



Pathogenicity assessment of breast cancer associated BRCA1/2 variants using explainable machine learning approach

Burak Kolukisa*^{id}

Department of Software Engineering, Faculty of Engineering, Architecture, and Design, Kayseri University, 38280, Kayseri, Türkiye

Highlights:

- ClinVar-derived dataset of BRCA1/2 variants
- Comparative analysis of classification models
- Accuracy and macro F1-score optimized via Bayesian optimization

Keywords:

- Machine Learning
- Bayesian Optimization
- Explainable AI
- Breast Cancer
- BRCA1/2 Variants
- Classification

Article Info:

Research Article

Received: 10.09.2025

Accepted: 10.04.2026

DOI:

10.17341/gazimmfd.1781575

Acknowledgement:

This research was supported by the TÜBİTAK TEYDEB program under Project Number 7241058

Correspondence:

Author: Burak Kolukisa
e-mail: burakkolukisa@kayseri.edu.tr
phone: +90 352 432 3838 / 21369

Graphical/Tabular Abstract

The proposed methodological framework for BRCA1/2 variant pathogenicity classification follows a stepwise pipeline, as illustrated in Figure A. The process begins with the integration of ClinVar database information and variant annotation using the GRCh38 reference genome, population frequency resources, and in silico prediction tools. Subsequently, preprocessing steps are applied to enhance model performance. Multiple classification algorithms are then employed and systematically compared. Bayesian hyperparameter optimization is conducted to maximize accuracy and macro F1-score. Finally, an explainability method is integrated to interpret feature contributions to model predictions, ensuring transparency and clinical interpretability.

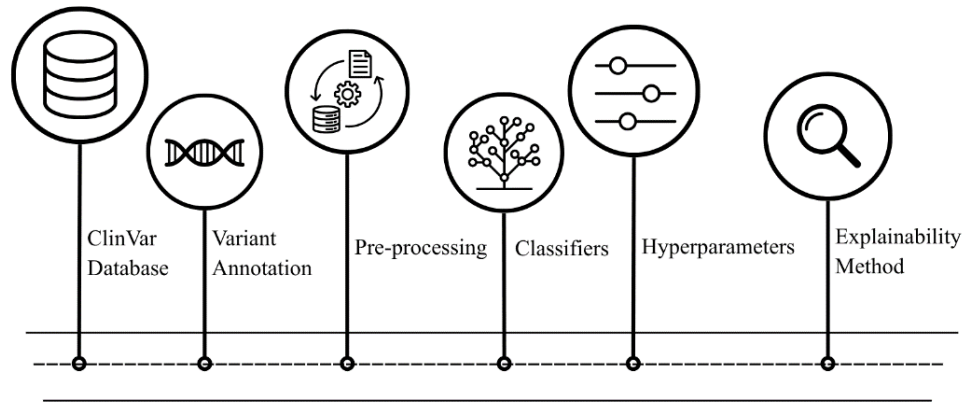


Figure A. Overview of the pathogenicity classification process, including data preprocessing, machine learning models, and benchmarking

Purpose: This study aims to develop and evaluate machine learning models for the pathogenicity classification of BRCA1/2 variants to improve prediction accuracy and provide scalable support for clinical genomic applications.

Theory and Methods: This study aims to classify the pathogenicity of BRCA1/2 variants using a gene-based machine learning framework. A dataset comprising 32,719 variants was obtained from the ClinVar database and annotated using the GRCh38 reference genome via Ensembl VEP, incorporating population frequency databases and multiple in silico predictors. After preprocessing and filtering, 22,474 variants were retained with a final feature set of 251 variables and classified into five classes. Seven supervised algorithms were implemented. Bayesian hyperparameter optimization and stratified 10-fold cross-validation were applied to maximize accuracy and macro F1-score. Model explainability was addressed using SHAP analysis to identify the most important features contributing to model predictions.

Results: XGBoost achieved the best predictive performance, with an accuracy of 96.12% and a weighted F1-score of 95.67%. Explainable AI analysis further demonstrated that specific features played a decisive role in classification outcomes, supporting the robustness and interpretability of the model.

Conclusion: These findings suggest that machine learning-based approaches offer a scalable and reliable strategy for variant interpretation and show strong potential for future integration into clinical genomics.



Açıklanabilir makine öğrenimi yaklaşımı kullanılarak meme kanseriyle ilişkili BRCA1/2 varyantlarının patojenite değerlendirilmesi

Burak Kolukisa*

Kayseri Üniversitesi, Mühendislik Mimarlık ve Tasarım Fakültesi, Yazılım Mühendisliği Bölümü, 38280, Kayseri, Türkiye

Ö N E Ç I K A N L A R

- ClinVar'dan türetilmiş BRCA1/2 varyantları veri seti
- Sınıflandırma modellerinin karşılaştırmalı analizi
- Bayes optimizasyonu yoluyla optimize edilmiş doğruluk ve makro F1 skoru

Makale Bilgileri

Araştırma Makalesi

Geliş: 10.09.2025

Kabul: 10.04.2026

DOI:

10.17341/gazimmfd.1781575

Anahtar Kelimeler:

Açıklanabilir yapay zekâ,
makine öğrenimi,
bayes optimizasyonu,
meme kanseri,
BRCA1/2 varyantları

ÖZ

Meme kanseri, dünya genelinde kadınlarda en sık görülen kanser türüdür; genetik ve çevresel faktörlerin etkileşimi sonucu ortaya çıkan heterojen bir hastalık grubunu temsil etmektedir. BRCA1/2 genlerindeki patojenik mutasyonlar, yaşam boyu meme kanseri riskini önemli ölçüde artırmakta ve bu durum, erken tanı ile önleyici stratejilerde genetik testlerin kritik rolünü vurgulamaktadır. Bu çalışmada, BRCA1/2 varyantlarının patojenitesinin makine öğrenimi algoritmaları kullanılarak değerlendirilmesi amaçlanmıştır. Model geliştirme sürecinde, BRCA1/2 genlerine ait 32.719 varyant içeren kapsamlı bir veri seti ClinVar veritabanından elde edilmiştir. Veri seti sistematik olarak hazırlanmış, ön işleme uygulanmış ve çoklu sınıflandırma (iyi huylu, muhtemel iyi huylu, belirsiz öneme sahip varyant, muhtemel patojenik, patojenik) çerçevesinde analiz edilmiştir. Yedi farklı denetimli sınıflandırma algoritması uygulanmış, en iyi performansı gösteren Aşırı Gradyan Artırma (AGA) modeli ise 10-katlı tabakalı çapraz doğrulama altında Bayes optimizasyonu ile çalıştırılmıştır. AGA modeli %96,12 doğruluk ve %95,67 ağırlıklı F1-skoru ile en yüksek başarıyı elde etmiştir. Açıklanabilir yapay zekâ yöntemiyle yapılan analizler, modelin sınıflandırma kararlarını hangi özelliklere dayandığını ve en yüksek katkırı sunan kriterlerin öne çıktığını göstermektedir. Bulgular, makine öğrenimi tabanlı yaklaşımların BRCA1/2 varyant yorumlamasında ölçeklenebilir ve klinik açıdan anlamlı araçlar olarak kullanılabileceğini göstermektedir.

Pathogenicity assessment of breast cancer associated BRCA1/2 variants using explainable machine learning approach

H I G H L I G H T S

- ClinVar-derived dataset of BRCA1/2 variants
- Comparative analysis of classification models
- Accuracy and macro F1-score optimized via Bayesian optimization

Article Info

Research Article

Received: 10.09.2025

Accepted: 10.04.2026

DOI:

10.17341/gazimmfd.1781575

Keywords:

Explainable artificial
intelligence,
machine learning,
bayesian optimization,
breast cancer,
BRCA1/2 variants

ABSTRACT

Breast cancer is the most common type of cancer in women worldwide and represents a heterogeneous group of diseases resulting from the interaction of genetic and environmental factors. Pathogenic mutations in the BRCA1/2 genes significantly increase lifetime breast cancer risk, highlighting the critical role of genetic testing in early diagnosis and preventive strategies. This study aimed to evaluate the pathogenicity of BRCA1/2 variants using machine learning algorithms. During model development, a comprehensive dataset comprising 32,719 BRCA1/2 variants was obtained from the ClinVar database. The dataset was systematically collected, preprocessed, and analyzed within a multi-class classification framework (benign, likely benign, variant of uncertain significance, likely pathogenic, pathogenic). Seven different supervised classification algorithms were applied, and the best-performing model, Extreme Gradient Boosting (XGBoost), was trained using Bayesian optimization under 10-fold stratified cross-validation. The XGBoost model achieved the highest performance, with an accuracy of 96.12% and a weighted F1-score of 95.67%. Furthermore, analyses using explainable artificial intelligence methods demonstrated which features the model relied on for classification decisions and highlighted the criteria with the greatest contribution. These findings suggest that machine learning-based approaches can serve as scalable and clinically meaningful tools for the interpretation of BRCA1/2 variants.

1. Giriş (Introduction)

Meme kanseri, süt kanallarını döşeyen epitel hücrelerinden kaynaklanan, genetik ve çevresel faktörlerin etkileşimi sonucu ortaya çıkan heterojen bir tümör grubunu temsil etmektedir [1, 2]. 2020 yılında dünya genelinde 2,3 milyondan fazla yeni vaka ve 685.000'den fazla ölüm gerçekleşmiş olup, bu sayıların 2040 yılına kadar nüfus artışı ve yaşlanmaya bağlı olarak önemli ölçüde artması beklenmektedir [3]. Meme kanseri, kadınlarda dünya genelinde en sık teşhis edilen kanser türü olup, kadınlarda kanser ölümlerinin de başlıca nedenidir [4, 5]. Erkeklerde meme kanseri son derece nadirdir ve tüm meme kanseri olgularının yalnızca yaklaşık %1'ini oluşturmaktadır. Erkeklerle özgü tarama programlarının bulunmaması nedeniyle genellikle geç evrede tanı almakta ve kadınlara kıyasla daha düşük sağkalım oranları göstermektedir [6, 7]. Meme kanseri için en önemli kalıtsal risk faktörü, BRCA1/2 genlerindeki mutasyonlarla ilişkilidir [8]. Bu genlerde patojenik mutasyon taşıyan bireyler, genel popülasyona kıyasla yaşam boyu meme kanseri riskinde anlamlı bir artış göstermekte olup, bu risk %72'ye kadar çıkabilmektedir [9]. Tümör baskılayıcı özellik taşıyan BRCA1/2 genleri, hasarlı DNA'nın onarımını sağlayarak genomik stabilitenin korunmasında rol oynamaktadır; söz konusu genlerdeki patojenik mutasyonlar DNA onarım mekanizmasını bozarak meme kanseri riskini artırmaktadır [10]. Genetik testler bu mutasyonları tanımlayarak risk değerlendirmesi, koruyucu önlemler ve erken müdahale açısından kritik bilgiler sağlamaktadır.

Genetik varyantların analizi, nadir görülen Mendeliyen bozukluklardan kalıtsal kanserlere kadar birçok hastalığın riskinin belirlenmesi açısından kritik öneme sahiptir [11]. Yeni nesil dizileme (NGS) teknolojilerindeki ilerlemeler, genetik varyasyonların keşfini ve tanımlanmasını önemli ölçüde hızlandırmış, binlerce varyantın tespit edilmesine olanak sağlamıştır [12, 13]. Bu varyantların yalnızca sınırlı bir kısmı, doğaları gereği zararlı olma potansiyeline sahiptir ve bu potansiyel varyantları tanımlamak amacıyla farklı özellik ve kapsamda çeşitli tahmin yöntemleri geliştirilmiştir [14, 15]. Bu yöntemler, bir varyantın olası zararlı etkisini tahmin etse de, ancak fonksiyonel bozulma her zaman hastalığa yol açmaz. Bu nedenle patojenite değerlendirmesinde güvenilirlikleri tek başına yeterli değildir. Nitekim, iyi huylu varyantların yaklaşık üçte birine kadar olan kısmı patojenik olarak yanlış sınıflandırılabilen [16] ve en gelişmiş in silico tahmin araçları dahi tek başına patojenite kanıtı olarak kullanılmamalıdır [17]. Bu nedenle, varyantların patojenite değerlendirmesinde daha güvenilir ve ölçeklenebilir yöntemlere ihtiyaç duyulmaktadır. ClinVar gibi kapsamlı veritabanlarında yer alan etiketli varyantlar, makine öğrenimi algoritmalarının eğitilmesi için önemli bir kaynak sunmakta olup, bu yaklaşımla geleneksel yöntemlere kıyasla daha tutarlı ve yüksek doğrulukta sonuçlar elde edilebileceği değerlendirilmektedir.

Sınıflandırma güvenilirliğini artırmak amacıyla araştırmacılar, genomik verilerdeki ilişkileri modellemek ve örüntüleri ortaya çıkarmak için giderek daha fazla makine öğrenimi yöntemlerine başvurmaktadır. Özellikle, denetimli öğrenme yöntemleri kullanarak mevcut veri kümeleri üzerinden varyant patojenitesini tahmin eden modeller geliştirilmektedir [18]. Yaklaşımlar genelde üç gruba ayrılmaktadır: genom çapında, hastalık bazlı ve gen bazlı. Genom çapında modeller geniş kohortlarla varyansı azaltabilse de, hastalık ya da gen düzeyine özgü örüntüleri yakalayamazlar. Hastalık bazlı modeller, çoğunlukla genom çapındaki modellere kıyasla daha iyi performans gösterirken, gen bazlı modeller ise genellikle hastalık bazlı yaklaşımları da geride bırakmaktadır [19]. Öte yandan, makine öğrenimi modelleri yüksek doğruluk oranlarına ulaşabilse de, karar süreçleri çoğu zaman doğrudan yorumlanabilir değildir. Bu durum, özellikle klinik genomik uygulamalarda, modelin hangi öz niteliklere dayanarak tahminde bulunduğu sorusunu kritik hale getirmektedir.

Son yıllarda, bu modellerin karar mekanizmalarını şeffaflastırarak hangi öz niteliklerin sınıflandırmada daha belirleyici olduğunu ortaya koymak önemli bir araştırma alanı haline gelmiştir [20, 21]. Böylece yalnızca yüksek doğruluk oranları elde edilmekle kalmamakta, aynı zamanda modelin hangi biyoinformatik kriterlere dayandığı ve hangi özelliklerin varyant sınıflandırmasında kritik rol oynadığı da anlaşılabilir. Mevcut literatür bilgisi doğrultusunda, bu çalışma Genome Reference Consortium Human Genome Build 38 (GRCh38) tabanlı anotasyon, evrensel ve yerel popülasyon verileri, in silico tahmin araçları ve American College of Medical Genetics and Genomics / Association for Molecular Pathology (ACMG/AMP) kriterlerini bütüncül bir yaklaşımla BRCA1/2 varyant sınıflandırmasına entegre etmektedir. Ayrıca makro F1-skoru ve doğruluk hedeflerine göre Bayesyen hiperparametre optimizasyonu uygulamakta ve SHapley Additive exPlanations (SHAP) tabanlı açıklanabilirlik sunarak ilgili alandaki ilk çalışmalar arasında yer almaktadır.

Bu çalışmada, BRCA1/2 varyantlarının sınıflandırılmasına yönelik gen temelli bir makine öğrenimi yaklaşımı uygulanmıştır. İlk aşamada varyantların anotasyon süreci; farklı bilgi kaynaklarının bir varyant yorumlama hattına entegrasyonu için standart bir protokol sunmak amacıyla yapılandırılmıştır. Varyantlar, GRCh38 referans genomu kullanılarak Ensembl Variant Effect Predictor (VEP) ile analiz edilmiş; popülasyon frekansı veritabanları ve çeşitli in silico tahmin araçları entegre edilmiştir. Elde edilen veriler birleştirilip uygun formata dönüştürülmüş, ardından sayısal ve kategorik öz nitelikler uygun ön işleme teknikleriyle işlenmiştir. Bu özellikler, yedi farklı denetimli sınıflandırma algoritmasına girdi olarak kullanılmıştır: Naive bayes (NB), lojistik regresyon (LR), karar ağaçları (KA), rassal ormanlar (RO), destek vektör makineleri (DVM), aşırı gradyan artırma (AGA) ve çok katmanlı algılayıcı (ÇKA). Böylece doğrusal ve doğrusal olmayan sınıflandırıcılar eğitilmiş ve doğrulanmıştır. En iyi sonuç veren AGA algoritmasına daha sonra Bayes hiperparametre optimizasyonu uygulanmıştır. AGA algoritması %96,12 doğruluk ve %95,67 ağırlıklı F1-skoru ile en yüksek performansı göstermiştir. Ayrıca SHAP tabanlı açıklanabilirlik analizi, modelin sınıflandırma kararlarının belirli öz nitelikler tarafından yönlendirildiğini ortaya koymuştur. Bu sayede modelin karar mekanizması daha şeffaf hale gelmiş ve sonuçların klinik bağlamda yorumlanabilirliği güçlenmiştir. Sonuç olarak, bu çalışma gen temelli öğrenmenin BRCA1/2 varyantlarının patojenite tahmininde uygulanabilir ve etkili bir yaklaşım olduğunu desteklemektedir. Mevcut literatürden farklı olarak, bu çalışmanın sağladığı temel katkılar aşağıda özetlenmiştir:

- GRCh38 referans genomu tabanlı kapsamlı bir anotasyon iş akışı geliştirilmiş; evrensel veritabanları, yerel popülasyon verileri, ACMG/AMP varyant yorumlama kriterleri ve çok sayıda in silico tahmin aracı bütünleştirilerek BRCA1/2 varyantlarının moleküler ve klinik düzeyde daha kapsamlı temsil edilmesi amaçlanmıştır.
- Yedi farklı makine öğrenimi algoritması için veri seti eğitim ve test olarak ayrılmış; modeller eğitim seti üzerinde 10-katlı tabakalı çapraz doğrulama ile eğitilmiş ve karşılaştırılmıştır. En iyi performansı gösteren algoritma Bayesyen hiperparametre optimizasyonu ile iyileştirilmiş, nihai performans test seti üzerinde raporlanmıştır.
- Geliştirilen model, yüksek tahmin başarısının yanı sıra SHAP analizi ile desteklenen açıklanabilirlik sunmuştur. Böylece model karar mekanizmasının biyolojik bulgularla tutarlı olabileceği gösterilmiş ve klinik yorumlanabilirliğe katkı sağlanmıştır.

Makalenin geri kalan bölümleri aşağıdaki şekilde organize edilmiştir: Bölüm 2, çalışmada kullanılan materyal ve yöntemleri kapsamaktadır. Bu bölümde sırasıyla varyant anotasyonu süreci, veri setinin tanımlanması, uygulanan ön işleme adımları, kullanılan sınıflandırma yöntemleri, hiperparametre optimizasyonu stratejisi, açıklanabilirlik

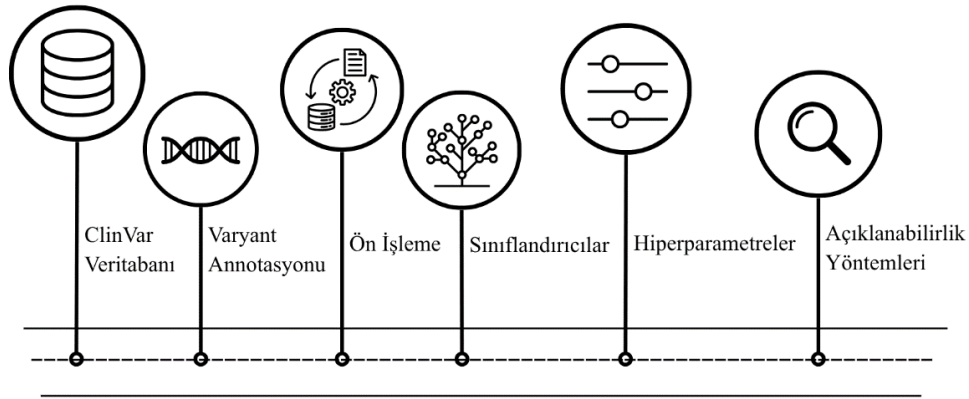
yöntemi ve deneysel düzen ayrıntılı olarak sunulmaktadır. Bölüm 3, deneysel çalışmalardan elde edilen sonuçları sunmaktadır. Son olarak, Bölüm 4, çalışmanın genel sonuçlarını ve çıkarımlarını özetlemekte ve gelecekteki araştırmalara yönelik olası yönelimlere işaret etmektedir.

2. Malzemeler ve Yöntemler (Materials and Methods)

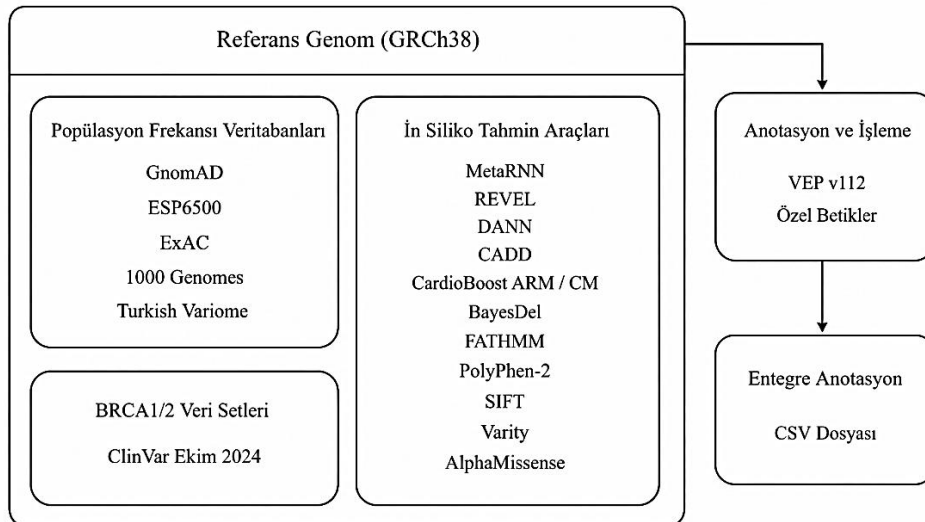
Bu bölümde, Şekil 1’de gösterilen yöntemsel çerçeve sistematik biçimde sunulmaktadır. İlk olarak, varyant anotasyon süreci, analizlerin güvenilirliğini ve tekrarlanabilirliğini sağlamak amacıyla referans genom, popülasyon frekansı veritabanları ve in silico tahmin araçlarının entegrasyonunu kapsamaktadır. Ardından, kullanılan veri setinin kapsamı ve uygulanan filtreleme adımları açıklanmıştır. Ön işleme süreci, modelin genel performansını artırmak ve aşırı uyum riskini azaltmak amacıyla ele alınmıştır. Sonraki aşamada, farklı öğrenme paradigmalarını temsil eden sınıflandırma algoritmalarının seçimi, yöntemler arası karşılaştırmaya olanak tanımaktadır. Hiperparametre optimizasyonu süreci, modelin optimum performansa ulaşması için sistematik bir iyileştirme yaklaşımını ortaya koyarken, açıklanabilirlik yöntemi, geliştirilen modellerin karar mekanizmalarının şeffaflaştırılmasına katkı sağlamaktadır. Son olarak, deneysel ayarlar sunularak tüm sürecin nasıl yürütüldüğü açıklanmakta ve çalışma tasarımının bütüncül bir şekilde anlaşılması sağlanmaktadır.

2.1. Varyant Anotasyonu (Variant Annotation)

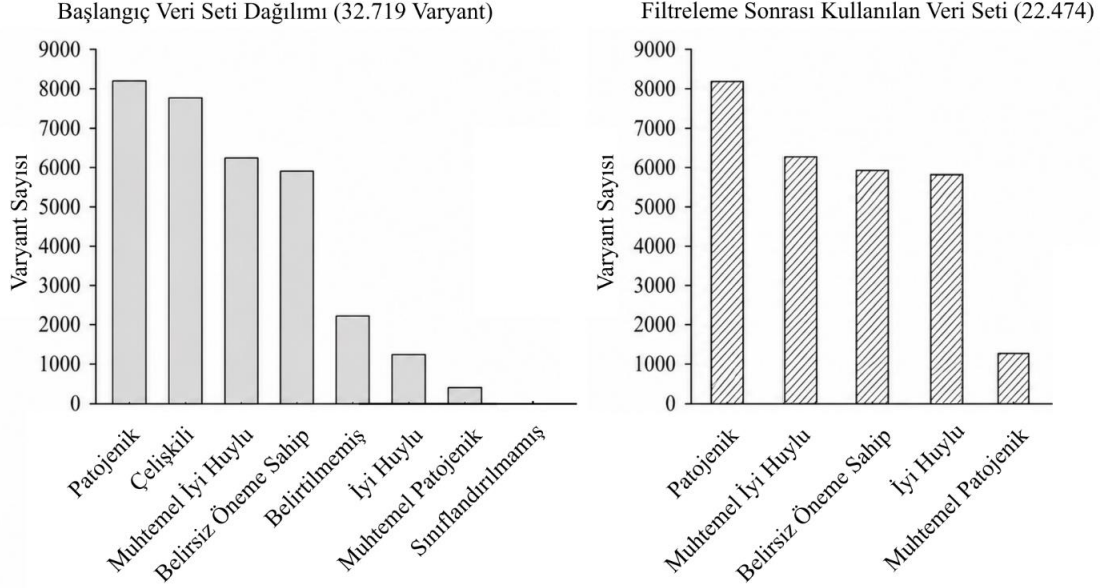
Varyant anotasyonu, GRCh38 referans genomu temel alınarak gerçekleştirilmiştir. Otomatik yorumlama sürecine Genome Aggregation Database (gnomAD v4.1.0) [22], Exome Sequencing Project v6500 (ESP6500), Exome Aggregation Consortium (ExAC), 1000 Genomes Project (1KGP) ve Turkish Variome gibi popülasyon frekansı veritabanları dâhil edilmiştir. Ayrıca MetaRNN [23], REVEL [24], DANN [25], CADD [15], CardioBoost ARM, CardioBoost CM, BayesDel, FATHMM [26], PolyPhen-2 [27], SIFT [28], Varsity [29] ve AlphaMissense [30] gibi çeşitli in silico tahmin araçları kullanılmıştır. ClinVar veritabanından (Ekim 2024) [31] elde edilen BRCA1/2 veri setinin Ensembl VEP (v112) [32] kullanılarak anotasyonu yapılmıştır. ClinVar, uzman kürasyonuna dayalı klinik varyant yorumları sunan ve klinik genomik çalışmalarda yaygın olarak kullanılan güvenilir bir referans kaynaktır [33]. ClinVar etiketlerinin zaman içinde revizyona uğrayabilme potansiyeli göz önünde bulundurularak, kullanılan sürüm açıkça tanımlanmış ve olası etiket değişimlerinin analize yansımaya riski dikkate alınmıştır. Bu doğrultuda, çok kaynaklı bir anotasyon stratejisi benimsenmiş ve modelin tek bir veritabanına özgü örüntüler öğrenmesinin önüne geçilerek klinik genellenebilirliğin artırılması hedeflenmiştir. Bu süreçte, klinik yorumu destekleyen çoklu biyolojik ve popülasyon temelli öznitelikler anotasyon sürecine birlikte dâhil edilmiştir. Bu anotasyon sürecinin şematik gösterimi Şekil 2’de sunulmaktadır.



Şekil 1. BRCA1/2 varyant analizine yönelik iş akışı (Workflow for analyzing BRCA1/2 variants)



Şekil 2. BRCA1/2 varyant anotasyonuna yönelik iş akışı (Workflow for BRCA1/2 variant annotation)



Şekil 3. BRCA1/2 varyant sınıflarının filtreleme öncesi ve sonrası dağılımı (Distribution of BRCA1/2 variant classes before and after filtering)

Tablo 1. Makine öğrenimi modellerinde kullanılan öznitelikler: temel varyant bilgileri, gen/transkript anotasyonları, in silico tahmin skorları, popülasyon frekansları, ek kaynaklar ve ACMG kriterleri (Features used in the machine learning models: basic variant information, gene/transcript annotations, in silico prediction scores, population frequencies, supplementary resources, and ACMG criteria)

No	Öznitelik	No	Öznitelik	No	Öznitelik
-	<i>Varyant Temel Bilgileri</i>	28	MetaRNN	-	<i>Popülasyon Frekansları</i>
1	Kromozom	29	DANN	54-56	Genom- AF, HoS, KK
2	Pozisyon	30	Fathmm-XF	57-59	Ekzom- AF, HoS, KK
3	Referans Alel	31	BayesDel	60-62	Turkish Variome- AF, HoS, HeS
4	Alternatif Alel	32	BayesDel AF	63-67	GenomAF- NFE, FIN, ASJ, AFR, EAS
5	Varyant Sınıf	33	Varity	68-72	GenomAF- SAS, AMR, Amish, MID, OTH
6	Etki Türü	34	PhyloP100Way	73-77	EkzomAF- NFE, FIN, ASJ, AFR, EAS
7	Etki Düzeyi	35	ADA	78-82	EkzomAF- SAS, AMR, Amish, MID, OTH
-	<i>Gen/Transkript Anotasyonları</i>	36	RF	83-87	GenomHo- NFE, FIN, ASJ, AFR, EAS
8	Gen	37	MaxEntScan	88-92	GenomHo- SAS, AMR, Amish, MID, OTH
9	Gen Sembolü	-	<i>ACMG Kriterleri</i>	93-97	EkzomHo- NFE, FIN, ASJ, AFR, EAS
10	Gen Transkript	38	PVS1	98-102	EkzomHo- SAS, AMR, Amish, MID, OTH
11	Kanonik Transkript	39	PS1	103	1K Genom Projesi AF
12	Ekzon	40	PM1	104-108	1K Genom Projesi AF -AMR, AFR, EUR, EAS, SAS
13	İntron	41	PM2	109-111	ExAC- AF, HoS, HeS
14	Komplementer DNA Pozisyonu	42	PM4	112-118	ExACHo- NFE, FIN, AFR, EAS, SAS, AMR, OTH
15	Kodlayan DNA Dizisi	43	PM5	119-125	ExACHE- NFE, FIN, AFR, EAS, SAS, AMR, OTH
16	Protein Pozisyonu	44	PP3	-	<i>Diğer Kaynaklar</i>
17	Amino Asitler	45	PP5	126	HUGO Gen Adlandırma Komitesi
18	Kodonlar	46	BA1	127	HGVS Pozisyon Kaydırması
19	Baz Mesafesi	47	BS2	128	Fenotip
20	Zincir Yönü	48	BP1	129	Somatik
21	Biyotip	49	BP3	130	Genom Ortalama Derinliği
22	Gen Sembol Kaynağı	50	BP4	131	Ekzom Ortalama Derinliği
23	Gen Anotasyon Kaynağı	51	BP6		
-	<i>in silico Tahmin Skorlar</i>	52	BP7		
24	CADD	53	ACMG Skoru		
25	SIFT				
26	PolyPhen				
27	MutationAssessor				

2.2. Veri Seti (Dataset)

Makine öğrenimi modellerinin eğitimi için BRCA1/2 veri seti oluşturulmuştur. Bu veri seti, ClinVar etiketlerine göre sekiz sınıfa ayrılmış toplam 32.719 varyant içermektedir. Anotasyon sonrası her bir varyant için 289 öznitelik elde edilmiştir. Sınıf dağılımı sırasıyla 8.259 patojenik (Pathogenic, P), 7.838 çelişkili sınıflandırma

(Conflicting Classifications of Pathogenicity), 6.188 muhtemelen iyi huylu (Likely Benign, LB), 5.971 belirsiz öneme sahip varyant (Variant of Uncertain Significance, VUS), 2.253 belirtilmemiş (Not Provided), 1.158 iyi huylu (Benign, B), 498 muhtemelen patojenik (Likely Pathogenic, LP) ve 10 adet sınıflandırılmamış varyant şeklindedir. Ancak çelişkili sınıflandırma, belirtilmemiş ve sınıflandırılmamış gruplarında yer alan varyantlar güvenilir etiket

sağlamadığından denetimli öğrenme sürecine dâhil edilmemiştir. Bu filtreleme sonrasında toplam 22.474 varyant, makine öğrenimi modellerinin eğitimi için kullanıma uygun hale getirilmiştir. Başlangıç veri seti ile filtreleme sonrası kullanılan veri setinin sınıf dağılımları Şekil 3'te karşılaştırmalı olarak sunulmaktadır. Ayrıca kullanılan öznelikler altı ana kategori altında sınıflandırılmıştır: (i) varyantın temel bilgileri; (ii) gen ve transkript anotasyonları; (iii) in silico öngörücü skorlar; (iv) popülasyon frekansları; (v) diğer kaynaklardan elde edilen ek klinik ve teknik bilgiler; ve (vi) ACMG kriterleri. Bu kategoriler, varyantın genomik konumu, biyolojik bağlamı, hesaplamalı tahminleri ve uluslararası varyant yorumlama standartlarına dayalı kanıt çerçevesini birlikte değerlendirmeyi amaçlamaktadır. Ön değerlendirme aşamasında sabit değerlere sahip, yüksek oranda eksik veri içeren veya yinelenen değişkenler veri setinden çıkarılmıştır. Bu işlemler sonucunda toplam 158 öznelik elenerek öznelik sayısı 289'dan 131'e indirgenmiştir. Nihai olarak seçilen öznelikler Tablo 1'de sunulmuştur.

2.3. Ön İşleme (Pre-processing)

Varyant anotasyonlarından elde edilen öznelikler, makine öğrenimi modellerine girdi olmadan önce türlerine ve içerik yapılarına uygun şekilde ön işleme tabi tutulmuştur. Sayısal değişkenler; genomik konumlar, cDNA/CDS/protein pozisyonları, baz mesafeleri, popülasyon frekansları ve in silico tahmin skorlarını kapsayacak biçimde ele alınmıştır. Veri setinde sayısal bir temsil olarak yer alan ACMG tabanlı değerler, ACMG/AMP yönergelerine dayalı eşiklere göre beş sınıflı klinik patojenite şemasına dönüştürülmüştür. Bileşik yapıya sahip somatik ve fenotip öznelikleri doğrudan modele dâhil edilmemiştir. Bunun yerine bilginin varlığını, tutarlılığını ve yoğunluğunu temsil eden üç türetilmiş değişkene ayrıştırılmıştır. Ekzon ve intron bilgileri gen-bazlı olarak yeniden yapılandırılmış ve yalnızca ilk konumsal bileşen korunmuştur. Benzer şekilde kodlayan DNA dizisi, tamamlayıcı DNA ve protein pozisyonları normalize edilerek sürekli sayısal öznelikler olarak yeniden yapılandırılmıştır. Kategorik değişkenler kardinalite düzeylerine göre seçici biçimde kodlanmıştır. Sınırlı kategori sayısına sahip biyolojik tanımlayıcılar one-hot encoding ile dönüştürülürken, yüksek kardinaliteli dizesel öznelikler frekans kodlama yaklaşımıyla sayısallaştırılmıştır. Son aşamada eksik değer gösterimleri standart biçime dönüştürülmüş; ölçek duyarlılığı yüksek sayısal öznelikler RobustScaler kullanılarak ölçeklenmiştir. Ön işleme adımları sonucunda öznelik sayısı, kategorik genişletmeler ve türetilmiş biyolojik değişkenlerin eklenmesiyle 131'den 251'e yükselmiştir.

2.4. Sınıflandırma Metodları (Classification Methods)

Sınıflandırma performansını değerlendirmek amacıyla, farklı öğrenme paradigmasını temsil eden çeşitli denetimli sınıflandırma algoritmaları kullanılmıştır: NB (olasılıksal modelleme), LR (doğrusal sınıflandırıcı), DVM (çekirdek uzantılarıyla maksimum marjın sınıflandırıcısı), KA (kural tabanlı ağaç modeli), RO (bagging temelli topluluk modeli), AGA (boosting temelli topluluk modeli) ve ÇKA (doğrusal olmayan ilişkileri modelleyen yapay sinir ağı mimarisi). Veri setindeki sınıf dengesizliğini yönetmek ve azınlık sınıfların öğrenilmesini teşvik etmek amacıyla, uygulanabilir tüm modellerde maliyet duyarlı öğrenme yaklaşımı benimsenmiş ve sınıf ağırlıkları sınıf frekanslarıyla ters orantılı olacak şekilde ayarlanmıştır. NB için standart Gauss yaklaşımı yerine Tamamlayıcı NB tercih edilmiştir. Lojistik Regresyon, L2 düzenleme ve liblinear çözücüsü ile yapılandırılmıştır. DVM'de karesel menteşe kaybını optimize eden doğrusal destek vektör sınıflandırıcısı kullanılmıştır. Ağaç tabanlı yöntemlerde KA, RO ve AGA sınıf ağırlıkları ile eğitilmiştir. Son olarak, doğrusal olmayan ilişkiler için ÇKA mimarisine sahip yapay sinir ağı modeli kullanılmıştır. Tüm analizler Scikit-learn kütüphanesi [34] ile gerçekleştirilmiştir. Bu

kapsamlı model seçimi, ek mimari veya veri düzeyinde müdahale yapılmaksızın olasılıksal, doğrusal, topluluk ve sinir ağı tabanlı yaklaşımların karşılaştırılmasına olanak sağlamıştır.

2.5. Hiperparametre Ayarlaması (Hyperparameter Tuning)

Hiperparametre ayarlaması, parametre uzayını keşfederken keşif ve sömürü dengesini kazanım fonksiyonu aracılığıyla sağlayan Gauss süreci tabanlı vekil modellerinden yararlanan Bayes optimizasyonu çerçevesi [35] kullanılarak gerçekleştirilmiştir. Bu yaklaşım, kapsamlı arama stratejilerine kıyasla önemli ölçüde hesaplama avantajı sağlamaktadır. Bu çalışmada söz konusu işlem 100 iterasyon boyunca yürütülmüştür. Tablo 2'de görüldüğü üzere AGA modeli için optimizasyon; öğrenme oranı (0.0001–0.1), maksimum ağaç derinliği (3–10), minimum çocuk ağırlığı (1–10), örnekleme oranı (0.5–1.0), her ağaç için sütun örnekleme oranı (0.5–1.0) ile hem L1 (0.1–10.0) hem de L2 (0.1–10.0) düzenleme terimlerini kapsamıştır.

Tablo 2. Bayesyen optimizasyon kapsamında AGA modeli için hiperparametre aralıkları
(Hyperparameter search space for the AGA model using Bayesian optimization)

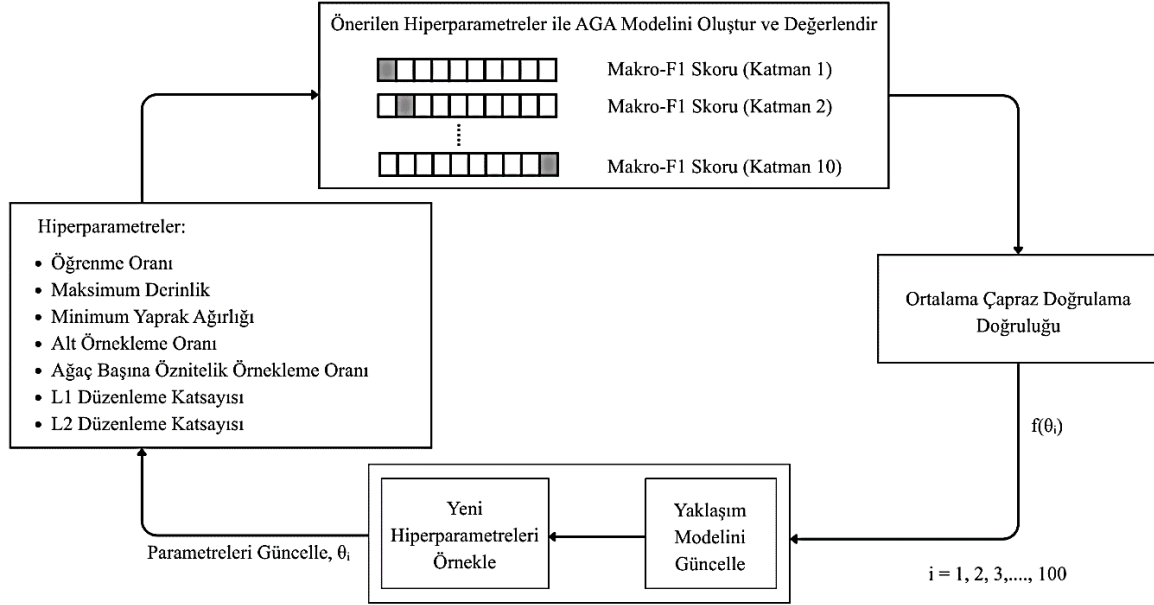
Parametre	Minimum	Maksimum
Öğrenme Oranı	1e-4	0,1
Maksimum Derinlik	3	10
Minimum Yaprak Ağırlığı	1	10
Alt Örnekleme Oranı	0,5	1
Ağaç Başına Öznelik Örnekleme Oranı	0,5	1
L1 Düzenleme Katsayısı	0,1	10
L2 Düzenleme Katsayısı	0,1	10

2.6. Açıklanabilirlik Yöntemi (Explainability Method)

Makine öğrenimi modelleri çoğu durumda yüksek doğruluk oranlarına ulaşabilse de, karar mekanizmaları genellikle doğrudan yorumlanabilir değildir. Bu durum, özellikle klinik genomik uygulamalarda modelin hangi özneliklere dayanarak tahminde bulunduğu sorusunu kritik hale getirmektedir. Bu nedenle, çalışmada açıklanabilir yapay zekâ yöntemlerinden SHAP kullanılmıştır [36]. SHAP, kooperatif oyun teorisine dayalı Shapley değerlerini kullanarak her bir özneliğin modele katkısını niceliksel olarak hesaplar. Bir öznelik katkısı, söz konusu öznelik eklendiğinde model tahmininde meydana gelen değişikliklerin tüm olası öznelik alt kümeleri üzerindeki ortalama marjinal etkisiyle tanımlanır. Bu yöntem, hem küresel düzeyde (model genelinde en etkili özneliklerin belirlenmesi) hem de yerel düzeyde (tek bir örneğe ilişkin katkılarının gösterilmesi) açıklama imkânı sağlamaktadır. Çalışmada SHAP analizi, modellerin karar mekanizmalarının daha iyi anlaşılması ve genomik varyantların sınıflandırılmasında hangi özneliklerin belirleyici olduğunun ortaya konulması amacıyla uygulanmıştır.

2.7. Deneysel Ayarlar (Experimental Settings)

Bu çalışmada, beş farklı klinik sınıfa ait varyantların ayrımını gerçekleştirmek üzere makine öğrenimi tabanlı sınıflandırma modelleri geliştirilmiştir. ClinVar veritabanından elde edilen BRCA1/2 varyantlarından oluşan veri seti kullanılarak deneysel süreç Python ortamında ve Scikit-learn kütüphanesi ile yürütülmüştür. İlk aşamada, yedi farklı makine öğrenimi algoritması (NB, LR, KA, DVM, RO, AGA ve ÇKA) uygulanmıştır. Veri seti %80 eğitim ve %20 test kümesi olacak şekilde ayrılmıştır. Eğitim sürecinde sonuçların güvenilirliğini artırmak ve aşırı uyumu azaltmak amacıyla, eğitim kümesi üzerinde 10-katlı tabakalı çapraz doğrulama stratejisi uygulanmış ve modellerin performansı bu katmanlardan elde edilen ortalama skorlar üzerinden değerlendirilmiştir. İkinci aşamada, eğitim



Şekil 4. Bayesyen optimizasyon ile AGA hiperparametre ayarlaması (Hyperparameter tuning for XGBoost using Bayesian optimization)

kümesi üzerinde en yüksek performansı gösteren AGA algoritması hiperparametre optimizasyonu için seçilmiştir. Şekil 4'te özetlendiği üzere model, Bayesyen hiperparametre optimizasyonu kapsamında tabakalı 10-katlı çapraz doğrulama altında 100 iterasyon boyunca değerlendirilmiş; böylece sağlamlık artırılmış ve rastgele bölünmelerden kaynaklanan varyans azaltılmıştır. Optimizasyon süreci iki senaryo kapsamında yürütülmüştür: (i) doğruluk ve (ii) makro F1-skorunun maksimize edilmesi. Nihai model performansı doğruluk, makro F1-skoru, ağırlıklı F1-skoru, kesinlik ve duyarlılık metrikleri kullanılarak raporlanmıştır.

3. Sonuçlar ve Tartışmalar (Results and Discussions)

Bu çalışmanın temel amacı, BRCA1/2 genlerine özgü karmaşık örüntülerin makine öğrenimi ile ortaya çıkarılması ve SHAP analizi ile bu öz niteliklerin biyolojik gerçeklerle uyumunun gösterilmesi ve model karar mekanizmasının şeffaflaştırılmasıdır. Geliştirilen sistem, ClinVar gibi veritabanlarında etiketi bulunmayan veya Belirsiz Önele Sahip Varyant olarak sınıflandırılan varyantların yorumlanmasında, biyolojik örüntülere dayalı güvenilir bir karar destek mekanizması sunmayı amaçlamaktadır. Bu doğrultuda, BRCA1/2 varyantlarının beşli klinik sınıflandırmasını gerçekleştirmek üzere olasılıksal, doğrusal, topluluk öğrenmesi ve derin öğrenme yaklaşımlarını temsil eden yedi farklı algoritma (NB, LR, DVM, KA, RO, AGA ve ÇKA) kapsamlı bir performans analizine tabi tutulmuştur. Analiz sürecinde modellerin başarıları doğruluk, makro F1-skoru, ağırlıklı F1-skoru, kesinlik ve duyarlılık metrikleri kullanılarak değerlendirilmiştir. Özellikle sınıf dengesizliğine karşı direnci temsil eden makro F1-skoru, kritik bir performans ölçütü olarak dikkate alınmıştır. Çalışmanın ayırt edici bir yönü olarak, in silico tahmin araçları ve ACMG/AMP varyant yorumlama kriterleri öz nitelik setine dâhil edilerek klinik standartlarla uyumlu bir entegrasyon sağlanmıştır. Ayrıca, literatürde kullanılan 13 farklı in silico tahmin aracının performansları ile önerilen AGA yönteminin sonuçları karşılaştırılmıştır. Bu karşılaştırma, in silico araçların sonuç üretmediği durumlarda geliştirilen nihai modelin tahmin üretebildiğini ve klinik açıdan belirsiz varyantların sınıflandırılmasına katkı sunabileceğini göstermiştir.

Tablo 3 ve Tablo 4 birlikte değerlendirildiğinde, hem eğitim hem de test süreçlerinde en tutarlı ve en yüksek performansı AGA algoritmasının sergilediği görülmektedir. Eğitim setinde %95,97

doğruluk ve %82,11 makro F1-skoruna ulaşan AGA, test setinde de bu başarısını koruyarak %95,82 doğruluk ve %83,16 makro F1-skoru elde etmiştir. Bu bulgular, modelin yalnızca eğitim verisine uyum sağlamadığını, aynı zamanda güçlü bir genelleme kapasitesine sahip olduğunu göstermektedir. Diğer algoritmalar incelendiğinde, RO ve KA yöntemlerinin eğitim setinde AGA'ya yakın doğruluk oranları (%95,42 ve %93,55) sergilediği; ancak test setinde makro F1-skorlarının AGA'nın belirgin biçimde gerisinde kaldığı görülmektedir. Doğrusal modellerden LR ve DVM ise her iki veri setinde de %87–88 doğruluk bandında kalarak görece sınırlı bir performans ortaya koymuştur. Derin öğrenme yaklaşımını temsil eden ÇKA modeli eğitim setinde nispeten düşük bir başlangıç yapmış olsa da (%82,06 doğruluk), test setinde %90,17 doğruluğa ulaşarak orta düzeyde bir başarı sergilemiştir. Bununla birlikte, ÇKA'nın makro F1-skorunun (%75,60) AGA'nın (%83,16) altında kalması; veri setinin ölçeği ve sınıf dengesizliği dikkate alındığında, artırma (boosting) tabanlı topluluk öğrenme algoritmalarının tabular verilerde daha dengeli performans gösterebildiğine işaret etmektedir. Bu bulgular doğrultusunda, genelleme kapasitesi ve SHAP analizi ile sağlanan yorumlanabilirlik göz önünde bulundurularak AGA algoritması Bayesyen hiperparametre optimizasyonu için temel model olarak seçilmiştir.

Tablo 5'te sunulan sonuçlar incelendiğinde, Bayesyen hiperparametre optimizasyonu uygulanan AGA modelinin hem doğruluk hem de Makro-F1 hedeflerinde tutarlı bir performans sergilediği görülmektedir. Makro-F1 odaklı optimizasyon, eğitim setinde %92,90 ve test setinde %83,38 değerlerine ulaşarak sınıf dengesizliğine sahip veri kümesinde azınlık sınıflarının daha dengeli biçimde ayırt edilmesine katkı sağlamıştır. Bu bulgular, modelin dengesiz veri kümeleri üzerindeki etkinliğini ve klinik açıdan önemli azınlık sınıflarını ayırt etme kapasitesini desteklemektedir. Tablo 6'da gösterildiği üzere, Makro-F1 metriğini maksimize eden optimum AGA hiperparametreleri; maksimum ağaç derinliği 9, minimum yaprak ağırlığı 1, öğrenme oranı 0.0773, alt örnekleme oranı 0.8222, öz nitelik örnekleme oranı 0.6730 ve düzenleme katsayıları (L1: 0.8549, L2: 1.6035) olarak belirlenmiştir. Bu hiperparametre konfigürasyonu, modelin karmaşık genetik örüntülerini öğrenmesini desteklerken aşırı öğrenmeye karşı daha dengeli bir yapı sunmuştur.

Şekil 5'deki Sankey diyagramı, AGA modelinin tahminleri ile ClinVar etiketleri arasındaki sınıf geçişlerini ve uyum örüntülerini görselleştirmektedir.

Tablo 3. Yedi sınıflandırma algoritmasının varsayılan hiperparametreler kullanılarak eğitim veri kümesi üzerinde elde edilen 10 katlı tabakalı çapraz doğrulama performansları
(Performance of seven classification algorithms evaluated on the training dataset using 10-fold stratified cross-validation with default hyperparameters)

Sınıflandırıcı	Doğruluk (%)	Makro-F1 (%)	Ağırlıklı-F1 (%)	Kesinlik (%)	Duyarlılık (%)
NB	88.09 ± 0.74	68.06 ± 0.76	87.07 ± 0.74	67.40 ± 0.79	68.79 ± 0.93
LR	88.76 ± 2.05	75.91 ± 2.63	88.70 ± 1.97	75.97 ± 2.81	76.04 ± 2.54
DVM	88.76 ± 1.27	75.47 ± 2.32	88.64 ± 1.36	75.58 ± 2.18	75.61 ± 2.43
KA	93.55 ± 0.52	79.34 ± 1.41	93.58 ± 0.52	79.29 ± 1.31	79.49 ± 1.62
RO	95.42 ± 0.45	79.52 ± 1.33	94.66 ± 0.49	87.22 ± 2.61	78.18 ± 0.95
AGA	95.97 ± 0.24	82.11 ± 1.42	95.46 ± 0.27	86.42 ± 2.07	80.47 ± 1.13
ÇKA	82.06 ± 6.18	64.79 ± 7.57	81.24 ± 7.97	71.73 ± 7.24	65.15 ± 7.47

Tablo 4. Yedi sınıflandırma algoritmasının varsayılan hiperparametreler kullanılarak test veri kümesi üzerinde elde edilen performansları (Performance of seven classification algorithms on the test dataset using default hyperparameters)

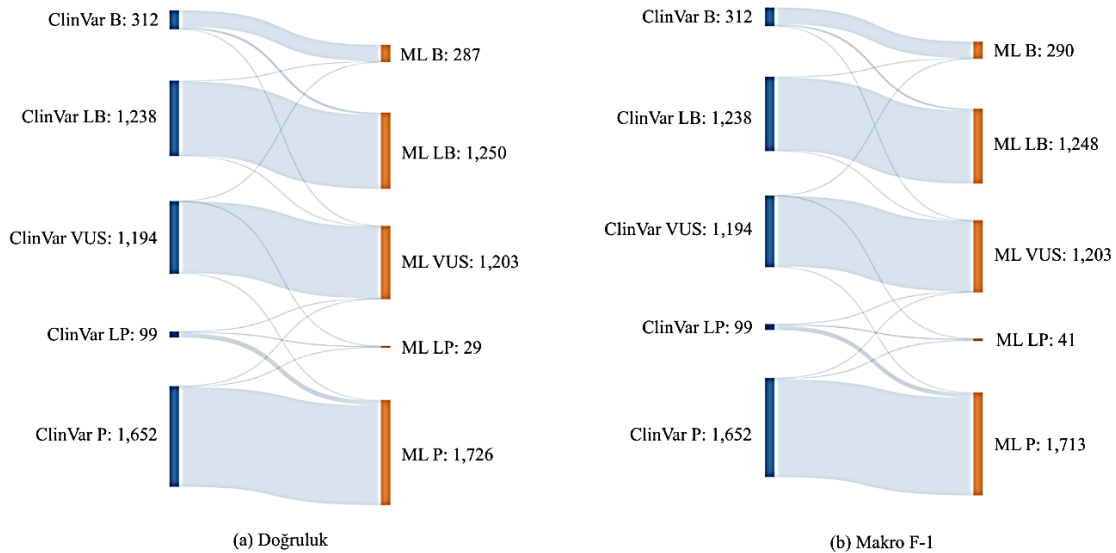
Sınıflandırıcı	Doğruluk (%)	Makro-F1 (%)	Ağırlıklı-F1 (%)	Kesinlik (%)	Duyarlılık (%)
NB	87.63	67.64	86.62	67.00	68.29
LR	88.25	75.60	88.05	76.17	85.23
DVM	87.43	74.17	87.24	74.32	74.24
KA	93.37	79.97	93.39	80.13	79.83
RO	95.28	80.24	94.62	86.98	78.60
AGA	95.82	83.16	95.44	86.73	81.33
ÇKA	90.17	75.60	90.12	76.23	75.30

Tablo 5. Bayesyen optimizasyonu kullanılarak elde edilmiş AGA sınıflandırıcısının eğitim ve test performans sonuçları
(Training and test performance results of the XGBoost classifier obtained using Bayesian optimization)

Hedef Metrik	Doğruluk (%)	Makro-F1(%)	Ağırlıklı-F1(%)	Kesinlik (%)	Duyarlılık (%)
Eğitim					
Doğruluk	97.17	86.28	96.72	97	97
Makro-F1	98.39	92.90	98.24	98	98
Test					
Doğruluk	96.08	82.42	95.51	96	96
Makro-F1	96.12	83.38	95.67	96	96

Tablo 6. Bayesyen optimizasyon kapsamında AGA modelinin optimum hiperparametreleri
(Optimal XGBoost hyperparameters obtained via Bayesian optimization)

Parametre	Doğruluk	Makro-F1
Öğrenme Oranı	0.0948	0.0773
Maksimum Derinlik	5	9
Minimum Yaprak Ağırlığı	3	1
Alt Örnekleme Oranı	0.7065	0.8222
Ağaç Başına Öznitelik Örnekleme Oranı	0.6644	0.6730
L1 Düzenleme Katsayısı	0.1216	0.8549
L2 Düzenleme Katsayısı	4.9311	1.6035

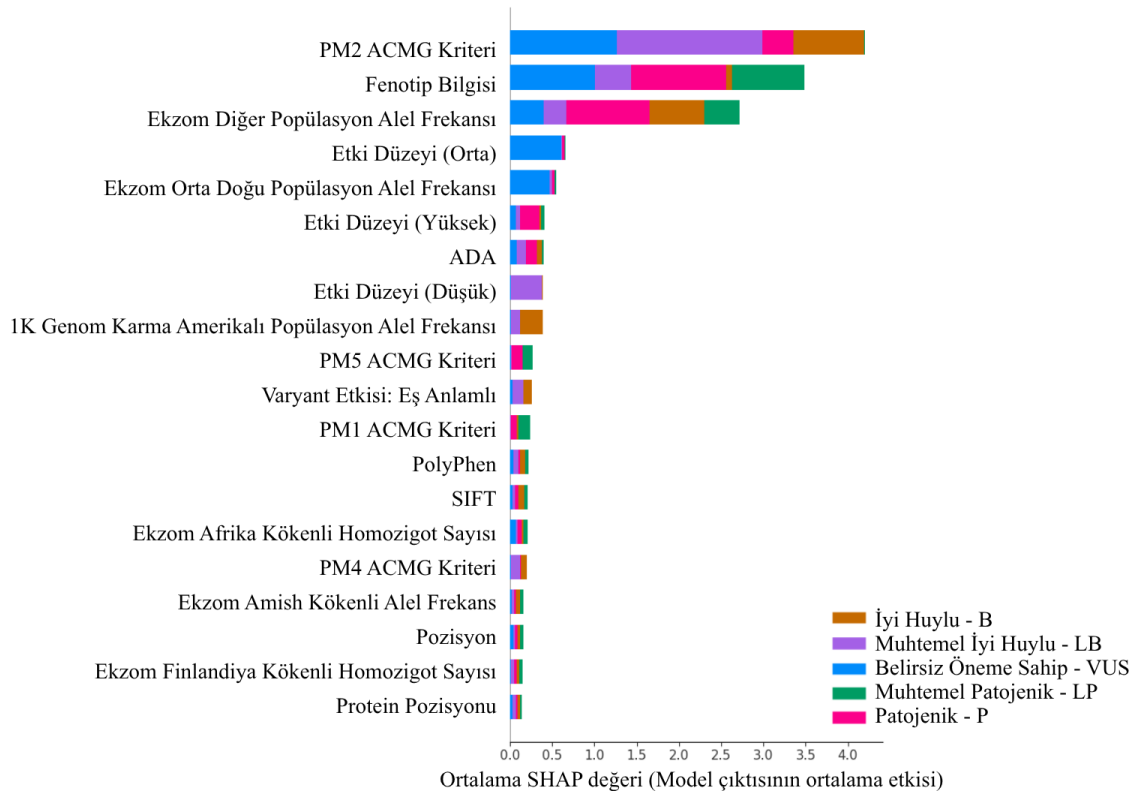


Şekil 5. (a) Doğruluk ve (b) makro F1-skoru hedeflerine göre optimize edilen AGA sınıflandırıcısının test veri kümesi üzerindeki sonuçları (Results on the test dataset for the AGA classifier optimized for accuracy and macro F1-score)

ClinVar tarafından İyi Huylu (B) olarak etiketlenen 312 varyantın 272'si model tarafından aynı sınıfta tahmin edilmiştir. Kalan varyantların 36'sı Muhtemel İyi Huylu (LB) ve 4'ü Belirsiz Öneme Sahip (VUS) olarak sınıflandırılmıştır. Modelin, ClinVar'da İyi Huylu olarak tanımlanan bazı varyantları VUS kategorisine kaydırması, model tahminleri ile klinik etiketler arasındaki sınır bölgelerini yansıtabilecek örnekler olarak değerlendirilebilir. Klinik pratikte VUS olarak raporlanan varyantlar belirsizlik içeren bir kategoriye temsil ettiğinden, bu tür geçişler yalnızca araştırma amaçlı yorumlanmalıdır. Benzer biçimde, ClinVar'da VUS olarak sınıflandırılan 1194 varyantın 1179'u model tarafından aynı kategoride tutulurken, 13 varyantın Patojenik (P) sınıfına yöneldiği gözlemlenmiştir. Bu bulgular, model tahminlerinin belirsiz varyantların önceliklendirilmesine yönelik araştırma bağlamında ek bir analiz katmanını sağlayabileceğini düşündürmektedir. Bununla birlikte, model tarafından önerilen sınıf değişimleri klinik yeniden sınıflandırma anlamına gelmemektedir. Bu çıktılar yalnızca hipotez üretimi ve ileri değerlendirme gerektirebilecek varyantların belirlenmesine yönelik araştırma bulgularıdır; dolayısıyla nihai yorumun uzman genetik danışmanlar veya hekimler tarafından yapılması gerekmektedir.

Şekil 6 ve Şekil 7, geliştirilen modelin varyant sınıflandırma sürecinde hangi özneliklerin model çıktısı üzerinde en yüksek etkiye sahip olduğunu nicel olarak göstermektedir. SHAP analizinde etki büyüklüğüne göre sıralanan ilk 20 öznelik, popülasyon frekansı, protein yapısal/fonksiyonel etkisine ilişkin öngörüler, topolojik konum ve etnik özgüllük gibi temel biyolojik parametrelerin, klinik sınıfların ayırımında nasıl bir ağırlıkla kullanıldığını açık biçimde göstermektedir. SHAP değerlerinin sınıflara göre dağılımı, AGA modelinin bir "kara kutu" olmanın ötesinde, her bir özneliği belirli klinik tahminleri destekleyen bir karar hiyerarşisi içinde kullandığını göstermektedir.

SHAP analizi, model kararlarında PM2 kriteri, popülasyon alel frekansı, fenotip varlığı ve etki düzeyi (orta/yüksek) özneliklerinin öne çıktığını göstermektedir. PM2 ve popülasyon alel frekansı özneliklerinin öne çıkması, modelin patojenite değerlendirmesinde nadirlik sinyalinin anlamlı biçimde kullandığını düşündürmektedir. Fenotip varlığı özneliğinin katkısı, sınıflandırma sürecinin fenotip anotasyonlarından gelen klinik bağlamla desteklendiğine işaret ederken; etki düzeyinin (orta/yüksek) rolü, varyantın işlevsel sonuçlarına ilişkin özet bilginin karar mekanizmasına sistematik biçimde yansıdığını göstermektedir. Sınıflandırma performansında öne çıkan üç in silico öznelik, varyant etkisini biyolojik sistemin farklı katmanları üzerinden değerlendiren tamamlayıcı sinyaller sunmaktadır. PolyPhen, amino asit değişimlerinin protein yapısı ve işlevi üzerindeki olası etkilerini değerlendirerek sınıflandırma sürecine katkı sağlamaktadır. SIFT ise filogenetik korunmuşluk ilkesine dayanarak evrimsel olarak tolere edilmeyen değişimleri işaret etmekte ve özellikle iyi huylu sınıfların ayırıştırılmasında ayırt edici bir filtre niteliği taşımaktadır. ADA, RNA olgunlaşma süreçlerine (özellikle splicing) ilişkin olası bozulmaları öngörerek patojenite ile ilişkili olabilecek varyantların belirlenmesine katkıda bulunmaktadır. Analizde ACMG/AMP kriterleriyle ilişkili özneliklerin öne çıkması, modelin karar mekanizmasının uluslararası varyant yorumlama standartlarıyla uyumlu kanıt sinyallerine duyarlı olduğunu desteklemektedir. Özellikle PM1 kriteri, mutasyonların kümелendiği fonksiyonel domenleri tanımlayarak bu bölgelerdeki değişimlerin patojenite ile ilişkili olabileceğine işaret etmektedir. PM5 özneliği ise belirli bir amino asit konumunda farklı bir değişimin klinik önem taşıdığı durumlarda, aynı pozisyona ilişkin ek bir kanıt sinyali sunabilmektedir. PM4 kriteri de protein düzeyindeki değişim tiplerinin (ör. uzunluk değişimleri) klinik yorumlamadaki önemini yansıtan tamamlayıcı bir kanıt sinyali olarak değerlendirilebilir. Sonuç olarak, elde edilen bulgular AGA modelinin frekans temelli sinyalleri ve klinik kanıt

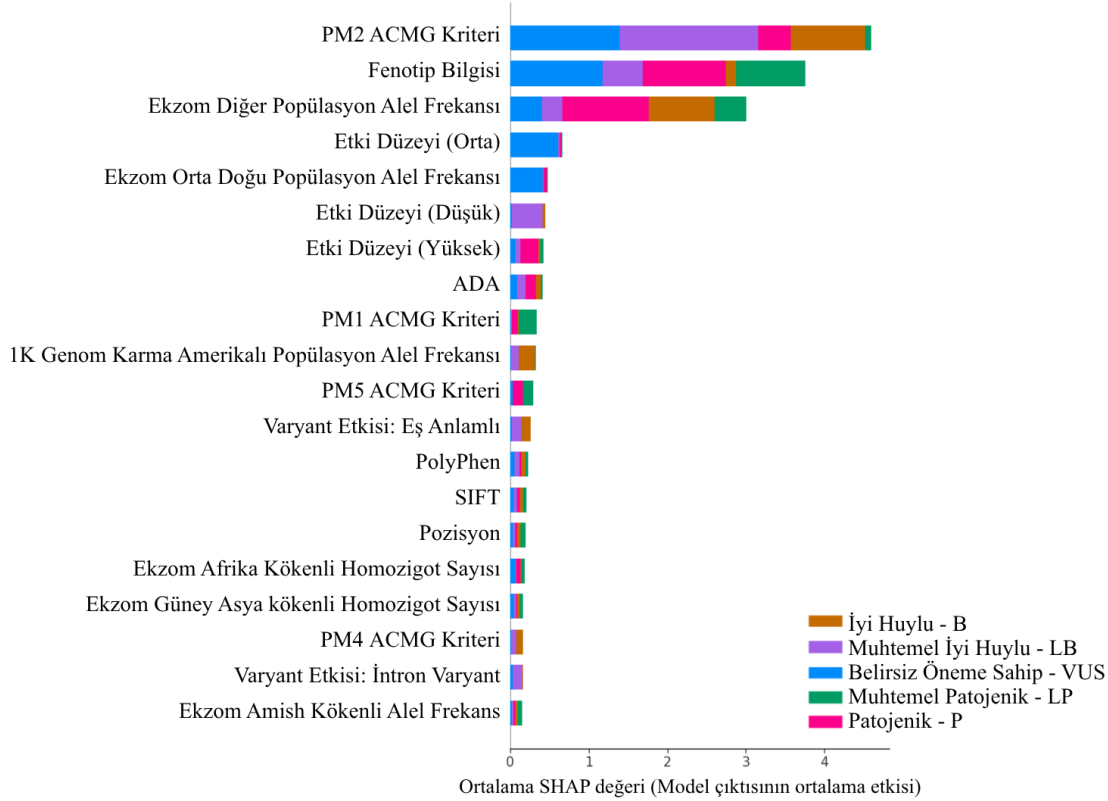


Şekil 6. Makro F1-skoru hedefiyle optimize edilen AGA sınıflandırıcısında özneliklerin sınıflandırma kararına etkisi
(Feature effects on the classification decision for the XGBoost classifier optimized for macro F1-score)

çerçevelerini birlikte değerlendirerek daha açıklanabilir bir karar mekanizması oluşturduğunu göstermektedir.

Tablo 7, BRCA1/2 test veri setinde AGA modeli ile literatürde yaygın kullanılan in silico tahmin araçlarının patojenik varyantları yakalama performansını karşılaştırmaktadır. İn silico tahmin araçları, genetik varyantların biyolojik sistemler üzerindeki potansiyel etkilerini bilgisayar tabanlı algoritmalar aracılığıyla modellemeyi amaçlamaktadır. Bu araçlar genellikle çok sınıflı klinik sınıflandırma üretmeyip eşik değerlerine dayalı ikili (patojenik/patojenik olmayan) değerlendirme yaptıkları için karşılaştırma patojenik yakalama performansı ile sınırlandırılmıştır. PolyPhen ve MutationAssessor

protein yapısal etkilerine, SIFT ve PhyloP100Way evrimsel korunmuşluğa, CADD, DANN ve BayesDel ise çoklu sinyalleri bütünleştiren meta-skora dayanmaktadır. Sonuçlar, bazı araçların sınırlı veri kapsamına sahip olduğunu (ör. SIFT ve PolyPhen %2,23; CADD %26,21) göstermektedir. Bu sınırlı kapsama oranlarının başlıca nedenleri arasında, araçların gereksinim duyduğu spesifik anotasyonlara bağımlılıkları, eksik referans verileri ve belirli biyolojik mekanizmalara odaklanmaları yer almaktadır. Buna karşın AGA modeli, farklı sinyalleri tek bir çerçevede birleştirerek daha geniş kapsama ve daha tutarlı performans sunmuş; ayrıca beşli klinik sınıflandırma üretebilmesiyle yalnızca risk skoru sağlayan yöntemlerin ötesinde karar destek potansiyeli göstermiştir.



Şekil 7. Doğruluk hedefiyle optimize edilen AGA sınıflandırıcısında özniteliklerin sınıflandırma kararına etkisi (Feature effects on the classification decision for the XGBoost classifier optimized for accuracy)

Tablo 7. AGA modeli ve in silico tahmin araçlarının patojenik varyant yakalama performanslarının karşılaştırılması (Comparison of the performance of the XGBoost model and in silico prediction tools in detecting pathogenic variants)

Tahmin Edici Yöntem	Geçerli Varyant Sayısı	Veri Kapsama Oranı (%)	Patojenik Olarak Sınıflandırılanlar	Hassasiyeti (%)
DANN	327	19.79	327	100
BayesDel AF	391	23.66	388	99.23
AGA	1652	100	1634	98.91
SIFT	37	02.23	36	97.30
CADD	433	26.21	414	95.61
MutationAssessor	35	02.11	33	94.29
ADA	92	05.56	86	93.48
RF	92	05.56	84	91.30
Varity	22	1.33	20	90.91
MaxEntScan	97	05.87	80	82.47
metaRNN	43	02.60	34	79.07
PolyPhen	37	02.23	24	64.86
PhyloP100Way	866	52.42	359	41.45
Fathmm-XF	187	11.31	15	08.02

Bu çalışmanın bazı sınırlılıkları bulunmaktadır. Öncelikle, kullanılan veri seti ClinVar veritabanındaki mevcut klinik etiketlere dayandığından, elde edilen sonuçlar bu kaynaktaki varyantlarla sınırlıdır ve etiketlerin zaman içinde değişebilme olasılığı göz önünde bulundurulmalıdır. Klinik genomik çalışmalarda sık karşılaşılan bir durum olarak, Belirsiz Öneme Sahip Varyantlar ve Çelişkili Klinik Yorumlara Sahip Varyantlar sınıflarının baskın olması veri setinde sınıf dengesizliğine yol açmakta ve nadir görülen Muhtemel Patojenik varyantların temsilini sınırlamaktadır. Geliştirilen model yalnızca BRCA1/2 genlerine özgü varyantlar üzerinde eğitilmiş olup, bulguların diğer genlere genellenebilirliği ileri çalışmalarla değerlendirilmelidir. Bununla birlikte, model çıktıları retrospektif anotasyonlar ve in silico tahminlere dayandığından, elde edilen yüksek doğruluk oranlarına rağmen performans henüz prospektif klinik testler veya fonksiyonel doğrulamalar ile sınanmamıştır. Bu nedenle önerilen yaklaşım klinik tanı amacıyla değil, araştırma amaçlı destekleyici bir analiz aracı olarak değerlendirilmelidir ve model çıktıları klinik bağlamda tek başına kullanılmamalı, nihai yorum uzman genetik danışmanlar veya hekimler tarafından yapılmalıdır. Gerçek hasta popülasyonlarında benzer performansın sürdürülebilirliği, prospektif çalışmalar ve çok merkezli kohort analizleri ile doğrulanmalıdır.

5. Sonuçlar (Conclusions)

Bu çalışmada, BRCA1/2 varyantlarının patojenite sınıflandırılması amacıyla açıklanabilir makine öğrenimi tabanlı bir yaklaşım geliştirilmiş ve farklı öğrenme paradigmasını temsil eden algoritmalar kapsamlı biçimde karşılaştırılmıştır. ClinVar kaynaklı klinik etiketler ile çoklu popülasyon veritabanları, ACMG/AMP varyant yorumlama kriterleri ve in silico tahmin araçlarından elde edilen özniteliklere dayalı olarak yürütülen analizler, sınıf dengesizliğine rağmen geliştirilen modellerin yüksek doğruluk ve güçlü genelleme kapasitesi sergilediğini göstermiştir. Elde edilen sonuçlar, özellikle AGA algoritmasının hem doğruluk hem de makro F1-skoru açısından en tutarlı performansı sunduğunu ve klinik açıdan kritik öneme sahip azınlık sınıfların ayırt edilmesinde belirgin bir üstünlük sağladığını ortaya koymuştur. SHAP tabanlı açıklanabilirlik analizleri, modelin sınıflandırma kararlarını biyolojik ve klinik açıdan anlamlı özniteliklere dayandığını ortaya koyarak, geliştirilen yaklaşımın yalnızca yüksek performanslı değil, aynı zamanda biyolojik gerçeklerle uyumlu ve yorumlanabilir bir karar mekanizması sunduğunu desteklemektedir. Bu bulgular, makine öğrenimi temelli ve açıklanabilir yaklaşımların BRCA1/2 varyant yorumlamasında ölçeklenebilir ve güvenilir karar destek araçları geliştirilmesine katkı sağlayabileceğine işaret etmektedir. Önerilen modelin çıktıları klinik tanı veya yeniden sınıflandırma amacıyla doğrudan kullanılmak üzere tasarlanmamış olup, hipotez üretimi ve varyant önceliklendirme süreçlerini destekleyen araştırma amaçlı çıktılar olarak değerlendirilmelidir. Gelecek çalışmalarda, bu çalışmada tanımlanan sınırlılıkların giderilmesi hedeflenmektedir. Bu kapsamda, modelin farklı gen panelleri ve bağımsız klinik kohortlar üzerinde değerlendirilerek genellenebilirliğinin artırılması planlanmaktadır. Bununla birlikte, fonksiyonel deneyler ve prospektif klinik doğrulama çalışmalarının modele entegre edilmesiyle, önerilen yaklaşımın klinik güvenilirliğinin ve gerçek hasta popülasyonlarındaki etkinliğinin daha güçlü biçimde ortaya konması amaçlanmaktadır.

Teşekkür (Acknowledgement)

Bu araştırma, TÜBİTAK TEYDEB programı kapsamında 7241058 numaralı proje ile desteklenmiştir.

Veri Erişilebilirliği Beyanı (Data Accessibility Statement)

Bu çalışmada kullanılan veri seti, makul talep üzerine paylaşılacaktır.

Kaynaklar (References)

- Xu H, Xu B. Breast cancer: Epidemiology, risk factors and screening. *Chin J Cancer Res.* 35(6), 565-583, 2023.
- Polyak K., Heterogeneity in breast cancer, *J. Clin. Invest.*, 121 (10), 3786-3788, 2011.
- Arnold M., Morgan E., Rumgay H., Mafra A., Singh D., Laversanne M., Vignat J., Gralow J.R., Cardoso F., Siesling S., Soerjomataram I., Current and future burden of breast cancer: Global statistics for 2020 and 2040, *Breast*, 66, 15-23, 2022.
- Sung H., Ferlay J., Siegel R.L., Laversanne M., Soerjomataram I., Jemal A., Bray F., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, 71 (3), 209-249, 2021.
- Etil, M., Dedetürk, B. K., Kolukisa, B., Bakır-Güngör, B., & Güngör, V. C., Breast Cancer Detection Using a New Parallel Hybrid Logistic Regression Model Trained by Particle Swarm Optimization and Clonal Selection Algorithms. *Concurrency and Computation: Practice and Experience*, 37 (12-14), e70107, 2025.
- Sanguinetti A., Polistena A., Lucchini R., Monacelli M., Galasse S., Avenia S., Triola R., Monacelli A., Avenia N., Male breast cancer, clinical presentation, diagnosis and treatment: Twenty years of experience in our Breast Unit, *Int. J. Surg. Case Rep.*, 20 (Suppl), 8-11, 2016.
- Huang A., Li D., Fan Z., Chen J., Zhang W., Wu W., Long-term trends in the incidence of male breast cancer and nomogram for predicting survival in male breast cancer patients: a population-based epidemiologic study, *Scientific Reports*, 15, 2027, 2025.
- Karami F., Mehdipour P., A comprehensive focus on global spectrum of BRCA1 and BRCA2 mutations in breast cancer, *Biomed. Res. Int.*, 2013 (1), 928562, 2013.
- Vatsyayan A., Scaria V., Machine learning approach to assess the pathogenicity of BRCA1/2 genetic variants: brca-NOVUS, *medRxiv*, 2023-10, 2023.
- Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., & Ding, W., A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1, *Science*, 266 (5182), 66-71, 1994.
- Richards S., Aziz N., Bale S., Bick D., Das S., Gastier-Foster J., Grody W.W., Hegde M., Lyon E., Spector E., Voelkerding K., Rehm H.L., 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.*, 17 (5), 405-423, 2015.
- Nicora G., Zucca S., Limongelli I., Bellazzi R., Magni P., A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization, *Sci. Rep.*, 12 (1), 2517, 2022.
- Vatsyayan A., Kumar M., Saikia B.J., Scaria V., BK B., WilsonGenAI: A deep learning approach to classify pathogenic variants in Wilson disease, *PLoS One*, 19 (5), e0303787, 2024.
- Adzhubei I., Jordan D.M., Sunyaev S.R., Predicting functional effect of human missense mutations using PolyPhen-2, *Curr. Protoc. Hum. Genet.*, 76 (1), 7-20, 2013.
- Rentzsch P., Witten D., Cooper G.M., Shendure J., Kircher M., CADD: Predicting the deleteriousness of variants throughout the human genome, *Nucleic Acids Res.*, 47 (D1), D886-D894, 2019.
- Niroula A., Vihinen M., How good are pathogenicity predictors in detecting benign variants?, *PLoS Comput. Biol.*, 15 (2), e1006481, 2019.
- Ernst C., Hahnen E., Engel C., Nothnagel M., Weber J., Schmutzler R.K., Hauke J., Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics, *BMC Med. Genomics*, 11 (1), 35, 2018.
- Bai K., Yang L., Xue J., Zhao L., Hao F., Pathogenicity classification of missense mutations based on deep generative model, *Comput. Biol. Med.*, 170, 107980, 2024.
- Kang M., Kim S., Lee D.B., Hong C., Hwang K.B., Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants, *Sci. Rep.*, 13 (1), 10478, 2023.
- Ansari, Z. A., Tripathi, M. M., Ahmed, R., The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models. *Discover Artificial Intelligence*, 5 (1), 75, 2025.
- Güler, H., Avcı, D., Ulaş, M., Omma, T., In-depth analysis of machine learning models and explainable artificial intelligence methods in

- diabetes diagnosis. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 40 (3), 1995–2012, 2025.
22. Karczewski K.J., Francioli L.C., Tiao G., Cummings B.B., Alfoldi J., Wang Q., Collins R.L., Laricchia K.M., Ganna A., Birnbaum D.P., ... MacArthur D.G., The mutational constraint spectrum quantified from variation in 141,456 humans, *Nature*, 581 (7809), 434–443, 2020.
 23. Li C., Zhi D., Wang K., Liu X., MetaRNN: Differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning, *Genome Med.*, 14 (1), 115, 2022.
 24. Ioannidis N.M., Rothstein J.H., Pejaver V., Middha S., McDonnell S.K., Baheti S., Musolf A., Li Q., Holzinger E., Karyadi D., ... Cooney K.A., REVEL: An ensemble method for predicting the pathogenicity of rare missense variants, *Am. J. Hum. Genet.*, 99 (4), 877–885, 2016.
 25. Quang D., Chen Y., Xie X., DANN: A deep learning approach for annotating the pathogenicity of genetic variants, *Bioinformatics*, 31 (5), 761–763, 2014.
 26. Shihab H.A., Gough J., Cooper D.N., Stenson P.D., Barker G.L., Edwards K.J., Day I.N.M., Gaunt T.R., Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models, *Hum. Mutat.*, 34 (1), 57–65, 2013.
 27. Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., Kondrashov A.S., Sunyaev S.R., A method and server for predicting damaging missense mutations, *Nat. Methods*, 7 (4), 248–249, 2010.
 28. Ng P.C., Henikoff S., SIFT: Predicting amino acid changes that affect protein function, *Nucleic Acids Res.*, 31 (13), 3812–3814, 2003.
 29. Wu Y., Liu H., Li R., Sun S., Weile J., Roth F.P., Improved pathogenicity prediction for rare human missense variants, *Am. J. Hum. Genet.*, 108 (10), 1891–1906, 2021.
 30. Cheng J., Novati G., Pan J., Bycroft C., Zemgulyte A., Applebaum T., Treiber M., Logan B., Yeager C., Leong J., Raghavan M., Taylor-Weiner A., Parker M., Esteban O., Salvatore M., Persikov A., Durbin R., Avsec Ž., Accurate proteome-wide missense variant effect prediction with AlphaMissense, *Science*, 381 (6664), eadg7492, 2023.
 31. Landrum M.J., Lee J.M., Benson M., Brown G., Chao C., Chitipiralla S., Gu B., Hart J., Hoffman D., Hoover J., Jang W., Katz K., Ovetsky M., Riley G., Sethi A., Tully R., Villamarin-Salomon R., Rubinstein W., Maglott D.R., ClinVar: Public archive of interpretations of clinically relevant variants, *Nucleic Acids Res.*, 44 (D1), D862–D868, 2016.
 32. McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R., Thormann A., Flicek P., Cunningham F., The Ensembl Variant Effect Predictor, *Genome Biol.*, 17 (1), 122, 2016.
 33. Landrum, M. J., Chitipiralla, S., Kaur, K., Brown, G., Chen, C., Hart, J., ... & Kattman, B. L., ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Research*, 53 (D1), D1313-D1321, 2025.
 34. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., VanderPlas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 12, 2825–2830, 2011.
 35. Snoek J., Larochelle H., Adams R.P., Practical Bayesian optimization of machine learning algorithms, *Adv. Neural Inf. Process. Syst.*, 25, 2951–2959, Lake Tahoe, Nevada, USA, 2012.
 36. Lundberg S.M., Lee S.I., A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.*, 30, 4765–4774, Long Beach, CA, USA, 2017.