LP	0.078809524	-0.186666667	0.289642857	0.33952381
Genel Toplam	**0.151845238**	**-0.013392857**	**0.203809524**	**0.266011905**
RT				
LB	0.188648649	0.179189189	0.082432432	0.120405405
LP	0.326914894	0.130319149	-0.088829787	0.269468085
Genel Toplam	**0.266011905**	**0.151845238**	**-0.013392857**	**0.203809524**
DAA				
LB	0.177755102	0.212653061	0.048061224	0.124489796
LP	0.389571429	0.066714286	-0.099428571	0.314857143
Genel Toplam	**0.266011905**	**0.151845238**	**-0.013392857**	**0.203809524**
DVM Polinomiyal				
LB	0.200948276	0.195258621	0.107068966	0.137068966
LP	0.411153846	0.055	-0.282115385	0.352692308
Genel Toplam	**0.266011905**	**0.151845238**	**-0.013392857**	**0.203809524**
DVM RBF				
LB	0.183932584	0.22505618	0.081573034	0.130224719
LP	0.358481013	0.069367089	-0.120379747	0.286708861
Genel Toplam	**0.266011905**	**0.151845238**	**-0.013392857**	**0.203809524**
DVM Doğrusal kernel				
LB	0.172804878	0.273292683	0.323170732	0.108414634
LP	0.354883721	0.036046512	-0.334302326	0.294767442
Genel Toplam	**0.266011905**	**0.151845238**	**-0.013392857**	**0.203809524**

Tablo 16. Makine öğrenmesi yöntemi prediksiyon sonuçlarının skorlara göre standart sapmalarının dağılımı

ML metodu		Tahmin skoru standart sapmaları		
	Revel	**FATHMM**	**FATHMM**	**MetalR**
RO				
LB	0.117361636	0.270638989	0.794849791	0.054745912
LP	0.142898504	0.162006773	0.740521137	0.174824878
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**
KNN				
LB	0.131642771	0.273594961	0.836385646	0.061030898
LP	0.145824658	0.151557415	0.702596447	0.179815477
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**
RT				**StdSapma MetalR**
LB	0.122310231	0.235533535	0.786347218	0.074179331
LP	0.154766745	0.228762224	0.787842919	0.176856463
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**
DAA				
LB	0.086947438	0.261248197	0.677920803	0.066690893
LP	0.149680197	0.148692089	0.921647683	0.183340523
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**
DVM polynomial				
LB	0.098510458	0.250964959	0.683854825	0.074577269
LP	0.166199194	0.144581885	0.937866449	0.194110494
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**
DVM RBF	**StdSapma Revel**	**StdSapma SIFT**	**StdSapma FATHMM**	**StdSapma MetalR**
LB	0.122823025	0.267811516	0.957088268	0.109398952
LP	0.139072833	0.14717286	0.528200727	0.166122611
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**
DVM Doğrusal kernel	**StdSapma Revel**	**StdSapma SIFT**	**StdSapma FATHMM**	**StdSapma MetalR**
LB	0.11635699	0.279483893	0.832052168	0.05532105
LP	0.138179288	0.061916055	0.592660086	0.17235208
Genel Toplam	**0.156901662**	**0.232347149**	**0.789444457**	**0.159162514**

5. TARTIŞMA

FMF tekrarlayan ateş, karın ağrısı ve seröz zarların tutulumu ile giden otoinflamatuar bir hastalıktır. Uygun tedavi alınmadığında amiloidoz gibi ölümcül olabilecek ciddi komplikasyonlara yol açmaktadır(65). Bu nedenle, FMF hastalığının tanısının konulması oldukça önemlidir. Ancak, hastalığın diğer otoinflamatuar hastalıklarla ayırıcı tanısını yapmak her zaman için kolay değildir. Bu anlamda genetikte etkisi bilinen mutasyonlara yönelik çalışılan birtakım testler bulunmaktadır. Bu testlerde sonucu tespit edilen mutasyonlar hem ayırıcı tanıda hem hastalığın seyrinde hem de tedaviye yanıtın değerlendirilmesinde büyük kolaylıklar sağlamaktadır. Genetik test sonuçları bu açıdan hastalığın takibini yapan klinisyenlere ayırıcı tanıda, hasta takibinde ve tedavinin düzenlenmesinde büyük kolaylıklar sağlamaktadır.

Ancak yeni nesil dizileme teknikleri yapılan otoinflamatuar panellerde saptanan MEFV varyantlarının önemli bir kısmının klinik önemi bilinmemeltedir(61). FMF hastalığının kalıtılabilir bir hastalık olduğu gözönünde bulundurulduğunda bu varyantların ayrımının doğru yapılması sadece hasta açısından değil aynı zamanda sonraki nesiller için de büyük bir öneme sahiptir. Çalışmamızda SIFT, FATHMM, REVEL ve MetaLR skorlarının (4 skor) eğitim setinde (LB ve LP) RO ve KA metotları kullanıldığında bilinen *MEFV* varyantlarını %100 doğrulukla sınıflandırıldığını tespit ettik. 7 makine öğrenmesi metodu kullanılarak yapılan basamaklı tahminleme algoritmamız önemi bilinmeyen ve sınıflandırılamayan 168 varyantın 139'unu (%82) sınıflandırmayı başardı. Çalışmamız, 4 tahmin skorunun beraber kullanılarak *MEFV v*aryantlarının makine öğrenmesi yöntemleri ile sınıflandırıldığı ilk çalışma özelliğini taşımaktadır.

Bu çalışmanın birçok güçlü yönü bulunmaktadır. Bunlardan ilki ve çalışmanın getirdiği en önemli yenilik sadece tek bir ML yöntemine odaklanılmamış 7 adet ML yönteminin hepsinin değerlendirildiği basamaklı bir değerlendirme stratejisinin kullanılmasıdır. İkincisi, algoritmasına tek bir skor değil 4 adet skor dahil etmiş olmasıdır. Bu sayının çok fazla olmamasına dikkat edilmiş az sayıda veri seti ile mümkün olan en doğru sonuç elde edilmeye çalışılmıştır. Skorlar seçilirken eksik veri olmamasına ve aralarında çoklu bağlantı sorunu olmamasına dikkat edilmiştir. İkincisi, sınıflandırma temel olarak iki gruba göre (LB/LP) yapılmıştır. Klasik ML algoritmaları en iyi iki grup üzerinden değerlendirme yaptığı için tahmin

edilen değişken iki kategoriden oluşmaktadır. Sadece iki kategorinin yer aldığı bu sınıflamayla çok boyutluluk lanetinin getireceği birtakım zorlukların da önüne geçilmesi sağlanmaktadır. Bu manada özgün bir bir skorlama sistemi kullanılmıştır.

Güncel olarak birçok metaprediktör ve tahmin skoru bulunmaktadır. Ancak varyantları değerlendirmede *in silico* tahmin programları tarafından kullanılan ve kabul görmüş veritabanlarını seçmeyi esas aldık. Bu nedenle hem varsome hem de franklin genooxun tahmin algoritmasında kullanılan tahmin programlarını seçtik. Ayrıca, kullanılacak skor seçiminde eksik veri olmaması şartını aradık. Son olarak, yapılan çalışmalarda nadir varyantları değerlendirirken hangi skorların en etkili sınıflandırma yaptığını değerlendirdik. Bu çalışmalara göre, Revel skorunun Allel frekansının (AF) %3 ve altında daha olan varyantlar için MetaLR skorunun ise AF'ının %3-5 olan varyantlar için en uygun prediksyonu yapmaktaydı(12). Bu nedenle *MEFV* geninde saptanan bu varyantların tahmininde bu iki skorun tahmin etkisinin en yüksek olacağını öngördük. Bu nedenle bu iki skoru tahmin algoritmasına dahil etmeyi planladık. Sonrasında her iki veriseti arasında çoklu uyum problemi olup olmadığını inceledik. Her iki metaprediktör arasında çoklu uyum sorunu olmadığını tespit ettik ve algoritmaya dahil ettik. Benzer bir şekilde Revel ve MetaLR skorları algoritmalarına dahil ettikleri tahmin skorları arasında SIFT ve FATHMM skorları da yer almaktaydı. Ancak dahil edilen metaprediktörler ve tahmin skorları arasında hiçbir çoklu bağlantı problemi tespit edilemedi. Sonrasında, uç değer analizi yaptıktan sonra sonuca etki edecek aşırı uç değer tespit edilemedi. Bu nedenle bütün değişkenler analize dahil edildi.

Verisetimizde az sayıda benign ve patojen varyant bulunmaktadır. ACMG kılavuzuna göre varyantların değerlendirilirken LP ve P varyantların değerlendilmesi açısından bir farklılık bulunmamaktadır(43,66). Yapılan çalışmalarda da B ve LB arasında klinik açıdan önemli bir farklılık bulunmamaktadır. Benzer bir şekilde P ve LP varyantlar arasında da klinik açıdan önemli bir farklılık bulunmamaktadır. Bu nedenle *MEFV* varyantlarının sınıflandırıldığı bir çalışmada dahil olmak üzere yapılan önceki birçok çalışmada B ve LB varyantlar LB, P ve LP varyantlar ise LP olarak değerlendirilmiştir(61,66). Biz de çalışmamızda LB/B varyantları LB ve LP/P varyantları ise LP olarak değerlendirdik. Yaptığımız bu boyut indirgeme işlemi bize hem klinik hem de istatistiki açından önemli iki avantaj sundu. İlki, eğitim setimiz için dengeli bir veriseti elde etmiş olduk. İkincisi, değişken sayısını mümkün olan en optimum düzeye indirdik. Benzer algoritmayı klinik olarak sınıflandırılamayan unresolved, not classified ve VUS varyantlara da uyguladık. Bu varyantları VUS başlığı altında topladık.

Ancak çalışmamızın birtakım kısıtlılıkları bulunmaktadır. Çalışmamıza dahil edilen varyant sayısı bildirilen varyant sayısı infevers veritabanında bildirilen varyant sayısı ile sınırlıdır. Ancak bu durum, kullanılan tahmin programı sayısı arttırılarak minimalize edilmeye çalışılmıştır. Tahmin programları seçilirken sayının optimum düzeyde tutulmasına dikkat edilmiş olup, bu anlamda 4 araç seçilmiştir. Kullanılan 4 adet aracın da çalışmaya dahil edilen bütün varyantları açısından eksik veri içermeyen tahmin programının olmasına dikkat edilmiştir.

Çalışmaya dahil edilen skorların kürete edilmesinde dbNSFP'nin yanısıra Franklin genoox veritabanını da kullanılmıştır. Ancak franklin genoox veritabanında kullanılan onlarca tahmin aracı ve metaprediktörden birçoğu herhangibir skor üretmemekte bu durum veritabanında yer alan birçok varyant klinik önemi bilinmeyen varyant olarak değerlendirilmesinde etkin bir rol oynamaktadır. VUS varyantlar klinisyenlerin nadir görülen varyantları yorumlamasında zorluk ortaya çıkarmaktadır. Bu veritabanlarının en büyük eksikliği varyantların gene özgü olarak değerlendirilmemesidir. Oysa skorların kullanılmasında gen spesifik bir yaklaşım denenebilir. Bu sebeple, biz bu çalışmaya eksik veri içermeyen iki metaprediktör ve iki aminoasit değişim skorunu dahil ettik.

MEFV gen varyantlarının sınıflandırılmasında iyileştirmeye yönelik olarak yapılan bir çalışmada varyantlar Revel skorları kullanılarak tekrar sınıflandırılmıştır. Bu çalışmaya göre, varryantlar dört gruba ayrılmış (LB, VUS, LP, Unsolved) , bu dört grubun ortalama ve standart sapma değerlerine göre eşik bir değer belirlenmiş, sonrasında varyantlar yeniden oluşturulan ortalama ve standart sapmalara göre üç gruba ayrılıp (LB, VUS, LP) yeniden değerlendirildiklerinde iyi skoru % 75 doğrulukla (sensitivite: %62,5, spesifite:%87,5) DDA'nın verdiğini tespit etmişlerdir. Çalışmamızla karşılaştırıldığında eğitim verisinin doğruluğu oldukça düşüktür. Bu çalışmada kullanılan 5 makine öğrenmesi üzerinden en iyi sınıflandırma yapan makine öğrenmesi seçilmiş ve sınıflandırma sadece Revel skorları üzerinden yapılmıştır. Diğer makine öğrenmesi metotlarının etkisi dahil edilmemiştir. Buna göre LB ve LP varyantlar üzerinden yeni bir cut off değeri seçilmiş. Sonraki sınıflamalar bu cutoff değeri kullanılarak yapılmıştır. Bu sınıflandırmanın başarılı olması için daha fazla varyantın çalışmaya dahil olması gerekmektedir. Ayrıca, her dahil edilen veriyle birlikte çalışmada tekrar bir eşik değer hesaplanması gerekmektedir. Bu nedenle, pratikte uygulanması kolay değildir. Çalışmamızda ise basamaklı bir değerlendirme algoritması kullanılmış olup kullanılan bütün makine öğrenmesi metotlarının etkisi dahil edilmiştir.Bu nedenle,

kullandığımız metotun pratikte uygulanması daha kolaydır ve her eklenen veriyle birlikte eşik değer hesaplamayı gerektirmez(61).

Sonuç olarak, Basamaklı sınıflandırma metodu *MEFV* varyantlarının sınıflandırılmasında mevcut algoritmalar arasında en iyi sonucu vermektedir. *MEFV* varyantlarının yaklaşık %60'ının önemi bilinmeyen varyant olarak değerlendirildiği göz önüne alındığında mevcut değerlendirme yöntemi klinik bilgilerin eşliğinde klinisyenlere varyant yorumu için daha kesin bir yorumlama imkanı ve kolaylık sağlayabilir. Ancak, basamaklı sınıflandırma algoritmasının diğer genler üzerindeki etkinliğinin belirlenebilmesi için daha fazla sayıda çalışmaya ihtiyaç vardır.

6. SONUÇ VE ÖNERİLER

- RO ve KA, eğitim verisinde yer alan *MEFV* gen varyantlarını %100 doğrulukla sınıflandırmıştır.
- Basamaklı sınıflandırma yaparak bütün makine öğrenmesi metotlarının etkisi sınıflandırmaya dahil edilmektedir.
- Birden fazla skor kullanılarak makine öğrenmesi metotlarının tahmin gücü arttırılabilir.
- Kullanılan 4 skor arasında *MEFV* varyantlarının sınıflandırılmasında en fazla etkiye sahip skor Revel'dir.
- Basamaklı sınıflandırma metodumuz klinik önemi bilinmeyen *MEFV* varyantlarının sınıflandırılmasında pratik ve kullanımı kolay bir metot olarak kullanılabilir.

KAYNAKLAR

1. Andresen SL. John McCarthy: father of AI. IEEE Intell Syst. 2002;17(5):84–5.

2. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017 Apr;69S:S36–40.

3. Haenlein M, Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. Calif Manage Rev. 2019;61(4):5–14.

4. Sipper M. Fifty years of research on self-replication: An overview. Artif Life. 1998;4(3):237–57.

5. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc [Internet]. 2020;92(4):807–12. Available from: https://www.sciencedirect.com/science/article/pii/S0016510720344667

6. Fradkov AL. Early History of Machine Learning. IFAC-PapersOnLine [Internet]. 2020;53(2):1385–90. Available from: https://www.sciencedirect.com/science/article/pii/S2405896320325027

7. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33.

8. Vargas R, Mosavi A, Ruiz R. Deep learning: a review. 2017;

9. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019 May;20(5):e262–73.

10. Shanthi D, Narsimha G, Mohanthy RK. Human intelligence vs. artificial intelligence: Survey. Int J Electron Commun Comput Eng. 2015;6(5):30–4.

11. Bellinger G, Castro D, Mills A. Data, information, knowledge, and wisdom. 2004.

12. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al.

REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct;99(4):877–85.

13. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform [Internet]. 2006 Mar 1;7(1):86–112. Available from: https://doi.org/10.1093/bib/bbk007

14. Albaradei S, Thafar M, Alsaedi A, Van Neste C, Gojobori T, Essack M, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. Comput Struct Biotechnol J. 2021;19:5008–18.

15. Olisah CC, Smith L, Smith M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput Methods Programs Biomed. 2022 Jun;220:106773.

16. Atilla E, Öztuna D, Alpar R. Ayırma(Diskriminant)Analizi. In: Alpar R, editor. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 5th ed. Ankara: Detay Yayıncılık; 2017. p. 671–730.

17. Karabulut E, Alpar R. Logistik Regresyon. In: Alpar R, editor. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 5th ed. Ankara: Detay Yayıncılık; 2017. p. 591–660.

18. Diao JA, Kohane IS, Manrai AK. Biomedical informatics and machine learning for clinical genomics. Hum Mol Genet. 2018 May;27(R1):R29–34.

19. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. Molecules. 2020 Nov;25(22).

20. Baştanlar Y, Ozuysal M. Introduction to machine learning. Methods Mol Biol. 2014;1107:105–28.

21. Suthaharan S. Support Vector Machine BT - Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. In: Suthaharan S, editor. Boston, MA: Springer US; 2016. p. 207–35. Available from: https://doi.org/10.1007/978-1-4899-7641-3_9

22. Wang T, Zhao D, Tian S. An overview of kernel alignment and its applications. Artif Intell Rev [Internet]. 2015;43(2):179–92. Available from: https://doi.org/10.1007/s10462-012-9369-4

23. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018;15(1):41–51.

24. Noble WS. What is a support vector machine? Nat Biotechnol [Internet]. 2006;24(12):1565–7. Available from: https://doi.org/10.1038/nbt1206-1565

25. García-Magariños M, López-de-Ullibarri I, Cao R, Salas A. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Ann Hum Genet. 2009 May;73(Pt 3):360–9.

26. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int J Med Inform. 2021 Jul;151:104484.

27. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med. 2016 Feb;44(2):368–74.

28. Langarizadeh M, Moghbeli F. Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. Acta Inform medica AIM J Soc Med Informatics Bosnia Herzegovina Cas Drus za Med Inform BiH. 2016 Oct;24(5):364–9.

29. Hall GH. The clinical application of Bayes' theorem. Lancet (London, England). 1967 Sep;2(7515):555–7.

30. Karaboga HA, Gunel A, Korkut SV, Demir I, Celik R. Bayesian Network as a Decision Tool for Predicting ALS Disease. Brain Sci. 2021 Jan;11(2).

31. Skansi S. Machine Learning Basics BT - Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. In: Skansi S, editor. Cham: Springer International Publishing; 2018. p. 51–77. Available from: https://doi.org/10.1007/978-3-319-73004-

2_3

32. Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Mach Learn [Internet]. 1997;29(2):103–30. Available from: https://doi.org/10.1023/A:1007413511361

33. Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008;26(9):1011–3.

34. Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Cent Eur J Oper Res [Internet]. 2018;26(1):135–59. Available from: https://doi.org/10.1007/s10100-017-0479-6

35. Boulesteix A, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2012;2(6):493–507.

36. Roohi A, Faust K, Djuric U, Diamandis P. Unsupervised Machine Learning in Pathology: The Next Frontier. Surg Pathol Clin. 2020 Jun;13(2):349–58.

37. Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015;61:85–117.

38. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49:107739.

39. Tufan A, Lachmann HJ. Familial Mediterranean fever, from pathogenesis to treatment: a contemporary review. Turkish J Med Sci. 2020 Nov;50(SI-2):1591–610.

40. Turkish FMF Study Group. Familial Mediterranean Fever (FMF) in Turkey: Results of a Nationwide Multicenter Study. Medicine (Baltimore) [Internet]. 2005;84(1). Available from: https://journals.lww.com/md-journal/Fulltext/2005/01000/Familial_Mediterranean_Fever__FMF__in_Turkey_.1.aspx

41. Booty MG, Chae JJ, Masters SL, Remmers EF, Barham B, Le JM, et al. Familial Mediterranean fever with a single MEFV mutation: where is the second hit? Arthritis

Rheum. 2009 Jun;60(6):1851–61.

42. Kırnaz B, Gezgin Y, Berdeli A. MEFV gene allele frequency and genotype distribution in 3230 patients' analyses by next generation sequencing methods. Gene. 2022 Jun;827:146447.

43. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. Genet Med Off J Am Coll Med Genet. 2015 May;17(5):405–24.

44. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho Y-Y, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet Med Off J Am Coll Med Genet. 2017 Oct;19(10):1105–17.

45. Davydov E V, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010 Dec;6(12):e1001025.

46. Ramani R, Krumholz K, Huang Y-F, Siepel A. PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. Bioinformatics. 2019 Jul;35(13):2320–2.

47. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul;31(13):3812–4.

48. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics [Internet]. 2015 Aug 15;31(16):2745–7. Available from: https://doi.org/10.1093/bioinformatics/btv195

49. Sandell L, Sharp NP. Fitness Effects of Mutations: An Assessment of PROVEAN Predictions Using Mutation Accumulation Data. Genome Biol Evol. 2022 Jan;14(1).

50. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20.

51. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8(11):R232.

52. Bhat GR, Sethi I, Rah B, Kumar R, Afroze D. Innovative in Silico Approaches for Characterization of Genes and Proteins. Front Genet. 2022;13:865182.

53. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Vol. 7, Nature methods. United States; 2010. p. 575–6.

54. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan;47(D1):D886–94.

55. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015 Mar;31(5):761–3.

56. Oluwole OG, Kuivaniemi H, Abrahams S, Haylett WL, Vorster AA, van Heerden CJ, et al. Targeted next-generation sequencing identifies novel variants in candidate genes for Parkinson's disease in Black South African and Nigerian patients. BMC Med Genet. 2020 Feb;21(1):23.

57. Knecht C, Mort M, Junge O, Cooper DN, Krawczak M, Caliebe A. IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants. Nucleic Acids Res. 2017 Feb;45(3):e13.

58. Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, et al. MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One. 2020;15(7):e0236962.

59. Li S, van der Velde KJ, de Ridder D, van Dijk ADJ, Soudis D, Zwerwer LR, et al. CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. Genome Med. 2020 Aug;12(1):75.

60. Favalli V, Tini G, Bonetti E, Vozza G, Guida A, Gandini S, et al. Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. Am J Hum Genet. 2021 Apr;108(4):682–95.

61. Accetturo M, D'Uggento AM, Portincasa P, Stella A. Improvement of MEFV gene variants classification to aid treatment decision making in familial Mediterranean fever. Rheumatology (Oxford). 2020 Apr;59(4):754–61.

62. Hart SN, Polley EC, Shimelis H, Yadav S, Couch FJ. Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. npj Breast Cancer [Internet]. 2020;6(1):13. Available from: https://doi.org/10.1038/s41523-020-0159-x

63. Rodrigues E da S, Griffith S, Martin R, Antonescu C, Posey JE, Coban-Akdemir Z, et al. Variant-level matching for diagnosis and discovery: Challenges and opportunities. Hum Mutat. 2022 Jun;43(6):782–90.

64. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med [Internet]. 2020;12(1):103. Available from: https://doi.org/10.1186/s13073-020-00803-9

65. Bustaffa M, Koné-Paut I, Ozen S, Amaryan G, Papadopoulou-Alataki E, Gallizzi R, et al. The impact of the Eurofever criteria and the new Infevers MEFV classification in real life: results from a large international FMF cohort. In: Seminars in Arthritis and Rheumatism. Elsevier; 2022. p. 151957.

66. Accetturo M, Bartolomeo N, Stella A. In-silico Analysis of NF1 Missense Variants in ClinVar: Translating Variant Predictions into Variant Interpretation and Classification. Int J Mol Sci. 2020 Jan;21(3).

EK-1

Çalışmadaki bütün varyantlara ait 7 makine öğrenmesi sonuçları aşağıda gösterilmiştir.

Varyantlar	LDA	RTree	SVM_poly	SVM_rbf	SVM_li	Rforest_x	KNN_x	Eski	LP	LB	puanlar
c.224G>A	LP	LP	LP	LP	LB	LP	LP	Unresolved	6	1	5
c.260A>G	LB	LP	LP	LP	LB	LP	LP	Unresolved	5	2	3
c.343C>A	LB	LB	LB	LP	LB	LB	LB	Unresolved	1	6	-5
c.359C>T	LB	LB	LB	LP	LB	LB	LP	Unresolved	2	5	-3
c.422G>T	LP	LP	LP	LP	LB	LB	LP	Unresolved	5	2	3
c.428G>C	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.476G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.505C>G	LB	LP	LB	LB	LP	LP	LP	Unresolved	4	3	1
c.530C>T	LB	LP	LB	LB	LB	LB	LB	Unresolved	1	6	-5
c.536G>A	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.581T>C	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.605G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.688G>C	LB	LB	LB	LB	LP	LB	LP	Unresolved	2	5	-3
c.694T>C	LB	LP	LB	LB	LB	LP	LP	Unresolved	3	4	-1
c.722G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.739A>G	LP	LP	LB	LP	LP	LP	LP	Unresolved	6	1	5
c.775A>G	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.796A>G	LP	LB	LB	LP	LP	LP	LB	Unresolved	4	3	1
c.803C>T	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.808A>G	LP	LP	LB	LP	LP	LP	LP	Unresolved	6	1	5
c.818C>T	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7

c.829C>A	LB	LP	LB	LB	LB	LB	LB	Unresolved	1	6	-5
c.833G>C	LB	LB	LB	LP	LB	LB	LP	Unresolved	2	5	-3
c.896A>G	LP	LP	LP	LB	LP	LP	LP	Unresolved	6	1	5
c.910G>A	LB	LP	LB	LB	LB	LB	LB	Unresolved	1	6	-5
c.949G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.986G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.1006G>A	LB	LP	LP	LB	LP	LP	LB	Unresolved	4	3	1
c.1127A>G	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.1171G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.1318C>G	LP	LB	LB	LP	LP	LP	LB	Unresolved	4	3	1
c.1370C>T	LB	LP	LB	LB	LP	LP	LP	Unresolved	4	3	1
c.1382G>A	LB	LP	LB	LB	LP	LP	LB	Unresolved	3	4	-1
c.1405G>T	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.1502G>A	LP	LP	LB	LB	LB	LB	LB	Unresolved	2	5	-3
c.1516A>G	LP	LP	LP	LB	LB	LP	LP	Unresolved	5	2	3
c.1744A>C	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.1764G>A	LP	LP	LP	LB	LB	LP	LB	Unresolved	4	3	1
c.1773T>G	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.1776C>T	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.1828A>G	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.1883G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.1895G>C	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.1996A>G	LP	LP	LB	LP	LP	LP	LP	Unresolved	6	1	5
c.2024G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.2068G>C	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.2210G>A	LP	LP	LB	LP	LP	LP	LP	Unresolved	6	1	5
c.2314A>G	LP	LB	LB	LP	LP	LP	LP	Unresolved	5	2	3
c.2338C>A	LB	LP	LB	LB	LB	LP	LP	Unresolved	3	4	-1
c.74A>G	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.115A>G	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.250G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7

c.265G>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.501G>C	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.623G>C	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.724A>G	LB	LB	LB	LB	LB	LB	LB	Unresolved	0	7	-7
c.726C>A	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.726C>G	LP	LP	LP	LP	LP	LP	LP	Unresolved	7	0	7
c.800C>T	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.938C>A	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.1060C>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1115T>C	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.1151T>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1166A>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1186C>T	LP	LB	LB	LP	LP	LB	LP	VUS	4	3	1
c.1207G>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1437C>G	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5
c.1501C>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1508C>G	LB	LB	LB	LB	LP	LB	LB	VUS	1	6	-5
c.1729A>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1729A>T	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.1730C>A	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.1730C>G	LB	LP	LB	LP	LP	LB	LB	VUS	3	4	-1
c.1949C>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1967A>C	LB	LB	LP	LB	LP	LB	LB	VUS	2	5	-3
c.1975G>T	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.2002G>A	LB	LP	LP	LP	LP	LP	LB	VUS	5	2	3
c.2038A>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2038A>G	LB	LP	LP	LB	LP	LP	LP	VUS	5	2	3
c.2040G>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2040G>C	LB	LB	LB	LB	LP	LB	LB	VUS	1	6	-5
c.2040G>T	LP	LB	LB	LP	LP	LB	LB	VUS	3	4	-1
c.2060G>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7

c.2063A>G	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5
c.2063A>T	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.2080A>T	LB	LB	LP	LB	LP	LB	LB	VUS	2	5	-3
c.2080A>G	LB	LB	LB	LB	LP	LB	LP	VUS	2	5	-3
c.2081T>A	LP	LP	LB	LP	LP	LP	LB	VUS	5	2	3
c.2082G>A	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5
c.2084A>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2085G>C	LB	LB	LB	LB	LP	LB	LP	VUS	2	5	-3
c.2177T>C	LP	LP	LP	LP	LP	LB	LP	VUS	6	1	5
c.2228T>A	LB	LB	LB	LB	LP	LB	LP	VUS	2	5	-3
c.2259A>T	LB	LP	LB	LB	LP	LP	LP	VUS	4	3	1
c.2281C>T	LP	LP	LB	LP	LP	LP	LB	VUS	5	2	3
c.2282G>A	LB	LP	LB	LP	LP	LP	LP	VUS	5	2	3
c.2296A>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2305C>G	LP	LB	LP	LP	LP	LP	LP	VUS	6	1	5
c.35C>T	LB	LP	LB	LB	LP	LP	LP	VUS	4	3	1
c.56A>G	LP	LB	LB	LP	LP	LB	LB	VUS	3	4	-1
c.101A>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.104A>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.124C>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.149C>T	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.184G>T	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.311C>G	LB	LP	LB	LP	LB	LP	LP	VUS	4	3	1
c.489G>T	LP	LP	LB	LP	LB	LB	LB	VUS	3	4	-1
c.560G>A	LP	LB	LP	LP	LP	LB	LB	VUS	4	3	1
c.578C>T	LP	LP	LP	LP	LB	LP	LP	VUS	6	1	5
c.611G>A	LB	LP	LB	LP	LB	LP	LP	VUS	4	3	1
c.675G>C	LB	LB	LB	LB	LP	LB	LB	VUS	1	6	-5
c.704C>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.707G>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.730G>A	LB	LP	LB	LB	LB	LP	LP	VUS	3	4	-1

c.941G>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.955G>A	LP	LP	LP	LP	LP	LP	LB	VUS	6	1	5
c.983T>C	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5
c.1016C>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1043G>A	LB	LP	LB	LP	LB	LB	LP	VUS	3	4	-1
c.1057C>G	LB	LP	LP	LP	LB	LB	LP	VUS	4	3	1
c.1099C>G	LB	LP	LB	LB	LB	LP	LB	VUS	2	5	-3
c.1211A>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1229G>A	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.1268T>C	LB	LB	LB	LB	LB	LB	LP	VUS	1	6	-5
c.1368A>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1450C>A	LB	LP	LB	LB	LB	LP	LP	VUS	3	4	-1
c.1459G>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1538T>C	LB	LB	LB	LB	LP	LB	LB	VUS	1	6	-5
c.1559A>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1675C>T	LP	LP	LB	LP	LP	LP	LB	VUS	5	2	3
c.1792G>A	LB	LP	LB	LB	LP	LP	LB	VUS	3	4	-1
c.1876T>A	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.1894G>A	LB	LB	LB	LB	LB	LB	LP	VUS	1	6	-5
c.1898C>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1907T>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.1910A>G	LB	LB	LB	LP	LB	LB	LB	VUS	1	6	-5
c.1925T>C	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.1945C>T	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.1981G>A	LB	LP	LB	LB	LB	LP	LB	VUS	2	5	-3
c.1981G>T	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2012A>T	LB	LB	LB	LP	LB	LP	LP	VUS	3	4	-1
c.2069T>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2072T>G	LP	LB	LB	LB	LB	LB	LB	VUS	1	6	-5
c.2078T>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2079G>C	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7

c.2084A>T	LB	LP	LB	LB	LB	LB	LP	VUS	2	5	-3
c.2110G>A	LB	LB	LP	LB	LP	LP	LB	VUS	3	4	-1
c.2113C>T	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.2126T>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2149C>A	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5
c.2189C>G	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.2198A>G	LB	LP	LB	LB	LP	LB	LB	VUS	2	5	-3
c.2229C>G	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.2230G>A	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.2261C>G	LP	LP	LB	LP	LP	LP	LP	VUS	6	1	5
c.2272C>T	LP	LB	LB	LP	LP	LP	LP	VUS	5	2	3
c.82C>G	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5
c.200T>A	LP	LB	LP	LP	LP	LP	LP	VUS	6	1	5
c.289C>A	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.290A>G	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.314C>A	LP	LB	LB	LP	LP	LB	LB	VUS	3	4	-1
c.322A>G	LB	LP	LB	LP	LB	LP	LP	VUS	4	3	1
c.322A>C	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.329T>C	LB	LP	LB	LB	LB	LP	LP	VUS	3	4	-1
c.331G>A	LB	LB	LB	LB	LP	LB	LP	VUS	2	5	-3
c.332G>A	LP	LP	LP	LP	LP	LP	LP	VUS	7	0	7
c.344C>G	LB	LB	LB	LB	LB	LB	LB	VUS	0	7	-7
c.369C>A	LB	LP	LB	LB	LB	LB	LB	VUS	1	6	-5

İNTİHAL RAPORU İLK SAYFASI

MEFV GEN VARYANTLARININ SINIFLANDIRILMASI ESNASINDA ORTAYA ÇIKAN BELİRSİZLİKLERİN YAPAY ZEKA ALGORİTMALARIYLA ÇÖZÜMLENMESİ

ORİJİNALLİK RAPORU

%17	%16	%10	%10
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Öğrenci Ödevi	%2
2	test.cl.ut.ee İnternet Kaynağı	%2
3	acikbilim.yok.gov.tr İnternet Kaynağı	%1
4	docplayer.es İnternet Kaynağı	%1
5	motto.tc İnternet Kaynağı	%1
6	archive.org İnternet Kaynağı	%1
7	Submitted to Hacettepe University Öğrenci Ödevi	%1
8	www.tibbigenetik2018.org İnternet Kaynağı	%1