T.C.

ERCİYES ÜNİVERSİTESİ

GEVHER NESİBE GENOM VE KÖK HÜCRE ENSTİTÜSÜ

MOLEKÜLER BİYOLOJİ VE GENETİK ANABİLİM DALI

OTOİNFLAMATUVAR HASTALIKLARDA KLİNİK ÖNEMİ BİLİNMEYEN PATOJENİK TEK NÜKLEOTİD VARYANTLARININ YAPAY ZEKA TEMELLİ ARAÇLARLA TESPİT EDİLMESİ

Hazırlayan

Tugay TOPAK

Danışman

Doç. Dr. Özkan Ufuk NALBANTOĞLU

Yüksek Lisans Tezi

Ocak 2026

KAYSERİ

T.C.

ERCİYES ÜNİVERSİTESİ

GEVHER NESİBE GENOM VE KÖK HÜCRE ENSTİTÜSÜ

MOLEKÜLER BİYOLOJİ VE GENETİK ANABİLİM DALI

OTOİNFLAMATUVAR HASTALIKLARDA KLİNİK ÖNEMİ BİLİNMEYEN PATOJENİK TEK NÜKLEOTİD VARYANTLARININ YAPAY ZEKA TEMELLİ ARAÇLARLA TESPİT EDİLMESİ

(Yüksek Lisans Tezi)

Hazırlayan

Tugay TOPAK

Danışman

Doç. Dr. Özkan Ufuk NALBANTOĞLU

Ocak 2026

KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Tugay TOPAK

YÖNERGEYE UYGUNLUK ONAYI

"**Otoinflamatuvar Hastalıklarda Klinik Önemi Bilinmeyen Patojenik Tek Nükleotid Varyantlarının Yapay Zeka Temelli Araçlarla Tespit Edilmesi**" adlı **Yüksek Lisans** Tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi' ne uygun olarak hazırlanmıştır.

Tezi hazırlayan	Danışman
Tugay TOPAK	Doç. Dr. Özkan Ufuk NALBANTOĞLU

Moleküler Biyoloji ve Genetik ABD Başkanı

Prof. DR. Zühal HAMURCU

KABUL ONAY SAYFASI

Doç. Dr. Özkan Ufuk NALBANTOĞLU danışmanlığında **Tugay TOPAK** tarafından hazırlanan "**Otoinflamatuvar Hastalıklarda Klinik Önemi Bilinmeyen Patojenik Tek Nükleotid Varyantlarının Yapay Zeka Temelli Araçlarla Tespit Edilmesi**" adlı bu çalışma jürimiz tarafından Erciyes Üniversitesi Gevher Nesibe Genom ve Kök Hücre Enstitüsü Moleküler Biyoloji ve Genetik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

……./……./20…….

JÜRİ **İmza**

Danışman : Doç. Dr. Özkan Ufuk NALBANTOĞLU

Üye : Dr. Öğr. Üyesi Müge Gülcihan ÖNAL

Üye : Doç. Dr. Burcu GÜNGÖR

ONAY

Bu tezin kabulü Ensitü Yönetim Kurulunun …….. tarih ve …….. sayılı kararı ile onaylanmıştır.

……./……./20…….

Prof. Dr. Zühal HAMURCU

Enstitü Müdürü

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmam süresince, bilgi birikimi, bilimsel rehberliği ve desteğiyle her aşamada yolumu aydınlatan ve çalışmamın olgunlaşmasında büyük katkısı olan değerli danışmanım Doç. Dr. Özkan Ufuk NALBANTOĞLU'na,

Tez çalışmam için gerekli akademik ortamı ve teknik altyapıyı sağlayan Erciyes Üniversitesi'ne, Gevher Nesibe Genom ve Kök Hücre Enstitüsü ve Betül-Ziya Eren Genom ve Kök Hücre Merkezi'nin yönetici ve çalışanlarına,

Tez sürecinde bana değerli fikirleri ile yol gösteren ekip arkadaşım ve manevi ağabeyim Mehmet HORA'ya,

Yüksek lisans eğitimim süresince manevi desteğini ve dostluğunu esirgemeyen tüm arkadaşlarıma,

Hayatımın her döneminde olduğu gibi bu süreçte de maddi ve manevi desteklerini esirgemeyen, her zaman yanımda olduklarını hissettiren sevgili aileme; özellikle bana her koşulda güvenen, emekleriyle bugünlere gelmemi sağlayan sevgili annem Tülay ÖRS'e,

Tez çalışmam boyunca gerek akademik gerekse manevi anlamda desteğini sürekli hissettiğim, fikir alışverişleriyle çalışmamın gelişmesine katkıda bulunan, tüm bu süreç boyunca yanımda olan sevgili arkadaşım Kevser ZADIKAN'a içtenlikle teşekkür ederim.

Tugay Topak

Ocak 2026, Kayseri

OTOİNFLAMATUVAR HASTALIKLARDA KLİNİK ÖNEMİ BİLİNMEYEN PATOJENİK TEK NÜKLEOTİD VARYANTLARININ YAPAY ZEKA TEMELLİ ARAÇLARLA TESPİT EDİLMESİ

Tugay TOPAK

Erciyes Üniversitesi, Gevher Nesibe Genom ve Kök Hücre Enstitüsü

Yüksek Lisans Tezi, Ocak 2026

Danışman: Doç. Dr. Özkan Ufuk NALBANTOĞLU

ÖZET

Otoinflamatuvar hastalıklar, doğuştan bağışıklık yanıtını düzenleyen moleküler mekanizmalardaki kalıtsal bozukluklar sonucu ortaya çıkan, ataklar halinde seyreden ve yüksek morbiditeye yol açabilen heterojen bir hastalık grubudur. Dizileme temelli moleküler testler bu hastalıklarla ilişkili genlerde çok sayıda varyant ortaya çıkarmakta; bunların önemli bir kısmının klinik önemi belirsiz (VUS) olarak raporlanması, yorum sürecini güçleştirmekte ve hesaplamalı patojenite tahmin araçlarının (Variant Effect Predictors, VEP) güvenilirliğini kritik hale getirmektedir. Bu bağlamda, VUS yorumunda sık kullanılan VEP araçlarının ayrıştırma performansı ve olasılık kalitesi ile, bu skorların makine öğrenmesi (ML) modelleri içinde birleştiğinde ek bir kazanç sağlayıp sağlamadığının değerlendirilmesi önem taşımaktadır.

Bu tezde, on iki monogenik otoinflamatuvar hastalığın çekirdek genlerinden oluşturulan bir panel üzerinde, üç yaygın VEP aracı (AlphaMissense, PolyPhen-2, SIFT) ile bu araçların kalibre edilmiş olasılık çıktılarından türetilen ML sınıflandırıcılarının (Logistic Regression, SVM, Random Forest, XGBoost, MLP) performansı sistematik olarak karşılaştırılmıştır. Panel, ClinVar veritabanından seçilen 287 patojenik/benign missense varyanttan oluşturulmuştur. Her varyant için elde edilen ham VEP skorları Platt scaling ile 0–1 aralığında ortak bir olasılık diline dönüştürülmüş ve bu üç olasılık, ML modelleri için üç boyutlu bir özellik vektörü olarak kullanılmıştır. Tüm yöntemler, veri sızıntısını önlemek amacıyla nested katmanlı 5-fold çapraz doğrulama altında değerlendirilmiştir. Ayrıştırma performansı ROC eğrileri, AUROC ve AUPRC ile, olasılık kalitesi ise Brier skoru, Expected

Calibration Error (ECE), kalibrasyon eğimi/kesişimi ve Hosmer–Lemeshow testi ile incelenmiştir. Modeller arası farklar, eşik-bağımsız düzlemde eşleştirilmiş bootstrap ΔAUC ve eşik-bağımlı düzlemde 0.50 karar eşiğinde McNemar exact testiyle istatistiksel olarak sınanmıştır.

Bulgular, ham skor evresinde AlphaMissense'in düşük yanlış-pozitif bölgede belirgin early lift ve yüksek AUROC/AUPRC ile PolyPhen-2 ve SIFT'e üstünlük sağladığını; Logistic Regression'ın yalnızca üç VEP olasılığını kullanmasına rağmen ayrıştırma bakımından AlphaMissense'i yakından izlediğini göstermiştir. Kalibrasyon sonrasında özellikle Logistic Regression, MLP ve SVM modelleri Brier skoru ve ECE açısından daha iyi olasılık kalitesi sunarken, AlphaMissense, Random Forest ve XGBoost için bazı olasılık aralıklarında sistematik sapmanın sürdüğü gözlenmiştir. Eşik-bağımsız (ΔAUC) ve eşik-bağımlı (McNemar) karşılaştırmalar, bu üstünlüklerin rastlantısal örnekleme gürültüsüyle açıklanamayacağını ve benzer AUROC değerlerine sahip modellerin dahi aynı varyantlarda farklı hata desenleri üretebildiğini ortaya koymuştur. Sonuç olarak bu çalışma, otoinflamatuvar gen paneli bağlamında yaygın kullanılan VEP araçları ve VEP-türevli ML sınıflandırıcılarını aynı etiketli veri üzerinde ayrıştırma, kalibrasyon ve karar düzeyi istatistikleri açısından birlikte değerlendirerek; hangi yöntemlerin hangi koşullarda daha yüksek ayrıştırma sağladığı ve hangi kalibrasyon ayarlarında klinik açıdan daha güvenilir olasılıklar ürettiğine dair kullanılabilir, veri temelli bir çerçeve sunmaktadır.

Anahtar Kelimeler: AlphaMissense; Makine Öğrenmesi; Otoinflamatuar Hastalıklar; Tek Nükleotid Varyantı; Yapay Zeka

DETECTION OF CLINICALLY AMBIGUOUS PATHOGENİC SINGLE NUCLEOTIDE VARİANTS IN AUTOINFLAMMATORY DISEASES USING ARTIFICIAL INTELLIGENCE TOOLS

Tugay TOPAK

Erciyes University Gevher Nesibe Genome and Stem Cell Institute

Master Thesis, January 2026

Advisor: Assoc. Prof. Dr. Özkan Ufuk NALBANTOĞLU

ABSTRACT

Autoinflammatory diseases are a heterogeneous group of disorders that arise from inherited defects in the molecular mechanisms regulating the innate immune response and are characterized by recurrent inflammatory attacks and substantial morbidity. Sequencing-based molecular testing reveals numerous variants in genes associated with these diseases, and a considerable proportion of them are reported as variants of uncertain significance (VUS). This complicates clinical interpretation and makes the reliability of computational pathogenicity predictors (Variant Effect Predictors, VEPs) a critical issue. In this context, it is important to evaluate both the discrimination performance and probability calibration of VEPs that are widely used in VUS interpretation, and to assess whether combining their scores within machine learning (ML) models provides additional predictive gain.

In this thesis, the performance of three widely used VEP tools (AlphaMissense, PolyPhen-2, SIFT) and ML classifiers (Logistic Regression, SVM, Random Forest, XGBoost, MLP) trained on their calibrated probability outputs was systematically compared on a panel constructed from the core genes of twelve monogenic autoinflammatory diseases. The panel was generated from 287 pathogenic/benign missense variants selected from the ClinVar database. For each variant, the raw scores of the VEP tools were transformed into a common 0–1 probability scale using Platt scaling, and these three probabilities were used as a three-dimensional feature vector for the ML models. All methods were evaluated under nested 5-fold cross-validation to prevent data leakage. Discrimination performance was examined using ROC curves, AUROC and AUPRC, whereas probability quality was assessed by

Brier score, Expected Calibration Error (ECE), calibration slope/intercept and the Hosmer–Lemeshow test. Differences between models were tested statistically using paired-bootstrap ΔAUC in the threshold-independent setting and the exact McNemar test at the 0.50 decision threshold in the threshold-dependent setting.

The results showed that, at the raw-score level, AlphaMissense achieved clear early lift in the low false-positive region and higher AUROC/AUPRC than PolyPhen-2 and SIFT, while Logistic Regression, despite using only three VEP probabilities as input, closely followed AlphaMissense in terms of discrimination. After calibration, Logistic Regression, MLP and SVM provided better probability quality in terms of Brier score and ECE, whereas AlphaMissense, Random Forest and XGBoost exhibited persistent systematic deviations in certain probability ranges. Threshold-independent (ΔAUC) and threshold-dependent (McNemar) comparisons indicated that these advantages cannot be attributed to random sampling noise and that even models with similar AUROC values can display different error patterns on the same variants. In conclusion, by jointly evaluating widely used VEP tools and VEP-derived ML classifiers on the same labelled dataset in terms of discrimination, calibration and decision-level statistics, this study provides a practical, data-driven framework for identifying which methods achieve higher discrimination under which conditions and which calibration settings yield clinically more reliable probabilities in the context of autoinflammatory gene panels.

Keywords: AlphaMissense; Artificial Intelligence; Autoinflammatory Diseases; Machine Learning; Single Nucleotide Variant

İÇİNDEKİLER

BİLİMSEL ETİĞE UYGUNLUK i
YÖNERGEYE UYGUNLUK ONAYI ii
KABUL ONAY SAYFASI iii
TEŞEKKÜR iv
ÖZET v
ABSTRACT vii
İÇİNDEKİLER ix
KISALTMALAR xii
TABLOLAR LİSTESİ xvi
ŞEKİLLER LİSTESİ xvii
1. GİRİŞ VE AMAÇ 1
2. GENEL BİLGİLER 6
2.1. Otoinflamatuvar Hastalıklar 6
2.2. CLINVAR Klinik Veri Tabanı 8
2.3. Patojenite Tahmin Araçları 9
2.3.1 AlphaMissense 10
2.3.2. PolyPhen-2 11
2.3.3. SIFT 12
2.4. Metrikler 14
2.4.1. Accuracy 15
2.4.2. AUROC 16
2.4.3. F1 Skoru 17
2.4.4. Matthews Correlation Coefficient (MCC) 19
2.4.5. Specificity (Özgüllük) 19
2.4.7. Precision 21
2.4.8. AUPRC 21
2.4.9. Brier Skoru 22
2.4.10. Expected Calibration Error (ECE) 23
2.5. Makine Öğrenmesi Algoritmaları 23
2.5.1. Random Forest 26
2.5.2. Multi-Layer Perceptron (MLP) 27

2.5.3. Extreme Gradient Boosting (XGBoost) 29
2.5.4. Support Vector Machine (SVM) 30
2.5.5. Logistic Regression 31
2.6. İstatistiksel Değerlendirme ve Karşılaştırma Yöntemleri 32
2.6.1. Bootstrap Testi 32
2.6.2. Hosmer–Lemeshow Kalibrasyon Testi 34
2.6.3. McNemar Testi 35
3. GEREÇ VE YÖNTEM 37
3.1 GEREÇLER 38
3.2. YÖNTEM 39
3.2.1. Araştırmanın Tipi 39
3.2.2. Araştırmada Kullanılacak Genlerin İncelenmesi 39
3.2.3. Araştırma Evreni ve Örneklem 41
3.2.4. Verilerin Elde Edilmesi 42
3.2.4.1. CLINVAR'dan Varyant Verilerinin Elde Edilmesi 42
3.2.4.2. Tahmin Araçlarından Skor Elde Edilmesi 43
3.2.4.2.1. AlphaMissense 43
3.2.4.2.2. SIFT 43
3.2.4.2.3. PolyPhen-2 44
3.2.5. CLİNVAR Verilerinin ve Araç Tahminlerinin Eşleştirilmesi 44
3.2.6. Direction Harmonization 45
3.2.7. Araçların Ham Skor Analizi 46
3.2.7.1. Youden's J Yöntemi İle Araçların Optimum Threshold Eşiğinin Belirlenmesi 46
3.2.7.2. Araç-Özel Eşik Değerinde Performans Metriklerinin Değerlendirilmesi 48
3.2.8. Araçların Post-Hoc Kalibrasyonu 50
3.2.9. Makine Öğrenmesi Algoritmalarını Eğitmek İçin Kalibre Edilmiş Veriler İle Özellik Seti Oluşturma 53
3.2.10. Hiperparametre Optimizasyonu (Tuning) 54
3.2.11. Makine Öğrenmesi Algoritmalarının Eğitimi 58

3.2.12. Makine Öğrenmesi Algoritmalarının Bireysel Başarılarının Değerlendirilmesi 59
3.2.12.1. Youden's J Yöntemi İle ML Algoritmalarının Optimum Threshold Eşiğinin Belirlenmesi 59
3.2.12.2. Modellerin Özel Optimum Eşik Noktasında Performanslarının Değerlendirilmesi 60
3.2.13. ML Çıktılarının Post-Hoc Olasılık Kalibrasyonu 61
3.2.14. Araç ve ML Algoritmalarının Ortak Eşikte Karşılaştırılması 63
3.2.15. İstatistiksel Analiz 64
4. BULGULAR 67
4.1. Modellerin Model-Spesifik Karar Eşiğinde Tekil Performansı 68
4.2. Modellerin Kalibrasyon Sonrası Ortak Eşikte Karşılaştırmalı Performansları 78
4.3. İstatistiksel Analiz Sonuçları 98
4.4. Genler Özelinde Kalibre Edilmiş Modellerin Metrik Performansları 113
5. TARTIŞMA 120
5.1. Ana Bulgular 124
5.2. Ayrıştırma Performansı Sonuçları 127
5.3. Olasılık Skorları ve Kalibrasyon Kalitesi 129
5.4. Eşik Bağımsız ve Eşik Bağımlı İstatistiksel Karşılaştırmalar 132
6. SONUÇ VE ÖNERİLER 135
7. KAYNAKLAR 138

KISALTMALAR

AA : Amyloid A

ACMG : American College of Medical Genetics and Genomics

AID : Autoinflammatory Disease

AM : AlphaMissense

AMP : Association for Molecular Pathology

AUC : Area Under the Curve

AUPRC : Area Under the Precision–Recall Curve

AUROC : Area Under the Receiver Operating Characteristic Curve

B/LB : Benign / Likely Benign

BH-FDR : Benjamini–Hochberg False Discovery Rate

BP4 : Benign Supporting Criterion 4

β5 : Constitutive proteasome β5 subunit

β5i : Immunoproteasome β5i subunit

CADD : Combined Annotation Dependent Depletion

CAPS : Cryopyrin Associated Periodic Syndromes

CARD15 : Caspase Recruitment Domain containing protein 15

CD2BP1 : Cluster of Differentiation 2 Binding Protein 1

CI : Confidence Interval

CINCA : Chronic Infantile Neurological Cutaneous and Articular syndrome

ClinVar : Clinical Variant Database

CRP : C-Reactive Protein

χ^2 : chi-square

DIRA : Deficiency of the Interleukin-1 Receptor Antagonist

DITRA : Deficiency of the Interleukin-36 Receptor Antagonist

ΔAUC : Difference in Area Under the Curve

ECE : Expected Calibration Error

ESR : Erythrocyte Sedimentation Rate

FAM105B : Family with Sequence Similarity 105 Member B

FCAS : Familial Cold Autoinflammatory Syndrome

FDR : False Discovery Rate

FMF : Familial Mediterranean Fever

FN : False Negative

FP : False Positive

FPR : False Positive Rate

Fβ : F-beta score

GGPP : Geranylgeranyl pyrophosphate

GPP : Generalized Pustular Psoriasis

HGVS : Human Genome Variation Society

HIDS : Hyperimmunoglobulinemia D with Periodic Fever Syndrome

HL : Hosmer–Lemeshow

IL : Interleukin

IL-1R : Interleukin-1 receptor

IL-1Ra : Interleukin-1 receptor antagonist

IL-1RAcP : Interleukin-1 receptor accessory protein

IL-1RL2 : Interleukin-1 receptor-like 2

IL1RN : Interleukin 1 Receptor Antagonist gene

IL-36 : Interleukin-36

IL-36R : Interleukin-36 receptor

IL36RN : Interleukin 36 receptor antagonist gene

ISG : Interferon-Stimulated Gene

LMP7 : Low Molecular Mass Polypeptide 7

LR : Logistic Regression

MAPK : Mitogen-Activated Protein Kinase

MCC : Matthews Correlation Coefficient

MDP : Muramyl Dipeptide

MEFV : Mediterranean fever gene

MKD : Mevalonate Kinase Deficiency

ML : Machine Learning

MLP : Multi-Layer Perceptron

MSA : Multiple Sequence Alignment

MVK : Mevalonate Kinase

MWS : Muckle–Wells syndrome

NCBI : National Center for Biotechnology Information

NF-κB : Nuclear Factor kappa B

NGS : Next-Generation Sequencing

NLR : NOD-like receptor

NLRP12 : NOD-Like Receptor Family Pyrin Domain Containing 12

NLRP3 : NOD-Like Receptor Family Pyrin Domain Containing 3

NOD2 : Nucleotide Binding Oligomerization Domain Containing 2

NOMID : Neonatal Onset Multisystem Inflammatory Disease

NPV : Negative Predictive Value

NSAID : Non Steroidal Anti-Inflammatory Drug

OOF : Out of Fold

OR : Odds Ratio

OTULIN : Ovarian Tumor Deubiquitinase With Linear Linkage Specificity

P/LP : Pathogenic / Likely Pathogenic

PAPA : Pyogenic Arthritis, Pyoderma Gangrenosum and Acne

POMP : Proteasome Maturation Protein

PP2 : PolyPhen-2

PP3 : Pathogenic Supporting Criterion 3

PPV : Positive Predictive Value

PR : Precision-Recall

PRAAS : Proteasome-Associated Autoinflammatory Syndrome

PSMB8 : Proteasome Subunit Beta Type-8

PSTPIP1 : Proline-Serine-Threonine Phosphatase Interacting Protein 1

PTPN12 : Protein Tyrosine Phosphatase Non-receptor Type 12

PYD : Pyrin Domain

RBF : Radial Basis Function

RCV : Reference ClinVar Variation

REVEL : Rare Exome Variant Ensemble Learner

RF : Random Forest

ROC : Receiver Operating Characteristic

ROS : Reactive Oxygen Species

SAA : Serum Amyloid A

SAID : Systemic Autoinflammatory Diseases

SCV : Submitted ClinVar Variation

SHAP : SHapley Additive Explanations

SIFT : Sorting Intolerant From Tolerant

SNV : Single Nucleotide Variant

SVC : Support Vector Classifier

SVM : Support Vector Machine

TN : True Negative

TNFR1 : Tumor Necrosis Factor Receptor 1

TNFRSF1 : Tumor Necrosis Factor Receptor Superfamily Member 1A

TNR : True Negative Rate

TP : True Positive

TRAPS : Tumor Necrosis Factor Receptor-Associated Periodic Syndrome

TRIPOD : Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

VCV : Variant Level Aggregate

VEP : Variant Effect Predictor

VUS : Variant of Uncertain Significance

XGB : Extreme Gradient Boosting

TABLOLAR LİSTESİ

Tablo 3.1. Hedef gen paneli ve örnek uzayı 42
Tablo 4.1. Modellerin Ham Skorlarının Model-Spesifik Eşikte Metrik Performansları 70
Tablo 4.2. Kalibre Edilmiş Modellerin Ortak Eşikte Metrik Performansları 80
Tablo 4.3. Hosmer–Lemeshow uyum testi sonuçları 102
Tablo 4.4. Eşleştirilmiş bootstrap ΔAUC ikili model karşılaştırmaları 104
Tablo 4.5. Ortak eşikte McNemar testi ikili model karşılaştırma sonuçları 112
Tablo 4.6. MEFV geninde kalibre edilmiş modellerin metrik performansları 114
Tablo 4.7. MVK geninde kalibre edilmiş modellerin metrik performansları 114
Tablo 4.8. NLRP3 geninde kalibre edilmiş modellerin metrik performansları ... 115
Tablo 4.9. NLRP12 geninde kalibre edilmiş modellerin metrik performansları . 115
Tablo 4.10. NOD2 geninde kalibre edilmiş modellerin metrik performansları 116
Tablo 4.11. TNFRSF1A geninde kalibre edilmiş modellerin metrik performansları 116
Tablo 4.12. PSTPIP1 geninde kalibre edilmiş modellerin metrik performansları . 117
Tablo 4.13. LPIN2 geninde kalibre edilmiş modellerin metrik performansları 117
Tablo 4.14. PSMB8 geninde kalibre edilmiş modellerin metrik performansları ... 118
Tablo 4.15. IL36RN geninde kalibre edilmiş modellerin metrik performansları .. 118
Tablo 4.16. IL1RN geninde kalibre edilmiş modellerin metrik performansları 119
Tablo 4.17. OTULIN geninde kalibre edilmiş modellerin metrik performansları . 119

ŞEKİLLER LİSTESİ

Şekil 4.1. (A) AlphaMissense'e ait ham skor kalibrasyon eğrisi ve (B) PolyPhen-2'ye ait ham skor kalibrasyon eğrisi 71

Şekil 4.2. (A) SIFT'e ait ham skor kalibrasyon eğrisi ve (B) Logistic Regression'a ait ham skor kalibrasyon eğrisi 72

Şekil 4.3. (A)Şekil 4.3 SVM'ye ait ham skor kalibrasyon eğrisi ve (B) Random Forest'e ait ham skor kalibrasyon eğrisi 73

Şekil 4.4. (A) XgBoost'a ait ham skor kalibrasyon eğrisi ve (B) MLP'ye ait ham skor kalibrasyon eğrisi 74

Şekil 4.5. (A) AlphaMissense'e ait ham skor ROC eğrisi ve (B) PolyPhen-2'ye ait ham skor ROC eğrisi 75

Şekil 4.6. (A) SIFT'e ait ham skor ROC eğrisi ve (B) Logistic Regression'a ait ham skor ROC eğrisi 76

Şekil 4.7. (A) SVM'ye ait ham skor ROC eğrisi ve (B) Random Forest'e ait ham skor ROC eğrisi 77

Şekil 4.8. (A) XgBoost'a ait ham skor ROC eğrisi ve (B) MLP'ye ait ham skor ROC eğrisi 78

Şekil 4.9.A. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) AUROC Değerleri 81

Şekil 4.9.B. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) AUPRC Değerleri 82

Şekil 4.9.C. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Brier Skoru Düzeyleri 83

Şekil 4.9.D. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) ECE Değerleri 84

Şekil 4.9.E. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Accuracy Skorları 85

Şekil 4.9.F. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) F1 Skorları 86

Şekil 4.9.G. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) MCC Skorları 87

Şekil 4.9.H. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Precision Skorları 88

Şekil 4.9.I. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Recall Skorları 89

Şekil 4.9J. Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Specificity Skorları 90

Şekil 4.10. (A) Kalibrasyon sonrası AlphaMissenseye ait kalibrasyon eğrisi, (B) PolyPhen-2'ye ait kalibrasyon eğrisi 91

Şekil 4.11. (A) Kalibrasyon sonrası SIFT'e ait kalibrasyon eğrisi, (B)Logistic Regression'a ait kalibrasyon eğrisi 92

Şekil 4.12. (A) Kalibrasyon sonrası SVM'ye ait kalibrasyon eğrisi, (B) Random Forest'e ait kalibrasyon eğrisi 93

Şekil 4.13. (A) Kalibrasyon sonrası XgBoost'a ait kalibrasyon eğrisi, (B) MLP'ye ait kalibrasyon eğrisi 94

Şekil 4.14. (A) Kalibrasyon sonrası AlphaMissense'e ait ROC eğrisi, (B) PolyPhen-2'ye ait ROC eğrisi 95

Şekil 4.15. (A) Kalibrasyon sonrası SIFT'e ait ROC eğrisi, (B) Logistic Regression'a ait ROC eğrisi 96

Şekil 4.16. (A) Kalibrasyon sonrası SVM'ye ait ROC eğrisi, (B) Random Forest'e ait ROC eğrisi 97

Şekil 4.17. (A) Kalibrasyon sonrası XGBoost'a ait ROC eğrisi, (B) MLP'ye ait ROC eğrisi 98

Şekil 4.18. Kalibre modeller arası eşleştirilmiş bootstrap analizi ısı haritası 107

Şekil 4.19. Ortak eşikte kalibre edilmiş modellerin McNemar testi ısı haritası ... 110

1. GİRİŞ VE AMAÇ

Otoinflamatuvar hastalıklar, doğuştan bağışıklık sisteminin düzenleyici mekanizmalarındaki kalıtsal bozukluklar sonucu ortaya çıkan, tekrarlayan ateş atakları, serözit, deri bulguları ve çeşitli organ sistemlerini tutan sistemik inflamasyon epizotları ile seyreden heterojen bir hastalık grubudur. Bu sendromlarda özellikle IL-1 ailesi sitokinleri, inflammazom bileşenleri (örneğin NLR ailesi sensörleri) ve pyrin kompleksi gibi moleküler düğümler, hastalığın patogenezinde merkezî bir rol oynamaktadır. Klinik düzeyde bakıldığında, özellikle Akdeniz havzası ve çevresinde sık görülen Ailevi Akdeniz Ateşi başta olmak üzere birçok monogenik otoinflamatuvar sendrom, çocukluk çağından itibaren ciddi morbiditeye, yaşam kalitesinde belirgin azalmaya ve tedavi edilmediğinde organ hasarına yol açabilmektedir. Tanının gecikmesi veya yanlış sınıflandırma ise hem gereksiz tedavilere hem de kalıcı komplikasyon riskinin artmasına neden olmaktadır.

Son yıllarda yeni nesil dizileme (NGS) teknolojilerinin klinik rutine girmesi, bu hastalıkların tanısında önemli bir dönüşüm sağlamıştır. Hedefli gen panelleri, ekzom ve genom dizileme uygulamaları sayesinde, inflammazom bileşenlerini ve ilgili düzenleyici genleri içeren panel testleri kısa sürede ve görece düşük maliyetle yapılabilmekte; ancak gen paneli taramalarında saptanan tek nükleotid değişimlerinin önemli bir bölümü "klinik önemi belirsiz varyant (VUS)" kategorisinde kalmaktadır. Varyantın nadirliği, çalışılan popülasyona özgü olması, fonksiyonel deney verilerinin sınırlılığı veya literatür bilgisinin yetersizliği gibi nedenlerle, bu varyantların patojenik mi yoksa benign mi olduğuna dair klinik olarak ikna edici bir kanıt çoğu zaman üretilememektedir.

Bu belirsizliği yönetmek için ACMG/AMP tarafından ortaya konan standart sınıflandırma çerçevesi, populasyon frekansı, fonksiyonel kanıt, segregrasyon bulguları, literatür raporları ve hesaplamalı tahminleyiciler gibi çok sayıda kanıt türünü bir araya getiren yapılandırılmış bir yaklaşım sunmaktadır. Bununla birlikte,

özellikle nadir ve popülasyona özgü missense değişimlerde, fonksiyonel deneyler ve güçlü segregrasyon verisi çoğu zaman mevcut değildir. Bu durumda, hesaplamalı patojenite kestiricileri yani variant effect predictor (VEP) araçları, ACMG/AMP sınıflandırmasında PP3/BP4 gibi kodlar üzerinden dolaylı da olsa karar sürecine dâhil edilen başlıca kanıt kaynaklarından biri haline gelmektedir. Bu nedenle, kullanılan VEP araçlarının ayrıştırma gücü kadar, ürettikleri olasılıkların güvenilirliği ve klinik bağlamda ne ölçüde kalibre oldukları da önemli bir metodolojik ve pratik problem olarak öne çıkmaktadır.

Literatürde SIFT, PolyPhen-2 gibi "klasik" kestiricilerin ve bunlardan türetilen bileşik skorların performansını değerlendiren çok sayıda çalışma bulunmaktadır. Bu çalışmaların bir kısmı, geniş ve heterojen veri üzerinde eğitilen, çoklu anotasyonu gömülü modellerle birleştiren CADD, REVEL gibi bileşik skorlar, genel olarak güçlü ayrıştırma metrikleri sunsalar da gen veya yolak düzeyinde spesifik panellerde her zaman aynı performansı tekrarlamayabilmektedir.

Öte yandan derin öğrenme tabanlı yeni nesil yaklaşımlar, bu bağlamda yeni bir imkân sunmaktadır. AlphaMissense gibi araçlar, protein yapı temsillerini, çok ölçekli evrimsel bilgiyi ve yerel bağlam ipuçlarını derin sinir ağları ile birleştirerek tüm proteom boyunca missense varyantlar için olasılık tabanlı patojenite skorları üretmektedir. Orijinal çalışmalarında, çok sayıda fonksiyonel deney seti üzerinde yüksek ayrıştırma gücü gösterdikleri ve belirli gen gruplarında önceki nesil kestiricilere göre anlamlı üstünlük sağladıkları bildirilmiştir. Ancak bu üstünlüğün, spesifik klinik panellerde, özellikle de otoinflamatuvar çekirdek genler için oluşturulmuş etiketli veri kümelerinde ne ölçüde korunduğu henüz sistematik olarak ortaya konmamıştır.

Hesaplamalı kestiricilerin tek tek performanslarının ötesinde, bu skorların birlikte kullanıldığı makine öğrenmesi sınıflandırıcıları da önemli bir araştırma eksenini oluşturmaktadır. VEP skorları yapısal, evrimsel ve biyofiziksel bilginin özetleri olarak düşünülebilir ve bu skorların kalibre edilerek ortak bir olasılık diline taşınması, Logistic Regression, Support Vector Machine, Random Forest, XGBoost ve Multi-Layer Perceptron gibi modeller için kompakt fakat bilgilendirici bir özellik uzayı oluşturma potansiyeli taşımaktadır. Bu tür modeller, tek tek araçların sunduğu

kısmi bilgileri öğrenerek daha dengeli performans sergileyebilir; ancak burada da metodolojik olarak dikkat edilmesi gereken iki temel nokta bulunmaktadır. Birincisi, eğitim ve değerlendirme süreçlerinde nested cross-validation ve out-of-fold (OOF) tahminler gibi veri sızıntısını sistematik biçimde engelleyen yapılar kullanılmadığında, ML modellerinin sergiledikleri performanslar klinik gerçeklikle uyuşmayabilir.

İkincisi, bu modellerin ürettiği olasılıkların kalibrasyon durumu ayrı bir boyut olarak ele alınmadığında, yüksek AUROC değerleri "iyi olasılık" anlamına gelmeyebilir ve yanlış eşik seçimleri klinik hata profilini büyütebilir.

Otoinflamatuvar alanına özgü literatür incelendiğinde, tekil genlere odaklanan veya sınırlı örneklemli yöntem denemeleri bulunmakla birlikte, bu tez çalışmasının hedefi olan on iki monogenik otoinflamatuvar hastalığın çekirdek genlerini kapsayan bir panel üzerinde, AlphaMissense, PolyPhen-2 ve SIFT gibi araçlar ile bu araçların kalibre edilmiş çıktılarından eğitilen ML sınıflandırıcılarını aynı etiketli veri matrisi üzerinde, hem ayrıştırma (ROC/AUPRC) hem kalibrasyon (Brier, ECE, slope/intercept, HL) hem de eşik bağımsız ve eşik bağımlı istatistiksel karşılaştırmalar (paired-bootstrap ΔAUC, McNemar) çerçevesinde eşzamanlı olarak değerlendiren kapsamlı bir çalışma bulunmamaktadır. Bu boşluk, otoinflamatuvar gen paneli bağlamında "hangi skorun nerede güçlü olduğu ve hangi ayarlarda güvenilir kabul edilebileceği" sorusunu açık bırakmakta; klinik moleküler tanıda kullanılan hesaplamalı araçların hem ayrıştırma hem de olasılık kalitesi açısından sistematik ve yöntemsel olarak sağlam bir zeminde karşılaştırılmasını zorunlu kılmaktadır.

Bu tez, söz konusu boşluğu hedefleyerek, on iki monogenik otoinflamatuvar hastalığın çekirdek genlerini içeren bir panel üzerinde, üç yaygın varyant etki kestiricisinin (AlphaMissense, PolyPhen-2, SIFT) ve bu araçlardan elde edilen kalibre olasılıkları girdi olarak kullanan makine öğrenmesi sınıflandırıcılarının (Logistic Regression, SVM, Random Forest, XGBoost, MLP) klinik etiketli veri üzerinde yöntemsel olarak karşılaştırılması ve bu karşılaştırmanın hem ayrıştırma hem de olasılık kalitesi hem de klinik karar düzeyi bakımından ayrıntılı biçimde ortaya konması amaçlanmıştır.

Bu genel amaç doğrultusunda, ilk olarak otoinflamatuvar çekirdek gen paneline ait patojenik ve benign olarak etiketlenmiş missense varyantlardan oluşan bir veri kümesi oluşturulmuştur. Bu veri kümesi, literatür ve ClinVar gibi klinik veri tabanları temel alınarak derlenmiş, kalite kontrol ve anotasyon adımlarından geçirildikten sonra analiz için kullanılabilir hale getirilmiştir. Her varyant için AlphaMissense, PolyPhen-2 ve SIFT skorları elde edilmiş; ardından bu skorlar, post-hoc olasılık kalibrasyonu ile ortak bir olasılık ölçeğine taşınmıştır. Böylece, her araç için 0,50 değeri yaklaşık "yüzde elli patojenlik olasılığı" anlamına gelecek şekilde hizalanmış ve araçların ham skor uzayındaki ölçek farklılıkları klinik açıdan yorumlanabilir tek bir ortak dile dönüştürülmüştür.

İkinci olarak, bu kalibre edilmiş olasılıklar üç boyutlu, kompakt bir özellik vektörü şeklinde kullanılarak Logistic Regression, RBF çekirdekli SVM, Random Forest, XGBoost ve MLP sınıflandırıcıları eğitilmiştir. Bu aşamada nested stratified 5-fold cross-validation yapısı benimsenmiş; iç döngüde hiperparametre optimizasyonu, dış döngüde ise seçilen konfigürasyonların out-of-fold tahminleri üretilmiştir. Ön-işleme adımları (özellik ölçekleme, sınıf ağırlıklandırma vb.) çapraz doğrulamadaki her bir fold için yalnızca o fold'un eğitim altkümesinden hesaplanmış; ardından aynı değerler, hiçbir değişiklik yapılmadan ilgili doğrulama altkümesine uygulanmıştır. Böylece doğrulama verisi bu hesaplama sürecine hiçbir aşamada dâhil edilmemiş ve veri sızıntısı olasılığı etkin biçimde ortadan kaldırılmıştır. Hiperparametre seçimi, öncelikle AUROC'u maksimize etmeyi hedefleyen bir kural çerçevesinde yürütülmüştür. Üçüncü olarak, hem VEP'lerin kendileri hem de bu VEP-türevli ML modelleri ayrıştırma performansı (ROC eğrileri, AUROC ve dengesizlik durumunda AUPRC) ile olasılık kalitesi (Brier skoru, Expected Calibration Error, kalibrasyon eğimi ve kesişimi, Hosmer–Lemeshow uyum testi) açısından değerlendirilmiştir. Bu değerlendirme iki düzeyde yapılmıştır. İlk düzeyde, her yöntem kendi model-özel optimum karar eşiğinde (Youden's J'i maksimize eden eşik) incelenmiş; böylece aracın veya modelin kendi doğasında en dengeli hata-fayda dengesine ulaştığı nokta tanımlanmıştır. İkinci düzeyde ise tüm yöntemler, kalibrasyon sonrasında ortak bir karar eşiğine (t=0.50) taşınarak aynı klinik yorum altında karşılaştırılmıştır. Bu yapı, bir yandan her modelin "kişisel optimum" performansını, diğer yandan da klinikte

sık kullanılan sabit bir eşik koşulunda görece üstünlük ve güvenilirliğini görünür kılmayı amaçlamaktadır.

Dördüncü olarak, modeller arasındaki farkların istatistiksel sağlamlığı eşik-bağımsız ve eşik-bağımlı çerçevelerde sınanmıştır. Eşik bağımsız düzlemde, aynı örnekler üzerinde elde edilen AUROC değerleri için eşleştirilmiş bootstrap ΔAUC analizi uygulanmış, fark dağılımlarından güven aralıkları ve p-değerleri hesaplanmış, çoklu karşılaştırmalar Benjamini–Hochberg yanlış keşif oranı kontrolü ile düzeltilmiştir. Eşik-bağımlı düzlemde ise, ortak 0.50 eşiğinde her model çifti için McNemar exact testi kullanılarak karar düzeyi hata simetrisi değerlendirilmiştir. Böylece, benzer AUROC değerlerine sahip iki modelin aynı hastalarda benzer mi yoksa farklı mı hata yaptıkları, sayısal ve istatistiksel olarak ortaya konmuştur.

Son olarak, elde edilen ayrıştırma, kalibrasyon ve karar düzeyi bulguları klinik bağlama tercüme edilerek otoinflamatuvar gen panellerinde pratik olarak uygulanabilir bir kullanım şeması önerilmiştir. Bu şema, özetle, yüksek ayrıştırma gücüne sahip bir skorun (örneğin AlphaMissense veya panelde iyi performans gösteren Logistic Regression modeli) triayj ve önceliklendirme amacıyla kullanılması; nihai kararın ise panel-özgü yeniden kalibrasyon uygulanmış bir ML modelinin olasılık çıktıları üzerinden, laboratuvarın hedeflediği duyarlılık/özgüllük ya da PPV/NPV profiline göre seçilmiş kalibre bir eşikte verilmesi esasına dayanmaktadır.

Bu yönleriyle tez, otoinflamatuvar çekirdek gen panelinde kullanılan VEP'ler ile bu araçlardan beslenen ML sınıflandırıcılarını aynı veri, aynı etiket seti ve aynı metodolojik iskelet altında ayrıştırma, kalibrasyon ve klinik karar boyutlarında birlikte ele alarak, hem literatürdeki boşluğu doldurmayı hem de klinik moleküler tanı pratiği için denetlenebilir, tekrarlanabilir ve veri temelli bir karar destek çerçevesi sunmayı amaçlamaktadır.

2. GENEL BİLGİLER

2.1. Otoinflamatuvar Hastalıklar

Otoinflamatuvar hastalıklar (systemic autoinflammatory diseases; SAID/AID), doğuştan gelen bağışıklık sisteminin tehdit algılama ve yanıt verme mekanizmalarındaki kalıtsal bozukluklar sonucunda, enfeksiyon kanıtı veya belirgin otoantikor/antijen-spesifik T hücre yanıtı olmaksızın gelişen, yineleyen ya da sürekli inflamasyon ataklarıyla seyreden heterojen bir bozukluk ailesini ifade eder. Bu tablolar klasik olarak otoimmün spektrumdan, adaptif yanıttan ziyade innate bağışıklık eksenindeki düzensizlikleriyle ayrılır; klinik olarak tekrarlayan ateş, serözit, deri döküntüleri, artralji/artrit, miyalji ve belirgin akut faz yanıtı ile karakterizedir (An ve ark., 2024; Kastner ve ark., 2010; McDermott ve Aksentijevich, 2002).

Moleküler düzeyde temel problem, doğuştan gelen bağışıklık yanıtının "eşik ayarının" bozulması ve negatif geri-besleme mekanizmalarının yetersiz kalmasıdır. Hücre içi sensörler (örneğin NLR ailesi reseptörleri), adaptor proteinler ve bunların oluşturduğu platformlar (inflamazom kompleksleri gibi) tehlike sinyallerini algılayarak IL-1 ailesi sitokinlerinin olgunlaşmasını, NF-κB ve diğer proinflamatuvar yolakların aktivasyonunu düzenler. Normalde kısa süreli ve kontrollü olması gereken bu süreçte, sensör, adaptör ve efektör ağının herhangi bir bileşeninde ortaya çıkan kazanım veya kayıp fonksiyon tipi mutasyonlar, inflamasyon eşiğini düşürerek yanıtın kolay tetiklenir ve zor söner hale gelmesine yol açar (Krainer ve ark., 2020; Schnappauf ve ark., 2019). Sonuçta IL-1 ekseni başta olmak üzere çeşitli sitokin ağlarında kronik uyarılabilirlik, tekrarlayan ateş atakları, yükselmiş akut faz yanıtı ve çoklu organ tutulumu gibi klinik olarak ortak fenotipler ortaya çıkar (Gutierrez ve Lapidus, 2022; Kastner ve ark., 2010).

Otoinflamatuvar hastalıkların klinik-genetik spektrumu monojenik sendromlardan (tek genli, yüksek etkili patojenik varyantlarla seyreden tablolar) kompleks, çok faktörlü hastalıklara kadar uzanır. Aynı gen içindeki farklı varyantlar birbirinden oldukça farklı fenotipler oluşturabildiği gibi, aynı yolaktaki farklı genlerdeki defektler de kısmen örtüşen klinik tablolar yaratabilir. Ayrıca otoinflamasyon, bazı olgularda otoimmünite veya immün yetmezlik bulguları ile iç içe geçerek "saf" kategoriler yerine süreklilik gösteren bir spektrum oluşturur (Gutierrez ve Lapidus, 2022; Krainer ve ark., 2020; Kim, 2023). Özellikle çocukluk çağı başvurularında tekrarlayan ateş, deri bulguları ve sistemik inflamasyonla gelen hastalarda, klinik değerlendirme ile genetik incelemenin birlikte yürütüldüğü algoritmik yaklaşımlar önerilmektedir (An ve ark., 2024; Meertens ve ark., 2024). Bu geniş spektrum içinde, genetik bulguların doğru ve zamanında yorumlanması klinik yönetim açısından merkezi öneme sahiptir. Hedefe yönelik biyolojik tedaviler (örneğin IL-1, IL-6 veya TNF inhibitörleri) çoğunlukla altta yatan patomekanizmanın aydınlatılmasına dayanır; prognoz öngörüsü, komplikasyon riski değerlendirmesi ve ailelere verilecek genetik danışmanlık da büyük ölçüde bu bilginin kalitesine bağlıdır (Kastner ve ark., 2010; Gutierrez ve Lapidus, 2022). Ancak dizileme temelli panellerle saptanan varyantların önemli bir bölümü, mevcut klinik ve fonksiyonel kanıtların yetersizliği nedeniyle klinik önemi belirsiz varyant (variant of uncertain significance; VUS) olarak sınıflandırılmaktadır. Bu durum, tanısal belirsizliği uzatmakta ve laboratuvarların in-silico patojenite tahminleyicilerine yardımcı kanıt olarak başvurma ihtiyacını artırmaktadır (Ghosh ve ark., 2017; Krainer ve ark., 2020).

ACMG/AMP kılavuzları, in-silico algoritma çıktılarının tek başına belirleyici olmayıp "destekleyici" kanıt düzeyinde kullanılması gerektiğini vurgulamakla birlikte, hangi araçların ne şekilde ve hangi eşiklerde kullanılacağı konusunda ayrıntılı bir standardizasyon sunmamaktadır. Farklı otoinflamatuvar gen panellerinde kullanılan VEP araçlarının sayısı, türü ve raporlama biçimi merkezler arasında değişmekte; bu da VUS yorumunda heterojenliğe ve klinik uygulamada belirsizliğin sürmesine yol açmaktadır (Ghosh ve ark., 2017; Meertens ve ark., 2024). Bu bağlamda, otoinflamatuvar hastalıklarla ilişkili genlerde kullanılan hesaplamalı patojenite tahmin araçlarının ayrıştırma gücü, kalibrasyon özellikleri ve klinik karara katkısının sistematik olarak değerlendirilmesi, hem laboratuvar içi standartların

geliştirilmesi hem de merkezler arası karşılaştırılabilirliğin sağlanması açısından kritik bir ihtiyaç olarak öne çıkmaktadır.

2.2. CLINVAR Klinik Veri Tabanı

ClinVar, insan genetik varyantlarının klinik önemine ilişkin yorumları (interpretation) ve bu yorumları destekleyen kanıtları tek bir açık erişimli arşiv altında birleştiren bir veritabanıdır. Temel amacı, klinik laboratuvarlar, araştırma grupları ve konsorsiyumlar arasında şeffaflık, izlenebilirlik ve uzlaşıyı güçlendirmektir. Farklı kurumların gönderdiği kayıtlar ortak bir arayüzde toplanır; her gönderim için kullanılan sınıflandırma kriterleri ve destekleyici belgeler arşivlenir. Böylece klinik sınıflandırmaların zaman içindeki değişimi izlenebilir hale gelir ve topluluk temelli kalite iyileştirmesi için bir zemin oluşturulur (Landrum ve ark., 2020; NCBI, 2024d).

ClinVar'da gönderimler öncelikle SCV (submitted record) düzeyinde tutulur; aynı varyanta ait birden fazla SCV kaydı, VCV (variant-level aggregate) başlığı altında birleştirilir. Belirli bir varyantın klinik hastalık/fenotip bağlamındaki yorumları ise RCV (variant–condition aggregate) kayıtlarında gruplanır. Bu yapı sayesinde hem "bir varyant genel olarak nasıl sınıflandırılıyor?" sorusu VCV üzerinden, hem de "hangi klinik bağlamda nasıl yorumlanıyor?" sorusu RCV üzerinden izlenebilir. Ayrıca kayıtlara sürüm numarası verilmesi, her güncellemede versiyon numarasının artması ve eski sürümlere erişimin korunması, zaman içindeki değişikliklerin geriye dönük olarak da incelenebilmesine imkân tanır (NCBI, 2024b; NCBI, 2024e).

Arşiv, gönderimlerin kanıt gücü ve uzlaşı düzeyi hakkında kullanıcıya hızlı bir fikir verebilmek için "review status/yıldız" sistemini kullanır. Tek göndericili ve sınıflandırma kriterleri açıkça belirtilmiş kayıtlar "tek yıldız" düzeyinde değerlendirilirken; birden fazla bağımsız göndericinin bulunduğu veya uzman paneli tarafından gözden geçirilen kayıtlar daha yüksek yıldız düzeylerine ulaşır. Böylece, aynı varyant için birbiriyle çelişen yorumlar söz konusu olduğunda, kullanıcının güvenilirlik ve konsensüs düzeyini ayırt edebilmesi kolaylaşır (NCBI, 2024a; NCBI, 2024c).

ClinVar'ın güçlü yönleri arasında açık veri politikası, gönderimlerin kaynak ve kriter düzeyinde izlenebilir olması ve kayıtların düzenli biçimde güncellenmesi öne çıkar. Nucleic Acids Research'te yayımlanan platform makaleleri ve yıllık veritabanı bültenleri, ClinVar'ın kapsamını ve yeni işlevlerini (örneğin, germline–somatik varyant sınıflandırmalarının teknik olarak ayrıştırılması, yardımcı dokümantasyonun güncellenmesi gibi) düzenli olarak rapor eder. Bu sürekli güncelleme döngüsü, klinik varyant yorumlarının topluluk ölçeğinde standardizasyonuna ve uyumlu bir dilde yürütülmesine katkı sağlamaktadır (Landrum ve ark., 2018).

Bununla birlikte ClinVar, ikincil bir kaynak niteliğindedir: kayıtlar, farklı laboratuvarların kullandığı yöntemler, kanıt yorumundaki nüanslar ve gönderim zamanlamasına bağlı olarak değişkenlik gösterebilir. Aynı varyant için birden fazla ve bazen birbiriyle çelişen klinik yorumların bulunması mümkündür; bu tür durumlarda yıldız düzeyi, göndericinin itibarı ve en güncel sürüm bilgisi mutlaka dikkate alınmalıdır. Ayrıca ClinVar'da yer alan klinik önem etiketleri, nihai klinik kararın yerini tutmaz; ACMG/AMP varyant yorumlama rehberleri ve hasta bağlamındaki klinik/biyokimyasal kanıtlarla birlikte değerlendirilmesi gerekir (NCBI, 2025; Richards ve ark., 2015).

Özetle ClinVar, klinik varyant sınıflandırmalarının paylaşıldığı merkezî bir referans arşivi olarak, yorum sürecinin standart bir çerçeve ve kanıt hiyerarşisi içinde yürütülmesini destekler. Varyant düzeyi agregasyon (VCV/RCV), sürümleme ve "review status" gibi özellikleri sayesinde hem izlenebilirlik hem de güncellenebilirlik sağlar ve bu yönüyle klinik ve araştırma amaçlı varyant yorum çalışmalarında temel bir başvuru kaynağı olarak konumlanır (NCBI, 2024e; NCBI, 2024a).

2.3. Patojenite Tahmin Araçları

Genomik varyantların klinik etkisini öngörmek amacıyla geliştirilmiş patojenite tahmin araçları, bir genetik değişikliğin hastalığa yol açıcı (patojenik) mı yoksa zararsız (iyi huylu) mı olabileceğini hesaplamaya çalışır. Özellikle tek bir amino asit değişimine yol açan missense varyantlar üzerinde yoğunlaşan bu yazılımlar, evrimsel korunmuşluk (bir amino asidin farklı türler boyunca ne kadar korunduğu) ve proteinin yapısal ve bağlamsal özellikleri gibi verileri kullanarak ilgili varyantın fonksiyonel etkisini tahmin eder. Bu tür araçlar, geniş ölçekli dizileme çalışmalarıyla

keşfedilen sayısız varyantın klinik yorumlanmasında kritik rol oynar, zira insan genomunda gözlemlenen milyonlarca missense varyantın büyük çoğunluğunun klinik önemi belirsiz olduğu bilinmektedir (Ghos ve ark., 2017; Minton, 2023; Vaser ve ark., 2016).

2.3.1 AlphaMissense

DeepMind (Google) araştırmacıları tarafından geliştirilen AlphaMissense, başlangıçta protein yapı öngörüsü için tasarlanan AlphaFold modelinin uyarlanmasıyla oluşturulmuş, missense varyantların patojenitesini tahmin etmeye odaklı bir yapay zekâ modelidir. Modelin geliştirilmesinde iki aşamalı bir eğitim stratejisi benimsenmiştir. İlk aşamada, AlphaFold'da olduğu gibi yalnızca referans (normal) protein dizileri kullanılarak ilgili proteinlerin üç boyutlu (3B) yapıları öngörülmüş; ikinci aşamada ise insan proteinlerindeki tek amino asit değişikliklerini (missense varyantlar) patojenite açısından sınıflandırmayı öğrenebilmesi için modele ince ayar (fine-tuning) uygulanmıştır (Cheng ve ark., 2023; Jumper ve ark., 2021). Bu süreçte, eğitim verisi olarak insan ve primat popülasyonlarına ait varyant sıklığı veritabanlarından yararlanılmıştır. Belirli bir pozisyondaki varyant, eğer sağlıklı popülasyonda yaygın olarak gözleniyorsa "iyi huylu (benign)", popülasyonda hiç gözlenmiyorsa "potansiyel olarak patojenik" kabul edilerek modele etiket olarak verilmiştir. Bununla birlikte, popülasyonda hiç bildirilmeyen bir varyantın gerçekte zararsız olabileceği olasılığına karşı iki aşamalı stratejiye ek bir filtreleme adımı dâhil edilmiştir. Bu adımda, modelin ilk tahminleri kullanılarak "gözlenmemiş" varyantlar arasından muhtemelen benign olanlar ayıklanmakta; ardından süzülmüş, daha güvenilir etiketlere sahip veri kümesi ile yeniden eğitim yapılarak gürültülü etiketlerin olumsuz etkisi azaltılmaktadır (Cheng ve ark., 2023).

AlphaMissense, eğitimin sonunda elde ettiği patojenite skoru için ek bir kalibrasyon aşamasından geçirilmiştir. Modelin ürettiği ham skorlar, güvenilir eşik değerleri belirleyebilmek amacıyla ClinVar veritabanında uzmanlarca sınıflandırılmış bilinen varyantlar üzerinde ayarlanmış; böylece patojenite skorları yaklaşık %90 kesinlik (precision) hedefiyle raporlanacak şekilde kalibre edilmiştir. Bu sayede AlphaMissense, bir varyantın yapısal bağlamını (proteindeki 3B ortam) ve evrimsel korunmuşluk bilgisini birlikte değerlendirerek, önceki pek çok araca kıyasla üstün

performans sergileyebilen bir sistem hâline gelmiştir (Cheng ve ark., 2023; Minton, 2023). Model, çeşitli genetik ve deneysel test setleri üzerinde güncel en iyi başarımlara yakın veya onları aşan bir performans düzeyine ulaşmış; ayrıca öğrenimini doğrudan ClinVar gibi kuratoryal veriler üzerinde yapmaksızın, genel anlamda yüksek doğruluk oranları elde etmiştir (Cheng ve ark., 2023). Buna ek olarak AlphaMissense, insan proteomundaki olası tüm tek amino asit değişimlerini tarayarak yaklaşık 71 milyon missense varyant için patojenite skoru üretmiştir. Bu kapsamlı tarama sonucunda varyantların yaklaşık %57'si "muhtemelen benign", %32'si "muhtemelen patojenik", geri kalan kısmı ise "belirsiz" kategoride raporlanmıştır (Cheng ve ark., 2023; Google DeepMind, 26.10.2025). Bu özellikleriyle AlphaMissense, AlphaFold'un yapısal öngörü kapasitesini klinik varyant yorumlamaya taşıyan yenilikçi bir araç olarak öne çıkmakta; özellikle nadir hastalık genetiğinde tanısal süreçlere ve yeni araştırma alanlarının açılmasına önemli katkılar sağlaması beklenmektedir (Minton, 2023; Cheng ve ark., 2023).

2.3.2. PolyPhen-2

PolyPhen-2 (Polymorphism Phenotyping v2), insan proteinlerindeki tek aminoasit değişimlerinin (missense varyantlarının) yapısal ve işlevsel etkisini tahmin etmeye yönelik, literatürde en yaygın kullanılan in silico araçlardan biridir. Yöntem, varyantın bulunduğu pozisyonun evrimsel korunmuşluk düzeyini, ilgili aminoasit ikamesinin biyofiziksel ve kimyasal özelliklerini ve uygun bir üç boyutlu (3B) yapı modeli mevcutsa bu ikamenin protein yapısı içindeki konumsal bağlamını bütüncül olarak değerlendirmektedir. Bu çok boyutlu bilgi çerçevesinde, varyantın zararlı olup olmadığı sorusuna olasılıksal bir yanıt üretilmekte ve temel sınıflayıcı olarak seçilmiş dizi ve yapı özniteliklerini birleştiren naif Bayes yaklaşımı kullanılmaktadır. Model, çıktı olarak hem kategorik bir sınıflandırma (benign, possibly damaging, probably damaging) hem de hasar olasılığını yansıtan posterior bir olasılık ölçütü sunmaktadır (Adzhubei ve ark., 2010).

PolyPhen-2'nin bilgi kaynakları ve kullandığı öznitelikler katmanlı bir yapıya sahiptir. Dizi temelli katmanda, konumsal korunmuşluk, aminoasit özellikleri, profil skorları ve motif ya da fonksiyonel alan bağlamı gibi parametreler dikkate alınmaktadır. Yapı temelli katmanda, çözümlenmiş ya da homoloji modellemesi ile

elde edilmiş protein yapılarından türetilen çözünebilirlik ve yüzey erişilebilirliği, ikamenin yakın çevresindeki sterik yerleşim ve olası etkileşimsel ipuçları değerlendirilmektedir. Buna ek olarak, fonksiyonel anotasyon katmanında UniProt ve Pfam gibi veritabanlarından alınan bölgeye özgü işlevsel açıklamalar kullanılmaktadır. Bu farklı kaynaklardan gelen özniteliklerin aynı çatı altında birleştirilmesi, yalnızca evrimsel korunuma ya da yalnızca yapısal bilgiye dayanan yaklaşımlara kıyasla ayrıştırma gücünü artırmakta; kararın belirli eşik değerleri ve yanlış pozitif oranı gibi ölçütler üzerinden kalibre edilmesi ise klinik yoruma uygun, yorumlanabilir bir skor uzayı sağlamaktadır (Adzhubei ve ark., 2010).

Uygulamada PolyPhen-2, yeni nesil dizileme çıktılarındaki missense varyantların hızlı önceliklendirilmesi için yaygın biçimde kullanılmış ve çok merkezli kıyaslama çalışmalarında SIFT, MutPred, MutationTaster, CADD ve benzeri araçlarla birlikte değerlendirildiğinde rekabetçi bir performans sergilemiştir. Bu çalışmalar, veri setinin bileşimine (gen panelleri, hastalık sınıfları, denge/önyargı durumları) duyarlı olmakla birlikte, PolyPhen-2'nin genel doğruluk ve ayırt ediciliğinin yüksek olduğunu; ancak iyi huylu varyantlarda bazı durumlarda gereğinden fazla "zararlı" uyarısı üretebildiğini (eşik ve kalibrasyona duyarlılık) göstermiştir. Dolayısıyla puanların tek başına klinik etiket yerine geçmediği, ACMG/AMP hiyerarşisi içinde destekleyici kanıt olarak ele alınmasının uygun olduğu vurgulanmıştır (Ghosh ve ark., 2017; Grimm ve ark., 2015; Thusberg ve ark., 2011).

Yöntemin ayrıntılı teknik tanımı ve kullanım rehberi, özgün kısa makaleyle birlikte yayımlanan ayrıntılı protokolde sunulmuştur. Bu kaynaklar, kullanılan öznitelik türlerini, sınıflayıcı mimarisini, çıktının nasıl yorumlanacağını (kategori sınırları, posterior olasılık, yanlış-pozitif/yanlış-negatif dengesi) ve yüksek hacimli verilerde toplu çalıştırma seçeneklerini ayrıntılandırır. Klinik genetikte yaygın kabul görmesinin temel nedeni, evrimsel, yapısal ve fonksiyonel ipuçlarını birlikte kullanarak missense etkisini pratik ve yorumlanabilir bir skora çevirmesidir (Adzhubei ve ark., 2013).

2.3.3. SIFT

SIFT (Sorting Intolerant From Tolerant), tek aminoasit ikamelerinin (missense) proteinin işlevini bozma olasılığını, ilgili konumdaki evrimsel korunmuşluk sinyali

üzerinden değerlendiren bir patojenite tahmin aracıdır. Temel ilke, yüksek derecede korunmuş bir pozisyondaki ikamenin daha büyük olasılıkla zararlı olacağı; değişken (tolerant) pozisyonlardaki ikamelerin ise genellikle tolere edileceği varsayımıdır. Bu amaçla SIFT, çoklu dizilimlerden (MSA) türetilen konumsal olasılıklar ve aminoasit ikameleri arasındaki tolerans profilleri ile çalışır. Çıktıyı 0–1 aralığında sürekli bir skor olarak verir ve 0.05 eşik değeri altı "zararlı (deleterious)" sınıfını, 0.05 üstü değerler ise "tolere (tolerated)" sınıfını ifade eder (Ng ve Henikoff, 2001; Ng ve Henikoff, 2003). Yöntem, dizi odaklı bir çerçevede özellik mühendisliği gerektirmeden doğrudan MSA tabanlı olasılıkları kullanır: belirli bir pozisyonda alternatif aminoasidin gözlenme olasılığı düşükse, bu ikame işlevsel olarak tolere edilmez kabul edilir. SIFT'in erken sürümleri araştırmacının kendi çoklu dizilimini üretmesini gerektirirken, güncel sürüm olan SIFT4G, UniRef90/Ensembl gibi büyük referans dizinlerinden ön hesaplanmış skor tabloları sağlayarak genom çapında hızlı yorum yapılmasına olanak tanır (Vaser ve ark., 2016). Bu tasarım, varyant çağrım boru hatlarına entegrasyonu kolaylaştırmış; klinik ve popülasyon çalışmalarında yüksek hacimli değerlendirmeler için pratik bir çözüm sunmuştur. SIFT'in güçlü yönü, evrimsel sinyalin doğrudan ve yorumlanabilir biçimde kullanılmasıdır; hesaplama maliyeti düşüktür, farklı organizmalar ve büyük gen panelleri üzerinde ölçeklenebilir. Bununla birlikte performans, hizalamanın kalitesi ve çeşitliliğine duyarlıdır: yakın akraba dizilerle sınırlı MSA'lar veya hatalı hizalanmış bölgeler, yanıltıcı skorlar üretebilir. Ayrıca yalnızca dizisel korunuma dayalı olduğundan, üç boyutlu yapı/alan bağlamı, bağlanma yüzeyleri veya post-translasyonel modifikasyon gibi bilgiler SIFT skoru içine doğrudan yansımaz; bu nedenle bazı yapısal-bağlamsal etkiler gözden kaçabilir (Thusberg ve ark., 2011). Karar eşik duyarlılığı ve kalibrasyon da pratikte önemlidir; farklı veri kümelerinde optimal eşik küçük sapmalar gösterebilir (Ghosh ve ark, 2017).

Klinik yorum bağlamında SIFT, destekleyici nitelikte hesaplamalı kanıt üretir ve çoğu zaman diğer tahmin araçları ile tamamlayıcı biçimde kullanılır. Özellikle yüksek korunmuşluk sergileyen pozisyonlarda güçlü negatif sinyal sağlarken, zayıf hizalama/düşük çeşitlilik koşullarında güven sınırlıdır. Bu nedenle SIFT skorları, popülasyon sıklığı, fenotip uyumu, fonksiyonel/segregasyon verileri ve diğer in-

silico çıktılarla birlikte, standart sınıflandırma çerçeveleri içinde değerlendirilmelidir (Ghosh ve ark., 2017).

2.4. Metrikler

Tahmin ve sınıflandırma modellerinin bilimsel veya klinik bağlamda kullanılabilmesi için, ürettikleri çıktının sistematik biçimde değerlendirilmesi gerekir. Bu amaçla kullanılan metrikler, modelin kararlarının nicel bir özetini sunarak farklı modellerin, ayarların veya veri kümelerinin karşılaştırılmasını mümkün kılan standart ölçütlerdir. Başka bir ifadeyle metrikler, bir modelin "ne kadar iyi çalıştığını" yalnızca sezgisel izlenimlerle değil, sayısal ve tekrarlanabilir göstergelerle tanımlamayı sağlar (Chen ve ark., 2024; Tharwat, 2021). Genel olarak metrikler, model çıktısının türüne ve hedef kullanım senaryosuna göre farklı ailelere ayrılabilir. İkili sınıflamada en temel yapı taşı, gerçek etiketler ile model tahminlerinin dört hücreli karışıklık matrisi (confusion matrix) çerçevesinde özetlenmesidir. Bu matristen türeyen oranlar, doğru–yanlış sınıflandırmaların dağılımını, hangi hata türünün ne ölçüde baskın olduğunu ve model kararlarının ne kadar dengeli olduğunu ortaya koyar (Tharwat, 2021). Bu gruba giren metrikler genellikle karar eşiğine bağımlıdır, yani modelin "pozitif" olarak etiketlediği sınır değeri değiştikçe, aynı model için farklı değerler üretirler. Klinik kullanımda bir metrik seçilirken, hangi hata türünün (yanlış pozitif mi, yanlış negatif mi) daha maliyetli olduğunun açıkça düşünülmesi bu nedenle kritik önem taşır (Chicco ve Jurman, 2020). İkinci bir metrik ailesi, karar eşiğini tek bir noktaya sabitlemek yerine, eşiğin tüm aralık boyunca "süpürülmesi" ile elde edilen eğri tabanlı özetlerdir. ROC ve precision–recall gibi eğriler, modelin sıralama gücünü ve farklı eşiklerdeki yakalama ve ayıklama dengesini bütüncül biçimde sergiler; bu eğrilerden türetilen alan ölçütleri (örneğin ROC eğrisi altındaki alan) eşiğe görece daha az bağımlı, özet performans göstergeleri sunar (Chen ve ark., 2024; Tharwat, 2021). Özellikle sınıf dengesizliğinin belirgin olduğu durumlarda, eğri tabanlı metrikler tek bir eşikte raporlanan oranlardan daha güvenilir bir genel davranış tasviri sağlayabilir. Olasılık çıktısı üreten tahmin modelleri söz konusu olduğunda, yalnızca "doğru–yanlış" ayrımına odaklanmak yeterli değildir; modelin verdiği olasılık skorlarının gerçek veri ile ne kadar uyumlu olduğunu değerlendiren kalibrasyon metrikleri de

önem kazanır. Kalibrasyon eğrileri, kalibrasyon eğimi ve kesişimi, Brier skoru veya beklenen kalibrasyon hatası (Expected Calibration Error, ECE) gibi ölçütler, modelin olasılıkları sistematik olarak fazla mı yoksa az mı tahmin ettiğini gösterir ve klinik ya da operasyonel karar eşikleri belirlenirken tamamlayıcı bilgi sağlar (Huang ve ark., 2020; Van Calster ve ark., 2019). Örneğin aynı ayrıştırma gücüne sahip iki modelden, olasılıkları gerçek risklerle daha iyi örtüşen modelin klinik karar desteği açısından daha güvenilir olması beklenir. Metriklerin önemli bir bölümü, örneklenen popülasyondaki sınıf prevalansına ve veri kümesinin özelliklerine duyarlıdır. Özellikle nadir olayların veya dengesiz sınıf dağılımlarının söz konusu olduğu senaryolarda, yalnızca "yüzde doğru sınıflama" gibi genel görünümlü ölçütlere dayanmak yanıltıcı sonuçlar doğurabilir (Chicco ve Jurman, 2020; Chen ve ark., 2024). Bu nedenle uygulayıcının, kullanacağı metriklerin hangi varsayımlar altında anlamlı olduğunu, hangi durumlarda yanıltıcı olabileceğini ve hedeflenen klinik/uygulamalı senaryonun hangi hata profiline daha duyarlı olduğunu bilmesi gerekir. Güncel metodoloji yazınında, dengesiz veri ve önyargılı örnekleme durumlarında hata dengesini gözeten ve hem pozitif hem negatif sınıftaki performansı simetrik biçimde özetleyen ölçütlerin tercih edilmesi önerilmektedir (Chicco ve Jurman, 2020).

Sonuç olarak metrikler, tek bir sayıya indirgenmiş "not"lar olmaktan çok, bir tahmin modelinin farklı yönlerini (doğru–yanlış dağılımı, sıralama gücü, olasılık kalitesi, sınıf dengesine duyarlılık vb.) farklı açılardan görünür kılan tamamlayıcı araçlardır. Sağlam bir değerlendirme için, tek bir metrik yerine birden fazla ve birbirini tamamlayan metrik ailesinin birlikte raporlanması; seçilen metriklerin de çalışma tasarımı, veri yapısı ve hedef kullanım bağlamıyla tutarlı gerekçelendirilmesi beklenir (Chen ve ark., 2024; Tharwat, 2021; Van Calster ve ark., 2019).

2.4.1. Accuracy

Accuracy (doğruluk), sınıflandırıcının doğru sınıflandırdığı örneklerin tüm örnekler içindeki oranını ifade eder ve karışıklık matrisine göre

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

“True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)” şeklinde tanımlanır. Değer 0–1 aralığındadır; 1’e yaklaştıkça modelin genel sınıflandırma isabetinin arttığı kabul edilir. Yorumu görece basittir, farklı modellerin veya veri toplama şemalarının sonuçlarının hızlıca karşılaştırılmasına imkân verir ve hata oranının (error rate) tamamlayıcısıdır; yani 1−Accuracy toplam hata payını verir. Klinik ve biyomedikal çalışmalarda, özellikle farklı örneklem yapılarının kıyaslandığı durumlarda doğruluk değerinin güven aralıklarıyla birlikte raporlanması ve örnekleme farklılıklarının göz önünde bulundurulması önerilir (Fawcett, 2006; Chicco ve Jurman, 2020). Buna karşın Accuracy, sınıf dengesizliğinde belirgin biçimde yanıltıcı olabilir. Pozitif sınıfın çok seyrek olduğu bir veri kümesinde tüm örneklere “negatif” diyen basit bir kural, klinik olarak tamamen değersiz olsa bile yüksek doğruluk değeri üretebilir; çünkü bu metrik, gerçek pozitifleri yakalama başarısını (Sensitivity/Recall) ve yanlış pozitif yükünü ayrı bir boyut olarak yansıtmaz (Saito ve Rehmsmeier, 2015; Chicco ve Jurman, 2020). Ayrıca Accuracy, sınıf prevalansındaki değişimlere doğrudan duyarlıdır ve karar eşiğine bağlı olarak anlamlı biçimde değişebildiği için, tek başına klinik karar eşiği seçimini yönlendirmek için yeterli kabul edilmez (Fawcett, 2006). Bu nedenle doğruluk değeri, hata türlerini ayıran oranlarla (örneğin Sensitivity/Recall ve Specificity) ve dengesiz veri kümelerinde daha dengeli özet sağlayan diğer metriklerle birlikte yorumlanmalıdır (Fawcett, 2006; Saito ve Rehmsmeier, 2015; Chicco ve Jurman, 2020).

2.4.2. AUROC

ROC (Receiver Operating Characteristic) eğrisi, bir skorlama/sınıflandırma yönteminin karar eşiği tüm aralık boyunca değiştirildiğinde elde edilen doğru pozitif oranı (True Positive Rate, TPR; Sensitivity/Recall) ile yanlış pozitif oranı (False Positive Rate, FPR = 1 − Specificity) ikililerinin oluşturduğu düzlemsel gösterimdir. Bu eğrinin altında kalan alan olan AUROC (Area Under the ROC Curve), yöntemin pozitif örneklere negatif örneklere kıyasla sistematik olarak daha yüksek skor verme kapasitesini tek bir sayısal özet halinde ifade eder (Fawcett, 2006). Kuramsal olarak AUROC, iki örnek uzayından rastgele seçilen bir pozitif ve bir negatif örneğin skorlarının karşılaştırılması üzerinden

$$AUROC = Pr[s(X^+) > s(X^-)] + \frac{1}{2} Pr[s(X^+) = s(X^-)]$$

şeklinde tanımlanabilir. Burada s(·), modelin verdiği skoru; X^+ ve X^- ise sırasıyla pozitif ve negatif sınıfa ait örnekleri temsil eder. Bu nedenle AUROC, yönteminin "sıralama temelli ayırt ediciliğini" (discrimination) ölçer ve skora uygulanan herhangi bir monotonik dönüşümden bağımsızdır (Hand, 2009). Değer 0–1 aralığındadır; 0.5'e yakın değerler rasgele sınıflandırmaya, 1.0'a yaklaştıkça değerler ise kusura yakın ayrıştırma gücüne işaret eder. Eşik seçimine bağlı tek noktalı metriklerden farklı olarak AUROC, model davranışını tüm olası karar eşikleri boyunca bütüncül biçimde yansıttığı için, özellikle olasılık yerine göreli risk/puan üreten skorlama yöntemlerinin karşılaştırılmasında standart bir referans ölçüt olarak kabul edilmektedir (Fawcett, 2006; Hand, 2009).

Bununla birlikte AUROC, yalnızca ayrıştırma boyutunu karakterize eder; modelin verdiği olasılıkların gerçek gerçekleşme olasılıklarıyla ne ölçüde örtüştüğünü yansıtan kalibrasyon hakkında doğrudan bir bilgi sağlamaz. Ayrıca pozitif sınıfın çok seyrek olduğu (dengesiz sınıf dağılımlı) veri kümelerinde, belirli bir FPR artışının pratikte yaratacağı yanlış alarm yükü dikkate alındığında, AUROC kimi zaman görece iyimser bir tablo çizebilir. Bu tip senaryolarda, özellikle nadir pozitiflerin yakalanmasının kritik olduğu durumlarda, duyarlılık ile kesinliği birlikte izleyen Precision–Recall uzayının daha ayırt edici ve bilgilendirici olabildiği gösterilmiştir (Saito ve Rehmsmeier, 2015). Dolayısıyla AUROC, yöntem karşılaştırmalarında güçlü ve yerleşik bir ölçüt olmakla birlikte, klinik/operasyonel bağlamda hata maliyetleri, sınıf dengesizliği ve olasılık kalitesi de dikkate alınarak, uygun tamamlayıcı metriklerle birlikte yorumlanmalıdır (Fawcett, 2006; Hand, 2009; Saito ve Rehmsmeier, 2015).

2.4.3. F1 Skoru

F1 skoru, Precision (PPV) ile Recall (Sensitivity, TPR)'ın harmonik ortalamasıdır ve pozitif sınıf için "yanlış alarm" (false positive, FP) ile "kaçırma" (false negative, FN) hatalarını eşit ağırlıkla dengeleyen tek sayısal bir özet sunar (Sokolova ve Lapalme, 2009).

$$F1 = \frac{2\text{Precision}\,.\,\text{Recall}}{\text{Precision} + \text{Recall}} \quad \left(\text{Precision} = \frac{TP}{TP + FP}\,, \text{Recall} = \frac{TP}{TP + FN}\right)$$

Değer aralığı 0–1 arasındadır; 1, hem Precision hem de Recall'un yüksek ve birbirine yakın olduğu durumu, 0 ise bu iki bileşenden en az birinin 0 olduğu durumu ifade eder. Harmonik ortalama yapısı gereği, bileşenlerden biri belirgin biçimde düşükse F1 skoru da keskin biçimde düşer ve performans büyük ölçüde "zayıf halka" tarafından belirlenir (Sokolova ve Lapalme, 2009). Bu özellik, pozitif sınıf açısından hem yanlış alarm (FP) hem de kaçırma (FN) hatalarını aynı derecede önemli gören senaryolarda F1 skorunu özellikle kullanışlı kılar.

F1 skoru, TN hücresine duyarlı değildir; yani negatif sınıftaki doğru sınıflamaların sayısındaki geniş dalgalanmalar F1 değerine yansımaz. Bu nedenle, pozitif sınıf odaklı izlemin ön planda olduğu tıp ve biyoenformatik uygulamalarında (örneğin patojenik varyant önceliklendirme çalışmalarında) yaygın biçimde kullanılmaktadır (Sokolova ve Lapalme, 2009). Literatürde F1'in, daha genel Fβ ailesinin β=1 için elde edilen özel bir durumu olduğu; β>1 değerleri kullanıldığında Recall'a, β<1 değerlerinde ise Precision'a görece daha fazla ağırlık verilebileceği belirtilmektedir (Sokolova ve Lapalme, 2009).

F1, eşik-ilişkili bir metriktir; yani olasılık veya skor çıktılarında pozitif sınıf için seçilen karar eşiğine doğrudan bağlıdır. Bu nedenle raporlanırken, kullanılan karar eşiği ve veri kümesindeki sınıf dengesi mutlaka belirtilmelidir (Sokolova ve Lapalme, 2009). Veri dengesizliği arttıkça, TN sayısındaki geniş değişim F1 skoruna yansımadığı için, genel bir "doğruluk" hissi vermek amacıyla tek başına kullanılmasının yanıltıcı olabileceği vurgulanmıştır. Karşılaştırmalı çalışmalar, F1'in pragmatik ve kolay yorumlanabilir bir özet sunduğunu; ancak konfizyon matrisinin tüm hücrelerini simetrik olarak kullanan Matthews korelasyon katsayısı (MCC) gibi metriklerin, özellikle dengesiz sınıf dağılımlarında daha sağlam ve dengeli bir performans tablosu sağlayabildiğini göstermektedir (Chicco ve Jurman, 2020).

Öte yandan Precision–Recall (PR) uzayının, özellikle pozitif sınıfın nadir olduğu dengesiz veri durumlarında ROC uzayına göre daha yüksek bilgi içeriği sunduğu bilinmektedir. Bu çerçevede F1, PR-uzayındaki tek-nokta özetlerinden biri olarak görülebilir ve PR eğrisi altındaki alan (AUPRC) gibi eğri-tabanlı özetlerle birlikte

yorumlandığında daha anlamlı ve bütüncül bir değerlendirme sağlar (Saito ve Rehmsmeier, 2015)

2.4.4. Matthews Correlation Coefficient (MCC)

Matthews Correlation Coefficient (MCC), ikili sınıflandırma problemlerinde karışıklık matrisi (confusion matrix) içindeki tüm hücreleri (TP, TN, FP, FN) simetrik biçimde kullanan, [−1, 1] aralığında değer alan korelasyon temelli bir performans ölçütüdür. MCC, gerçek sınıf etiketleri ile modelin ürettiği ikili tahminler arasındaki ilişkiyi özetleyen bir katsayı olup, tanımı aşağıdaki eşitlikte verilmiştir.

$$\mathrm{MCC} = \frac{\mathrm{TP}.\mathrm{TN} - \mathrm{FP}.\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Bu aralıkta 1, gerçek etiketlerle tahminler arasında kusursuz uyumu; 0, rastlantısal performansı; −1 ise sistematik biçimde ters sınıflandırmayı ifade eder. MCC, biçimsel olarak gerçek ve tahmin etiketleri arasındaki Pearson korelasyon katsayısına eşdeğer olduğundan, yalnızca doğru sınıfları değil, yanlış alarm (FP) ve kaçırma (FN) örüntülerini de birlikte hesaba katar. Ölçüt ilk olarak Matthews tarafından protein ikincil yapı kestirimlerinin değerlendirilmesinde kullanılmış ve sonraki yıllarda sınıflandırma literatüründe standart ölçütlerden biri hâline gelmiştir (Chicco ve Jurman, 2020).

MCC'nin önemli bir avantajı, sınıf dengesizliği koşullarında dahi bilgi verici kalabilmesidir. Accuracy veya F1 gibi ölçütler sınıf yaygınlığına daha duyarlı olup iyimser görünebilirken, MCC tüm hücreleri birlikte değerlendirdiği için dengesiz dağılımlarda daha sağlam bir özet sunar. Örneğin pozitif sınıfın seyrek olduğu durumlarda Accuracy yüksek çıkabilse bile model pozitifleri yakalayamayabilir; MCC ise FP ve FN'yi de içerdiğinden bu tabloyu daha gerçekçi yansıtır (Chicco ve Jurman, 2020). Bu yönü, klinik genetik ve moleküler tanı gibi pozitif sınıfın nadir olduğu alanlarda MCC'yi özellikle değerli kılar (Chicco ve Jurman, 2020).

Bununla birlikte MCC eşik bağımlı (threshold-dependent) bir ölçüttür; olasılık ya da sürekli skor üreten modellerde ikili karar için seçilen eşik MCC değerini doğrudan etkiler. Bu nedenle MCC raporlanırken kullanılan eşik değerin belirtilmesi uygundur.

Ayrıca paydada yer alan çarpanlardan herhangi birinin sıfıra düştüğü uç durumlarda (örneğin modelin yalnızca tek sınıfı tahmin ettiği eşikler) ifade matematiksel olarak tanımsız hâle gelebilir; pratikte MCC 0 olarak raporlanabilir veya değerlendirme daha anlamlı bir eşik düzeyine taşınabilir (Chicco ve ark., 2021).

Genel olarak MCC, karışıklık matrisinin tüm bileşenlerini dengeli biçimde kullanan ve dengesiz veri setlerinde görece daha dayanıklı bir özet sunan ölçüttür; bu nedenle Accuracy ve F1 gibi ölçütlere tamamlayıcı, bazı durumlarda ise daha güvenilir bir alternatif olarak önerilmektedir (Chicco ve Jurman, 2020; Chicco ve ark., 2021).

2.4.5. Specificity (Özgüllük)

Specificity, testin veya modelin gerçekte negatif olan örnekleri negatif olarak tanımlama başarısını ölçen orandır ve genellikle "özgüllük" ya da "true negative rate (TNR)" olarak adlandırılır. İkili sınıflama bağlamında:

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

şeklinde tanımlanır. Yüksek Specificity, modelin sağlıklı veya "non-event" grubunda yanlış alarm (FP) üretme olasılığının düşük olduğu anlamına gelir. Klinik bağlamda, özellikle yanlış pozitif sonuçların gereksiz ileri tetkik, maliyet veya anksiyete yaratabildiği durumlarda yüksek Specificity tercih edilir. Bununla birlikte Specificity, yalnız başına model performansının genel resmini vermez; Sensitivity (Recall), PPV ve diğer ölçütlerle birlikte yorumlanması beklenir (Sokolova ve Lapalme, 2009; Chicco ve Jurman, 2020).

2.4.6. Recall

Recall, Sensitivity ya da true positive rate (TPR) olarak da adlandırılır ve gerçekte pozitif olan örnekler içinde modelin pozitif olarak yakalayabildiği kısmı ifade eder:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Burada TP, doğru pozitif; FN ise yanlış negatif sayısıdır. Yüksek Recall, modelin pozitif vakaları kaçırma olasılığının (FN) düşük, yani "duyarlılığının" yüksek olduğunu gösterir. Özellikle pozitif sınıfın klinik açıdan kritik olduğu, bir vakayı kaçırmanın ağır sonuçlara yol açtığı durumlarda (örneğin ciddi hastalık taramaları,

patojenik varyant saptama) Recall öncelikli bir performans ölçütü hâline gelir. Ancak Recall, yanlış alarm (FP) oranını dikkate almadığı için, Precision ve diğer metriklerle birlikte değerlendirilmelidir (Sokolova ve Lapalme, 2009; Chicco ve Jurman, 2020).

2.4.7. Precision

Precision, klinik literatürde sık kullanılan adıyla Positive Predictive Value (PPV), modelin pozitif olarak etiketlediği örnekler içinde gerçekte pozitif olanların oranını ifade eder. İkili sınıflandırmada

$$\text{Precision} = \frac{TP}{TP + FP}$$

şeklinde tanımlanır. Bu oran, sezgisel olarak "Model 'pozitif' dediğinde ne kadar haklı?" sorusuna yanıt verir. Precision'ın yüksek olması, modelin pozitif sınıf kararı verdiği örneklerin büyük kısmının gerçekten pozitif olduğu, yani yanlış pozitif (FP) sayısının görece düşük olduğu anlamına gelir. Bu nedenle özellikle pozitif sonucun invazif, maliyetli veya geri dönüşü zor bir müdahale ile sonuçlandığı klinik senaryolarda (örneğin ileri görüntüleme, cerrahi girişim, toksik kemoterapi vb.) PPV/Precision, karar desteği açısından merkezi öneme sahiptir. Precision, değerlendirme yapılan veri kümesindeki pozitif sınıf prevalansına ve FP sayısına duyarlıdır; bu nedenle tek başına kullanıldığında model performansının sınıflar arası dengesini tam olarak yansıtmaz. Pozitifleri ne kadar doğru işaret ettiğini gösterirken, pozitifleri ne kadar eksiksiz yakaladığını (Sensitivity/Recall) ya da negatifleri ne kadar doğru dışladığını (Specificity) söylemez. Bu yüzden, klinik veya biyomedikal sınıflandırma problemlerinde Precision'ın, özellikle Recall/Sensitivity, Specificity ve gerektiğinde bu ikisini birleştiren F1 skoru gibi ölçülerle birlikte yorumlanması, hata deseninin daha dengeli ve güvenilir biçimde değerlendirilmesini sağlar (Sokolova ve Lapalme, 2009; Chicco ve Jurman, 2020).

2.4.8. AUPRC

Precision–Recall (PR) eğrisi, ikili bir sınıflayıcının farklı karar eşiği değerleri boyunca Precision ile Recall arasındaki değişimi gösteren eğridir. Olasılık veya skor üreten bir modelde karar eşiği 0 ile 1 arasında süpürüldüğünde, her eşikte elde edilen Recall, Precision ikilileri PR-uzayında bir eğri oluşturur. Bu eğrinin altında kalan

alan, Area Under the Precision–Recall Curve (AUPRC) olarak adlandırılır ve modelin, özellikle pozitif sınıfı önceliklendirmedeki genel ayrıştırma gücünü özetler (Saito ve Rehmsmeier, 2015).

AUPRC, özellikle sınıf dengesizliğinin belirgin olduğu veri kümelerinde ROC eğrisi altındaki alandan (AUROC) daha bilgilendirici kabul edilmektedir. ROC eğrisi, true positive rate ile false positive rate'i birlikte değerlendirse de, negatif sınıfın aşırı baskın olduğu durumlarda yüksek AUROC değerleri, pozitif sınıfın pratikte zayıf yakalanmasına rağmen "iyimser" bir izlenim verebilir. Buna karşılık PR-eğrisi doğrudan pozitif sınıf performansına (Recall) ve pozitif tahminlerin doğruluğuna (Precision) odaklandığından, nadir olayların ve düşük prevalanslı pozitif sınıfların söz konusu olduğu senaryolarda daha duyarlı ve açıklayıcı bir özet sunar (Saito ve Rehmsmeier, 2015; Chicco ve Jurman, 2020).

2.4.9. Brier Skoru

Brier skoru, olasılık çıktısı üreten tahmin modellerinin performansını değerlendirmek için kullanılan, ortalama karesel hataya (mean squared error, MSE) dayalı bir ölçüttür. İkili bir sonlanım için, her biri için tahmin edilen olasılık değeri $p\hat{}i$ ve gerçek sonuç $y_i \in \{0,1\}$ olmak üzere, N gözlem içeren bir veri kümesinde Brier skoru

$$\text{Brier skoru} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$$

şeklinde tanımlanır. Bu ifade, modelin verdiği olasılık tahminleri ile gerçekleşen sonuçlar arasındaki ortalama karesel farkı gösterir. Değer 0 ile 1 arasında yer alır; 0'a yaklaştıkça tahmin edilen olasılıkların gözlenen sonuçlarla daha iyi örtüştüğü, 1'e yaklaştıkça ise tahminlerin sistematik olarak daha hatalı olduğu kabul edilir. Brier skoru bu yönüyle, hem modelin ayrıştırma (yüksek ve düşük riskli bireyleri birbirinden ayırabilme) kapasitesini hem de kalibrasyonunu (bildirilen olasılıkların gerçek gerçekleşme oranlarıyla uyumu) birlikte yansıtan bileşik bir ölçüt olarak değerlendirilir (Huang ve ark., 2020).

İstatistiksel literatürde Brier skorunun, "kalibrasyon" ve "refinement/keskinlik" bileşenlerine ayrıştırılabildiği; böylece modelin olasılık tahminlerinin hangi boyutta

(örneğin, genel kalibrasyon hatası mı yoksa risk grupları arasındaki ayrışma gücü mü) iyileştirilmeye ihtiyaç duyduğuna dair daha ayrıntılı geri bildirim sağlayabildiği gösterilmiştir. Klinik öngörü ve risk tahmin modellerinde Brier skorunun, AUROC ve kalibrasyon grafikleri ile birlikte raporlanması, özellikle olasılık temelli karar destek sistemlerinde tahmin kalitesinin nicel olarak değerlendirilmesi açısından önemli bir tamamlayıcı rol üstlenir (Huang ve ark., 2020).

2.4.10. Expected Calibration Error (ECE)

Expected Calibration Error (ECE), olasılık çıktısı üreten modellerde kalibrasyon hatasını tek bir değerde özetleyen yaygın bir ölçüttür. Temel yaklaşım, tahmin edilen olasılık aralığını 0–1 arasında bantlara ayırmak; her bantta modelin ortalama tahmin olasılığı ile gözlenen pozitiflik oranını hesaplamak ve bu iki değer arasındaki mutlak farkları bantların örnek sayıları ile ağırlıklandırarak toplamaktır. Böylece ECE, model olasılıklarının gerçek gerçekleşme oranlarına kıyasla sistematik olarak yüksek mi yoksa düşük mü kaldığını nicel biçimde ifade eder. Sonuç 0 ile 1 arasında değişir; 0'a yakın değerler iyi kalibrasyona, daha yüksek değerler ise belirgin kalibrasyon hatalarına işaret eder.

ECE, özellikle modern makine öğrenmesi ve derin öğrenme sınıflayıcılarında, ayrıştırma gücü yüksek modellerin dahi olasılık çıktılarında "aşırı kendinden emin" (overconfident) olabildiğini göstermek için sıklıkla kullanılmaktadır. Klinik öngörü ve risk tahmininde ECE'nin Brier skoru ve kalibrasyon eğrileri ile birlikte yorumlanması, modelin hangi olasılık aralıklarında tutarlı tahmin ürettiğini ve hangi aralıklarda sistematik sapma gösterdiğini daha görünür kılar; bu nedenle ayırt ediciliği (örneğin AUROC) tamamlayan bir olasılık güvenilirliği ölçütü olarak önerilmektedir (Huang ve ark., 2020).

2.5. Makine Öğrenmesi Algoritmaları

Makine öğrenmesi, bilgisayarların veriyle etkileşimleri sonucunda belirli görevlerdeki performanslarını zaman içinde iyileştirmesini amaçlayan; kuramsal temelleri istatistik, bilgisayar bilimi ve optimizasyon kuramının kesişiminde yer alan bir yöntemler bütünüdür (Jordan ve Mitchell, 2015). Bu alanda geliştirilen modeller, veriden örüntüler ve ilişkiler çıkararak genelleme yapar; temel hedef, yalnızca

eğitimde görülen örneklerde değil, daha önce karşılaşılmamış yeni örneklerde de tahmin hatasını en aza indirmektir. Modern makine öğrenmesinin yükselişi, bir yandan büyük veri hacimlerinin ve hesaplama gücünün artışı, diğer yandan kuramsal ve yöntemsel ilerlemelerle birlikte ivme kazanmıştır (Jordan ve Mitchell, 2015). Öğrenme süreci, veriden uygun bir temsil üretme, parametrik ya da parametrik olmayan karar yüzeyleri tanımlama ve bu karar yapısını bir kayıp fonksiyonuna göre optimize etme adımlarının bileşiminden oluşur.

Makine öğrenmesi uygulamada farklı problem türlerini hedefler. Sınıflandırma ve regresyon görevlerinde amaç, sırasıyla ayrık etiketlerin veya sürekli niceliklerin olabildiğince doğru tahmin edilmesidir. Kümeleme ve boyut indirgeme gibi denetimsiz görevlerde ise, verinin iç yapısını daha anlaşılır hale getiren temsillerin elde edilmesi ve doğal alt grupların ortaya çıkarılması hedeflenir. Başarılı genelleme için iki temel ilke öne çıkar: modelin ifade/temsil gücü ile düzenleme (regularization) arasındaki denge ve veri miktarı ile model karmaşıklığı arasındaki uyum. Aşırı esnek modeller eğitim verisine çok iyi uyum sağlayabilse de, genelleme hatasını artırabilir; düzenleme stratejileri ise parametreleri ve karar sınırlarını kısıtlayarak aşırı uyum riskini azaltmayı amaçlar (Jordan ve Mitchell, 2015). Yüksek boyutlu verilerde, özellikle derin öğrenme (deep learning) yaklaşımları ham girdilerden çok katmanlı ve giderek soyutlaşan temsil düzeyleri öğrenerek hem doğruluk hem de farklı veri kümelerine aktarılabilirlik açısından önemli avantajlar sağlayabilmektedir (LeCun ve ark., 2015).

Model kurma süreci genel olarak veri hazırlama, özellik temsili, öğrenme ve değerlendirme basamaklarından oluşur. Veri hazırlama; eksik ve hatalı kayıtların düzeltilmesi, değişkenlerin uygun ölçeklere getirilmesi, uç değerlerin yönetilmesi ve sınıf dağılımlarının gözetilmesi gibi işlemleri içerir. Öğrenme aşamasında parametrelerin güncellenmesi, çoğu yöntemde diferansiyellenebilir bir kayıp fonksiyonunun sayısal olarak minimize edilmesiyle, olasılıksal çerçevelerde ise model parametreleri için tanımlanan olasılık fonksiyonunu maksimize eden değerlerin bulunmasıyla gerçekleştirilir. Değerlendirme aşaması yalnızca tek bir doğruluk ölçüsüne indirgenmemeli; görev türü, sınıf dengesi ve hedeflenen klinik/uygulamalı bağlama uygun metriklerin birlikte ele alınması gerekmektedir (Jordan ve Mitchell, 2015). Özellikle dengesiz veri kümelerinde alıcı–işletim

karakteristiği (ROC) eğrileri sezgisel olarak "iyi" görünen ancak pozitif sınıfı zayıf yakalayan modeller için yanıltıcı olabilir; aynı koşullarda, kesinlik–duyarlılık (precision–recall) eğrileri olumlu sınıfın öngörülebilirliğine daha doğrudan odaklandığından daha bilgilendirici kabul edilmektedir (Saito ve Rehmsmeier, 2015). Sağlam bir değerlendirme için veri sızıntısını engelleyen, sınıf oranlarını koruyan ve hiperparametre seçimini eğitim verisi dışında kalan bir doğrulama kümesiyle yapan bölme stratejilerinin (örneğin uygun çapraz doğrulama şemalarının) kullanılması, nitelikli öngörü yeteneği açısından temel ilkeler arasında yer alır (Jordan ve Mitchell, 2015; Chicco ve Jurman, 2020).

Makine öğrenmesi, biyomedikal araştırma ve klinik uygulamalarda geniş bir yelpazede kullanılmaktadır. Klinik karar destek sistemlerinden tıbbi görüntü analizine, genomik varyantların önceliklendirilmesinden bireyselleştirilmiş risk öngörüsü modellerine kadar pek çok alanda, insan uzmanlığı ile algoritmik kestirimlerin bir araya getirilmesinin iş akışlarını hızlandırma, hata oranlarını azaltma ve ölçeklenebilirliği artırma potansiyeline sahip olduğu bildirilmektedir (Topol, 2019). Ancak bu potansiyelin güvenilir biçimde kullanılabilmesi, veri kaynaklarındaki önyargıların tanımlanmasını ve yönetilmesini, modellerin açıklanabilirlik düzeyinin gözetilmesini, kişisel verilerin gizliliğine ilişkin gerekliliklerin karşılanmasını ve geliştirilen modellerin bağımsız dış kohortlarda doğrulanmasını zorunlu kılar (Topol, 2019). Güncel yaklaşım, model başarımının yalnızca tek bir metrikle özetlenmesi yerine, sınıf dengesine ve kullanım bağlamına duyarlı birden çok ölçütün birlikte raporlanmasını, bu ölçütler arasındaki olası gerilimlerin tartışılmasını ve mümkün olduğunda bağımsız veri kümelerinde tekrarlanabilirliğin gösterilmesini önermektedir (Saito ve Rehmsmeier, 2015; Chicco ve Jurman, 2020).

Özetle makine öğrenmesi, veriden genelleme yapabilen hesaplamalı modellerin kuramsal ve uygulamalı çerçevesini sağlar. Başarı ölçütü, yalnızca eğitim verisine uyum derecesiyle değil, yeni örneklerde güvenilir performans sergileyebilme ve bu performansı hedeflenen klinik ya da biyolojik bağlamla tutarlı ölçütlerle nesnel olarak değerlendirebilme kapasitesiyle belirlenir. Yöntemsel özen; temsil gücünü regularizasyon ilkeleriyle dengeleme, veri dengesizliğine duyarlı metrikler kullanma ve bağımsız doğrulama pratiklerini yerleştirme üzerinden sağlanmaktadır (Jordan ve

Mitchell, 2015; LeCun ve ark., 2015; Saito ve Rehmsmeier, 2015; Chicco ve Jurman, 2020; Topol, 2019).

2.5.1. Random Forest

Random Forest, karar ağaçlarını temel alan ve çok sayıda ağacı bir araya getirerek oy çokluğuyla karar veren gözetimli bir topluluk (ensemble) yöntemidir. Yöntemde iki temel rastgelelik kaynağı bulunur: (i) her ağacın eğitim verisinden bootstrap ile seçilmiş rastgele bir alt örneklem üzerinde kurulması ve (ii) her düğümde bölme kararı verilirken tüm değişkenler yerine rastgele seçilmiş sınırlı sayıdaki özellik arasından en iyi bölmenin aranmasıdır. Örnekler ve özellikler düzeyindeki bu çift rastgelelik, ağaçlar arasındaki benzerliği azaltarak topluluğun varyansını düşürür ve tekil karar ağaçlarına kıyasla genelleme performansını artırır. Bu çeşitlilik mekanizmasının kuramsal temeli ve ampirik sonuçları literatürde ayrıntılı biçimde ele alınmıştır (Biau ve Scornet, 2016; Scornet ve ark., 2015).

Random Forest'in pratik güçlü yanları birkaç noktada özetlenebilir. Öncelikle doğrusal olmayan ve etkileşim içeren karar sınırlarını doğal biçimde temsil edebilir; karmaşık değişken ilişkileri ağaç dallanmaları içinde yerel kararlar olarak modellenir. Ayrıca katı bir ölçekleme zorunluluğu olmadan farklı tipte değişkenler görece sınırlı ön işleme ile kullanılabilir. Her ağaç bootstrap örneklemesiyle eğitildiğinden, bazı gözlemler ilgili ağacın eğitim setinin dışında kalır; bu dışarıda bırakılan örnekler (out-of-bag, OOB) üzerinden hesaplanan hata, modelin genelleme performansına ilişkin içsel bir kestirim sağlar. Aynı çerçevede, permütasyon temelli değişken önem ölçüleri de elde edilebilir.

Modelin ayarlanması gereken hiperparametre sayısı görece sınırlıdır. Özellikle ağaç sayısı ve her düğümde değerlendirilen özellik sayısı uygun seçildiğinde, modelin kararlılığı ve başarımı güvenilir biçimde iyileştirilebilir; ağaç sayısı arttıkça test hatasının çoğu uygulamada belirli bir düzeyde plato yaptığı ve tek, derin karar ağaçlarına göre aşırı uyum riskinin daha düşük olduğu gösterilmiştir (Biau ve Scornet, 2016; Probst ve ark., 2019; Wager ve Athey, 2018). Bununla birlikte sınıf dengesizliği belirgin veri kümelerinde model çoğunluk sınıfa eğilim gösterebilir; bu durumda yeniden örnekleme stratejileri (örneğin SMOTE ve türevleri), karar eşiğinin dikkatli ayarlanması ve dengesizliğe duyarlı performans metriklerinin birlikte

kullanılması önerilmektedir (Fernández ve ark., 2018). Öte yandan, safsızlık (impurity; çoğunlukla Gini) temelli önem ölçütlerinin belirli değişken türleri lehine yanlı olabildiği bildirildiğinden, değişken öneminin permütasyon önemi ve özellikle koşullu permütasyon önemi (conditional permutation importance, CPI) ile desteklenmesi daha güvenilir kabul edilmektedir (Nembrini ve ark., 2018; Debeer ve Strobl, 2020).

Random Forest tek bir karar ağacına göre daha az şeffaf kabul edilse de, model davranışını anlamak için permütasyon temelli önemler, kısmi bağımlılık grafikleri (partial dependence), bireysel koşullu beklenti eğrileri (ICE) ve örnek düzeyinde açıklayıcı yöntemler bir arada kullanılabilir (Molnar, 2022). Bu özellikleriyle Random Forest, uygun ön işleme, dengeli değerlendirme metrikleri ve dikkatli yorumlama stratejileriyle birleştirildiğinde, biyomedikal sınıflandırma ve risk öngörüsü çalışmalarında sık kullanılan temel yöntemlerden biri hâline gelmiştir.

2.5.2. Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP), giriş ve çıkış katmanları arasında en az bir gizli katman bulunan, tamamen bağlantılı (fully connected) beslemeli yapay sinir ağı ailesidir. Her nöronda doğrusal ağırlıklı toplam, doğrusal olmayan bir aktivasyon fonksiyonundan geçirilir; bu yapı doğrusal modellerin yakalayamadığı karmaşık karar sınırlarının öğrenilmesini mümkün kılar (Schmidhuber, 2015; LeCun ve ark., 2015). Kuramsal olarak, yeterli büyüklükte gizli katmanlar ve doğrusal olmayan aktivasyonlar kullanıldığında MLP'lerin geniş bir fonksiyon sınıfını istenen doğrulukta yaklaştırabildiği (universal approximation) gösterilmiştir; ReLU ailesi için bu sonuç modern derin öğrenme bağlamında ayrıntılı biçimde ele alınmaktadır (Yarotsky, 2017). Öğrenme, seçilen kayıp fonksiyonuna göre parametrelerin gradyan inişi ve geri yayılım (backpropagation) ile güncellenmesine dayanır; pratikte bu süreç uyarlamalı optimizasyon yöntemleri ve eğitim protokolleriyle desteklenir (Takeshima, 2022; Nag ve ark., 2022).

Eğitimin kararlı ilerleyebilmesi için başlatma, normalizasyon ve optimizasyon adımlarının uygun seçimi önemlidir. ReLU ile kullanılan He initialization, sinyalin katmanlar arasında sönümlenme ya da taşma yaşamadan taşınmasına katkı sağlayan yaygın bir yaklaşımdır (He ve ark., 2015). Batch Normalization, aktivasyon

dağılımlarını yeniden ölçekleyerek gradyan akışını iyileştirir ve daha yüksek öğrenme oranlarıyla hızlı yakınsamayı kolaylaştırır (Ioffe ve Szegedy, 2015). Adam/AdamW gibi uyarlamalı optimizasyon algoritmaları geniş hiperparametre aralıklarında kararlı eğitim sağlayabilir; ağırlık çürümesinin (weight decay) optimizasyon dinamiğinden ayrıştırılması regularizasyonun daha net denetlenmesine yardımcı olur (Takeshima, 2022; Nag ve ark., 2022).

Aşırı uyumu (overfitting) sınırlamak ve genelleme kapasitesini artırmak için farklı düzenleme teknikleri kullanılabilir. Erken durdurma (early stopping), doğrulama hatasına göre eğitimi uygun bir noktada sonlandıran pratik bir yaklaşımdır (Schmidhuber, 2015). Dropout, eğitim sırasında bazı nöronları geçici olarak devre dışı bırakarak belirli nöron kombinasyonlarına aşırı bağımlılığı azaltır ve genelleme performansını iyileştirir (Gal ve Ghahramani, 2016). Label smoothing ise hedef dağılımı yumuşatarak çıktı olasılıklarının gereksiz yere keskinleşmesini engelleyebilir (Müller ve ark., 2019). Bununla birlikte MLP'lerin olasılık çıktıları zaman zaman "fazla kendinden emin" (overconfident) olabilir; temperature scaling gibi post-hoc yaklaşımlar ECE ve Brier skorunda iyileşme sağlayarak olasılık kalitesini artırabilir (Nag ve ark., 2022).

MLP'nin tablo verilerindeki göreli başarımı, veri tipine ve temsil biçimine güçlü biçimde bağlıdır. Heterojen özelliklerin bulunduğu, etkileşim paternlerinin seyrek olduğu klasik tablo verilerinde gradient-boosting ağaçları pek çok senaryoda güçlü bir karşılaştırma standardı olmaya devam etmekte ve kimi durumlarda derin MLP'lerden daha iyi sonuçlar verebilmektedir (Grinsztajn ve ark., 2022). Bunun temel nedenlerinden biri, ağaç tabanlı yöntemlerin eşik davranışlarını ve doğrusal olmayan etkileşimleri doğrudan yakalayabilmesine karşın, MLP'lerin benzer etkileşimleri öğrenebilmesi için daha dikkatli ölçekleme, mimari tasarım ve düzenleme gerektirmesidir (Borisov ve ark., 2024). Buna rağmen uygun ön işleme, hiperparametre araması, regularizasyon ve kalibrasyon stratejileri uygulandığında MLP'lerin tablo verilerinde de rekabetçi olabildiği; ayrıca TabNet gibi tablo verisine özgü dikkat ve "sparse" özellik seçimi kullanan mimarilerin bazı veri kümelerinde klasik MLP'lere üstünlük sağlayabildiği bildirilmektedir (Arik ve Pfister, 2021; Borisov ve ark., 2024).

2.5.3. Extreme Gradient Boosting (XGBoost)

XGBoost, gradient boosting yaklaşımını karar ağaçları üzerinde uygulayan; her öğrenme döngüsünde modele yeni bir karar ağacı ekleyerek toplam hatayı kademeli biçimde optimize eden eklentisel bir yöntemdir. Bu süreçte amaç fonksiyonu, gradient (birinci türev) ve Hessian (ikinci türev) bilgisiyle optimize edilir; böylece yeni eklenen ağaç, o ana kadarki hataları hedef alıp düzeltmeye odaklanır (Wiens ve ark., 2025). Model karmaşıklığı, amaç fonksiyonuna eklenen L1 (lasso) ve L2 (ridge) türü regularizasyon terimleriyle denetlenir; L1 gereksiz dallanmaları ve büyük ağırlıkları seyrelterek yapıyı sadeleştirirken, L2 aşırı büyüyen ağırlıkları bastırarak daha dengeli bir çözüm üretir (Boldini ve ark., 2023; Wiens ve ark., 2025). Öğrenme hızı düşük seçildiğinde her yeni ağacın toplam modele katkısı sınırlanır ve hata daha yavaş ama kontrollü biçimde azalır. Çeşitliliği artırıp varyansı düşürmek amacıyla, her öğrenme döngüsünde gözlemlerin yalnızca bir bölümünün kullanılması (satır altörnekleme) ve her bölme kararında özelliklerin yalnızca bir altkümesinin denenmesi (sütun altörnekleme) birlikte uygulanır. Anlamsız dallanmaları önlemek için bölme kazancı eşiği kullanılır; veriye yeterince dayanmayan küçük yaprakların oluşmasını engellemek için de asgari yaprak örnek ağırlığı sınırı konur. Doğrulama kümesi izlenerek erken durdurma yapılması, iyileşme durduğunda ağaç eklemeyi sonlandırarak aşırı uyumu azaltır (Boldini ve ark., 2023; Wiens ve ark., 2025).

XGBoost, satırlarda bireylerin/gözlemlerin, sütunlarda ölçülen değişkenlerin yer aldığı tablo tipi (tabular) verilerde güçlü bir karşılaştırma ölçütü olarak kabul görür; bunun temel nedeni, karar ağaçlarının değişkenler arasındaki etkileşimleri doğrudan temsil edebilmesi ve düzenlileştirilmiş kaybın ikinci türev bilgisiyle kararlı biçimde optimize edilmesidir (Borisov ve ark., 2024; Wiens ve ark., 2025). Sınıf dengesizliği bulunan durumlarda, pozitif sınıfın ağırlığını artıran ayarlamalar ile karar eşiğini (örneğin F1'i en çoklayan eşik) birlikte belirlemek, duyarlılık–özgüllük dengesini anlamlı ölçüde iyileştirebilir (Boldini ve ark., 2023). Model çıktılarını olasılık olarak güvenle kullanabilmek için ikili sınıflamada temperature scaling, çok sınıflı sorunlarda ise Dirichlet kalibrasyonu önerilir; her iki yaklaşım da Expected Calibration Error ve Brier skoru gibi ölçütlerde belirgin iyileşmeler sağlamıştır (Dawood ve ark., 2023; Mehrtash ve ark., 2020).

Yöntemin sınırlılıkları da vardır. Ağaçların gereğinden fazla derinleşmesi veya regularizasyonun zayıf kalması aşırı uyum riskini artırır; bu durumda azami derinliği sınırlamak, asgari yaprak örnek ağırlığı ile bölme kazancı eşiğini yükseltmek ve satır/sütun altörnekleme oranlarını artırmak yararlıdır (Boldini ve ark., 2023; Wiens ve ark., 2025). Kategorik değişkenler XGBoost'ta doğal hâliyle işlenmediğinden genellikle tek-sıcak (one-hot) ya da hedef güdümlü kodlama gibi ön-işlemler gerekir; bu bakımdan CatBoost kategorik değişkenleri gözetimli şemalarla yerleşik olarak ele alırken, LightGBM yaprak-odaklı büyüme ve örnek seçme teknikleriyle hız ve ölçeklenebilirlik avantajları sunar (Hancock ve Khoshgoftaar, 2020; Peng ve ark., 2023; Ramalingam ve ark., 2024). Açıklanabilirlik için SHAP ailesi özellikle ağaç topluluklarına özgü TreeSHAP hem küresel (özellik önem dağılımı, etkileşim haritaları) hem de tekil gözlem düzeyinde tutarlı katkı değerleri üretir (Lundberg ve ark., 2020).

2.5.4. Support Vector Machine (SVM)

Support Vector Machine (SVM), doğrusal ayırmayı temel alan ve maksimum marjini (decision boundary ile en yakın örnekler arasındaki mesafe) büyütmeyi amaçlayan bir sınıflandırma yöntemidir; hataya tolerans tanıyan soft-margin formülasyonu, verideki sınıflar kısmen örtüştüğünde dahi genelleme gücünü koruyacak şekilde regularizasyon ile denge kurar. Doğrusal ayrımın yetersiz kaldığı durumlarda, kernel yöntemiyle örnekler örtük olarak daha yüksek boyutlu bir öznitelik uzayına (feature space) eşlenir ve bu uzayda doğrusal bir sınır öğrenilir; uygulamada RBF ve polinom kernel'leri yaygındır (Greener ve ark., 2022; Jovel ve Greiner, 2021). SVM'in amaç fonksiyonu, yanlış sınıflandırmaları "marjin ihlali" üzerinden cezalandıran hinge loss ile kurulur; böylece marjini koruyan, ama gereksiz karmaşıklığı sınırlayan bir denge elde edilir (Greener ve ark., 2022; Landeros ve Lange, 2022).

Model davranışını belirleyen başlıca ayarlar iki eksende toplanır: regularizasyon gücünü belirleyen C ve kernel'in esnekliğini belirleyen parametreler (örneğin RBF için gamma). Bu parametreler, marjin genişliği ile eğitim hatası arasındaki değiş tokuşu (bias–variance dengesi) doğrudan etkiler ve genellikle çapraz doğrulama ile seçilir. Özellik ölçeklerinin farklı olduğu tablo tipi (tabular) verilerde SVM'in mesafe temelli doğası nedeniyle standardizasyon/normalizasyonu doğru yapmak

kritiktir; çok sayıda karşılaştırmalı çalışmada SVM'in performans ve kararlılığının özellik ölçeklemeden belirgin biçimde etkilendiği gösterilmiştir (Greener ve ark., 2022; Jovel ve Greiner, 2021). Sınıf dengesizliği bulunan veri kümelerinde SVM, çoğunluk sınıfa eğilim gösterebilir; bu durumda pozitif sınıfa daha yüksek ceza veren sınıf ağırlıklandırması veya yanlış sınıflandırma maliyetlerini sınıflara göre farklılaştıran maliyet-duyarlı SVM varyantları, dengeyi düzeltmek için kullanılabilir. Güncel derlemeler, dengesiz veri için ağırlık, yeniden örnekleme ve karar eşiği birlikte ele alındığında duyarlılık–özgüllük dengesinin belirgin ölçüde düzeldiğini göstermektedir (Megahed ve ark., 2021). Karar desteğinde olasılık isteyen uygulamalarda SVM'in ham karar fonksiyonu güvenilir olasılıklara doğrudan karşılık gelmez; bu nedenle Platt scaling ve izotonik regresyon gibi kalibrasyon yöntemleriyle ECE/Brier gibi ölçütlerde anlamlı iyileştirmeler sağlanır (Ojeda ve ark., 2023).

Açıklanabilirlik açısından çekirdekli (kernel) SVM'lerde karar geometrisini doğrudan görmek güçtür; bununla birlikte son yıllarda geliştirilmiş yöntemler, tekil tahminlerin ve küresel örüntülerin görselleştirilmesine olanak sağlamaktadır. Örneğin, farklı kernel türleri için aktivasyon örüntülerine dayalı açıklamalar sunan yaklaşımlar, hangi özelliklerin marjin kararına ne ölçüde katkı verdiğini ortaya koyabilmektedir (Chaddad ve ark., 2023).

2.5.5. Logistic Regression

Logistic Regression, ikili sonuçları logit bağ fonksiyonu (log-olasılık oranı) üzerinden modelleyen, olasılığı 0–1 aralığına eşleyen genelleştirilmiş doğrusal model çerçevesidir; parametreler maksimum olasılık (maximum likelihood) yaklaşımı ile tahmin edilir ve katsayılar, uygun kodlama altında ilgili değişkendeki bir birimlik artışın olasılık oranını (odds ratio) nasıl değiştirdiği şeklinde yorumlanabilir Kayıp fonksiyonunun konveks yapısı, uygun koşullarda çözüm uzayında tek bir küresel en iyi noktaya güvenilir biçimde ulaşılmasını sağlar; bu nedenle lojistik regresyon, klinik ve biyomedikal öngörü modellerinde "başlangıç" ve aynı zamanda "karşılaştırma" modeli olarak yaygın biçimde kullanılmaktadır (Harris, 2021; Christodoulou ve ark., 2019).

Model esnekliği ve genelleme başarımı, regularizasyon ile artırılabilir: L1 (lasso) cezası değişken seçimini teşvik ederek daha seyrek bir çözüm üretirken, L2 (ridge) büyük katsayıları baskılayarak kararlılığı güçlendirir; uygulamada iki cezanın bir arada kullanıldığı elastic net yaklaşımı özellikle çoklu doğrusal ilişki ve yüksek boyut koşullarında yararlıdır (Fridgeirsson ve ark., 2024; Friedrich ve ark., 2022). Aşırı uyumu sınırlamak için erken durdurma, çapraz doğrulama ile hiperparametre seçimi ve özellik ölçekleme kritik adımlardır; özellikle tablo tipi (tabular) veride standartlaştırma/normalizasyon, mesafe ve gradyan temelli çözücülerde öğrenme dinamiklerini belirgin biçimde iyileştirir (Harris, 2021; Efthimiou ve ark., 2025).

Sınıf dengesizliği durumunda, lojistik model yalnızca olasılık eşiği 0.50'de değerlendirilirse çoğunluk sınıfa eğilim gösterebilir; bu nedenle karar eşiğinin kullanım amacına göre ayarlanması ve ağırlıklandırılmış öğrenme stratejilerinin kullanılması önerilir. Performans değerlendirmesinde, özellikle dengesiz veri bağlamında Precision–Recall eğrileri ve PR-AUC, ROC/ AUROC'a göre daha ayırt edici ve anlaşılır bir tablo sunabilir (van den Goorbergh ve ark., 2022; Saito ve Rehmsmeier, 2015).

Lojistik regresyonun önemli bir üstünlüğü, uygun tanımlama altında olasılık kalibrasyonunun güçlü olmasıdır; bununla birlikte örneklem kayması, özellik dönüşümleri veya sınıf dengesizliği gibi etkenlerle kalibrasyon bozulabilir ve o durumda temperature scaling ya da izotonik regresyon gibi basit kalibrasyon adımları Brier skoru ve ECE gibi ölçütlerde iyileşme sağlayabilir (Van Calster ve ark., 2019). Seyrek olay (rare events) ve ayrışma/separation durumlarında klasik maksimum olabilirlik tahminleri sapabilir ya da sonsuza gidebilir; bu koşullarda Firth sapma-azaltıcı (bias-reducing) cezalı olabilirlik yaklaşımı daha güvenilir katsayılar ve olasılık tahminleri üretir (Puhr ve ark., 2017).

2.6. İstatistiksel Değerlendirme ve Karşılaştırma Yöntemleri

2.6.1. Bootstrap Testi

Bootstrap, örneklemden yeniden örnekleme (resampling) yaparak istatistiklerin örnekleme dağılımını yaklaşık biçimde tahmin etmeye yarayan, dağılım varsayımından görece bağımsız bir yöntemdir. Temel fikir, elimizdeki veriyi

popülasyonun ampirik bir temsili olarak kabul etmek ve bu veriden, her seferinde aynı büyüklükte, yerine koyarak çok sayıda yeni örneklem çekmektir. Her bootstrap örnekleminde ilgilenilen istatistik (örneğin AUROC, Brier skoru veya iki model arasındaki ΔAUC farkı) yeniden hesaplanır; böylece bu istatistik için ampirik bir dağılım, güven aralıkları ve belirsizlik ölçütleri elde edilir (Efron, 1979).

Klasik (non-parametrik) bootstrap yaklaşımı, özellikle kapalı form çözümü olmayan karmaşık ölçütler (örneğin ayrıştırma ve kalibrasyon metrikleri) için kullanışlıdır; çünkü varyans, yanlılık (bias) ve güven aralıklarını türetmek için ek parametreik varsayımlar gerektirmez. Klinik öngörü ve sınıflandırma modellerinde bootstrap, hem iç geçerlilik (internal validation) hem de performans karşılaştırmalarının belirsizliğini incelemek amacıyla yaygın olarak önerilmektedir. TRIPOD rehberi ve onu izleyen metodolojik derlemeler, özellikle ROC eğrisi, AUROC, kalibrasyon eğrileri gibi karmaşık performans ölçütleri için bootstrap tabanlı belirsizlik tahminini, basit veri bölme (split-sample) stratejilerine kıyasla daha verimli ve tercih edilir bir yöntem olarak vurgulamaktadır (Moons ve ark., 2015; Steyerberg ve Harrell, 2016).

Tahmin modellerinde bootstrap kullanımının iki ana ekseni öne çıkar. Birincisi, model performansının ve performans farklarının güven aralıklarıyla değerlendirilmesidir. Aynı birey kümesi üzerinde geliştirilen veya kıyaslanan iki yöntem için, bootstrap örneklemleme her turda aynı bireyleri birlikte yeniden örnekleyerek eşleştirilmiş (paired) bir yapı kurar; bu sayede ΔAUC gibi fark istatistikleri için örnekleme dağılımı tahmin edilerek hem güven aralığı hem de anlamlılık sınaması (örneğin "farkın sıfırdan istatistiksel olarak farklı olup olmadığı") yapılabilir. İkincisi, özellikle model geliştirme çalışmalarında, bootstrap ile iyimserlik (optimism) düzeltmesi yapılarak eğitim verisi üzerinde görülen aşırı iyimser performans tahminleri (overfitting etkisi) düzeltilir; böylece modelin beklenen dış-geçerlilik performansı için daha gerçekçi tahminler üretilebilir (Efron, 1979; Steyerberg ve Harrell, 2016).

Özetle bootstrap yeniden örnekleme, bu tez kapsamında da olduğu gibi, AUROC ve benzeri performans ölçütleri ile modeller arası performans farklarının örnekleme belirsizliğini nicel olarak değerlendirmek ve eşleştirilmiş karşılaştırmalarda ΔAUC

gibi istatistikler için güven aralığı ve anlamlılık testi üretmek amacıyla kullanılan, modern klinik öngörü literatüründe yerleşik bir araçtır. Yöntemin hem tarihsel temeli (Efron, 1979) hem de klinik öngörü modellerine özgü uygulamaları (Moons ve ark., 2015; Steyerberg ve Harrell, 2016) güncel metodoloji rehberleriyle uyumlu bir çerçeve sunmaktadır.

2.6.2. Hosmer–Lemeshow Kalibrasyon Testi

Hosmer–Lemeshow (HL) kalibrasyon testi, özellikle lojistik regresyon ve benzeri olasılık üreten modellerde, modelin ürettiği tahmini olasılıkların gözlenen olay sıklıklarıyla ne derece uyumlu olduğunu değerlendiren klasik bir "iyi–uyum" (goodness-of-fit) testidir. Temel yaklaşım, tahmin edilen riskleri genellikle yaklaşık 10 gruba (desil) ayırmak; her grupta ortalama tahmin olasılığını ve gözlenen olay oranını hesaplamak ve bu iki büyüklük arasındaki farkın istatistiksel olarak anlamlı olup olmadığını ki-kare temelli bir istatistikle sınamaktır. Teknik olarak HL istatistiği, her risk diliminde beklenen olay sayısı ile gözlenen olay sayısı arasındaki farkların normalize edilerek özetlenmesiyle elde edilir ve grup sayısına bağlı serbestlik dereceli bir ki-kare dağılımı ile karşılaştırılır. Elde edilen p-değerinin 0.05'in altında olması belirgin bir uyumsuzluğa işaret edebilirken, yüksek p-değeri yalnızca testin "uyumsuzluk lehine güçlü kanıt bulamadığını" gösterir; bu nedenle "yüksek p = mükemmel kalibrasyon" biçiminde yorumlanmamalıdır (Huang ve ark., 2020; Van Calster ve ark., 2019).

HL testinin önemli sınırlılıkları, örneklem büyüklüğüne ve gruplandırma şemasına duyarlı olmasıdır. Büyük örneklemlerde klinik açıdan küçük sapmalar bile istatistiksel olarak anlamlı çıkabilirken, küçük örneklemlerde belirgin kalibrasyon bozuklukları test gücü yetersizliği nedeniyle kaçırılabilir. Ayrıca risk dilimlerinin sayısı ve nasıl oluşturulduğu p-değerini etkileyebildiğinden, test sonuçlarının yorumunda dikkatli olunmalıdır (Huang ve ark., 2020; Van Calster ve ark., 2019).

Bu nedenlerle güncel rehberler, HL testinin tek başına kalibrasyon değerlendirmesi için yeterli olmadığını; kullanılacaksa kalibrasyon eğrileri, kalibrasyon eğimi ve kesişimi gibi özetler ile Brier skoru gibi bütünleşik ölçütlerle birlikte tamamlayıcı bir gösterge olarak raporlanmasını önermektedir. Van Calster ve ark. (2019), kalibrasyonun öngörü modellerinde kritik bir boyut olduğunu vurgulayarak, HL p-

değerinin tek başına model kalitesi hakkında karar vermek için kullanılmaması gerektiğini belirtmektedir.

Özetle Hosmer–Lemeshow testi, lojistik regresyon ve benzeri olasılık tahmin modelleri için tarihsel olarak yaygınlaşmış, uygulaması kolay bir kalibrasyon testi olmakla birlikte; örneklem büyüklüğü, gruplama şeması ve p-değeri odaklı yorumun kısıtları nedeniyle modern literatürde daha çok grafiksel kalibrasyon analizleri ve sürekli kalibrasyon metrikleriyle birlikte, ikincil/yardımcı bir araç olarak konumlandırılmaktadır (Huang ve ark., 2020; Van Calster ve ark., 2019).

2.6.3. McNemar Testi

McNemar testi, aynı bireyler üzerinde elde edilen iki bağımlı ikili sonucun (paired binary outcomes) marjinal oranlarını karşılaştırmak için kullanılan klasik bir ki-kare tabanlı yöntemdir. Tipik kullanım, aynı örneklemde iki tanısal testin, iki sınıflandırma modelinin ya da müdahale öncesi–sonrası durumun karşılaştırılması olup, amaç "iki yöntemin pozitif olma olasılığı eşit mi?" sorusunu sınamaktır (Fagerland, Lydersen ve Laake, 2013). Bu çerçevede veri, A'nın ve B'nin ikili çıktılarının yer aldığı 2×2'lik eşleştirilmiş bir tabloya indirgenir. Testin ayırt edici yönü, yalnızca uyumsuz (discordant) hücreleri "A pozitif/B negatif ile A negatif/B pozitif durumlarını" kullanmasıdır; iki yöntemin aynı sonucu verdiği hücreler marjinal oranların eşitliği hakkında ek bilgi taşımaz (Fagerland ve ark., 2013).

Klasik gösterimde uyumsuz hücre sayıları b ve c ile ifade edilir ve sıfır hipotezi altında b ile c'nin beklenen değerlerinin eşit olduğu varsayılır. Büyük örneklemlerde kullanılan ki-kare yaklaşımında $X^{2}=\frac{(b-c)^{2}}{b+c}$ istatistiğinin 1 serbestlik dereceli ki-kare dağılımına yaklaşık uyduğu kabul edilir; b+c küçük olduğunda ise binomiyal temelli "exact" ya da "mid-p" türevleri önerilir (Fagerland ve ark., 2013). Sıfır hipotez reddedildiğinde, iki yöntemin marjinal performanslarının istatistiksel olarak farklı olduğu yorumu yapılır.

Tahmin modellerinin karşılaştırılmasında McNemar testi, aynı gözlem kümesinde iki modelin ikili kararlarının doğruluk açısından karşılaştırılmasında kullanılır. Her birey için "model 1 doğru/yanlış" ve "model 2 doğru/yanlış" bilgisi elde edilerek yalnızca "yalnızca model 1 doğru" ve "yalnızca model 2 doğru" hücreleri üzerinden test

gerçekleştirilir; böylece iki modelin hata/başarı desenlerinin simetrik olup olmadığı sınanır (van den Goorbergh ve ark., 2022). Testin güçlü yanı, eşleştirilmiş tasarımı doğrudan kullanması ve bu sayede bağımsız oran karşılaştırmalarına göre daha yüksek güç sağlayabilmesidir. Ancak McNemar testi yalnızca seçilen tek bir karar eşiği altında tanımlıdır; sürekli skorları/olasılık çıktılarının tamamını kullanmaz ve etki büyüklüğünü tek başına kapsamlı biçimde özetlemez. Bu nedenle, eşleştirilmiş farkın güven aralıkları ile ayrıştırma ve kalibrasyon ölçütleriyle birlikte raporlanması daha bütüncül bir değerlendirme sağlar (Fagerland ve ark., 2013; van den Goorbergh ve ark., 2022).

3. GEREÇ VE YÖNTEM

Bu tez çalışmasının yazım sürecinde, Yükseköğretim Kurulu Başkanlığı'nın (YÖK) "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Faliyetlerinde Üretken Yapay Zeka Kullanımına Dair Etik Rehber (2024)" inde belirtilen ilkeler doğrultusunda, yapay zeka aracı OpenAI ChatGPT (GPT-5.1) sınırlı bir şekilde kullanılmıştır. Yapay zeka aracı, yalnızca araştırmacı tarafından yazılan metinlerin dilbilgisi ve yazım kontrolünün yapılması, ifadelerin bilimsel ve akademik bir üsluba uyarlanması, kaynakça stilinin düzenlenmesinde ve kısmen literatür taramasına destek sağlaması amaçları ile kullanılmıştır. Araştırma hipotezinin oluşturulması, çalışma ve yöntem tasarımı, veri setinin oluşturulması, istatistiksel ve hesaplamalı analizlerin yürütülmesi aşamalarının tamamı, YÖK'ün rehberinde de belirtildiği gibi tamamen araştırmacı tarafından gerçekleştirilmiştir. Yapay zeka tarafından üretilen metinler, rehberde vurgulandığı üzere araştırmacı tarafından dikkatle gözden geçirilmiş, olgusal ve yorumsal hatalar düzeltilmiş ve bağlamsal tutarlılık sağlanmıştır (OpenAI, 22,10,2025).

Tez kapsamındaki tüm veri ön işleme, istatistiksel analiz ve makine öğrenmesi modellemeleri Python programlama dili kullanılarak gerçekleştirilmiştir. Varyant tablolarının oluşturulması, birleştirilmesi ve filtrelenmesinde pandas kütüphanesinden yararlanılmıştır. Sayısal işlemler, dizi hesaplamaları, paired bootstrap testi adımlarında numpy kütüphanesinden yararlanılmıştır. Logistic Regression, Support Vector Machine, Random Forest, Multi Layer Preceptron gibi sınıflandırma modellerinin eğitimi, çapraz doğrulama döngülerinin yürütülmesi, olasılık kalibrasyonu ve Accuracy, F1 skoru, MCC, AUROC, AUPRC, Brier skoru, ECE gibi metriklerin hesaplanması için scikit-learn paketi kullanılmıştır. XGBoost

modelinin kurulmasında xgboost paketi tercih edilmiştir. AUROC eğri grafikleri, kalibrasyon (reliability) eğrileri grafikleri ve diğer grafik çıktılar matplotlib kütüphanesi ile üretilmiştir. Tüm analizler Anaconda dağıtımı altında Spyder geliştirme ortamı altında yürütülmüştür (Anaconda, 16,07,2025; Python, 16,07,2025).

3.1 GEREÇLER

İşletim sistemi ve donanım

Bilgisayar türü / model: MONSTER ABRA A5 V15.5 dizüstü bilgisayar

İşletim sistemi: Microsoft Windows 10 Pro 64-bit

İşlemci (CPU): Intel® Core™ i5-10300H, 4 çekirdek / 8 iş parçacığı, ~2.50 GHz

Bellek (RAM): 16 GB DDR4

Grafik işlem birimleri (GPU):

Harici GPU: NVIDIA GeForce GTX 1650, 4 GB ayrılmış VRAM (3949 MB)

Dahili GPU: Intel® UHD Graphics, paylaşımlı bellek 8 GB (8111 MB)

Geliştirme ortamı ve programlama dili

Dağıtım ve ortam yöneticisi: Anaconda Navigator [sürüm 2.6.6]

Geliştirme ortamı: Spyder IDE [sürüm 6.0.7]

Programlama dili: Python [sürüm 3.13.5]

Python kütüphaneleri

Numpy [sürüm 2.1.3]

Pandas [sürüm 2.2.3]

Scikit-Learn [sürüm 1.6.1]

XGBoos [sürüm 3.1.1]

Matplotlib [sürüm 3.10.0]

Scipy [sürüm 1.15.3]

Joblib [sürüm 1.4.2]

Varyant ve skor veri kaynakları

Çalışmada klinik sınıflandırmalar NCBI ClinVar veritabanından (erişim tarihi: 28.06.2025) elde edilmiştir (Landrum ve ark., 2018).

Çalışmada kullanılan gen paneline ait missense varyantların patojenite tahmin skorlarını elde etmek için AlphaMissense tahmin aracı (erişim tarihi: 23.07.2025) (Cheng ve ark., 2023), SIFT tahmin aracı (erişim tarihi: 02.07.2025) (Ng ve Henikoff, 2003), PolyPhen-2 tahmin aracı (erişim tarihi: 22.07.2025) kullanılmıştır (Adzhubei ve ark., 2013).

3.2. YÖNTEM

3.2.1. Araştırmanın Tipi

Bu tez çalışması, açık erişimli web tabanlı veritabanlarından elde edilmiş ikincil veriler üzerinde yürütülen retrospektif ve tamamen in silico bir metodolojik doğrulama ve karşılaştırma (benchmark) araştırmasıdır. Çalışma, monogenik otoinflamatuvar hastalıklarla ilişkili genlerdeki missense varyantlara odaklanmakta ve bu varyantlar için ClinVar veritabanında yer alan ikili sınıflandırma etiketlerini referans olarak almaktadır. Araştırmanın temel amacı iki yönlüdür. Öncelikle patojenite tahmininde kullanılan biyoinformatik araçlarının seçilmiş gen panelindeki varyantlar üzerindeki tahmin performanslarını sayısal olarak tanımlamak ve kıyaslamaktır. İkinci olarak, bu araçlardan elde edilen tahmin skorlarını kullanarak makine öğrenmesi algoritmalarını eğiterek tahmin performansına ek katkısını aynı veri kümesi üzerinde sistematik biçimde değerlendirmektir. Tasarımda, veri sızıntısından kaçınmaya, varyant örnekleminin hedeflenen gen -hastalık bağlamını temsil etmesine ve kullanılan çapraz doğrulama şemaları sayesinde elde edilen performans tahminlerinin istatistiksel olarak güvenilir ve yeniden üretilebilir olmasına özellikle dikkat edilmştir.

3.2.2. Araştırmada Kullanılacak Genlerin İncelenmesi

İlk önce çalışmanın biyolojik bağlamı netleştirilerek varyantlar üzerindeki değerlendirme hedefli ve yorumlanabilir bir çerçeveye oturtuldu. Yapılan sistematik literatür taramasında varyant patojenitesi alanındaki çalışmaların genellikle tüm

genom ölçeğinde agnostik yaklaşımlarla tasarlanıldığı (Alirezaie ve ark., 2018; Dong ve ark., 2015; Ioannidis ve ark., 2016; Li ve ark., 2018) ya da otoinflamatuvar hastalıkları hedef almayan çeşitli gen panelleri üzerinde yürütüldüğü görülmüştür (Barbosa ve ark., 2022; Khandakji ve Mifsud, 2022; Kang ve ark., 2023).

Otoinflamatuvar hastalıklar özelinde, patojenite tahmin araçlarının ve makine öğrenmesi algoritmalarının aynı biyolojik eksen içinde sistematik kıyasını yapan çalışmaların sınırlı olduğu görülmüştür. Ayrıca tüm genom ölçeğinde veya heterojen gen kümelerinde yürütülmüş çalışmalarda bir aracın genel veri üzerinde yüksek performans göstermesi, belirli bir gen panelinde aynı başarıyı elde edemeyebileceği; çünkü modelin eğitimindeki kanıt yoğunluğu ve biyolojik bağlam bazı genlerde daha zengin, bazılarında daha sınırlı olabildiğinden, otoinflamatuvar çekirdek gen paneli özelinde, aynı etiketli veri matrisi üzerinde hem ayrıştırma hem de olasılık kalitesi boyutlarını birlikte ele alan sistematik bir değerlendirme ihtiyacı doğmaktadır. Bu boşluk, hem biyolojik gerekçeler hem de yöntemsel berraklık bakımından odağı otoinflamatuvar hastalıklar üzerine kurmayı motive etmiştir.

Gen paneli belirlenirken hastalıklar monofaktörlü olacak biçimde sınırlandırıldı. Bunun sebebi monofaktörlü hastalıklarda gendeki missense değişikliğin fenotiple ilişkisi çok daha doğrudan kurulur. Bu, varyantların patojenitesine dair model çıktılarının daha güvenilir bir şekilde sınanmasına olanak tanır. Multifaktörlü hastalıklarda patojenite, tek bir varyantın etkisinden ziyade birden fazla genetik değişken, epigenetik durum ve çevresel etkileşimlerin birleşik etkisi ile ortaya çıkar. Bu durumda, herhangi bir aracın verdiği skor ile gözlenen klinik sonucun arasına, doğrudan modellenemeyen çok sayıda aracı değişken girer. Bir varyantın tek başına patojenitesini değerlendirmek hem biyolojik hem de istatistiksel açıdan güçleşir. Çalışmanın hedefi, araçların ve makine öğrenmesi modellerinin tekil varyant sinyallerini ne ölçüde ayırt edebildiğini görmek olduğu için, monofaktörlü genler üzerinde çalışmak tercih edilmiştir. Bu doğrultuda, literatürde otoinflamatuvar monogenik sendromlarla nedensel ilişkisi güçlü biçimde gösterilmiş genler taranmış ve klinik uygulamada tanısal karşılığı olan ve ClinVar üzerinde missense SNV temsili yeterli düzeyde bulunan genler tercih edilmiştir.

3.2.3. Araştırma Evreni ve Örneklem

Araştırmanın evreni, literatürde monogenik otoinflamatuvar sendromlarla nedensel ilişkisi güçlü biçimde desteklenmiş 12 hedef gene ait, protein düzeyinde tanımlanan missense tek nükleotid varyantlarını (SNV) kapsar. Örneklem, ClinVar kayıtları temel alınarak ve çalışmanın ikili sınıflandırma çerçevesiyle uyumlu olacak biçimde oluşturulmuştur (Tablo 3.1).

Çalışma, monogenik otoinflamatuvar sendromlarla ilişkili MEFV, IL36RN, PSMB8, IL1RN, TNFRSF1A, PSTPIP1, OTULIN, NOD2, NLRP12, NLRP3, MVK ve LPIN2 genlerine ait, ClinVar veritabanında klinik önemi Pathogenic/Likely Pathogenic ya da Benign/Likely Benign olarak sınıflandırılmış 287 adet missense tek nükleotid varyantları dahil edilirken klinik önemi belirsiz ve çelişkili olan varyantlar kapsam dışı bırakılmıştır (Tablo 3.1). Missense dışı olan nonsense, frameshift, splice site, gibi değişiklikler ve UTR bölgelerindeki varyantlar örnekleme alınmamıştır. Tüm varyantlar, GRCh38 insan referans genomu üzerinde tanımlanmış ve konumlandırılmış kayıtlar arasından seçilmiştir. Çalışmanın hedefi, panel kapsamına giren ClinVar'daki uygun kayıtları mümkün olduğunca eksiksiz biçimde dahil etmek olduğu için, klasik anlamda olasılıklı bir örneklem yapılmamış, bunun yerine ilgili kriterleri taşıyan bütün kayıtların alınmasına dayalı "tam sayım" (census) yaklaşımı benimsenmiştir. Bu sayede, ikili sınıflandırma analizlerinde kullanılan veri kümesinin hem kapsayıcılığının hem de etiket tutarlılığının artırılması amaçlanmıştır. Bununla birlikte, klinik önemi belirsiz (VUS) ya da "conflicting interpretations" etiketli varyantların dışlanması, sınıf etiketlerinin netliğini güçlendirirken, evrenin yalnızca "klinik olarak daha kesin sınıflanmış" alt kümesine odaklanma yönünde bilinçli bir daralmaya yol açmaktadır.

Örneklem, ilerleyen makine öğrenmesi adımlarında katmanlı (stratified) çapraz doğrulama yapılabilmesine uygun şekilde düzenlenmiştir. Patojenik/olası patojenik (P/LP) ve benign/olası benign (B/LB) sınıflarının oranları, her bir katmanda (fold) tüm veri kümesindeki oranlara mümkün olduğunca yakın olacak biçimde korunmuştur. Bu yapı sayesinde, değerlendirme şemasında elde edilen out-of-fold (OOF) tahminlerin hem kendi aralarında hem de modeller arası karşılaştırılabilirliği güvence altına alınmaktadır.

Tablo 3.1 Hedef gen paneli ve örnek uzayı

Gen	Hastalık	Benign Varyant	Patojenik Varyant	Toplam
MEFV	FMF	4	7	11
IL36RN	DITRA	1	2	3
PSMB8	PRAAS	7	4	11
IL1RN	DIRA	2	-	2
TNFRSF1A	TRAPS	11	27	38
PSTPIP1	PAPA Sendromu	10	5	15
OTULIN	Otulinopeni	4	6	10
NOD2	Blau Sendromu	40	16	56
NLRP12	FCAS2	22	1	23
NLRP3	CAPS	8	59	67
MVK	HIDS	1	40	41
LPIN2	Majeed Sendromu	8	2	10
		118	169	287

3.2.4. Verilerin Elde Edilmesi

3.2.4.1. CLINVAR'dan Varyant Verilerinin Elde Edilmesi

Bu çalışmada varyant verileri, hedeflenen 12 monogenik otoinflamatuvar gen için ClinVar üzerindeki kayıtların sistematik taranmasıyla derlenmiştir. Her gen için ayrı sorgu yapılmış, yalnızca protein dizisini etkileyen tek nükleotid değişimleri seçilmiştir. Çalışmanın ikili sınıflandırma tasarımıyla uyumlu bir referans zemin oluşturmak amacıyla, klinik önemi Pathogenic/Likely pathogenic (P/LP) veya

Benign/Likely benign (B/LB) olarak bildirilen kayıtlar dâhil edilmiş; klinik önemi belirsiz (VUS) ya da "conflicting interpretations of pathogenicity" etiketli varyantlar, etiket kararsızlığı nedeniyle kapsam dışı bırakılmıştır (Li ve ark., 2022).

Ayrıca referans protein izoformu dışında raporlanan veya protein izoformu ile pozisyon eşleştirmesi güvenilir bir biçimde doğrulanamayan kayıtlar veri setinden çıkarılmıştır. Bu adımların sonucunda, panel genleri için GRCh38 uyumlu, missense SNV odaklı ve patojenik/muhtemelen patojenik ve benign/muhtemelen benign ikili klinik sınıflandırmasına sahip 287 tekil protein düzeyi değişiminden oluşan bir çalışma tablosu elde edilmiştir.

3.2.4.2. Tahmin Araçlarından Skor Elde Edilmesi

3.2.4.2.1. AlphaMissense

Tez çalışmasının odağındaki tüm genler için AlphaMissense veritabanından sorgulama yapılmış ve genlerle ilgili bütün varyantlar için geliştiricinin sağladı patojenite skorları ve patojenite kategorileri analiz yapılacak veri setine değişken olarak eklenmiştir. Verilerin edinimi sırasında, AlphaMissense çıktılarının referans protein izoformuyla uyumlu olup olmadığı kontrol edilmiş, izoform uyuşmazlığı, pozisyon kayması veya referans-varyant dizisi arasında tutarsızlık bulunan kayıtlar işaretlenmiş ve güvenilir eşleştirme saplanamayan varyantlar veri setinden çıkarılmıştır. Aynı protein değişimi için birden fazla sürüm ya da kayıt bulunduğu durumlarda, tekil satır ilkesi korunmuş, güncel ve izoformla uyumlu olanlar esas alınmıştır.

3.2.4.2.2. SIFT

SIFT aracından veri edinimi sırasında tüm varyantlar, HGVS p. Düzeyindeki protein değişimi ve referans alınan protein izoformu bağlamında doğrulanmıştır. Referans izoform dışında hesaplandığı anlaşılan skorlar ya da izoform-pozisyon eşleşmesi güvenilir bir biçimde teyit edilemeyen kayıtlar, veri bütünlüğünü ve sonuçların tutarlılığını korumak amacıyla çalışma setine dahil edilmemiştir. Aynı aminoasit değişimi için birden fazla SIFT kaydı bulunduğunda, güncellik ve referans izoformla uyum ölçütleri gözetilerek tekilleştirme yapılmış ve HGVS p. Değişimi için tek bir

SIFT skoru bırakılmıştır. Ayrıca SIFT'in sağladığı patojenite etiketleri de çalışma tablosuna eklenmiştir.

3.2.4.2.3. PolyPhen-2

PolyPhen-2 için, her HGVS p. Düzeyindeki varyantlar, ilgili gen ve referans protein izoformu dikkate alınarak sorgulanmış ve geliştiricinin raporladığı patojenite olasılık skoru çalışma tablosuna sayısal değişken olarak eklenmiştir. Ek olarak PolyPhen-2 çıktılarında mevcut olan patojenite etiketleri de sayısal skora ek olarak değişken olarak kaydedilmiştir.

Veri edinimi sırasında, PolyPhen-2 skorlarının referans kabul edilen izoformla uyumu birincil kalite ölçütü olarak ele alınmış; referans izoform dışında raporlandığı anlaşılan veya izoform–pozisyon eşleşmesi güvenilir şekilde doğrulanamayan kayıtlar çalışma seti dışında bırakılmıştır. Aynı protein değişimi için birden fazla PolyPhen-2 girdisi ile karşılaşıldığında, güncellik ve izoform uyumu dikkate alınarak tek satır prensibi uygulanmış, her HGVS p. değişimi için tek bir patojenite skoru ve eşlik eden kategori bilgisi bırakılmıştır.

3.2.5. CLİNVAR Verilerinin ve Araç Tahminlerinin Eşleştirilmesi

ClinVar veritabanından elde edilen varyant datası ile patojenite tahmin araçlarından elde edilen patojenite skorları ve etiketlerini doğru şekilde eşleştirerek makine öğrenmesi algoritmalarının eğitimi için kullanılacak hale getirecek bir eğitim datası oluşturmak amaçlanmıştır. Bu doğrultuda bütün veriler aynı referans protein izoformu ve aynı aminoasit pozisyonunda birleştirilmek üzere HGVS p. notasyonu kullanılarak her bir protein düzeyi değişim, ilgili ClinVar sınıf etiketi ve üç tahmina racının ham patojenite skorları ile kategorik açıklama yani sınıf etiketi eşleştirilmiştir.

Pratikte, ClinVar ve araç çıktıları arasında yazım farklılıkları, izoform referansındaki değişiklikler veya sürümler arası pozisyon kaymaları görülebildiği için, her varyant için izoform ve pozisyon doğrulaması yapılmıştır. Referans alınan izoformla birebir uyuşmayan, GRCh38 koordinatlarıyla uyumlu biçimde eşleştirilemeyen veya protein düzeyinde güvenilir karşılık üretilemeyen kayıtlar çalışma seti dışında bırakılmış,

ClinVar Variation ID ve ilgili araçların sürüm bilgileri izlenebilirlik amacıyla tabloda tutulmuştur.

3.2.6. Direction Harmonization

Tez çalışmasında, farklı tahmin araçlarından elde edilen ham skorların anlamsal yönünü ortak bir eksende birleştirerek, tüm değişkenler için geçerli tek bir yorum kuralı tanımlamak amaçlanmıştır. Bu doğrultuda tüm araçlarda, tahmin skorları arttıkça varyantın patojen olma olasılığının arttığı anlamının geçerli olması hedeflenmiştir. Bu durum başlangıçta her aracın kendine özgün tasarımları gereği farklılık göstermekte, Alphamissense ve Polyphen-2'de tahmin skoru arttıkça varyantın patojen olma ihtimali artarken, SIFT aracında ise tahmin skoru arttıkça varyantın patojen olma ihtimali azalıyor yorumu taşımaktaydı. Bu nedenle, SIFT skoruna basit bir doğrusal dönüşüm uygulanmış ve $s' = 1 - \text{SIFT}$ şeklinde yeniden tanımlanmıştır. AlphaMissense ve PolyPhen-2 için ek bir yön düzeltmesi gerekmemiştir. Böylece tüm araçların skorları 0'a yakın değerlerin benign olma ihtimali daha olası, 1'e yakın değerlerin ise patojen olma ihtimalinin daha olası şeklinde yorumlandığı ortak bir semantik düzleme taşınmıştır (Adzhubei ve ark., 2010; Cheng ve ark., 2023; Ng ve Henikoff, 2003).

Yön birliğinin sağlanması, yalnızca raporlamayı kolaylaştıran bir düzenleme değil, aynı zamanda ölçme mantığı ve karar metrikleri açısından da gereklidir. Eşik bağımlı karar ölçütleri olan Accuracy, Precision (PPV), Recall (Sensitivity), Specificity, F1 skoru ve Matthews korelasyon katsayısı (MCC) matematiksel olarak ikili sınıflandırmaya dayanarak hesaplanır. Çalışmada kullanılan skorların yön birliği olmadığında aynı eşik değeri için raporlanan bu ölçütlerin klinik anlamı araçtan araca ters düşer ve bu durum hem tabloların okunmasını hem de sonuçların yorumlanmasını güçleştirir.

Yön birliği sağlandığında, kalibrasyon eğrileri (reliability diagram) ile Expected Calibration Error (ECE) ve Brier skoru gibi olasılık doğruluğu ölçütleri, giriş skorunun artan bir risk sinyali taşıdığı varsayımı altında daha stabil ve yorumlanabilir sonuçlar verir. Ters yönlü bir değişken, kalibrasyon modelinin özellikle olasılık çıktısı üreten yöntemlerin davranışını gereksiz yere karmaşıklaştırabilir ve yanlılık-hata ayrımını belirsizleştirebilir. Bu nedenle yön

birliği, kalibrasyonun sayısal güvenilirliği açısından uygulanmıştır (Niculescu-Mizil ve Caruana, 2005; Zadrozny ve Elkan, 2002).

Ayrım (discrimination) ölçütleri açısından bakıldığında da skor yönünün standart hâle getirilmesi, hem sonuçların sunumunu hem de yorumunu sadeleştirmiştir. Eşik bağımsız bir ölçüt olan AUROC, skor yönünün değişmesinden etkilenmemiştir. Buna karşın, eşik bağımlı karar metrikleri ile kalibrasyon göstergeleri, skorların olasılık olarak nasıl yorumlandığıyla doğrudan ilişkilidir. Discrimination ile karar/olasılık katmanlarının kavramsal olarak birbirine karışmaması ve kalibrasyon sonrasında kullanılacak ortak klinik eşiğin tüm araçlar için aynı anlamı taşıması için, yön birliği adımı analiz akışının erken safhasında uygulanmış, ancak ham verinin izlenebilirliğini bozmayacak şekilde tasarlanmıştır.

Makine öğrenmesi modelleri açısından bakıldığında da skor yönlerinin uyumlu olması pratik bir avantaj sağlar. Tüm araç tahminlerinin skor arttıkça patojen olma ihtimali artar şeklinde aynı yönde anlam taşıması, model çıktılarının yorumlanmasını kolaylaştırır, değişkenler arası etkileşimlerin yorumlanmasını sadeleştirir ve farklı yönlere sahip özelliklerin aynı karar eşiği etrafında bir araya getirilmesiyle oluşabilecek anlamsal kaymaların önüne geçer.

Sonuç olarak yön birliği, bilgi içeriğini değiştirmeyen, ancak skorların anlamını standardize eden basit bir adımdır. Bu düzenleme, takip eden olasılık kalibrasyonu, eşik bağımlı karar metriklerinin hesaplanması ve makine öğrenmesi modellerinin eğitimi gibi süreçlerin hem teknik doğruluğunu hem de klinik açıdan okunabilirliğini güvence altına almayı hedeflemektedir.

3.2.7. Araçların Ham Skor Analizi

3.2.7.1. Youden's J Yöntemi İle Araçların Optimum Threshold Eşiğinin Belirlenmesi

Tekil tahmin araçlarının ham skorları üzerinde eşik bağımlı karar metriklerini yani Accuracy, Precision, Recall, Specificity, F1 skoru ve MCC değerlerini raporlayabilmek için araç bazında en dengeli karar noktasının (threshold) sızıntısız bir biçimde belirlenmesi amaçlanmıştır. Buradaki hedef, her aracın kendi skor uzayında pozitif ve negatif yani patojenik ve benign varyant ayırımının en dengeli

yapıldığı noktayı bulmak ve bu noktadaki performansını nesnel olarak raporlayabilmektir (Dong ve ark., 2015).

Öncelikle bütün tahmin araçlarında yön birliği sağlandı ve bütün bütün araçlarda skor arttıkça patojen olma ihtimali artar anlamı taşır hale getirildi. Böylece aynı eksende eşik taraması mümkün hale geldi. Eşik tayini, stratified 5 fold cross validation yöntemi ile yapılmıştır. Her fold'un eğitim kısmında aracın ham skorları 0-1 aralığında taranmış ve Youden indeksini maksimize eden değer eşik olarak seçilmiştir. Böylece seçicilik ve duyarlılık arasındaki dengeyi en iyi noktaya taşımak hedeflenmiştir. Seçilen bu eşik değeri, aynı fold'un validation bölümüne aynen uygulanmış ve bu doğrulama kümesinde eşik bağımlı performans ölçütleri (Accuracy, Precision, Recall, Specificity, F1 skoru, MCC) hesaplanmıştır. Tüm fold döngüleri tamamlandığında, validation bölümlerine ait out of fold (OOF) tahminleri birleştirilerek her araç için sızıntısız bir performans kestirimi elde edilmiştir. Eşik değeri belirlenirken, eşiğin öğrenildiği veri ile test edildiği veri birbirinden ayrı tutulmuş ve böylece aynı veri üzerinde hem eşik optimizasyonu hem de değerlendirme yapılmasından kaynaklanan iyimserlik yanlılığı (optimism bias) en aza indirilmiş ve OOF çerçevesi korunmuştur (Fluss ve ark., 2005).

Pratik uygulamada şu ilkeler izlenmiştir. Birincisi, Youden indeksi için birden fazla eşik değeri aynı maksimum J değerini veriyorsa, bunlar arasından medyan eşik seçilmiştir. Bu tercih, modelin gereksiz yere yüksek yanlış pozitif oranına kaymasını ve aşırı agresif pozitif sınıflandırma eğilimini sınırlamayı amaçlamaktadır. İkincisi, sınıf dengesizliğinin belirgin olduğu gen altkümelerinde, sınıf prevalansından görece bağımsız bir ölçüt olan J indeksinin kullanılması özellikle tercih edilmiştir. Buna ek olarak, karşılaştırmaların daha anlaşılır olması için, J indeksinin maksimum olduğu noktadaki F1 skoru ve MCC gibi sınıf dengesizliğine görece dayanıklı özet ölçütler de aynı out of fold (OOF) çıktılar üzerinden raporlanmıştır.

Bu prosedür sonucunda, her araç için kendi skor uzayında Youden indeksini maksimize eden araç özel optimum işletim eşiği (threshold) ve bu eşiğe karşılık gelen eşik bağımlı performans ölçütlerinin veri sızıntısından arındırılmış performansları elde edilmiştir. Burada belirlenen eşik değerleri yalnızca araçların ham performans profillerini betimlemek ve "aracın kendi doğasında en dengeli

ayrımı nerede yaptığını göstermek amacıyla kullanılmıştır. Böylece, her aracın kendi skorlama sisteminde patojenik ve benign varyantları en dengeli biçimde ayırdığı nokta görünür kılınırken, aynı zamanda sınıf dengesizliğinden ve eşik optimizasyonundan kaynaklanabilecek yanlılıkların en aza indirilmesi hedeflenmiştir (Attwood ve Tian, 2022; Shan., 2015; Thiele ve Hirschfeld, 2021).

3.2.7.2. Araç-Özel Eşik Değerinde Performans Metriklerinin Değerlendirilmesi

Tez çalışmasında, her bir tahmin aracının ham skorları kullanılarak eşik bağımsız ayırım gücünü (discrimination) ve araçların özel optimum eşiğinde (threshold) eşik bağımlı karar metrikleri değerlendirilmiştir. Bu analizde ClinVar'dan elde edilen varyant etiketleri mutlak doğru (ground truth) kabul edilmiş ve araç tahmin performansları değerlendirilirken ClinVar etiketleri referans alınmıştır.

Amaç, bir yandan aracın pozitif örnekleri negatif örneklerden sistematik biçimde daha yüksek sıralama kapasitesini herhangi bir eşik seçimine bağlı olmadan sayısallaştırmak, diğer yandan da klinik karar mantığına karşılık gelen bir işletim noktasında (Youden İndeksi) tarafsız performans ölçümleri elde etmektir. Bu iki düzlemin birlikte ele alınması hem yöntemsel hem de klinik yorum açısından gereklidir. Eşik bağımsız ayrım, modelin temel sinyal gücünü, eşik bağımlı metrikler ise belirli bir karar kuralı altında beklenen hata/fayda dengesini ortaya koyar.

Eşik bağımsız metrikler değerlendirilirken stratified 5 fold cros validation şeması altında elde edilen out of fold skor ve etiket verileri kullanılmıştır. Her araç için OOF skorları ve ClinVar etiketleri üzerinden AUROC değerleri elde edilmiş ve ROC eğrisi grafiği çizilmiştir. AUROC, 0-1 aralığında değişen tüm olası eşiklerin taranmasıyla oluşan doğru pozitif oranı (TPR) ile yanlış pozitif oranı (FPR) ilişkisini alan olarak özetleyen ölçü olduğu için SIFT skorunun yönünün ters çevrilmesi AUROC değerini etkilemez. Sınıf dengesizliğinin belirgin olduğu gen altkümelerinde, ROC eğrisinin yorum gücü azalabileceği için, ek olarak AUPRC de raporlanmıştır (Saito ve Rhemsmeier, 2015).

Eşik bağımlı metrikler değerlendirilirken ise araçların ham skorları ile elde edilen her aracın kendi özelinde en iyi youden indeksini veren eşik değerleri kullanılarak eşik bağımlı Accuracy, Precision, Recall, Specificity, F1 skoru ve Matthews Korelasyon

Katsayısı (MCC) metriklerinin performansları hesaplanmıştır. Araçların skorları, 0-1 aralığında Youden indeksi ile belirlenen eşik değerlerinin üstünde ise patojenik, eğer eşik değerin altında ise benign olarak değerlendirilmiştir. MCC'nin bu çerçeveye dahil edilmesinin gerekçesi, sınıf dengesizliğine görece dayanıklı olması ve karışıklık matrisinin tüm hücrelerini (TP, TN, FP, FN) simetrik olarak hesaba katmasıdır. Araç-özel optimum eşiklerin belirlenmesi ve geçerliliğinin sınanması, "cut-point" tahmini literatüründeki yaklaşımlarla uyumludur (Thiele ve Hirschfeld, 2021).

Her fold döngüsünde karar eşiği yalnızca o fold döngüsünün eğitim setinden skor ve etiket verileri kullanılarak seçilmiş, ROC eğrileri ve tüm karar metrikleri ise yalnızca fold döngüsünün doğrulama (OOF) setinden elde edilmiştir. Böylece, eşik seçimi ile performans ölçümü hiçbir aşamada aynı veri üzerinde yapılmamış, optimistik yanlılık azaltılmıştır.

Performans metriklerindeki istatistiksel değişkenliği değerlendirebilmek için, OOF örnekleri üzerinde bootstrap yöntemiyle %95 güven aralıkları hesaplanmıştır. Böylece hem veri setinin farklı eğitim/doğrulama ayrımlarına (fold yapısına) duyarlılık hem de metriklerin örnekleme kaynaklı dalgalanma payı nicel olarak ortaya konmuştur (Shan., 2015).

Bu ikili çerçevenin sebebi, tek başına sınıf ayrımın klinik karar kalitesini tam olarak yansıtmamasıdır. Teorik olarak mükemmel AUROC'a sahip bir araç, yanlış kalibrasyon veya klinik bağlamla uyumsuz bir eşik seçimi altında düşük F1 veya MCC değerleri üretebilir. Yani iyi sıralama yapan ama yanlış yerde kesen bir sistem klinik düzeyde zayıf performans gösterebilir. Tersine, yalnızca eşi bağımlı metriklere odaklanmak da yanıltıcıdır. Çünkü bu metrikler seçilen işletim noktasına, sınıf prevalansına ve hedeflenen duyarlılık/seçicilik dengesine duyarlıdır ve yanlış pozitif ve yanlış negatif gibi çeşitli hata tiplerini farklı ağırlıklarla yansıtır.

Bu çalışmada izlenen yaklaşım, önce eşik bağımsız düzlemde araçların bireysel ayırt gücünü hesaplamak, ardından araç özel optimum eşikte klinik karar kalitesini tarafsız biçimde göstermek ve bu kişisel optimum performansı gözlemlemektir (Berrar, 2019; Chicco ve Jurman, 2020).

3.2.8. Araçların Post-Hoc Kalibrasyonu

Bu çalışmada olasılık kalibrasyonunun temel amacı, AlphaMissense, PolyPhen-2 ve yön birliği sağlanmış SIFT skorlarını ham skorlama uzayından çıkararak, aynı sayısal değerin her araçta aynı klinik anlamı taşıdığı ortak bir olasılık ölçeğine dönüştürmektir. Böylece her araç için 0.5 değeri, yaklaşık olarak "%50 patojen olma olasılığı" eşiğini ifade eder hale gelmekte ve bu sayede Accuracy, Precision, Recall/Sensitivity, Specificity, F1 skoru ve MCC gibi eşik temelli performans ölçütleri, araçlar arasında aynı klinik anlam zemininde karşılaştırılabilir hale gelmiştir. Bu çalışmada kullanılan olasılık kalibrasyonu, modelin eğitiminden sonra uygulanan, yani "post-hoc" bir düzeltme adımıdır. Başka bir deyişle, aracın ürettiği skor ile gerçek patojenlik durumunun yani ClinVar'dan elde edilen etiketlerin ilişkisini, ek bir veri ya da özellik kullanmadan, yalnızca skorların dağılımına ve gözlenen gerçek etiketlere bakarak yeniden ölçeklendirmeyi amaçlar. Bu işlem sırasında her araç için ham skorlar, bu skora sahip bir varyantın patojen olma olasılığı nedir sorusuna yaklaşık olarak cevap veren olasılık değerlerine dönüştürülür. Kullanılan dönüşüm fonksiyonu, skorların kendi içindeki sıralamasını değiştirmeyecek şekilde tasarlandığı için, örneğin hangi varyantın diğerine göre daha yüksek riskli göründüğünü belirleyen göreli sıralama ve dolayısıyla ROC eğrisi ve AUROC gibi ayrım gücü ölçütleri korunurken hem de aynı skor değerlerinin, veri kümesinde gözlenen gerçek patojenlik oranları ile daha iyi uyumlu hale gelmesi sağlanır. Böylece araçların çıktıları hem klinik olarak daha doğrudan yorumlanabilir bir ölçeğe kavuşmakta, hem de farklı araç ve modeller arasında adil ve anlamlı performans karşılaştırmaları yapılabilmektedir (Pejaver ve ark., 2022).

Kalibrasyon işlemi, veri sızıntısını önlemek veperformans metriklerinin iyimserlik yanlılığından arındırılmasını sağlamak amacıyla stratified 5 fold cros validation düzeninde yürütülmüştür. Her bir araç için veriler, pozitif/negatif sınıf oranı korunacak şekilde beş katmana (fold) ayrılmıştır. Her döngüde bu beş katmandan dördü eğitim kümesi olarak kullanılırken, kalan bir katman ise doğrulama kümesi olarak kullanılmıştır. Bu işlem beş döngü halinde uygulanmış ve her döngüde doğrulama kümesi dönüşümlü olarak değiştirilmiş böylece bütün kümelerdeki varyant bilgileri hem eğitim hem de doğrulama datası olarak kullanılmıştır. Bu

sayede her gözlem, hiçbir aşamada kendi etiketi kullanılarak eğitilmemiş bir model tarafından tahmin edilen olasılık değeri ile temsil edilmiştir. Öncelikle eğitim kümesindeki skor ve etiket dataları kullanılarak o araca özgü bir kalibrasyon modeli kurulmuştur. Ardından, aynı döngüdeki doğrulama kümesinde yer alan örneklerin skorları bu modelden geçirilerek out of fold kalibre olasılıkları elde edilmiştir. Böylece hem kalibrasyon başarımı hem de bu olasılıklara dayalı eşik bağımlı performans metrikleri, aynı gözlemlerin hem model öğreniminde hem de değerlendirmede kullanıldığı durumlarda ortaya çıkabilecek yapay performans artışından olabildiğince arındırılmış biçimde hesaplanmıştır. Tüm veri kümesi üzerinde tek seferde bir kalibrasyon modeli kurmaya kıyasla bu prosedür metodolojik olarak daha güçlü kabul edilmektedir. Zira hiçbir aşamada test örneklerine ait bilgi kalibrasyon modeline geri beslemede kullanılmamakta, literatürde varyant skorlayıcılarının kalibrasyonunda önerilen katmanlı çapraz doğrulama temelli stratejilerle uyumlu bir yol izlenmiştir (Benevenuta ve ark., 2021).

Kalibrasyon yönteminin seçiminde veri kümesinin büyüklüğü, sınıf dağılımı ve araç skorlarının yapısı birlikte göz önünde bulundurulmuştur. Bu çalışmada tüm araçlar için tek tip bir yaklaşım benimsenmiş ve parametrik, düzgün (smooth) yapısı ile sınırlı örneklem koşullarında aşırı uyuma görece dayanıklı olması nedeniyle Platt scaling, yani tek öznitelikli lojistik regresyon temelli kalibrasyon yöntemi kullanılmıştır. Platt scaling, her bir aracın ham skorları ile gerçek patojenlik durumu arasındaki ilişkiyi, S-eğrisi (sigmoid) biçiminde monoton bir fonksiyonla modelleyerek, düşük skorları 0'a yakın, yüksek skorları ise 1'e yakın olasılık değerlerine çekerken ara bölgelerde yumuşak bir geçiş tanımlar. Bu yapı, skorların kendi içerisindeki göreli sıralamasını koruduğu için ROC eğrisi ve AUROC gibi ayrım gücü ölçütleri bozulmadan kalmakta, buna karşılık verilen bir skor düzeyinde gözlenen gerçek patojenlik oranlarıyla daha uyumlu bir olasılık ölçeği elde edilmektedir. Veri büyüklüğü ve stratified 5-fold bölünme şeması birlikte değerlendirildiğinde, Platt scaling'in sınırlı sayıda parametreye sahip olması, genellenebilirlik ile model esnekliği arasında dengeli bir çözüm sunması ve küçük katman boyutlarında dahi kararlı tahminler üretebilmesi nedeniyle kalibrasyon için uygun bir yöntem olduğu kanaatine varılmıştır. Ayrıca, tüm araçlarda sağlanan yön birliği sayesinde, Platt scaling kapsamında kullanılan lojistik bağlantı fonksiyonunun

varsaydığı monotonluk koşulunun verilerde sağlandığını güvence altına alarak, kalibrasyon modelinin daha tutarlı ve yorumlanabilir olmasını desteklemektedir (Benevenuta ve ark., 2021; Niculescu-Mizil ve Caruana, 2005).

Kalibrasyonun başarımı iki boyutta değerlendirilmiştir. Nicel değerlendirme için Brier skoru ve Expected Calibration Error (ECE) hesaplanmıştır. ECE, tahmin edilen olasılıkların [0,1] aralığının bu çalışmada n = 10 olmak üzere eşit genişlikte dilimlere ayrılması, her dilim için modelin ürettiği ortalama tahmin olasılığı ile aynı dilimde gözlenen pozitiflik oranı arasındaki farkın bulunması ve bu farkların dilimlerdeki örnek sayılarıyla ağırlıklandırılarak tek bir özet değerde toplanması şeklinde tanımlanmıştır. ECE değerinin düşük olması, modelin ürettiği olasılıkların fiili pozitiflik oranlarıyla daha iyi örtüştüğünü ve dolayısıyla daha iyi kalibrasyonu gösterir. Nicel ölçütlere ek olarak, görsel değerlendirme için kalibrasyon (reliability) eğrileri çizilmiş, farklı olasılık aralıklarında tahmin edilen ortalama olasılık ile gözlenen pozitiflik oranı karşılaştırılarak, eğrinin ideal 45° referans doğrusunun etrafında sistematik biçimde üstünde kalmasının aşırı güven (over-confidence), altında kalmasının ise yetersiz güven (under-confidence) örüntüsüne karşılık gelip gelmediği incelenmiştir. Ayrım (discrimination) ile kalibrasyonun kavramsal olarak farklı boyutlar olduğu özellikle vurgulanmış, yüksek AUROC değerine sahip bir aracın aynı anda zayıf kalibrasyon gösterebildiği nedeni ile AUROC gibi ayrım ölçütlerine ek olarak Brier skoru, ECE ve kalibrasyon eğrileri ayrıca raporlanmıştır. Bu metriklerin istatistiksel güvenilirliğini değerlendirmek amacı ile, out-of-fold (OOF) örnekler üzerinden bootstrap (5.000 tekrar) uygulanmış ve her araç ile model için AUROC, Brier skoru ve ECE'ye ait %95 güven aralıkları elde edilmiştir (Huang ve ark., 2020).

Kalibrasyonun uygulanmasında, araçlara özgü ayrıntılar izlenebilirlik ve yeniden üretilebilirlik amacıyla açık biçimde tanımlanmıştır. Tahmin araçlarına ait tüm ham skor sütunları veri setinde korunmuş, kalibrasyon sonrasında elde edilen olasılık değerleri ise ayrıca vektörlere kaydedilmiştir. Eksik veya uyumsuz kayıtlar kalibrasyon modellerinin eğitimine dahil edilmemiştir. Bu adımın metodolojik çıktısı iki yönlüdür. İlk olarak, her araç için out-of-fold (OOF) kalibre olasılık vektörü elde edilmiştir. Bu vektör, tüm araçlar için ortak eşik değerinde yani threshold = 0.5 eşiğinde hesaplanacak karar metriklerinin araçlar arasında aynı klinik anlam

temelinde karşılaştırılabilmesine olanak sağlamaktadır. İkinci olarak, bu kalibre olasılıklar, makine öğrenmesi sınıflandırıcılarının eğitiminde kullanılacak özellik uzayının çekirdeğini oluşturmaktadır. Böylece modeller, klinik anlamı sabitlenmiş ve araçlar arasında anlamsal olarak karşılaştırılabilir bir girdiye dayanmaktadır. Daha önce Youden's J ölçütü kullanılarak belirlenen araç-özel optimum threshold yalnızca modellerin bireysel başarısını değerlendirmek için kullanılmış, ne kalibrasyon modellerinin kurulmasında ne de makine öğrenmesi algoritmalarının eğitiminde girdi olarak kullanılmamıştır. Bu ayrım, bir yandan her aracın kendi doğasına özgü en iyi performans gösterdiği işletim noktasını görünür kılarken, diğer yandan ortak bir eşik etrafında yapılacak adil ve karşılaştırılabilir değerlendirmelerin yöntemsel bütünlüğünü korumaktadır.

3.2.9. Makine Öğrenmesi Algoritmalarını Eğitmek İçin Kalibre Edilmiş Veriler İle Özellik Seti Oluşturma

Kalibrasyon adımının ardından, her varyant için AlphaMissense, PolyPhen-2 ve yön birliği sağlanmış SIFT'e ait out of fold (OOF) kalibre olasılık değerleri elde edilmiştir. Özellik seti, bu tez çalışmasında kullanılan Logistic Regression, Support Vector Machine, Random Forest, XgBoost ve Multi Layer Preceptron makine öğrenmesi sınıflandırıcılarının eğitimi boyunca aynı klinik anlamı taşıyan, veri sızıntısından arındırılmış ve yeniden üretilebilir bir girdi matrisi sağlamak amacıyla bu kalibre olasılıklar üzerinden oluşturulmuştur. Veri yapısı, her satırın tek bir HGVS p. düzeyinde amino asit değişimini temsil edeceği şekilde düzenlenmiş, hedef değişken olarak ise ClinVar'ın ikili sınıf etiketleri kullanılmıştır. Bu düzenleme sonucunda, her araç için oluşturulan bir özellik matrisi ve bu matrise karşılık gelen ikili etiket vektörü, tüm makine öğrenmesi modelleri için ortak ve tutarlı bir öğrenme uzayı sağlamıştır (Ioannidis ve ark., 2016).

Özelliklerin OOF kökenli olması, veri sızıntısını engellemek açısından bilinçli ve kritik bir tercihtir. Kalibre olasılıklar, her fold'da yalnızca o fold'un eğitim diliminde öğrenilen kalibrasyon modeli kullanılarak, ilgili doğrulama dilimindeki skorların dönüştürülmesiyle üretilmiştir. Böylece hiçbir gözlem, kendi etiket bilgisine doğrudan ya da dolaylı biçimde maruz kalmadan bir olasılık değeri almaktadır. İzleyen aşamada makine öğrenmesi modelleri de stratified 5-fold çapraz doğrulama

şeması altında eğitilmiştir. Her döngüde dört fold eğitim, bir fold ise doğrulama için kullanılmıştır. Bu sayede modellerin doğrulama katmanında ürettiği tahmin olasılıkları da OOF mantığına uygun biçimde elde edilmiş ve raporlanan ayrım ve karar metrikleri bu sızıntısız tahminler üzerinden hesaplanmıştır. Makine öğrenmesi algoritmalarının eğitiminde tahmin araçlarına ait ham skorlar kullanılmamış sadece kalibre edilmiş olasılık verileri kullanılmıştır.

Sonuç olarak ortaya çıkan özellik matrisi, klinik anlamı ortak bir eşikte birleştirilmiş, OOF ile veri sızıntısından arındırılmış biçimde üretilmiş, model agnostik ve denetlenebilir niteliktedir. Bu yapı, makine öğrenmesi algoritmalarının hem ayrım hem de kalibrasyon boyutlarında adil, tutarlı ve klinik yorumla uyumlu biçimde değerlendirilmesine doğrudan zemin sağlamaktadır.

3.2.10. Hiperparametre Optimizasyonu (Tuning)

Bu çalışmada hiperparametre optimizasyonunun temel amacı, sınıflandırıcıların kalibre edilmiş araç olasılıkları üzerinde mümkün olan en yüksek genellenebilir ayrım gücüne ulaşmasını ve ürettikleri olasılıkların iyi kalibre edilmesini sağlamaktır. hiperparametre seçiminin performans değerlendirmesine sızmasını engellemek için nested cross-validation yapısı kullanılmıştır (Varma ve Simon, 2006).

Nested cross-validation düzeninde, dış halkada kurulan outer stratified 5-fold yapı, her yinelemede veri kümesini eğitim (outer-train) ve doğrulama (outer-validation) olmak üzere iki alt kümeye ayırmaktadır. Her outer-train kümesinin içinde, bağımsız bir inner stratified 5-fold döngüsü kurulmuş ve tüm aday hiperparametre kombinasyonları yalnızca bu iç döngünün eğitim ve doğrulama parçalarında değerlendirilmiştir. Inner döngüde, belirlenen seçim kuralına göre en yüksek performansı veren hiperparametre seti seçilmiş; ardından ilgili model bu ayarlarla outer-train kümesinin tamamında baştan eğitilmiş ve son aşamada yalnızca bir kez outer-validation kümesi üzerinde puanlanmıştır. Bu yapı sayesinde, dış halkadan elde edilen her tahmin, kendi örnek grubunun dışında seçilmiş hiperparametrelerle eğitilmiş bir modelden geldiği için out-of-fold (OOF) niteliği taşımakta ve böylece hiperparametre optimizasyonu ile performans değerlendirmesinin iç içe geçmesinden kaynaklanabilecek yanlılık önemli ölçüde kontrol altına alınmaktadır.

Bu çerçevede inner döngü, farklı hiperparametre ayarlarını denemek ve belirlenen seçim ölçütü altında en uygun ayarı belirlemek için kullanılan model seçimi katmanını; outer döngü ise, seçilmiş hiperparametrelerle eğitilen modelin performansını tarafsız biçimde ölçen asıl değerlendirme katmanını temsil etmektedir. Başka bir deyişle, hiperparametre araması yalnızca inner katmanda yürütülmekte, outer katmanda ise bu aramanın sonucunda seçilen nihai yapı sınanmaktadır. Nested cross-validation'ın, özellikle model seçimi ve tuning adımlarının çapraz doğrulama döngüsünün dışında tutulduğu senaryolara kıyasla hata tahminlerindeki sistematik yanlılığı azalttığı daha önce de gösterilmiştir (Varma ve Simon, 2006).

Hiperparametre seçiminde birincil ölçüt olarak AUROC (Area Under the ROC Curve) kullanılmıştır. AUROC'un tercih edilmesinin iki temel gerekçesi bulunmaktadır. İlk olarak AUROC, karar eşiğinden bağımsız bir ölçüttür. Bu özellik, farklı model aileleri (Logistic Regression, RBF çekirdekli SVC, Random Forest, XGBoost, MLPClassifier) ve aynı model içinde denenmiş farklı hiperparametre kombinasyonları arasında, belirli bir eşik değerine bağlı kalmadan, salt skor sıralamasına dayalı adil bir karşılaştırma yapılmasına imkân tanımaktadır. İkinci olarak AUROC, sınıf dağılımının hafif veya orta derecede dengesiz olduğu durumlarda dahi, pozitif sınıf örneklerinin negatiflere göre sistematik olarak daha yüksek skor alıp almadığını görece kararlı biçimde özetlemektedir. Bununla birlikte yalnızca sıralama başarısını esas almanın, olasılık kalitesine ilişkin sorunları perdelemesini önlemek amacıyla, aynı model için denenen hiperparametre kombinasyonlarının AUROC değerlerinin birbirine çok yakın olduğu durumlarda ikincil bağlayıcı ölçütler devreye alınmıştır. Bu bağlamda, Brier skorunun daha düşük ve MCC'nin (Matthews Correlation Coefficient) daha yüksek olduğu konfigürasyonlar tercih edilmiştir. Böylece hem ayrım gücü hem de modelin sınıf dengesizliği koşullarında ürettiği olasılıkların gerçekten gözlenen patojenlik oranlarıyla ne ölçüde uyumlu olduğu birlikte gözetilmiştir.

Ön işleme ve hiperparametre arama adımları, veri sızıntısını sistematik olarak engelleyecek biçimde tasarlanmıştır. Ölçek duyarlılığı bulunan model ailelerinde (Logistic Regression, SVC, MLP Classifier) özellikler, her katman için yalnızca o katmanın eğitim verisi kullanılarak standartlaştırılmıştır. Bu işlem sırasında eğitim kümesinin ortalama ve standart sapması hesaplanmış, daha sonra aynı katsayılar ilgili

dış doğrulama verisine uygulanmıştır. Böylece doğrulama verisi hiçbir aşamada ölçekleme parametrelerinin hesaplanmasına dâhil edilmemiş, her katmandaki gözlemler kendi iç istatistikleri ile değil, sadece eğitim grubunun istatistikleri ile dönüştürülmüştür. Ağaç tabanlı yöntemler olan Random Forest ve XGBoost ise, bu modellerin ölçek değişimlerine görece duyarsız olmaları nedeniyle ek bir ölçekleme uygulanmamıştır. Sınıf dengesizliğini dengelemek amacıyla Logistic Regression ve SVC için class weight parametresi balanced olarak ayarlanmış, XGBoost için ise pozitif/negatif örnek oranına göre belirlenen scale pos weight parametresi kullanılmış, bu ayarların nihai değerleri nested cross-validation'ın iç döngüsünde değerlendirilerek seçilmiş ve dış eğitim aşamasında aynı biçimde uygulanmıştır. Tüm kat bölünmeleri stratified olarak kurulmuş, random_state parametresi sabitlenerek hem inner hem outer döngülerin tekrar edilebilirliği güvence altına alınmıştır.

Arama uzayı, her model ailesinin istatistiksel özellikleri dikkate alınarak "makul" sınırlar içinde tanımlanmıştır. Logistic Regression için düzenleme düzeyini kontrol eden C parametresi logaritmik ölçekte artan değerler üzerinden taranmıştır. Bu modelde L2 cezası kullanılmış ve çözücü olarak liblinear tercih edilmiştir. Böylece hem aşırı büyük C değerlerinin yol açabileceği yüksek varyanslı, hem de aşırı küçük C değerlerinin neden olabileceği gereğinden fazla düzenlenmiş ve zayıf ayırt edici çözümlerden kaçınılması hedeflenmiştir. SVC (RBF kernel) modeli için C ve γ (gamma) parametreleri yine logaritmik olarak artan değerler üzerinde aranmıştır. İnner döngüde hiperparametre seçimi, doğrudan AUROC değerleri üzerinden yürütülmüştür. Random Forest için n_estimators, max_depth ve min_samples_leaf parametreleri sınırlı ve dengeli bir ızgara üzerinde sınanmış; çok derin fakat az sayıda ağaç içeren ve bu nedenle oynaklığa daha açık kurulumlar yerine, görece sığ ama daha fazla ağaçtan oluşan ve istikrarı yüksek yapıların tercih edilmesine özen gösterilmiştir. XGBoost modelinde learning_rate, n_estimators, max_depth, subsample, colsample_bytree, reg_lambda ve reg_alpha parametreleri birlikte optimize edilmiştir. Outer eğitim aşamasında inner döngüde seçilen hiperparametre setiyle tek seferde tam eğitim yapılmıştır. MLPClassifier için hidden_layer_sizes, alpha (L2 düzenleme katsayısı) ve max_iter parametreleri, mevcut örneklem büyüklüğünde istikrarlı bir öğrenme süreci sağlayacak aralıklarda sınırlandırılmış;

aşırı geniş veya çok katmanlı mimariler özellikle arama uzayının dışında bırakılmıştır. Arama stratejisinde, gereksiz derecede geniş ve ince taneli parametre ızgaralarından (grid search) kaçınılarak hem istatistiksel istikrar hem de hesaplama maliyeti açısından dengeli bir yaklaşım benimsenmiştir. Inner çapraz doğrulama döngüsü, farklı hiperparametre ayarları arasındaki rastlantısal dalgalanmaların ortalanmasına ve en tutarlı yapıların seçilmesine hizmet ederken, outer döngü her gözlem için yalnızca kendi dışında seçilmiş hiperparametrelerle eğitilmiş modellerden out-of-fold (OOF) olasılık tahmini üretmiş ve böylece raporlanan performans ölçütlerinin tarafsız biçimde hesaplanması için sağlam bir temel oluşturmuştur.

Tuning süreci tamamlandığında, varyantlar için out of fold olasılık vektörü elde edilmiştir. Bu vektör, yöntemsel açıdan iki nedenle önem taşımaktadır. İlk olarak, değerlendirme aşamasında raporlanacak ayrım ölçütleri (AUROC, AUPRC) ile olasılık kalitesini yansıtan ölçütlerin (Brier skoru, ECE ve kalibrasyon eğrileri) tarafsız biçimde hesaplanması için doğrudan bir temel sağlar. Her gözlem, kendi katmanı dışındaki verilerle eğitilmiş bir modelden elde edilen tahmin olasılığı ile temsil edildiği için, performans ölçütlerinin hesaplanmasında eğitim sürecinde kullanılan bilginin değerlendirme aşamasına sızması engellenmiş olur. İkinci olarak, eşik seçimi gibi karar düzeyi adımlar, tuning sürecinden bağımsız tutulur. Bu sayede modelin yapısal kapasitesi ve hiperparametre seçimi, herhangi bir özel eşik ya da ek kalibrasyon tercihinden bağımsız olarak veriye en uygun noktaya yerleştirilir. Bu ayrım hem klinik yorumun net kalmasına katkıda bulunur hem de kullanılan nested cross-validation iskeletinin scikit-learn araçları (GridSearchCV, StratifiedKFold) ve XGBoost API'si ile kod ortamında bire bir yeniden üretilebilir olmasını sağlar.

Sonuç olarak hiperparametre optimizasyonu, kalibre edilmiş araç olasılıklarını girdi olarak kullanan sınıflandırıcıların nested cross-validation altında, AUROC'un birincil ölçüt olduğu ve Brier skoru ile MCC'nin bağlayıcı ikincil ölçütler olarak dikkate alındığı bir seçim kuralı ile ayarlanmasını sağlamıştır. Böylece, farklı veri örneklemlerinde de benzer ayrım gücü sergilemesi beklenen ve olasılık davranışı tutarlı modeller elde edilmiştir. Ölçeklendirme, sınıf ağırlıklarının ayarlanması ve erken durdurma gibi yardımcı mekanizmalar, her zaman yalnızca ilgili eğitim verisinden öğrenilen bilgilere dayanarak uygulanmış; değerlendirme verisinin tuning

sürecine herhangi bir aşamada dâhil olmaması için mimari titizlikle korunmuştur. Bu yaklaşım sayesinde elde edilen OOF olasılıkların güvenilirliği artırılmış ve tez boyunca raporlanan performans ölçümlerinin yöntemsel olarak tutarlı ve yorumlanabilir bir temele oturması sağlanmıştır.

3.2.11. Makine Öğrenmesi Algoritmalarının Eğitimi

Makine öğrenmesi algoritmalarının eğitiminde üç tahmin aracının olasılık diline dönüştürülmüş ve kalibre edilmiş çıktıları kullanılmıştır. Her varyant için AlphaMissense, PolyPhen-2 ve yön birliği sağlanmış SIFT'ten elde edilen kalibre patojenlik olasılıklarını içeren üç boyutlu bir özellik vektöründen oluşmaktadır. Bu temsil sayesinde her sütun, 0,5 değerinin yaklaşık olarak "yüzde elli patojenlik olasılığı" anlamına geldiği ortak bir olasılık ölçeğini paylaşmakta ve böylece sınıflandırıcılar, araçlar arasında klinik anlamı tutarlı tek bir sayısal uzay üzerinde öğrenim yapmak mümkün hale gelmiştir. Araçların geliştiricileri tarafından tanımlanan kategorik sınıflar (örneğin SIFT için "tolerated/deleterious", PolyPhen-2 için "benign/possibly damaging/probably damaging", AlphaMissense için "ambiguous" gibi etiketler) bu aşamada öz nitelik olarak kullanılmamıştır. Bu etiketler, farklı karar kurallarına dayandıkları ve sürekli olasılık bilgisinin çözünürlüğünü düşürdükleri için yalnızca yorumlayıcı yardımcı bilgi olarak veri setinde korunmuş, ancak model girişine doğrudan taşınmamıştır. Modellerin eğitimi nested stratified 5-fold çapraz doğrulama çerçevesi altında yürütülmüştür. dış halkada kurulan katmanlı 5 katlı yapı, her yinelemede veri setini sınıf oranları korunacak şekilde bir eğitim ve bir doğrulama altkümesine ayırmıştır. Her dış eğitim altkümesi içinde bağımsız bir iç 5 katlı çapraz doğrulama döngüsü kurulmuş, hiperparametre adayları yalnızca bu iç döngünün eğitim ve doğrulama parçalarında değerlendirilmiş ve seçilen en iyi yapılandırma, ilgili dış eğitim kümesinin tamamında yeniden eğitilmiştir. Daha sonra model, dış doğrulama kümesi üzerinde sınanmış ve bu kümedeki her varyant için, o varyant dışarıda bırakılarak eğitilmiş bir modelden elde edilen tahmin olasılığı kaydedilmiştir.

3.2.12. Makine Öğrenmesi Algoritmalarının Bireysel Başarılarının Değerlendirilmesi

3.2.12.1. Youden's J Yöntemi İle ML Algoritmalarının Optimum Threshold Eşiğinin Belirlenmesi

Makine öğrenmesi sınıflandırıcılarının karar verme davranışını her bir model özelinde, sayısal ve yeniden üretilebilir biçimde tanımlayabilmek için, önce modellerin out-of-fold (OOF) olasılık çıktıları üzerinde model-özel optimum karar eşikleri belirlendi. Böylece her model için kendi olasılık çıktısı üzerinde, hangi eşik değeri çevresinde yanlış pozitif ve yanlış negatifler arasındaki dengenin en uygun biçimde kurulduğunu ve bireysel başarı gücünü ölçmk mümkün hale geldi. Uygulamada, her model için 0 ile 1 arasındaki eşik aralığı, ince bir ızgara üzerinde taranmıştır. Her aday eşik değeri için, OOF patojenlik olasılıkları pozitif ve negatif sınıf etiketlerine dönüştürülmüş ve bu tahminlere karşılık gelen karar metrikleri hesaplanmıştır. Pozitif sınıf kararı için $p \geq t$ ise 1 (patojen/olası patojen), $p < t$ ise 0 (benign/olası benign) kabul edilmiştir. Her bir aday eşiğe karşılık duyarlılık (Sensitivity/Recall), özgüllük (Specificity), kesinlik (Precision), F1 skoru, MCC (Matthews Correlation Coefficient) ve doğruluk (Accuracy) değerleri hesaplanmış, böylece her model için eşik–performans ilişkisi ayrıntılı biçimde haritalanmıştır.

Modellerin en iyi eşik değeri iki farklı ölçüte göre ayrı ayrı tanımlanmış ve her ikisi de raporlanmıştır. Birinci yaklaşımda, Youden's J istatistiğini maksimize eden eşik seçilmiştir. Ve böylece ROC uzayında, yanlış pozitif ve yanlış negatif hatalara simetrik ağırlık veren, "en dengeli ayrım noktası"nı yakalanmış olur. İkinci yaklaşımda ise F1 skorunu en çok artıran eşik belirlenmiştir. F1 skoru, Precision ve Recall'un harmonik ortalamasıdır ve özellikle dengesiz sınıf dağılımlarında, "pozitif sınıfı bulma başarısı" ile "yanlış pozitif üretmeme" arasında daha duyarlı bir denge ölçütü sunar. Önyargıyı azaltmak için eşik seçimi, yalnızca OOF tahmin skorları kullanılarak gerçekleştirilmiştir. Youden's J maksimum değerini veren birden fazla eşik değeri tespit edildiğinde özgüllüğü daha yüksek olan eşik tercih edilmiş, F1 açısından eşit performans veren eşikler arasında MCC değeri daha yüksek olan tercih edilmiştir.

Bu “kişisel optimum” eşikler, yalnızca her modelin kendi olasılık çıktısı üzerinde ulaşabildiği en dengeli karar noktasını tanımlamak için kullanılmıştır. Modeller arası karşılaştırmaların adil biçimde yapılabilmesi için, bu eşiklere dayalı doğrudan bir model kıyası yapılmamış; farklı modeller aynı sabit eşik altında ya da kalibrasyon sonrası ortak bir olasılık ölçeğinde ayrıca değerlendirilmiştir. Böylece, bir yandan her modelin kendi doğasında yanlış pozitif ve yanlış negatif dengesini en iyi nerede kurduğu görünür kılınmış, diğer yandan da ortak bir karar kuralı altında yürütülen kıyaslamalarla modeller arası adaletli değerlendirme ayrı bir zeminde korunmuştur (Hassanzad ve Hajian-Tilaki, 2024).

3.2.12.2. Modellerin Özel Optimum Eşik Noktasında Performanslarının Değerlendirilmesi

Her makine öğrenmesi modeli, out-of-fold (OOF) olasılık çıktıları üzerinde tanımlanan model-özel optimum işletim eşiğinde değerlendirilmiştir. Karar kuralı tüm modellerde üretilen patojenlik olasılığı seçilmiş eşik değerine eşit veya ondan büyük ise ilgili varyant patojen, eşik değerinden düşük ise benign olarak sınıflandırılmıştır. Böylece her modelin, kendi olasılık ölçeğinde en dengeli ayrımı sağladığı noktadaki hata deseni ve fayda profili sayısallaştırılmış olur. Performans değerlendirmesi, klinik karar süreciyle doğrudan ilişkili eşik-bağımlı ölçütler üzerinden yapılmıştır. Birincil göstergeler F1 skoru, Matthews korelasyon katsayısı (MCC) ve doğruluk (Accuracy) olarak seçilmiş; bunlara ek olarak duyarlılık (Sensitivity/Recall), özgüllük (Specificity) ve kesinlik (Precision) raporlanmıştır. Tüm bu metrikler, her model için ilgili eşikte elde edilen OOF tahminler kullanılarak oluşturulan karışıklık matrisi (confusion matrix) üzerinden hesaplanmıştır. Karışıklık matrisine eşlik eden gerçek pozitif, yanlış pozitif, gerçek negatif ve yanlış negatif sayıları, her modelin patojen varyantları atlama (false negative) ve benign varyantları yanlış pozitif olarak işaretleme eğilimini açık biçimde görünür kılmıştır.

Aynı modeller için, karar eşiğine bağlı olmayan ayrım ölçütleri de ayrıca sunulmuştur. Özellikle AUROC (Area Under the ROC Curve) ve sınıf dengesizliğinin belirgin olduğu durumlarda AUPRC (Area Under the Precision–Recall Curve), modelin tüm olası eşikler boyunca pozitif sınıf örneklerine sistematik olarak daha yüksek skor verip vermediğini özetleyen tamamlayıcı göstergeler olarak

raporlanmıştır. Bu eşik-bağımsız ölçütler, belirli bir karar eşiğinde elde edilen F1, MCC, duyarlılık ve özgüllük gibi metriklerin yerine geçmemekte; tersine, modellerin hem olasılık sıralama başarısını hem de seçilen çalışma noktasındaki hata dengesini birlikte yorumlayabilmek için tamamlayıcı bir bağlam sunmaktadır.

Yalnızca tek bir noktasal değere bağlı yorum yapılmaması için, performans ölçümü out-of-fold (OOF) tahminler üzerinden gerçekleştirilmiş ve her gözlemin daima kendi dışında eğitilmiş bir modelden gelen tahminle temsil edilmesi sağlanmıştır. Böylece değerlendirme, katmanlı çapraz doğrulama iskeleti içinde sızıntısız biçimde yürütülmüştür

Tüm ön işleme adımlarının yalnızca ilgili eğitim altkümeleri üzerinde öğrenilmesi ve bu adımlarda kullanılan parametrelerin hiçbir değişiklik yapılmadan aynı katmanın doğrulama altkümelerine uygulanması, ayrıca her gözlem için OOF tahminlerin daima o gözlemin yer almadığı katmanda eğitilmiş modellerden üretilmesi, sonuçların veri sızıntısından arınmış olmasını sağlamıştır. Bu sayede her model, kendi özel optimum eşik değerinde ortaya koyduğu karar performansını, eğitim ve değerlendirme aşamalarındaki sızıntıdan kaynaklanabilecek yanlılıkları büyük ölçüde engellemektedir. Böylece, modelleri birbirleriyle doğrudan kıyaslamaktan çok, her bir modelin kendi olasılık ölçeğinde yanlış pozitif ve yanlış negatif hatalar arasında kurabildiği en uygun denge noktasını nesnel ve sayısal olarak belgelemektir (Zhang ve ark., 2025).

3.2.13. ML Çıktılarının Post-Hoc Olasılık Kalibrasyonu

Makine öğrenmesi algoritmaları, ürettikleri olasılık skorlarını her araçta aynı klinik anlamı taşıdığı ortak bir olasılık ölçeğine dönüştürmek için Post-Hoc olarak kalibre edilmiştir. Bu sayede bütün modeller için 0.5 gibi sabit bir karar eşiği aynı klinik anlamı taşır hale getirilmiştir. Bu durum bütün modellerin ortak eşikte adil şekilde kıyaslanmasını mümkün hale getirir. Kalibrasyon aynı zamanda Brier skoru ve beklenen kalibrasyon hatası gibi göstergelerde iyileşme sağlayarak olasılıkların klinik karar desteği olarak güvenilirliğini artırır.

Kalibrasyon süreci, veri sızıntısını önlemek için daha önce kurulan katmanlı çapraz doğrulama iskeleti üzerine oturtulmuştur. Her outer fold için, ilgili makine öğrenmesi

modeli yalnızca o katmana ait outer-train kümesi kullanılarak eğitilmiştir. Ardından aynı küme üzerinde elde edilen tahmin olasılıkları ve ClinVar etiketleri kullanılarak Platt Scaling temelli bir kalibrasyon modeli kurulmuştur. Bu kalibratör, parametreleri sabit tutularak aynı dış katmanın doğrulama kümesinde modelin ürettiği olasılıklara uygulanmış ve yalnızca doğrulama kümesindeki gözlemler için kalibre edilmiş olasılık değerleri elde edilmiştir. Tüm dış katmanlar tamamlandığında, veri kümesindeki her varyant için out of fold (OOF) kalibre olasılık değerleri elde edilmiştir.

Kalibrasyonun başarısını değerlendirmek için modelin ürettiği olasılıkların genel hata düzeyi Brier skoru ile nicel olarak incelenmiştir. Ayrıca kalibrasyon (reliability) eğrileri ile tahmin edilen ve gerçek pozitif oranları arasındaki ilişki görselleştirilmiştir. ğrilerin ideal diyagonal çizgiye yakınlığı ile sistematik aşırı güven (over-confidence) veya yetersiz güven (under-confidence) örüntüleri nitel olarak yorumlanmıştır. Ek olarak, olasılık aralığının eşit genişlikte dilimlere ayrılmasıyla beklenen kalibrasyon hatası (Expected Calibration Error, ECE) hesaplanmış ve her dilimdeki ortalama sapmaların, dilimlerdeki örnek sayılarıyla ağırlıklandırılmış ortalaması tek bir özet ölçüt olarak raporlanmıştır.

Kalibrasyon sonrası tüm modeller ortak eşikte kıyaslanması mümkün hale gelmiştir. Buna göre modellerin ürettiği patojenite skoru 0,5'e şit veya büyük ise ilgili varyant patojen, eğer 0,5'ten düşük ise varyant benign olarak sınıflandırılmıştır. Kalibrasyon adımı sayesinde bu eşik, her modelde yaklaşık olarak "yüzde elli patojenlik olasılığı" anlamını taşımakta ve böylece farklı modellerin çıktıları, aynı klinik yoruma sahip tek bir olasılık ölçeği üzerinde değerlendirilebilmektedir. Bu ortak eşik kullanılarak F1 skoru, MCC, doğruluk (Accuracy), duyarlılık (Sensitivity/Recall), özgüllük (Specificity) ve kesinlik (Precision) gibi eşik-bağımlı karar metrikleri tüm modeller için hesaplanmış ve bu metrikler, modeller arası karşılaştırmada anlamsal olarak eşdeğer bir zeminde kullanılmıştır. Noktasal metrik değerlere, katmanlı yapıyı koruyan blok bootstrap yaklaşımıyla elde edilen %95 güven aralıklarıyla birlikte raporlanmıştır. Böylece hem kalibrasyon modelinin kurulmasına hem de ortak eşik seçimine bağlı istatistiksel değişkenlik nicel olarak ortaya konmuştur.

Teknik açıdan, özellik ölçeğine duyarlı Logistic Regression, SVC ve MLP model ailelerinde standartlaştırma, her dış katmanın yalnızca eğitim kümesi üzerinde gerçekleştirilmiş; bu kümeden hesaplanan ortalama ve standart sapma değerleri aynı katmanın doğrulama kümesine aynen uygulanmış, doğrulama verisinin bu parametrelerin elde edilmesine hiçbir aşamada dâhil olmasına izin verilmemiştir. Benzer şekilde, kalibrasyon modelleri de yalnızca eğitim temelli tahmin–etiket çiftleri kullanılarak kurulmuş ve daha sonra doğrulama kümesinde üretilen olasılıklara uygulanmıştır. Bu sızıntısız kurgu sayesinde elde edilen out-of-fold kalibre olasılıklar, hem model ayrım gücünü (örneğin AUROC, AUPRC) hem de olasılık doğruluğunu (örneğin Brier skoru, ECE ve ortak eşiğe dayalı karar metrikleri) birlikte ve güvenilir biçimde değerlendirmek için sağlam bir temel oluşturmuştur. Bu bölümde raporlanan kalibre olasılıklar ve bunlara bağlı metrikler, modellerin sabit bir klinik eşiğe göre karşılaştırılmasına hizmet etmekte; modelin kendi içinde en iyi performansı verdiği "kişisel optimum eşik" değerlendirmelerinden kavramsal olarak ayrı tutulmaktadır (Flevaris ve ark., 2024).

3.2.14. Araç ve ML Algoritmalarının Ortak Eşikte Karşılaştırılması

Kalibrasyon sonrasında, hem tahmin araçlarının hem de makine öğrenmesi modellerinin ürettiği olasılık değerleri ortak bir ölçek üzerinde yorumlanabilir hâle getirilmiştir. Bu çerçevede, tüm yöntemler için tek ve sabit bir karar kuralı kullanılmıştır: modelin ya da aracın ürettiği patojenlik olasılığı 0.5'e eşit veya büyük ise ilgili varyant patojenik, 0,5'ten küçük ise benign olarak sınıflandırılmıştır. Kalibrasyonun sağladığı hizalanma sayesinde 0,5 düzeyi, her bir model için yaklaşık olarak yüzde elli patojenlik olasılığı anlamına karşılık gelen bir eşik noktası olarak kabul edilebilmekte, böylece F1 skoru, MCC, Accuracy, Recall, Specificity ve Precision gibi eşik-bağımlı göstergeler, tüm araç ve modeller için aynı işletim noktasında ve aynı klinik yorum çerçevesinde karşılaştırılabilir hâle gelmektedir (Kia ve ark., 2020).

Karşılaştırma, tüm yöntemler için elde edilen out-of-fold (OOF) kalibre olasılık değerleri üzerinden yapılmıştır. Eşik-bağımlı karar ölçütleri olarak F1 skoru, Matthews korelasyon katsayısı (MCC), doğruluk (Accuracy), duyarlılık (Sensitivity/Recall), özgüllük (Specificity) ve kesinlik (Precision) ortak eşiğe göre

hesaplanmış, her yöntem için ek olarak Confusion Matrix sunularak yanlış pozitif ve yanlış negatif yükü sayısal olarak ortaya konmuştur. Olasılık kalitesini değerlendirmek amacıyla, kalibrasyon sonrası Brier skoru ve beklenen kalibrasyon hatası (ECE) değerleri raporlanmış, tahmin edilen ve gerçek pozitif oranları arasındaki ilişki kalibrasyon eğrileri üzerinden yorumlanmıştır. Ayrım gücü açısından AUROC ve sınıf dengesizliğine dayanıklı AUPRC, eşik bağımlı ölçütleri tamamlayan, karar eşiğinden bağımsız referans göstergeler olarak kullanılmıştır. İstatistiksel belirsizlik, katman yapısını koruyan fold-temelli blok bootstrap yöntemiyle nicel hâle getirilmiş ve başlıca tüm ölçütler için %95 güven aralıkları hesaplanmıştır. Bu yaklaşım, hem OOF düzenini hem de pozitif/negatif sınıf oranlarını sabit tutarak, performans ölçümlerinde rastlantısal değişimin etkisini daha gerçekçi biçimde ortaya koymuştur. Sonuçların yorumunda, kalibrasyonla hizalanmış ortak eşiğin sağladığı ortak zemin esas alınmış; farklı araç ve modellerin karar davranışları, klinik olarak aynı anlamı taşıyan bir patojenlik eşiği altında karşılaştırılmıştır. Eşik-bağımsız ayrım ölçütleri ve kalibrasyon göstergeleri ise bu tabloya ek bağlam sağlayarak, hem tekil tahmin araçlarının hem de bu araçlardan öğrenen sınıflandırıcıların aynı veri, aynı doğrulama kurgusu ve aynı karar kuralı altında yöntemsel olarak adil biçimde değerlendirilmesine imkân tanımıştır.

3.2.15. İstatistiksel Analiz

Modellerin karşılaştırılması, aynı varyantlar üzerinde elde edilen olasılık ve sınıf çıktıları kullanılarak eşleştirilmiş (paired) bir çerçevede yapılmıştır. Tüm istatistiksel işlemler Python ortamında, Anaconda/Spyder dağıtımı altında; scikit-learn 1.6.1, xgboost 3.1.1, numpy, pandas ve matplotlib paketleri kullanılarak gerçekleştirilmiştir. Rastgelelik içeren tüm adımlarda random_state = 42 değeri kullanılarak analizlerin tekrarlanabilirliği sağlanmıştır. Testler iki yönlü olarak uygulanmış ve istatistiksel anlamlılık eşiği $\alpha = 0.05$ olarak alınmıştır.

Öncelikle modellerin eşik-bağımsız ayrım gücü ve olasılık kalitesi birlikte değerlendirilmiştir. Bu amaçla her model için AUROC ve AUPRC hesaplanmıştır. Olasılıkların genel doğruluğunu değerlendirmek için Brier skoru; beklenen ve gözlenen pozitiflik oranları arasındaki sapmayı özetlemek için ise Expected Calibration Error (ECE, [0–1] aralığının on eşit olasılık dilimine ayrılmasıyla)

kullanılmıştır. Ardından, klinik karar düzeyini yansıtmak üzere ortak 0.50 eşiğinde Accuracy, F1 skoru, Matthews korelasyon katsayısı (MCC), Precision (PPV), Recall (Sensitivity) ve Specificity raporlanmıştır. Brier skoru ile ECE'nin "değeri düştükçe performansın iyileştiği" ölçütler olduğu metin ve şekillerde açıkça belirtilmiş; MCC'nin −1 ile 1 arasında değiştiği dikkate alınarak grafiklerde uygun eksenlendirme tercih edilmiştir.

İki modelin ayrım gücü arasındaki fark, aynı örnekler üzerinde elde edilen AUROC değerlerine dayalı eşleştirilmiş bootstrap yaklaşımı ile sınanmıştır. Her bootstrap yinelemesinde (B = 5000), N gözlem, tüm modeller için ortak indeks kümesi üzerinden yeniden örneklenmiş; her yinelemede her model için AUROC hesaplanmış ve iki model arasındaki fark $\Delta AUROC = AUROC_A - AUROC_B$ şeklinde tanımlanmıştır. Bu farkların bootstrap dağılımından %95 güven aralığı (persentil yöntemiyle) ve iki yönlü p-değeri elde edilmiştir. Çoklu ikili karşılaştırmalar söz konusu olduğunda yanlış keşif oranını (false discovery rate) kontrol etmek amacıyla Benjamini–Hochberg (FDR) düzeltmesi uygulanmış ve düzeltilmiş q-değerleri raporlanmıştır (Robin ve ark., 2011).

Karar düzeyindeki hata örüntülerinin benzerliği, ortak 0.50 eşiğinde McNemar testi ile değerlendirilmiştir. Her model çifti için, her bir varyantın iki model tarafından doğru/yanlış sınıflandırılmasına göre 2×2 bir tablo oluşturulmuş; bu tabloda "Model A doğru–Model B yanlış" ve "Model A yanlış–Model B doğru" hücreleri üzerinden McNemar testi uygulanmıştır. Küçük hücre frekanslarının bulunduğu durumlarda kesin (exact) binom temelli McNemar yaklaşımı tercih edilmiş; böylece modellerin klinik olarak birbirinin yerine kullanılabilirliği ve farklı türde hata yapma eğilimleri istatistiksel olarak sınanmıştır (Fagerland ve ark., 2013).

Olasılık çıktılarının güvenilirliği, hem kalibrasyon eğrileri (reliability diagram) hem de doğrusal kalibrasyon katsayıları ile incelenmiştir. Her model için kalibrasyon eğimi (ideal değer 1) ve kesişim terimi (intercept, ideal değer 0) tahmin edilmiş; bu katsayılar için bootstrap temelli %95 güven aralıkları hesaplanmıştır. Ek olarak, on gruba dayalı Hosmer–Lemeshow uyum testi uygulanmış; her model için test istatistiği (χ^2), serbestlik derecesi ve p-değeri raporlanmıştır (Steyerberg ve ark., 2010).

Bulguların görsel sunumunda, AUROC/AUPRC ile Accuracy, F1, Precision, Recall ve Specificity için modeller ekseninde tek tip tasarıma sahip sütun grafikler kullanılmıştır. MCC değerleri, −1 ile 1 aralığında, sıfıra göre merkezlenmiş sütunlar hâlinde gösterilmiş; Brier skoru ve ECE için ise 0–1 aralığında, "düşük değer daha iyi performansı gösterir" notunu içeren ayrı grafikler hazırlanmıştır. İkili model karşılaştırmalarına ait AUROC ve McNemar p-değerleri, model×model ekseninde düzenlenmiş ısı haritalarıyla özetlenmiş ve FDR düzeltmesi sonrası $q < 0.05$ eşiğini geçen hücreler görsel olarak vurgulanmıştır. Bootstrap yineleme sayısı $B = 5000$ olarak seçilmiş; duyarlılık analizlerinde $B \geq 2000$ değerlerinden itibaren AUROC ve türetilen ölçütler için elde edilen p-değeri ve güven aralıklarının kararlı bir desen sergilediği doğrulanmıştır. Bu bütüncül yaklaşım, aynı veri kümesi üzerinde değerlendirilen sekiz modelin ayrım gücü, sabit klinik eşiğe göre karar davranışı ve olasılık kalitesi açılarından sistematik ve istatistiksel bakımdan güvenilir bir karşılaştırmasını mümkün kılmıştır.

4. BULGULAR

Tez çalışmasında, ClinVar veritabanından elde edilen ve 12 monogenik otoinflamatuvar hastalığın çekirdek genlerini kapsayan varyant kümesi üzerinde, literatürde yaygın kullanılan üç tahmin aracı (AlphaMissense, PolyPhen-2, SIFT) ile bu araçlardan elde edilen veriler ile eğitilen beş makine öğrenmesi modeli (Logistic Regression, SVM, Random Forest, XGBoost, MLP) tarafından üretilen olasılık çıktıları sistematik olarak karşılaştırılmış ve istatistiksel olarak sınanmıştır. Tüm kıyaslamalar aynı örneklem üzerinde ve sabit rassal tohumla (seed=42) yürütülmüş, araçlar ve ML modelleri Platt Scaling ile kalibre edildikten sonra 0.5 ortak eşik üzerinde değerlendirilmiştir.

İlk basamakta, araçların kalibrasyon öncesi "ham skor" performansı model-özgü optimum eşiklerde değerlendirilmiş, bu amaçla Youden's J maksimumu ile belirlenen eşiklerde eşik bağımlı ölçütler (Accuracy, Precision/PPV, Recall/Sensitivity, Specificity, F1, MCC) raporlanmıştır. Eşik bağımsız ayrım gücü ise AUROC ve sınıf dengesizliğine duyarlı AUPRC ile özetlenmiştir. Bu adım, araçların bireysel (idiosyncratic) eşik koşullarındaki başarımını ortaya koymuştur.

İkinci basamakta, tahmin araçlarının ham skorları, model eğitimi tamamlandıktan sonra uygulanan bir post-hoc olasılık kalibrasyonu adımıyla, Platt scaling yöntemi kullanılarak olasılık ölçeğine dönüştürülmüştür. Bu sayede hem araçlar hem de makine öğrenmesi modelleri için eşik bağımsız ayrım gücü ve olasılık kalitesi birlikte değerlendirilebilmiştir. Her yöntem için AUROC ve AUPRC'nin yanı sıra, Brier score ve expected calibration error (ECE; [0,1] aralığının on eşit olasılık dilimine ayrılmasıyla) hesaplanmıştır. Klinik karar düzeyini temsil eden ortak 0.50 eşiğinde ise Accuracy, F1, Matthews Correlation Coefficient (MCC),

Recall/Sensitivity, Specificity ve Precision/PPV değerleri raporlanmıştır. Bu yapı, kalibrasyon sonrasında araçlar ile makine öğrenmesi modellerinin aynı karar eşiği altında doğrudan ve anlamlı biçimde karşılaştırılmasına olanak sağlamıştır.

Modeller arasındaki istatistiksel farklar iki tamamlayıcı çerçevede değerlendirilmiştir. Eşik bağımsız karşılaştırmalarda, her model çifti için eşleştirilmiş paired bootstrap yaklaşımı ile AUROC farklarının dağılımı elde edilmiş; bu dağılımlardan iki yönlü p-değerleri hesaplanmıştır. Çoklu karşılaştırma yapılması nedeni ile elde edilen p-değerleri Benjamini–Hochberg yanlış keşif oranı (FDR) yöntemiyle düzeltilmiş ve buna karşılık gelen BH-FDR q-değerleri raporlanmıştır. Bu sonuçlar, model×model ekseninde düzenlenmiş p-değeri matrisi ısı haritalarıyla görselleştirilmiştir.

Eşik bağımlı karşılaştırmalarda ise, ortak 0,5 karar eşiğinde modellerin sınıflandırma sonuçları üzerinden exact McNemar testi uygulanmıştır. Her model çifti için hesaplanan p-değerleri ve BH-FDR q-değerleri hem tablo biçiminde hem de model×model ısı haritaları şeklinde sunulmuştur. Böylece eşik-bağımsız bootstrap çerçevesi temel olarak ayrım gücündeki (ROC alanı) farkları, McNemar testi ise aynı eşiğe sahip modellerin karar düzeyindeki hata örüntülerini ve birbirlerinin yerine kullanılabilirliklerini ortaya koymaktadır.

Olasılık çıktılarının güvenilirliği ayrıca kalibrasyon eğrileri (reliability diagrams) ve doğrusal kalibrasyon katsayıları kullanılarak değerlendirilmiştir. Her model için kalibrasyon eğimi (ideal değer 1) ve kesişim terimi (ideal değer 0) tahmin edilmiş ve bu katsayılar için bootstrap temelli %95 güven aralıkları raporlanmıştır. Ek olarak, Hosmer–Lemeshow uyum testi uygulanmış; her model için test istatistiği, serbestlik derecesi ve p-değeri sonuç kısmında sunulmuştur.

4.1. Modellerin Model-Spesifik Karar Eşiğinde Tekil Performansı

Modellerin kendi spesifik karar eşiğinde tekil performanslarının test edildiği aşamada, tahmin araçlarının ham olasılık çıktıları kalibrasyon uygulanmadan, her model için Youden’s J maksimumu ile belirlenen araç-özel optimum eşikte değerlendirilmiştir. Amaç, her modelin kendi optimum karar eşiğinde ulaştığı ayırım gücünü ve sınıflama profilini ortaya koyarak bireysel başarılarını değerlendirmektir.

Eşik bağımsız ölçütler incelendiğinde, AlphaMissense'in belirgin bir üstünlük sağladığı görülmüştür. Araç, AUROC ≈ 0.87 ve AUPRC ≈ 0.91 değerleriyle hem klasik tahmin araçlarının hem de makine öğrenmesi modellerinin tamamının üzerinde bir ayrım gücüne ulaşmıştır. Makine öğrenmesi tarafında özellikle Logistic Regression (AUROC ≈ 0.85; AUPRC ≈ 0.87) ve SVM (AUROC ≈ 0.82; AUPRC ≈ 0.81) rekabetçi performans sergilemiştir. Olasılık kalitesini özetleyen Brier skoru ve ECE açısından da ham aşamada en düşük hata düzeyleri SVM (Brier ≈ 0.155; ECE ≈ 0.057) ve Logistic Regression'da (Brier ≈ 0.160; ECE ≈ 0.065) gözlenmiştir. Bu bulgu, söz konusu modellerin kalibrasyon uygulandığında güvenilir olasılık tahminleri üretme potansiyelini desteklemektedir. Buna karşılık, tahmin araçlarına ait ham skorların Brier ve ECE değerleri göreli olarak daha yüksek kalmış ve kalibrasyon işleminin gerekliliğine işaret etmiştir (Tablo 4.1).

Eşik bağımlı sınıflama metrikleri değerlendirildiğinde, Accuracy bakımından AlphaMissense ≈ 0.83 düzeyiyle ilk sırada yer aldı. Random Forest'ın (≈ 0.82), Logistic Regression ve SVM (≈ 0.81) değerleri ile AlphaMissense'ye yakın seyrettiği görülmüştür. F1 skoru açısından AlphaMissense (≈ 0.86) ve Random Forest (≈ 0.85) öne çıkarken, MCC değerlerinde AlphaMissense (≈ 0.65) birinci, Random Forest (≈ 0.63) ve Logistic Regression (≈ 0.62) onu izleyen modeller olmuştur. Karar profilleri incelendiğinde, Random Forest'ın yüksek Recall (≈ 0.87) ile daha duyarlı bir çalışma rejimine sahip olduğu; buna karşılık MLP'nin yüksek Specificity (≈ 0.85) ile yanlış pozitifleri sınırlayan, ancak daha düşük Recall (≈ 0.66) ile daha seçici bir davranış sergilediği dikkat çekmiştir. Logistic Regression ise yüksek Precision değeri (≈ 0.87) ile pozitif sınıf olarak işaretlenen varyantlar arasında doğru pozitif oranını maksimize etmiştir. Bu genel örüntü, aynı veri kümesi üzerinde farklı yöntemlerin farklı klinik önceliklere uygun işletim noktaları sunduğunu ve izleyen bölümlerde uygulanacak kalibrasyon ile ortak eşik altında yapılacak karşılaştırmanın, yöntemleri "eşit koşullarda" değerlendirebilmek için gerekli olduğunu ortaya koymaktadır (Tablo 4.1).

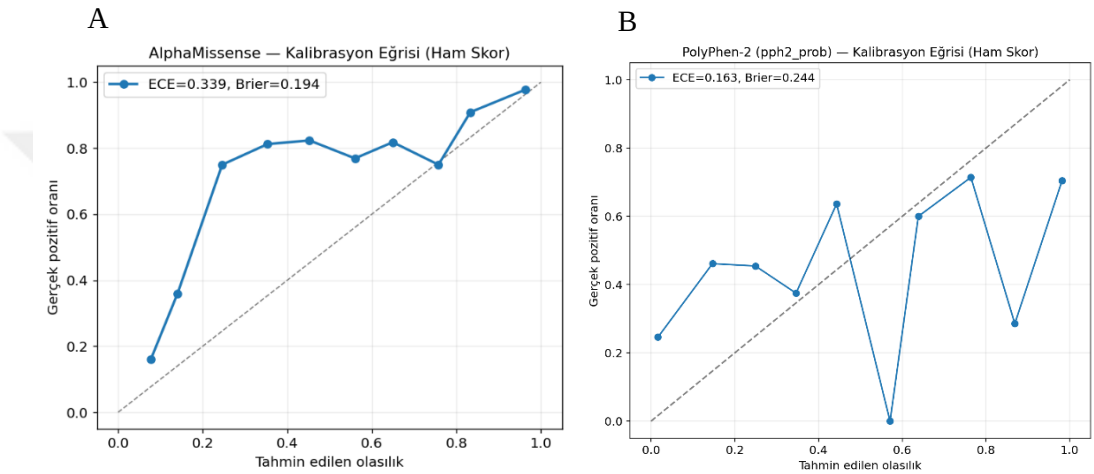
Tablo 4.1. Modellerin Ham Skorlarının Model-Spesifik Eşikte Metrik Performansları

	AMissense	Polyphen-2	SIFT	Logistic R	SVM	Random F	XgBoost	MLP
Youden's J Max	0.1684	0.932	0.92	0.3874	0.4404	0.3781	0.3708	0.5107
AUROC	**0.870**	0.761	0.718	0.851	0.820	0.810	0.776	0.811
AUPRC	**0.906**	0.784	0.770	0.873	0.812	0.830	0.770	0.867
Brier	0.194	0.244	0.270	0.160	0.155	0.166	0.201	0.185
ECE	0.339	0.163	0.198	0.065	0.057	0.110	0.160	0.097
Accuracy	**0.832**	0.718	0.666	0.810	0.810	0.820	0.771	0.739
F1	**0.857**	748	0.688	0.829	0.833	0.850	0.805	0.748
MCC	**0.652**	0.433	0.342	0.620	0.613	0.628	0.528	0.505
Precision	0.852	0.789	0.763	**0.873**	0.854	0.829	0.802	0.859
Recall	0.862	0.710	0.627	0.789	0.810	**0.873**	0.807	0.663
Specificity	0.788	0.729	0.720	**0.839**	0.805	0.746	0.720	0.847

Alphamissense' ye ait kalibrasyon eğrisi panelinde (Şekil 4.1A), olasılık aralığının büyük bölümünde referans hattının üstünde seyretmektedir. Bu görünüm, genelleşmiş bir under-confidence örüntüsüne işaret eder. Model, özellikle orta ve yüksek olasılık bölgelerinde gerçek pozitif oranını olduğundan düşük tahmin etmektedir. ECE=0.339 ve Brier=0.194 değerleri bu gözlemi nicel olarak desteklemekte, orta bantta çizgiye yeterince yaklaşmaması olasılıkların dar bir bantta sıkıştığını göstermektedir. Uç bölgelerde de (0.8–1.0) büyük oranda çizginin altında kalmaması, yüksek güvenli örneklerde bile temkinli (konservatif) bir olasılık ölçeği kullandığını düşündürür.

PolyPhen-2'ye ait kalibrasyon eğrisinde (Şekil 4.1B) özellikle orta olasılık bölgelerinde belirgin bir zikzak örüntüsü görülmektedir. Yaklaşık 0.4–0.5 aralığında eğri referans hattının hafif üzerinde yer alarak under-confidence eğilimi gösterirken, 0.6 civarında keskin biçimde hattın altına inmesi bu bölgede belirgin bir over-confidence olduğunu düşündürmektedir. Daha yüksek olasılık aralıklarında (≈0.7 ve üzeri) kalibrasyon eğrisi referans hattının altında kalmaya devam etmekte; model,

yüksek olasılık verdiği gruplarda dahi gerçekleşen pozitiflik oranını olduğundan yüksek tahmin etmektedir. Bu bölgeye bağımlı zikzaklı sapma profili, eğim hatası ve olası non-lineer bozulmaların rol oynadığı bir kalibrasyon sorununa işaret etmektedir. ECE=0.163 orta düzeyde bir kalibrasyon hatasına, Brier=0.244 ise araçlar arasında görece yüksek bir genel tahmin hatasına karşılık gelmekte; bu durum PolyPhen-2'nin ayrıştırma gücündeki sınırlılıklar ile uyumlu bir kalibrasyon eksikliğini yansıtmaktadır.

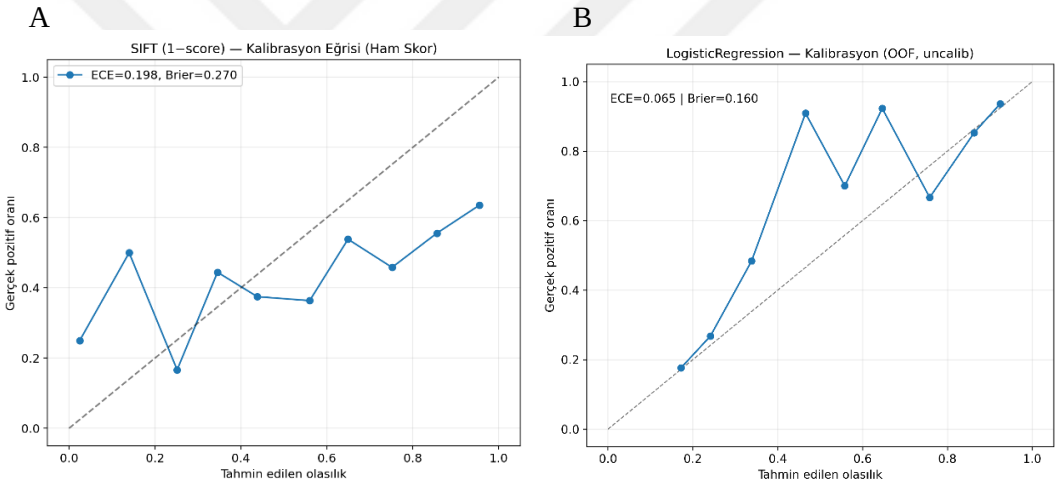


Şekil 4.1 (A) AlphaMissense'e ait ham skor kalibrasyon eğrisi ve (B) PolyPhen-2'ye ait ham skor kalibrasyon eğrisi

SIFT'e ait kalibrasyon eğrisinde (Şekil 4.2A), olasılık aralığının alt kısmında eğri çoğunlukla referans hattının üzerinde yer almakta, orta ve yüksek olasılık dilimlerine doğru ilerledikçe ise hattın altına düşen noktalar baskın hâle gelmektedir. Bu desen, düşük olasılık bölgelerinde under-confidence eğiliminde yani modelin verdiği olasılığın gerçek pozitif olasılığına göre düşük kaldığı, orta ve yüksek olasılık dilimlerinde ise over-confidence eğiliminde olduğu yani tahmin edilen olasılığın, gerçek pozitiflik oranını aştığı görülmektedir. Eğrinin olasılık ekseni boyunca belirgin zikzaklar yapması, özellikle orta bantta kalibrasyon eğrisinin kararlı bir doğrusal yapı sergilemediğini, uç dilimlerde ise sınırlı örnek sayısına bağlı varyans artışının rol oynadığını göstermektedir. Nicel olarak ECE = 0.198 ve Brier = 0.270 değerleri, SIFT için genel kalibrasyon

hatasının diğer yöntemlere kıyasla daha yüksek olduğunu ve bu görsel örüntüyü sayısal olarak da desteklediğini göstermektedir.

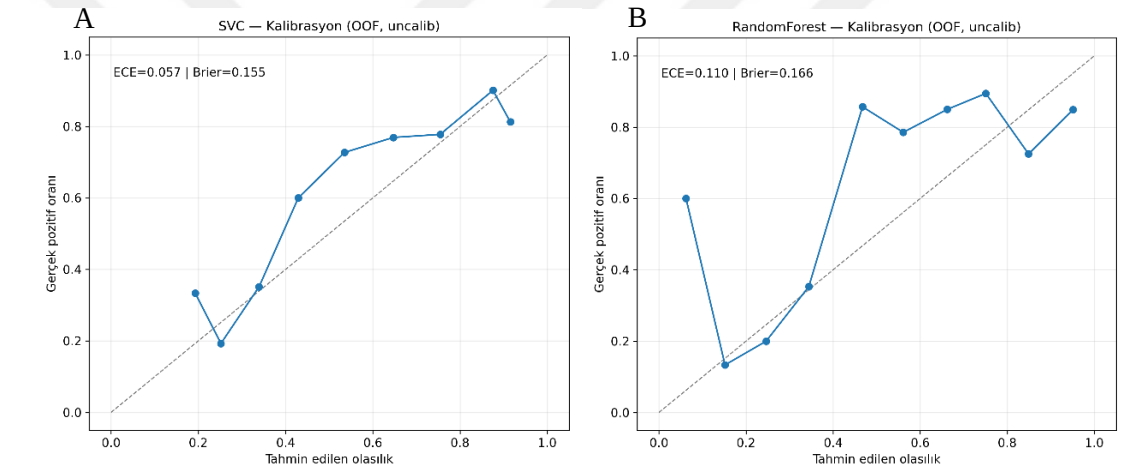
Logistic Regression'a ait kalibrasyon eğrisinde (Şekil 4.2B), modelin yaptı tahminlerin uç kısımlarda gerçek pozitif oranına yakın seyrettiği ama 0,3-0,7 aralığında tahminlerin gerçek pozitif oranına göre düşük kaldığı gözlemlenmiştir. 0,8 civarında ise gerçek pozitif oranından daha yüksek pozitiflik oranına sahip olduğu gözlemlenmiştir. ECE = 0.065 ve Brier = 0.160 değerleri, hem olasılık kalitesinin hem de genel tahmin hatasının diğer yöntemlere göre daha düşük düzeyde olduğunu ortaya koymakta; bu da Logistic Regression modelinin ek bir kalibrasyon adımı uygulanmadan önce dahi güçlü bir başlangıç kalibrasyonuna sahip olduğunu düşündürmektedir.



Şekil 4.2 (A) SIFT'e ait ham skor kalibrasyon eğrisi ve (B) Logistic Regression'a ait ham skor kalibrasyon eğrisi

Support Vector Machine'e ait kalibrasyon eğrisi grafiğinde (Şekil 4.3A) özellikle yaklaşık 0.35–0.8 olasılık aralığında noktalar referans hattının hemen üzerinde ve hatla oldukça yakın bir konumda yer almaktadır. Bu desen, modelin out-of-fold ham olasılık çıktılarının genel olarak düzgün bir yapı sergilediğini ve hafif bir under-confidence eğilimi taşıdığını göstermektedir. ECE=0.057 ve Brier=0.155 değerleri modeller arasında en iyi kalibrasyon–hata kombinasyonlarından birine sahip olduğunu göstermektedir.

Random Forest'a ait kalibrasyon eğrisinde (Şekil 4.3B) ise, olasılık aralığının sol ucunda belirgin bir dalgalanma göze çarpmaktadır. Çok düşük olasılık dilimlerinde (≈0.0–0.2) eğri önce referans hattının oldukça üzerine çıkarak yüksek bir under-confidence, hemen ardından hattın altına inerek over-confidence örüntüsü sergilemektedir. Model orta ve üst sınırlarda 0,4-0,7 arasında ise under-confidence, 0,8-1,0 aralığında ise over-confidence sergilemektedir. ECE = 0.110 ve Brier = 0.166 değerleri, düşük olasılık dilimlerindeki bu dalgalanma ile orta ve yüksek banttaki sistematik sapmanın birlikte oluşturduğu heterojen kalibrasyon hatasının sayısal karşılığını sunmaktadır. Random Forest'ın ham olasılık çıktılarının SVC ve Logistic Regression'a kıyasla daha belirgin düzeltmeye ihtiyaç duyduğunu göstermektedir.

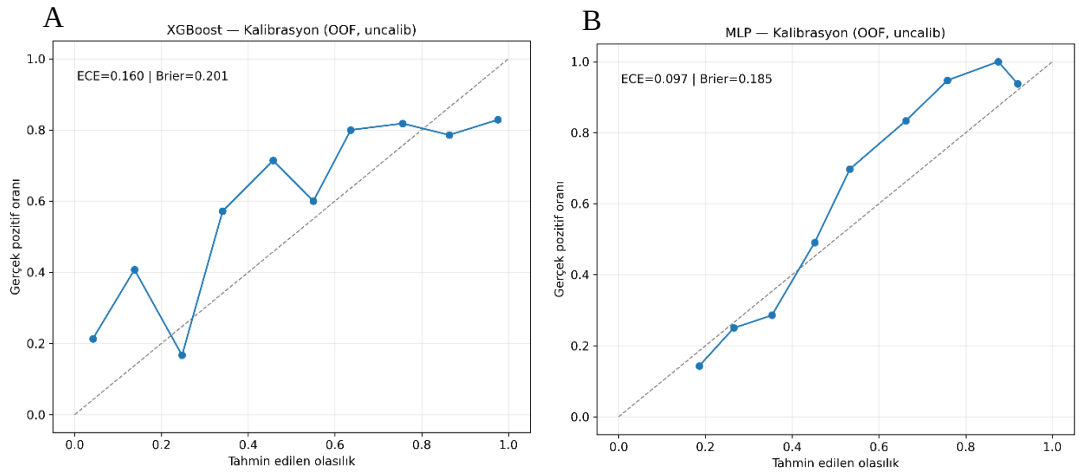


Şekil 4.3 (A) SVM'ye ait ham skor kalibrasyon eğrisi ve (B) Random Forest'e ait ham skor kalibrasyon eğrisi

XGBoost'a ait kalibrasyon eğrisinde (Şekil 4.4A) 0,3 civarında ve 0,8-1,0 aralığında modelin over-confidence sergilediği görülmektedir. 0-0,2 ve 0,3-0,7 aralıklarında aracın under-confidence eğiliminde olduğu gözlemlenmiştir. ECE = 0.160 ve Brier = 0.201 değerleri, XGBoost'un ham out-of-fold olasılık çıktılarında orta düzeyde bir kalibrasyon hatası ve genel tahmin hatası bulunduğunu, dolayısıyla sonraki post-hoc kalibrasyon adımının bu model için özellikle anlamlı bir iyileştirme potansiyeli taşıdığını göstermektedir.

MLP modeline ait kalibrasyon eğrisinde (Şekil 4.4B) 0,2-0,4 aralığında düşük bir düzeyde over-confidence gözlemlenmiş olsada referans hattına çok yakın ve paralel bir düzlem izlemiştir. 0,5-1,0 aralığında ise hafif düzeyde under-confidence

eğiliminde olduğu gözlemlenmiştir. Eğrinin genel olarak düzgün ve tek yönlü artan olması, belirgin non-lineer bozulmaların sınırlı olduğunu göstermektedir. ECE = 0.097 ve Brier = 0.185 değerleri, kalibrasyon hatasının XGBoost ve bazı diğer yöntemlere göre daha düşük, ancak tamamen ihmal edilebilir düzeyde olmadığını; modelin ham hâliyle bile nispeten düzenli bir olasılık davranışı sergilediğini ve kalibrasyon sonrası daha da iyileşmeye açık olduğunu ortaya koymaktadır.

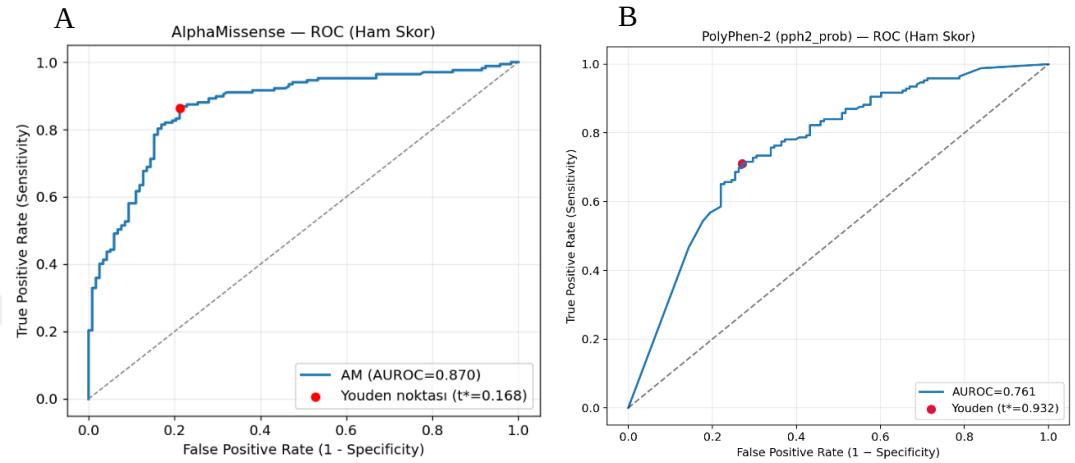


Şekil 4.4 (A) XgBoost'a ait ham skor kalibrasyon eğrisi ve (B) MLP'ye ait ham skor kalibrasyon eğrisi

AlphaMissense'nin Roc eğrisi (Şekil 4.5A), özellikle düşük yanlış pozitif oranlarında (FPR ≈ 0.0–0.15) hızlı bir yükseliş göstererek duyarlılığı kısa sürede yaklaşık 0.80 düzeyine taşımaktadır. Bu bölgedeki belirgin erken kaldırma (early lift) AlphaMissense'in patojenleri benignlerden ayırmada kuvvetli bir sinyal taşıdığını göstermektedir. Eğri genel olarak grafik alanının sol üst köşesine yakın seyretmekte ve AUROC = 0.870 değeri elde edilmektedir. Youden indeksine göre en iyi işletim noktası, 0.168 eşiğinde bulunmuştur. Eğrinin orta-yüksek FPR aralığında da (≈0.2–0.5) konveks formunu koruması, farklı klinik tercih profillerinde de avantajın sürdüğünü düşündürmektedir

PolyPhen-2'ye ait ROC eğrisi (Şekil 4.5B), başlangıçta Alphamissense'e kıyasla daha sınırlı bir early lift sergilemekte ve genel olarak diyagonale daha yakın bir seyir izlemektedir. Bütünleşik ayrıştırma gücü AUROC = 0.761 seviyesinde ve orta düzeyde bir performansa işaret etmektedir. Youden indeksine göre optimum eşik 0.932 olarak bulunmuştur. Bu eşikteki ROC noktası, düşük yanlış pozitif oranına

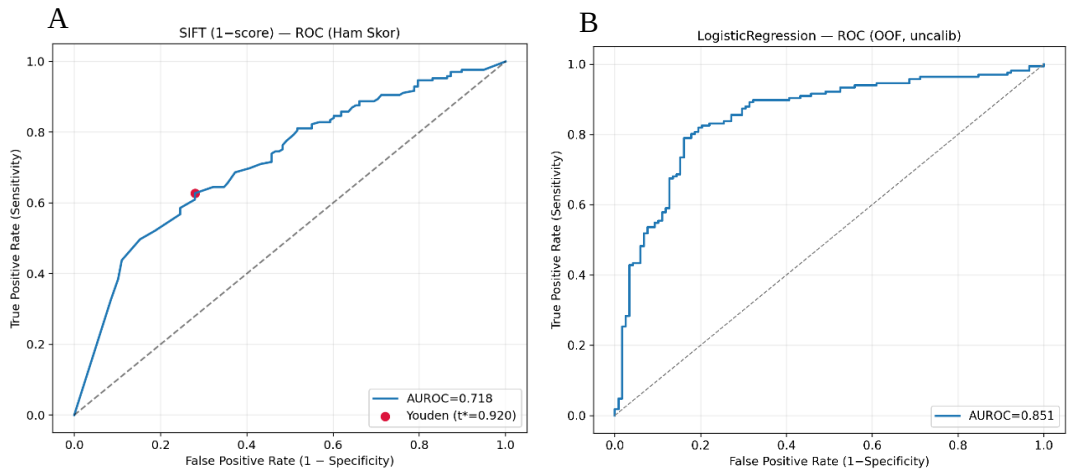
karşılık gelen görece yüksek bir özgüllük sağlarken duyarlılığın sınırlı kalmasına yol açmaktadır; başka bir deyişle, model patojen sınıfına çok yüksek skor verme eğilimindedir ve kabul edilebilir özgüllük elde edebilmek için oldukça yüksek bir karar eşiği gerekir.



Şekil 4.5 (A) AlphaMissense'e ait ham skor ROC eğrisi ve (B) PolyPhen-2'ye ait ham skor ROC eğrisi

SIFT'e ait ROC eğrisi (Şekil 4.6A), PolyPhen-2'ye benzer şekilde diyagonale daha yakın seyretmiştir ve AUROC değeri 0,718'dir. Youden noktası t≈0.920 olarak hesaplanmıştır; bu durum, ham skorların patojen sınıfa doğru ağır bastığını ve yanlış pozitifleri sınırlamak için çok yüksek bir karar eşiğinin olduğunu göstermektedir.

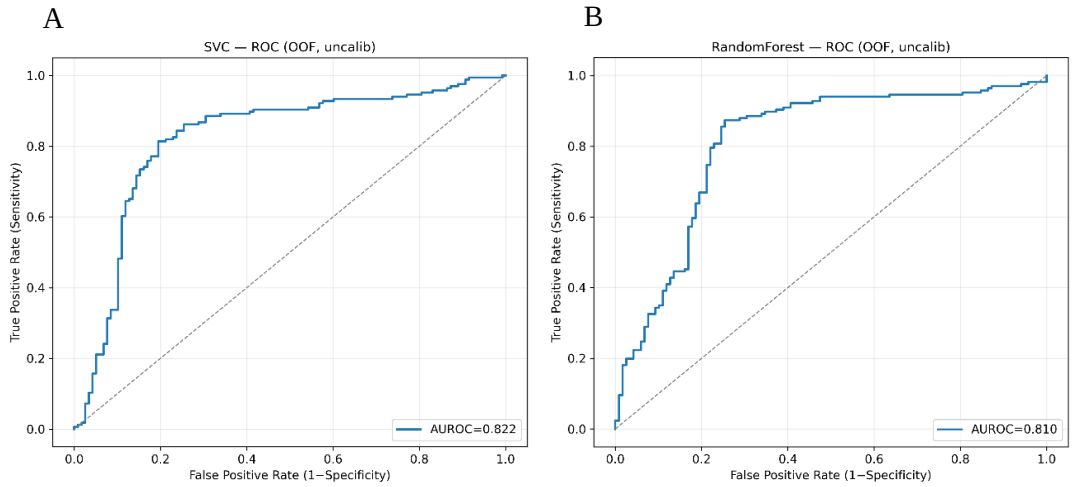
Logistic Regression'un ROC eğrisi (Şekil 4.6B), çok belirgin bir early lift sergilemekte ve 0,851 AUROC değerine ulaşarak güçlü bir ayırım gücü sergilemektedir. Eğrinin düşük FPR aralığında (≈0–0.1) duyarlılığı hızlıca 0.80'lere taşıması, modelin istikrarlı özellik kullanımıyla erken aşamada belirgin kazanç sağladığını göstermektedir. Bu profil, aynı model için Brier ve ECE değerlerinin görece düşük bulunmasıyla da tutarlıdı. Yani Logistic Regression yalnızca iyi ayırmakla kalmamakta, skoru olasılık olarak yorumlandığında da makul bir güvenilirlik sergilemektedir.



Şekil 4.6 (A) SIFT'e ait ham skor ROC eğrisi ve (B) Logistic Regression'a ait ham skor ROC eğrisi

SVM'nin Roc eğrisi (Şekil 4.7A), düşük yanlış-pozitif oranlarında (FPR≲0.10) belirgin bir "erken kaldırma" göstererek duyarlılığı hızla orta-yüksek değerlere taşımakta ve sonrasında 0.20–0.60 FPR bandında konveks formunu korumaktadır. Eğrinin diyagonale uzak seyri, modelin geniş bir işletim aralığında duyarlılık-özgüllük dengesini görece istikrarlı biçimde sürdürebildiğine işaret eder; bu görsel örüntü, AUROC=0.822 değeriyle uyumludur.

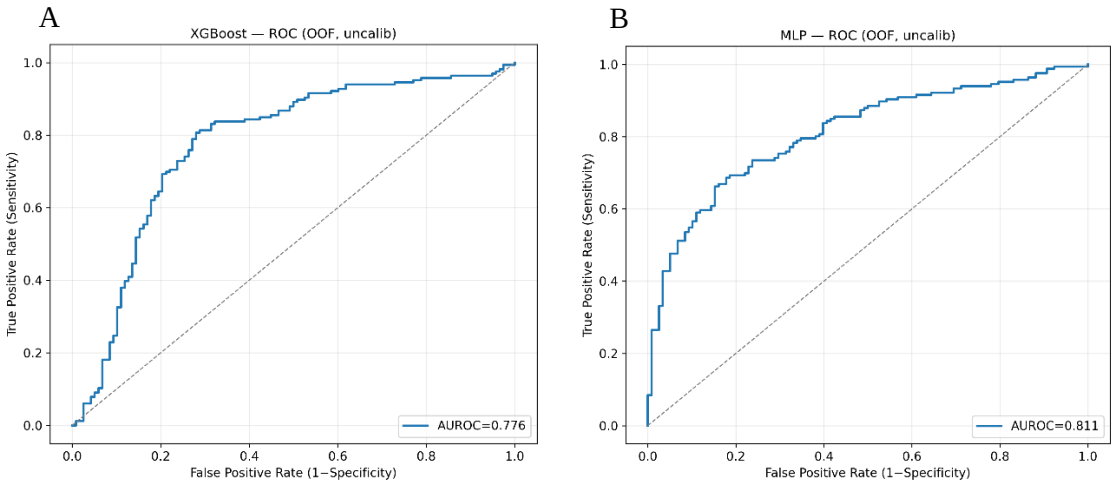
Random Forest'in ROC eğrisi (Şekil 4.7B), FPR≈0–0.2 aralığında hızlı yükselerek 0.85 civarında bir duyarlılık bandına early lift göstermekte, ardından orta FPR bölgesinde (≈0.15–0.25) kısa bir düşüş sergilemektedir. Bu desen, ağaç ensemble'ının bazı alt eşiklerde duyarlılığı hızla kazanabildiğini, fakat özgüllük arttırılırken duyarlılık artışının bir süre marjinal kaldığını gösterir.



Şekil 4.7 (A) SVM'ye ait ham skor ROC eğrisi ve (B) Random Forest'e ait ham skor ROC eğrisi

XgBoost'un ROC eğrisi (Şekil 4.8A), diyagonale görece daha yakın bir profil izlenmektedir yani görece AUROC değeri düşüktür (0,776). Düşük FPR bölgesinde early lift varlığını korumakla birlikte eğri, orta FPR bantlarında daha belirgin bir iç bükey doğrultusunda hareket etmekte, bu da sınıflar arası ayrımın sürdüğünü fakat SVM ve Random Forest kadar keskin olmadığını düşündürmektedir.

MLP'nin ROC eğrisi (Şekil 4.8B), düşük FPR oranlarında hızla yükselerek kısa sürede yüksek duyarlılık düzeylerine ulaşmakta ve 0,1-0,5 FPR bandında daha dengeli bir artışla devam etmektedir. Eğrinin genel olarak grafiğin sol-üst bölgesine yakın seyretmesi ve AUROC=0.811 değerine ulaşması, MLP'nin Random Forest ile benzer düzeyde bütünleşik ayrım gücüne sahip olduğunu göstermektedir. Yaklaşık 0.80 üzeri duyarlılık düzeylerine çıkarken dahi diyagonale belirgin biçimde uzak kalması, ağın sınıflar arasındaki ayrım sinyalini FPR ekseni boyunca tutarlı biçimde koruduğunu düşündürmektedir.



Şekil 4.8 (A) XgBoost'a ait ham skor ROC eğrisi ve (B) MLP'ye ait ham skor ROC eğrisi

Özetle, ROC eğrilerinin görsel analizi ham skor evresinde AlphaMissense ve Logistic Regression'ın belirgin biçimde diyagoale uzak seyrettiğini, SVM, Random Forest ve MLP'nin bu iki modele yakın, dengeli bir ayrım gücü sunduğunu, XGBoost'un ise daha mütevazi bir performans profili çizdiğini göstermektedir. PolyPhen-2 ve SIFT'te Youden noktalarının çok yüksek eşiklere kayması ve eğrilerin diyagonale görece yakın seyretmesi, özgüllük-duyarlılık dengesinin hassaslaştığını ve ham skor düzeyindeki kalibrasyon yanlılıklarının karar davranışını etkileyebildiğini düşündürmektedir.

4.2. Modellerin Kalibrasyon Sonrası Ortak Eşikte Karşılaştırmalı Performansları

Platt Scaling yöntemi ile yapılan kalibrasyon ve model içi hiperparametre ayarları tamamlandıktan sonra, tüm araçlar ile makine öğrenmesi modelleri ortak karar eşiği olan t = 0.50'de yeniden değerlendirildi. Böylece her yöntemin ürettiği sınıf olasılıkları aynı ölçeğe taşınmış, eşik-bağımlı ölçütlerin doğrudan ve yöntemler arası adil biçimde karşılaştırılabilmesi sağlanmıştır. Ayrım gücünü özetleyen AUROC ve AUPRC açısından bakıldığında, kalibrasyon sonrasında da AlphaMissense (AUROC = 0.861; AUPRC = 0,896) ile Logistic Regression (AUROC = 0,852; AUPRC = 0,880) en yüksek düzeyde performans göstermeye devam etmiş; MLP (AUROC = 0,838; AUPRC = 0,865) ve SVM (AUROC = 0,820; AUPRC = 0,811) bu grubu yakından izlemiştir. Random Forest ve XGBoost, sırasıyla AUROC = 0,804 / 0,792 ve AUPRC = 0.830 / 0.800 değerleriyle orta–yüksek bir ayrım gücü bandında yer

alırken, PolyPhen-2 (AUROC = 0,743; AUPRC = 0,761) ve SIFT (AUROC = 0,716; AUPRC = 0,780) kalibre edilmiş ortak eşiğe göre değerlendirildiğinde ayrım performansı bakımından daha alt sırada konumlanmıştır (Tablo 4.2).

Olasılık kalitesi göstergelerinde kalibrasyon sonrasında modellerin değerlerinin birbirine yaklaştığı ve farkların daraldığı görülmüştür. Brier score bakımından en düşük değer Logistic Regression'da (0.156) elde edilmiş, SVM ve MLP (0.158), AlphaMissense (0.160) ve Random Forest (0.162) çok yakın bir aralıkta yer almıştır. ECE (Expected Calibration Error) sonuçları da kalibrasyon etkisini yansıtmaktadır. SIFT (0.038) ve PolyPhen-2 (0.045) ile SVM (0.048) ve MLP (0.049) en düşük sapmaları vermiş, Logistic Regression (0.077) ve AlphaMissense (0.130) ise biraz daha yüksek fakat klinik açıdan kabul edilebilir düzeylerde kalmıştır. Bu görünüm, ayrım gücü görece zayıf olan modellerde dahi (örneğin SIFT) kalibrasyonla olasılık güvenilirliğinin belirgin biçimde iyileştirilebildiğini, buna karşılık ayrım gücü yüksek modellerde örneğin AlphaMissense kalibrasyon hatasının kısmen daha yüksek seyredebildiğini göstermektedir (Tablo 4.2).

Sınıflama metriklerine bakıldığında Random Forest özellikle öne çıkmıştır. Ortak 0,5 eşiğinde F1 = 0,839 ve MCC = 0,607 değerlerine ulaşarak, pozitif ve negatif sınıflar arasında en dengeli ayrımı sağlayan model konumuna gelmiştir. Precision açısından bakıldığında AlphaMissense (0,874) ve Logistic Regression (0,869) en yüksek seçiciliği sunmakta, SVM (0,845) ve MLP (0,832) de bu iki modeli yakından izleyerek yanlış pozitif oranını sınırlayan bir karar profili ortaya koymaktadır. Recall/Sensitivity boyutunda Random Forest (0,849) kalibrasyon sonrasında en yüksek duyarlılığı elde etmiş, SIFT ise ayrım gücü görece düşük olmasına rağmen eşiğin 0,5'te sabitlenmesiyle pozitif sınıfı daha cömert yakalayan bir model hâline gelmiştir (Recall ≈ 0,874). Specificity en yüksek AlphaMissense'de (0,847) gözlenmiş, Logistic Regression (0,839) ve MLP (0,771) bu modeli izlemiştir. Accuracy açısından zirvede Random Forest (0,810) bulunmaktadır, Logistic Regression(0,792), SVM (0,792) ve MLP (0,792) aynı düzeyde, AlphaMissense ise 0,789 düzeyinde performans sergilemiştir (Tablo 4.2).

Genel bir değerlendirme yapıldığında, kalibrasyon sonrasında AlphaMissense ve Logistic Regression'ın ayrım gücünü koruyarak yüksek Precision ve Specificity ile

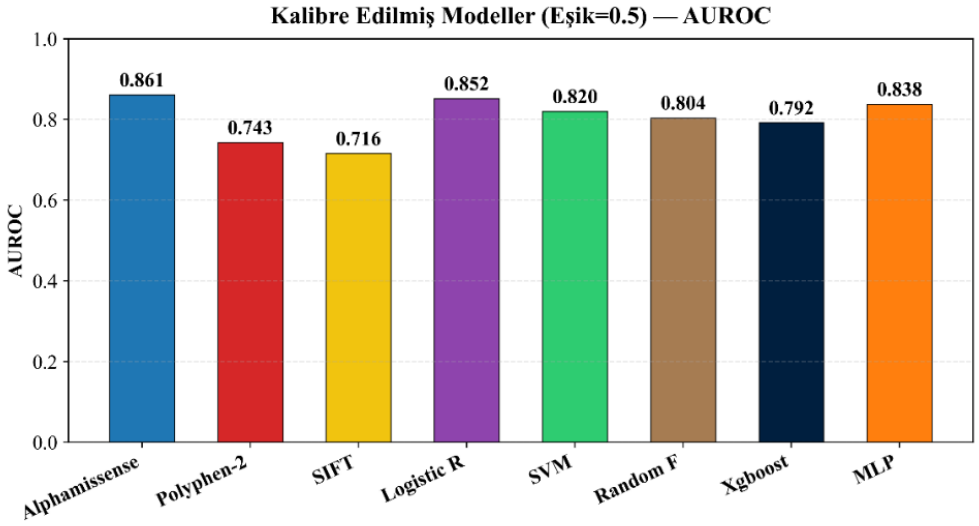
yanlış pozitif riskini sınırlayan bir işletim noktasına yerleştikleri görülmektedir. Buna karşılık, Random Forest en yüksek F1 ve MCC değerleriyle 0,5 ortak eşiğinde yanlış pozitif ve yanlış negatiflerin görece en dengeli dağılımını sağlayan ve pozitif/negatif sınıfları en istikrarlı şekilde ayıran model olarak öne çıkmaktadır. SVM ve MLP, ayrım gücü, Precision/Recall dengesi ve Brier/ECE gibi olasılık kalitesi ölçütleri birlikte ele alındığında, orta-üst düzeyde dengeli bir performans bandı oluşturmaktadır. PolyPhen-2 ve SIFT, AUROC bakımından alt sırada yer alsalar da, özellikle ECE değerlerindeki düşüş ile kalibrasyon sonrasında olasılık güvenilirliğinde belirgin bir iyileşme sergilemişlerdir. Bu bütün tablo, kalibrasyonun yalnızca eşik-bağımlı metrikleri karşılaştırılabilir kılmakla kalmadığını, aynı zamanda modellerin ürettiği olasılıkların klinik açıdan yorumlanabilirliğini ve güvenilirliğini anlamlı ölçüde artırdığını göstermektedir (Tablo 4.2).

Tablo 4.2 Kalibre Edilmiş Modellerin Ortak Eşikte Metrik Performansları

	Alphamissense	Polyphen-2	SIFT	Logistic R	SVM	Random F	Xgboost	MLP
AUROC	**0.861**	0.743	0.716	0.852	0.820	0.804	0.792	0.838
AUPRC	**0.896**	0.761	0.780	0.880	0.811	0.830	0.800	0.865
Brier	0.160	0.199	0.220	0.156	0.158	0.162	0.174	0.158
ECE	0.130	0.045	0.038	0.077	0.048	0.095	0.090	0.049
Accuracy	0.789	0.702	0.653	0.792	0.792	**0.810**	0.768	0.792
F1	0.806	0.756	0.747	0.810	0.816	**0.839**	0.799	0.820
MCC	0.587	0.376	0.256	0.590	0.580	**0.607**	0.524	0.575
Precision	**0.874**	0.725	0.652	0.869	0.845	0.829	0.809	0.832
Recall	0.749	0.790	0.874	0.759	0.789	**0.849**	0.789	0.807
Specificity	**0.847**	0.576	0.339	0.839	0.797	0.754	0.737	0.771

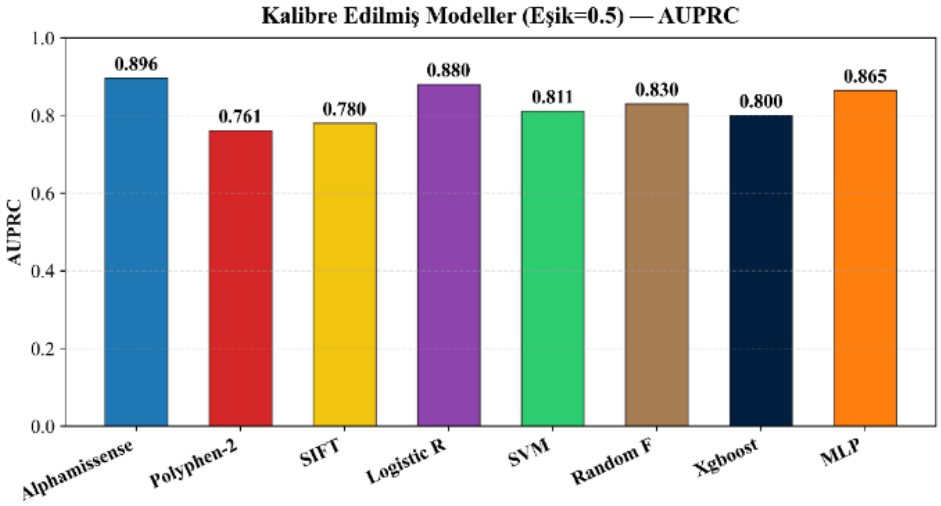
Kalibre edilmiş olasılıklar ortak eşikte (0.5) değerlendirilirken modellerin eşik-bağımsız ayrıştırma gücü AUROC ile özetlenmiştir. Sıralama AlphaMissense (0.861) ve Logistic Regression'ı (0.852) üst kümeye yerleştirirken, MLP (0.838) bu ikiliye

yakın performans göstermektedir; SVM (0.820) üst-orta bantta, Random Forest (0.804) ve XGBoost (0.792) orta bantta, PolyPhen-2 (0.743) ile SIFT (0.716) ise alt bantta konumlanmaktadır. Böylece kalibrasyon sonrası ortamda dahi ayrıştırma bakımından belirgin bir hiyerarşi korunmaktadır (Şekil 4.9A).



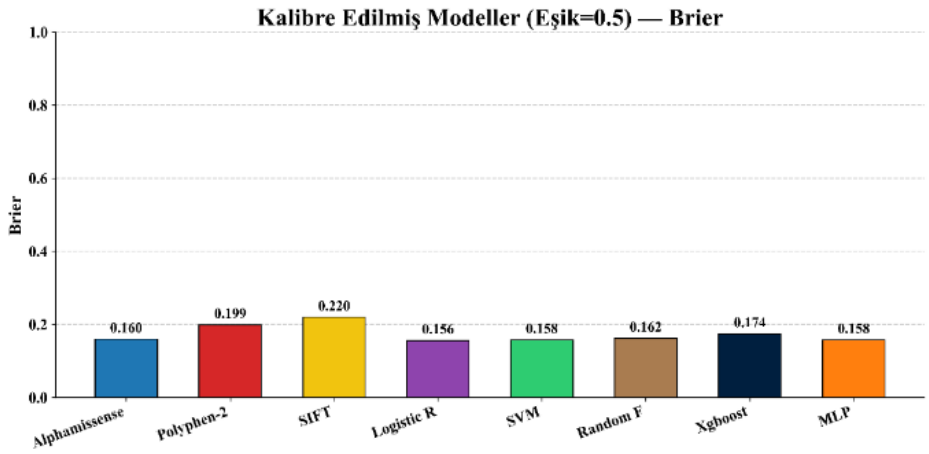
Şekil 4.9A Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) AUROC Değerleri

Kalibre edilmiş modellerin AUPRC performansları bu şekilde sıralanmaktadır. Sınıf dengesizliğine duyarlı olan bu metrik açısından AlphaMissense (0,896) ve Logistic Regression (0,880) değerleri ile en iyi performansı gösterdi. Bu iki modelin hemen ardından MLP (0,865) gelmektedir. Orta bantta Random Forest (0,830) ve SVM (0.811) konumlanırken, XGBoost (0,800) bunların hemen altında yer almaktadır. SIFT (0,780) ve PolyPhen-2 (0,761) ise AUPRC değerleri bakımından en alt grupta kalmıştır. Bu dağılım, AUROC sonuçlarıyla da uyumludur ve kalibrasyon sonrasında özellikle AlphaMissense, Logistic Regression, MLP üçlüsünün pozitif sınıfı yüksek Precision düzeylerinde yakalama kapasitesinin daha güçlü olduğunu, buna karşılık PolyPhen-2 ve SIFT'in ise aynı duyarlılık düzeylerinde daha fazla yalancı-pozitif ürettiklerini düşündürmektedir (Şekil 4.9B).



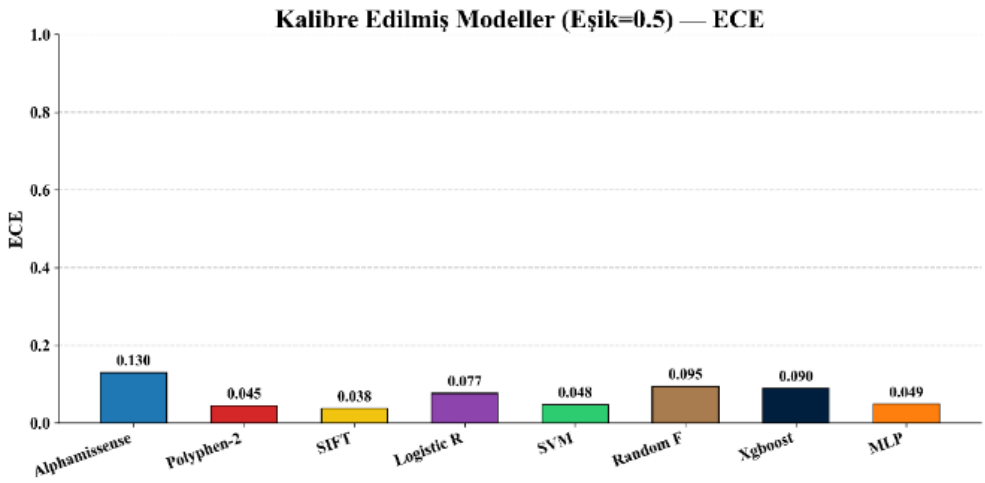
Şekil 4.9B Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) AUPRC Değerleri

kalibrasyon sonrasında ortak eşikte (t=0.5) hesaplanan Brier skorlarına ilişkin sütun grafiği, olasılık tahminlerinin toplam hata yükünü hem kalibrasyon sapması hem de keskinlik eksikliğini nicel olarak karşılaştırmaktadır. Brier skoru 0-1 arasında değerlendirilir ve skor 0'a yaklaştıkça olasılık kalitesi artar. Bu ölçekte en düşük hatayı Logistic Regression (0.156) üretmiş; onu SVM ve MLP (her ikisi de 0.158) ile AlphaMissense (0.160) çok dar bir aralıkla izlemiştir. Random Forest 0.162 ile bu üst kümeye yakın seyretmiştir. XGBoost'un 0.174'lük skoru orta bantta yer alırken, PolyPhen-2 (0.199) ve özellikle SIFT (0.220) daha yüksek hata düzeyleri göstermiştir. Bu tablo, Logistic Regression ve AlphaMissense gibi ayrıştırma gücü yüksek olan modellerin yalnızca sınıfları ayırma başarısı açısından değil, ürettikleri olasılıkların gerçek veriler ile uyumu bakımından da daha güvenilir olduklarını göstermektedir. Kısaca, ortak karar eşiğinde en iyi olasılık kalitesi Logistic Regression ile SVM ve MLP üçlüsünde yoğunlaşmakta; AlphaMissense de yüksek ayrım gücünü görece düşük bir Brier skoru ile destekleyerek güçlü bir genel performans profili sunmaktadır(Şekil 4.9C).



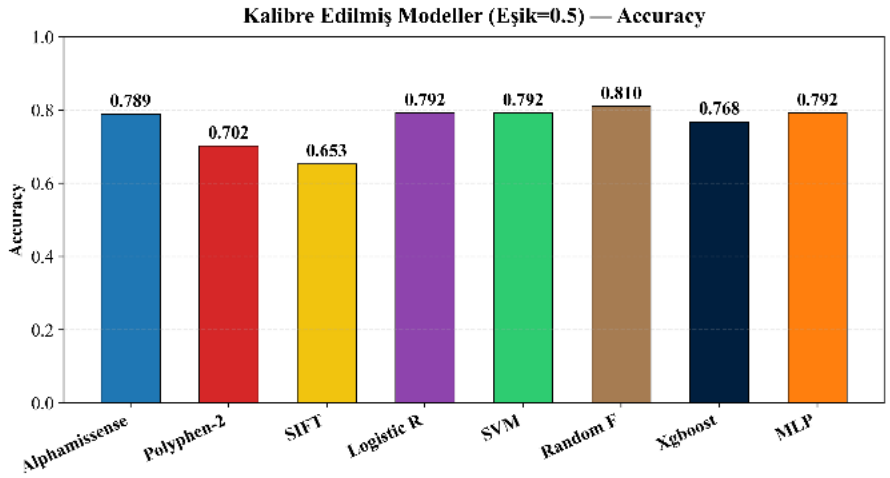
Şekil 4.9C Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Brier Skoru Düzeyleri

Kalibre edilmiş modeller için Expected Calibration Error (ECE) dağılımı, modellerin ürettiği olasılıkların ClinVar etiketlerine göre gözlenen pozitiflik oranlarıyla ne ölçüde örtüştüğünü özetlemektedir.ECE değeri 0'a ne kadar yakın olursa o kadar iyi kalibrasyon anlamına gelir. BU ölçüte göre, en düşük değer SIFT'te (0,038) elde edilmiştir. bunu PolyPhen-2 (0,045), SVM (0,048) ve MLP (0,049) izlemiştir. Logistic Regression orta düzeyde bir sapma göstermiş (0,077), Random Forest (0,095) ve XGBoost (0,090) ise görece daha yüksek ECE değerleri vermiştir. AlphaMissense'in ECE değeri 0.130 olup, kümenin geri kalanına kıyasla yüksektir, bu da modelin AUROC ve AUPRC bakımından güçlü ayrım gücüne sahip olmasına rağmen, olasılık ölçeğinde özellikle bazı aralıklarda sistematik aşırı veya yetersiz güven örüntülerinin sürdüğüne işaret etmektedir. Öte yandan SIFT ve PolyPhen-2'nin düşük ECE değerleri, bu araçların ayrıştırma gücündeki sınırlılıklarını örneğin ROC ve AUPRC sonuçlarını tek başına telafi etmemekte, ancak kalibrasyon sonrasında verdikleri olasılıkların, ClinVar'da gözlenen frekanslarla daha uyumlu hâle geldiğini göstermektedir. Genel olarak ECE sonuçları, kalibrasyon kalitesine ilişkin göreli sıralamayı ortaya koymakt, Brier skorları ve ayrım metrikleriyle birlikte değerlendirildiğinde Logistic Regression, SVM ve MLP üçlüsünün kalibrasyon ile performans arasında en dengeli profili sunduğunu, AlphaMissense'in ise yüksek ayrım gücü ve orta düzey kalibrasyon kombinasyonuyla ayrı bir konumda yer aldığı gözlemlenmektedir (Şekil 4.9D).



Şekil 4.9D Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) ECE Değerleri

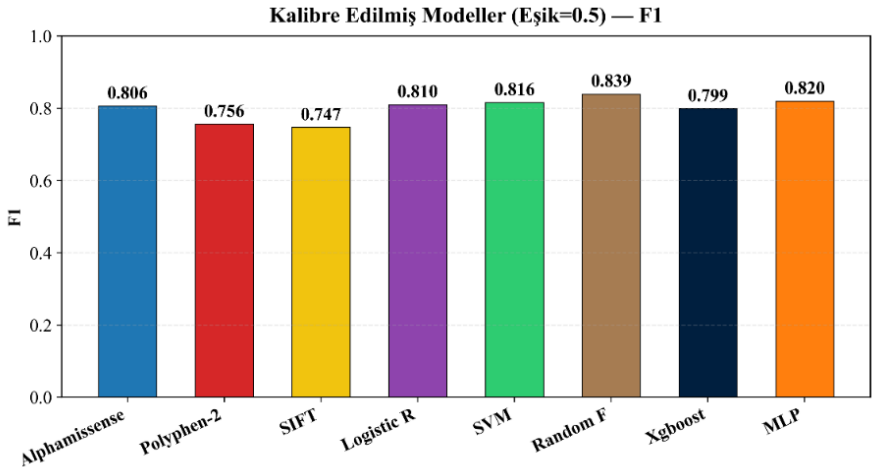
Ortak eşikte değerlendirilen modeller Accuracy açısından değerlendirildiğinde en yüksek performansın 0,810 değeri ile Random Forest olduğu gözlemlenmiştir. Ardından Logistic Regression, SVM ve MLP'nin her birinin 0,792 ile aynı düzeyde performans gösterdiği görülmektedir. AlphaMissense 0,789 ile bu kümeye çok yakın performans sergilerken, XGBoost 0,768, PolyPhen-2 0,702 ve SIFT 0,653 ile daha geride kalmıştır. Accuracy değerlerinin bu sıralaması diğer karar metrikleriyle, özellikle Precision, Recall/Sensitivity ve Specificity değerleriyle uyumludur. Örneğin Random Forest'ın yüksek doğruluğu, Recall = 0,849 ve Precision = 0,829 değerleriyle desteklenmektedir. Buna karşılık SIFT'in düşük doğruluk değeri (0,653), pozitif sınıfta yüksek duyarlılık (Recall = 0,874) elde etmesine rağmen, bunu düşük özgüllük (Specificity = 0,339) ile dengeleyememesinden kaynaklanmaktadır. Bununla birlikte doğruluğun, sınıf dağılımına duyarlı bir özet istatistik olduğu göz önünde bulundurulmalıdır. "Hangi model daha fazla gözlemi doğru sınıflandırdı?" sorusuna ilişkin ikili karşılaştırmaların istatistiksel olarak değerlendirilmesi için, 0,5 ortak eşiğinde McNemar testleri uygulanmış ve sonuçlar Tablo 4.5'te sunulmuştur. Genel olarak, kalibrasyon sonrası ortak eşikte elde edilen doğruluk değerleri, AlphaMissense, Logistic Regression, SVM ve MLP'nin benzer düzeyde yüksek performans gösterdiğini, Random Forest'ın bir miktar avantajlı olduğunu, PolyPhen-2 ve SIFT'in ise bu grup modellerine kıyasla belirgin biçimde daha düşük doğruluk sergilediğini ortaya koymaktadır (Şekil 4.9E).



Şekil 4.9E Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Accuracy Skorları

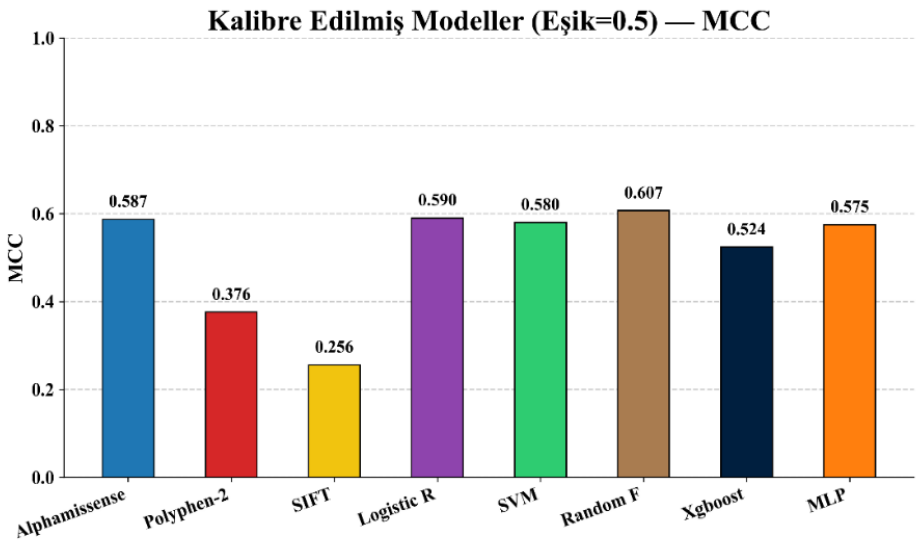
Kalibre edilmiş modeller arasında em yüksek F1 değerini Random Forest'in 0,839 ile sağladığı görülmektedir. Bunu MLP (0,820), SVM (0,816) ve Logistic Regression'ın (0,810) birbirine çok yakın değerlerle izlemiştir. AlphaMissense 0,806 ile bu grubun hemen gerisinde yer alırken, XGBoost 0,799, PolyPhen-2 0,756 ve SIFT 0,747 ile daha düşük F1 değerleri üretmiştir. F1 skoru, Precision ve Recall'un harmonik ortalaması olduğundan, Random Forest'ın üstünlüğü yüksek duyarlılık (Recall = 0,849) ile görece yüksek kesinliğin (Precision = 0,829) birlikte sağlanmasına dayanmaktadır. Benzer şekilde MLP (Recall = 0,807; Precision = 0,832) ve SVM (Recall = 0,789; Precision = 0,845) de yanlış pozitif ve yanlış negatifleri birlikte dikkate alan, görece dengeli bir hata profili sergilemektedir. Buna karşılık SIFT'in düşük F1 değeri (0,747), pozitif sınıfta yüksek duyarlılık (Recall = 0,874) elde etmesine rağmen, bunu çok düşük özgüllük ve daha düşük kesinlikle (Precision = 0,652) dengeleyememesinden kaynaklanmaktadır. PolyPhen-2 için de benzer bir tablo söz konusudur, görece düşük Recall (0,790) ve daha düşük Precision (0,725) ile bir araya gelince F1 skorunu sınırlı düzeyde tutmaktadır.

Genel olarak, kalibrasyon sonrasında ortak eşikte elde edilen F1 sıralaması, klinik senaryolarda yanlış pozitif ve yanlış negatif maliyetlerinin birlikte önem taşıdığı durumlarda Random Forest, MLP, SVM ve Logistic Regression'ın daha elverişli bir denge sunduğunu ortaya koymaktadır (Şekil 4.9F).



Şekil 4.9F Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) F1 Skorları

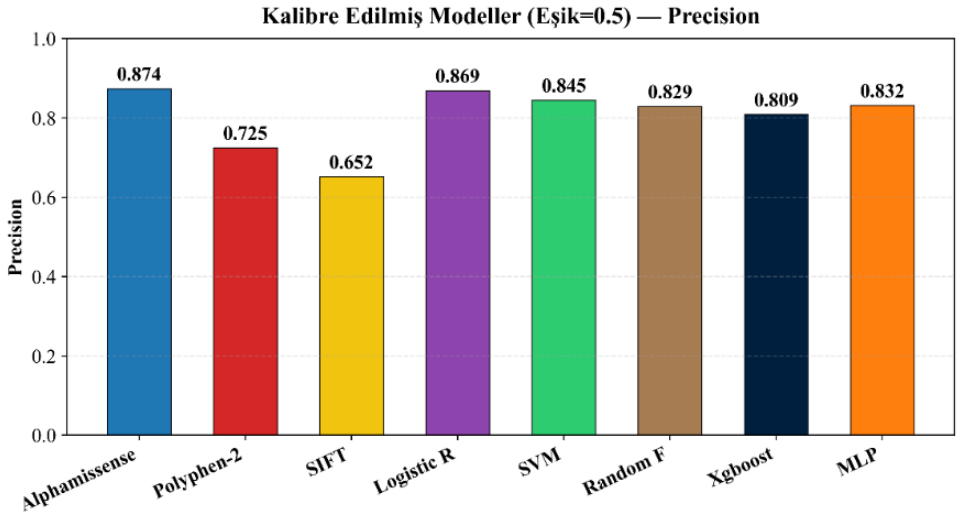
Kalibre edilmiş modellerden elde edilmiş Matthews korelasyon katsayısı (MCC) sonuçları, sınıf dengesizliğine duyarlı tek bir özet ölçü olarak modellerin genel tahmin kalitesini karşılaştırmaya olanak sağlamaktadır. Şekilde görüldüğü üzere en yüksek MCC değerini 0,607 ile Random Forest elde etmiştir. Bunu, birbirine oldukça yakın değerlerle Logistic Regression (0,590), AlphaMissense (0,587), SVM (0,580) ve MLP (0,575) takip etmektedir. XGBoost orta düzeyde bir performans sergilemiş (0,524), PolyPhen-2 (0,376) ve SIFT (0,256) ise belirgin biçimde daha düşük MCC değerleri vermiştir. Bu görünüm, önceki ayrıntılı metriklerle uyumludur. Random Forest'ın hem yüksek Recall (duyarlılık) hem de makul düzeyde Precision (kesinlik) sağlaması ve yanlış pozitif ile yanlış negatif hatalarını görece dengeli dağıtması MCC değerini yukarı taşımaktadır. Buna karşılık SIFT ve PolyPhen-2'de gözlenen özgüllük ve Precision kayıpları, bu modellerin hata desenini daha dengesiz hâle getirerek MCC'nin düşük kalmasına yol açmaktadır. Genel olarak MCC sonuçları, ortak 0,5 karar eşiğinde en dengeli hata profilinin Random Forest ve Logistic Regression, SVM, MLP tarafından sağlandığını, araç tabanlı yöntemlerden AlphaMissense'in de bu gruba oldukça yakın bir performans sergilediğini, buna karşılık PolyPhen-2 ve SIFT'in, hata örüntülerinin dengesizliği nedeniyle görece zayıf kaldığını göstermektedir (Şekil 4.9G).



Şekil 4.9G Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) MCC Skorları

Kalibrasyon sonrası modeller arasında en yüksek precision skoruna sahip model 0,874 ile AlphaMissense olmuştur. İkinci olarak Logistic Regression 0,869 değeri ile AlphaMissense yakın bir performans sergileyerek bu iki model, yanlış pozitif oranını sınırlama açısından belirgin bir üstünlük sergilemiştir. SVM (0,845), MLP (0,832), Random Forest (0,829) ve XGBoost (0,809) değerleri ile iyi sayılabilecek bir Positive Predictive Value (PPV) bandında gruplanırken, PolyPhen-2 (0,725) ve özellikle SIFT (0,652) skorları ile belirgin biçimde daha düşük pozitif öngörü değerleri sergilemiştir.

Bu tablo, specificity metrikleriylede uyumludur. Örneğin SIFT'te yüksek duyarlılığa (Recall = 0,874) karşın specificity değerinin oldukça düşük olması (0,339), beklendiği üzere PPV'nin de düşmesine yol açmıştır. PolyPhen-2'de ise benzer fakat daha ılımlı bir duyarlılık–özgüllük dengesi gözlenmektedir (Recall = 0,790; Specificity = 0,576). Buna karşılık AlphaMissense 0,847 ve Logistic Regression 0,839 specificity değerlerini görece yüksek düzeyde tutarken Recall değerini klinik olarak kabul edilebilir sınırlar içinde koruyarak yanlış pozitif yükünü etkin biçimde azaltmaktadır. Precision değerlerine bakıldığında en iyi performansa sahip modellerin AlphaMissense ve Logistic Regression olduğu görülmektedir. (Şekil 4.9H).



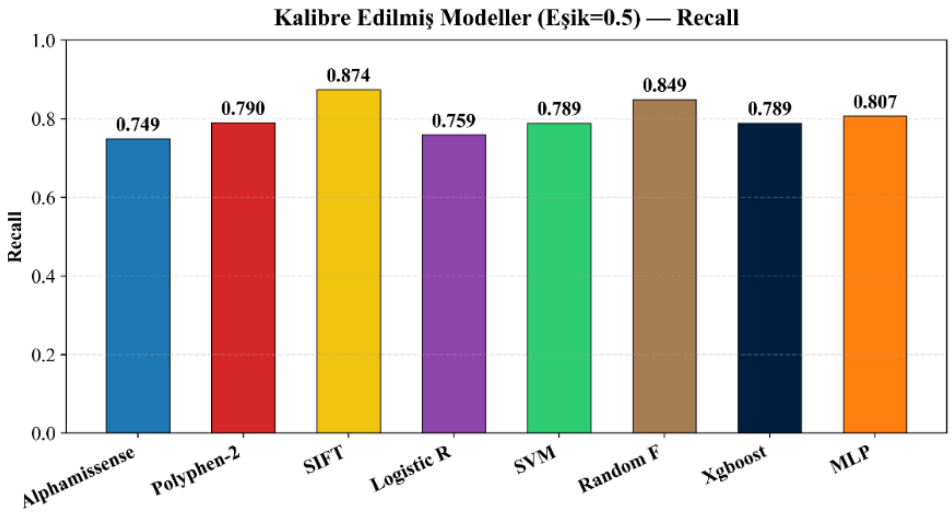
Şekil 4.9H Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Precision Skorları

Tüm modeller kalibre edildikten sonra ortak eşikte elde edilen Recall değerleri incelendiğinde, SIFT'in belirgin biçimde en yüksek değere sahip olduğu görülmektedir (Recall = 0,874). SIFT'i Random Forest (0,849) ve MLP (0,807) izlemektedir. PolyPhen-2 (0,790), SVM (0,789) ve XGBoost (0,789) değerleri ile iyi denebilecek bir düzeyde bulunurken, Logistic Regression (0,759) ve AlphaMissense (0,749) duyarlılık açısından görece daha düşük düzeyde kalmaktadır. Ancak bu tablo yorumlanırken, duyarlılığın tek başına değil, kesinlik (Precision) ve bileşik metrikler (F1, MCC) ile birlikte değerlendirilmesi gerekir. SIFT'in pozitif sınıfı yakalama oranı yüksek olmakla birlikte, bu durum Precision = 0,652 ve MCC = 0,256 hesaba katıldığında SIFT'in belirgin bir yanlış pozitif yüküne sahip olduğu sonucu çıkmaktadır. . Bu profil, klinik iş akışında daha çok, geniş tarama/eleme amaçlı kullanım için uygundur. Buna karşılık doğrulayıcı adımlarda ek filtreleme ve ikincil inceleme gerektiren bir karaktere işaret etmektedir. Random Forest ise yüksek duyarlılığı (Recall = 0,849) görece yüksek kesinlik (Precision = 0,829) ile birlikte sunduğu için, F1 = 0,839 ve MCC = 0,607 değerleriyle daha dengeli bir hata dağılımı üretmektedir.

AlphaMissense, duyarlılık sıralamasında alt grupta yer almakla birlikte, Precision = 0,874 ve güçlü AUROC = 0,861 değeriyle, yanlış pozitiflerden kaçınmaya daha fazla ağırlık veren, daha seçici (özgüllüğü yüksek) bir çalışma biçimi sergilemektedir. bu

nitelik, hatalı pozitiflerden kaçınmanın öncelikli olduğu doğrulayıcı klinik bağlamlarda avantaj yaratabilir.

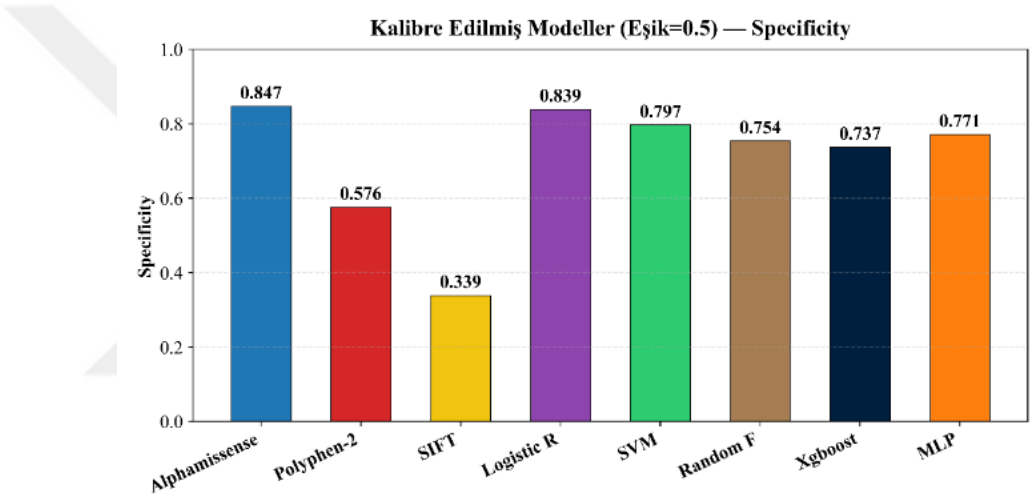
Genel olarak bu grafiğe bakıldığında, pozitif sınıfın olabildiğince az kaçırılmasının istendiği durumlardaki tercihin SIFT ve Random Forest eksenine kayabileceği, buna karşılık yanlış pozitif kontrolünün daha kritik olduğu senaryolarda AlphaMissense ve Logistic Regression çizgisinin öne çıkabileceği anlaşılmaktadır (Şekil 4.9I).



Şekil 4.9I Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Recall Skorlar

Kalibre edilmiş modellerin specificity değerlerini gösteren sütun grafiği, 0,5 ortak eşik değerinde her modelin gerçekte benign olan varyantları ne ölçüde doğru dışlayabildiğini ortaya koymaktadır. Grafiğe göre en yüksek specigicity AlphaMissense'te (0,847) ve onu çok yakın izleyen Logistic Regression'da (0,839) görülmektedir. Bu iki model benign varyantları yanlışlıkla patojenik etiketleme olasılığını en düşük düzeyde tutmaktadır. SVM (0,797) ve MLP (0,771) specificity bakımından görece yüksek düzeyde seyrederken, andom Forest (0,754) ve XGBoost (0,737) biraz daha alt bantta yer almaktadır. PolyPhen-2 (0,576) ve özellikle SIFT (0,339) ise benignleri doğru dışlama konusunda belirgin zayıflık göstermektedir. Bu örüntü, aynı eşikteki duyarlılık (Recall) değerleri ile birlikte okunduğunda modellerin karar profillerini daha net ortaya koyar. SIFT, Recall=0,874 ile patojen sınıfını çok yüksek oranda yakalarken özgüllüğünün 0,339'a düşmesi, çok sayıda benign varyantın yalancı-pozitif olarak işaretlendiğini göstermektedir. Random

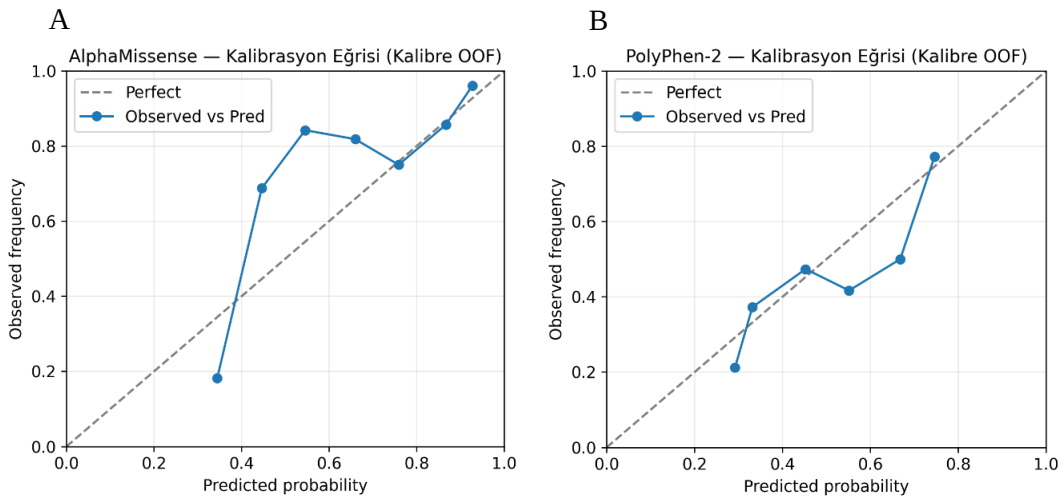
Forest, Recall=0,849 ile yüksek yakalama oranını özgüllüğü 0,754 düzeyinde tutarak desteklemekte; yani benignleri makul ölçüde korurken patojenleri kaçırma riskini de sınırlamaktadır. AlphaMissense (Recall=0,749) ve Logistic Regression (Recall=0,759) ise duyarlılıktan kısmen feragat ederek özgüllüğü yüksek tutar; bu durum, yanlış pozitiflerin en aza indirilmeye çalışıldığı doğrulayıcı klinik bağlamlar için daha temkinli bir karar davranışına işaret eder. Genel olarak şekil, kalibrasyon sonrasında ortak eşikte benign sınıfı açısından en "temiz" ayrımın AlphaMissense ve Logistic Regression tarafından sağlandığını; PolyPhen-2 ve özellikle SIFT'in yalancı-pozitifleri sınırlamada belirgin dezavantaj taşıdığını göstermektedir (Şekil 4.9J).



Şekil 4.9J Kalibre Edilmiş Modellerin Ortak Eşikte (0,5) Specificity Skorları

Kalibrasyon sonrasında kalibrasyon eğrilerine gelindiğinde AlphaMissense, yaklaşık 0,35 olasılık tahmini civarında over confidence sergilemiş, yani gerçek veriden daha yüksek olasılık tahmini verdiği gözlemlenmiştir. 0,4-0,7 tahmin bandında ise under confidence sergilediğini, yani verdiği olasılık tahminlerinin, gözlenen pozitif oranını (empirik prevalans) sistematik olarak olduğundan düşük yansıttığını göstermektedir. Buna karşılık 0.7–0.9 bantlarında noktasal sapmalar azalmakta ve eğri diyagonale yaklaşmaktadır. Bu durum bölgede anlamlı bir iyileşme sağladındığı ve tahmin edilen olasılıkların gözlenen frekanslarla uyumunun arttığını düşündürür. En yüksek olasılık sınıfında (≈0.9–1.0) gözlenen hafif yukarı sapma, bu dilimde gerçek pozitif oranının tahmin edilen olasılıktan daha yüksek olabildiğini, yani ihtiyatlı (conservative) bir kalibrasyon çıktığını gösterir (Şekil 4.10A).

PolyPhen-2'ye bakıldığında ise, tüm olasılık aralığında diyagonale daha yakındır. 0.3–0.8 bantlarında çizginin etrafında simetrik, küçük genlikli salınımlar görülmektedir. Bu profil, kalibrasyon sonrası artık sistematik bir aşırı-güven (over-confidence) ya da aşırı-ihtiyat (under-confidence) önyargısının kalmadığına işaret eder. Özetle PolyPhen-2, kalibrasyon sonrasında olasılıkların yorumsal doğruluğunu kayda değer ölçüde iyileştirmiş; fakat ayrıştırma kapasitesi sınırlı olduğu için düşük/orta olasılık dilimlerinde duyarlılık–seçicilik dengesi Alphamissense kadar avantajlı değildir (Şekil 4.10B).

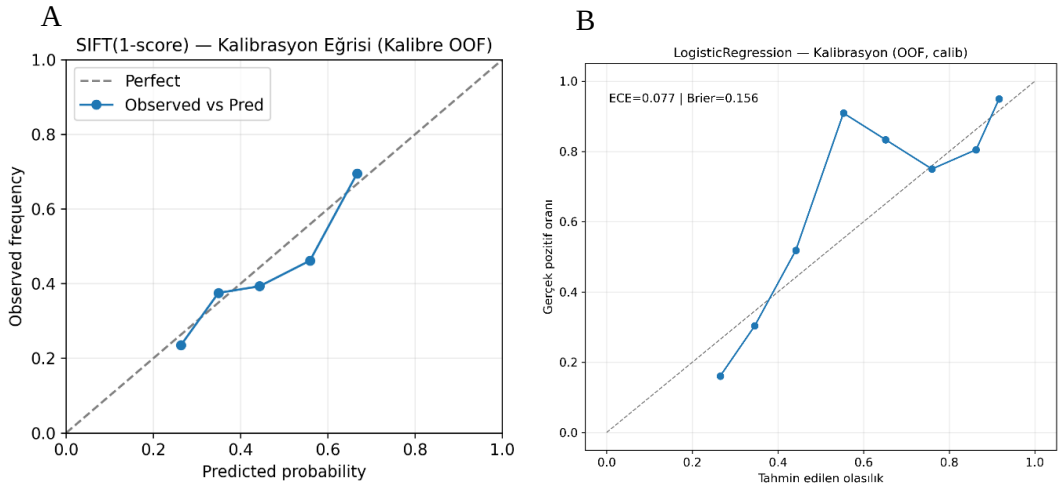


Şekil 4.10 (A) Kalibrasyon sonrası AlphaMissenseye ait kalibrasyon eğrisi, (B) PolyPhen-2'ye ait kalibrasyon eğrisi

SIFT'in kalibrasyon sonrası güvenilirlik eğrisi, ideal diyagonali genel olarak takip etmektedir.0,25 noktasında ve 0,4-0,6 aralığında model over confidence sergilemiş, 0,35 ve 0,7 noktalarında under confidence sergilemiştir. Uç olasılık bölgelerinde veri seyrekliği nedeniyle nokta belirsizliği artmakla birlikte, genel eğilim düz ve düzenlidir. Bu örüntü, kalibrasyon hatasının sayısal özetleriyle de uyumludur: SIFT'in ECE'si düşük (≈0.038) ve Brier skoru görece yüksek (≈0.220) kalmıştır; yani olasılıklar ortalamada iyi hizalanırken, ayrıştırma gücü ve kalan sınıflandırma hatası toplam karesel hatayı nispeten yukarıda tutmaktadır (Şekil 4.11A).

Logistic Regression'ın kalibre edilmiş güvenilirlik eğrisi, diyagonale genel olarak yakın olmakla birlikte zikzak çizmiştir. Genelde uç kısımlarda diyagonal çizgiye daha yakın seyrederken 0,4-0,7 aralığında under confidence gözlemlenmektedir. Bu

davranış, modelin klinik olarak anlamlı olasılık aralıklarında gerçek pozitif oranlarını büyük ölçüde doğru tahmin ettiğini gösterir. Nicel olarak, ECE=0.077 ve Brier=0.156 değerleri kalibrasyon uyumunun iyi, toplam hata bileşenlerinin yani hem ayrım hem kalibrasyon ise SIFT'e kıyasla daha düşük olduğunu teyit etmekedir (Şekil 4.11B).

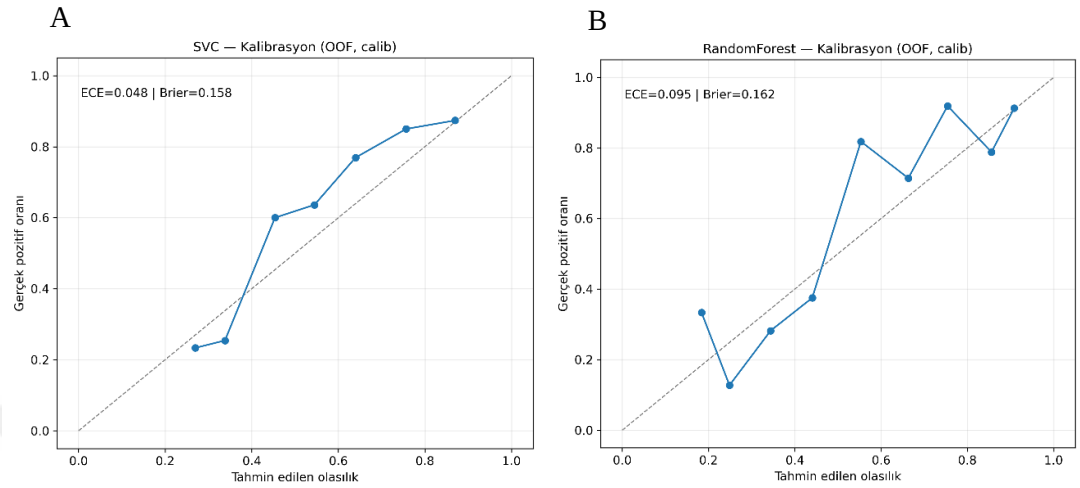


Şekil 4.11 (A) Kalibrasyon sonrası SIFT'e ait kalibrasyon eğrisi, (B) Logistic Regression'a ait kalibrasyon eğrisi

SVM'nin kalibre edilmiş güvenilirlik eğrisi, referans diyagonale genel olarak yakın seyrederek olasılıkların iyi hizalandığını göstermektedir. 0,25-0,4 aralığında eğri diyagonalin biraz altında kalmakta; bu durum modelin bu bantta temkinli davrandığını yani hafif under confidence sergilediğini göstermektedir. Olasılık 0.4–0.9 bandına çıktığında eğri diyagonalin üzerine geçmekte daha sonra ve 0.75–0.85 civarında yeniden diyagonale yaklaşmaktadır. Bu da 0,4-0,9 aralığında düşük over confidence eğiliminin ardından hizalanmanın düzeldiğini işaret eder. Sayısal göstergeler bu görsel yorumu desteklemektedir: ECE=0.048 ve Brier=0.158 değerleri, 0.5 eşikte SVC'nin kalibrasyon kalitesinin tatmin ve sınıflandırma hatasının düşük düzeyde olduğunu göstermektedir (Şekil 4.12A).

Random Forest'ın kalibre edilmiş eğrisi, düşük olasılıklarda yani 0,25-0,5 bandında over confidence sergilemektedir. 0,5-0,8 bandında eğri diyagonalin üzerinde kalarak pozitif oranları olduğundan düşük öngörme eğilimi yani under confidence gösterirken, 0.8 civarında yeniden diyagonale yaklaşmakta ve 0.85–0.9 bandında

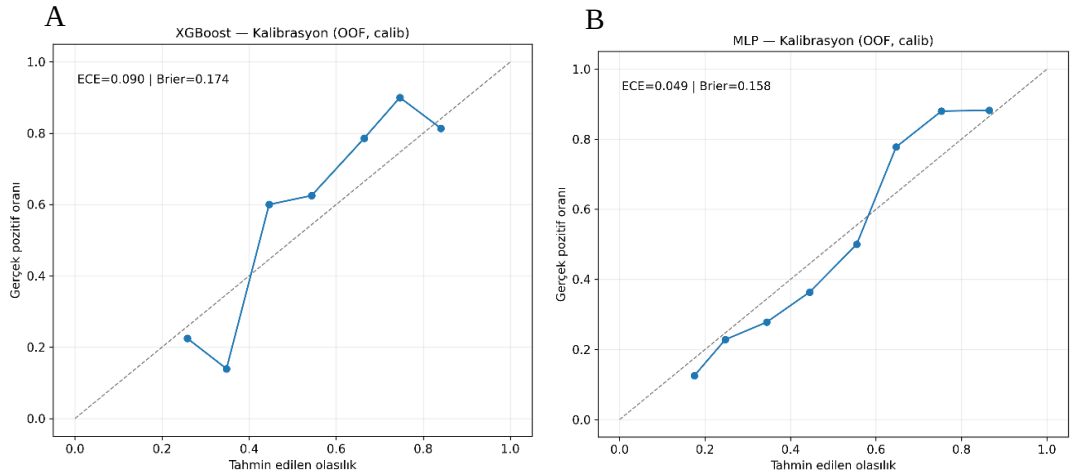
yeniden over confidence sergilemektedir. Bu dalgalı yapı, SVM'ye göre daha büyük ECE=0.095 ve Brier=0.162 ile uyumludur (Şekil 4.12B).



Şekil 4.12 (A) Kalibrasyon sonrası SVM'ye ait kalibrasyon eğrisi, (B) Random Forest'e ait kalibrasyon eğrisi

XGBoost'a ait kalibrasyon eğrileri genel olarak diyagonalin yakınında ilerlese de 0,4-0,8 bandında under confidence sergilediği görülmektedir. model bu aralıkta düşük olasılık verirken gözlenen pozitif oranı daha yüksektir. 0,2-0,4 arasında ise over confidence sergilemiştir. bu karışık örüntü ECE'nin görece yüksek kalmasına (ECE≈0.090) karşılık Brier skorunun orta düzeyde (≈0.174) seyretmesiyle uyumludur. (Tablo 4.2). Özetle XGBoost kalibrasyon sonrasında geniş aralıkta makul uyum gösterse de genel olarak dalgalı bir seyir izlemektedir (Şekil 4.13A).

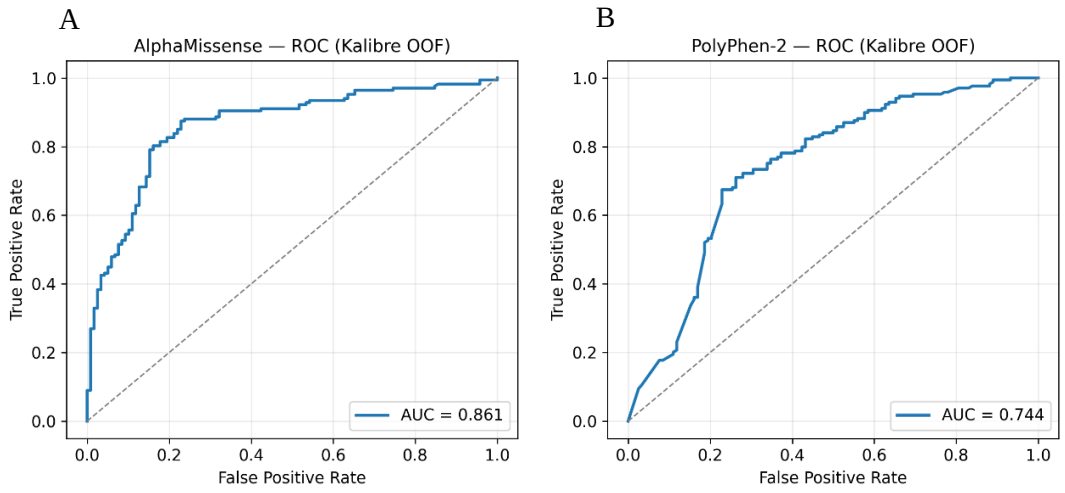
MLP'nin kalibrasyon eğrisi düşük-orta olasılıklarda diyagonalin altında başlayıp 0.6'dan sonra hızla üstüne çıkar; bu, düşük olasılık dilimlerinde kısmi aşırı güvene, yüksek olasılık dilimlerinde ise ihtiyatlılığa işaret eden, fakat neticede hataların büyük bölümünü telafi eden dengeli bir profil sunar. Eğrinin 0.7–0.9 bandında diyagonale tekrardan yaklaşması ve 0,9 noktasında neredeyse bire bir hizalanması, bu olasılık noktasında üretilen pozitif etiketlerin güvenilirliğinin yüksek olduğunu gösterir. Bu görsel bulgu, ECE'nin düşük (≈0.049) ve Brier'in elverişli (≈0.158) değerleriyle tutarlıdır (Tablo 4.2). Kısacası, MLP kalibrasyon sonrasında olasılıkların isabetliliğini artırmış ve istikrarlı bir güvenilirlik düzeyi elde etmiştir (Şekil 4.13A).



Şekil 4.13 (A) Kalibrasyon sonrası XgBoost'a ait kalibrasyon eğrisi, (B) MLP'ye ait kalibrasyon eğrisi

Kalibrasyon sonrası AlphaMissense'nin ROC eğrisine bakıldığında sol-üst köşeye yakın, belirgin bir early lift profili sergilemektedir. Yanlış pozitif oranı (FPR) ≈0–0.15 aralığında duyarlılık hızla 0.85 bandına yükselmekte, eğri orta bantta da (FPR ≈0.2–0.5) konveksitesini korumaktadır. AUROC değeri 0,861'dir. Kalibrasyon, ROC alanını teorik olarak değiştirmese de olasılıkların monoton dönüşümü, buradaki eğrinin düzgünleşmesi, modelin pozitif sınıfı düşük FPR bölgelerinde tutarlı biçimde yakalayabildiğini göstermektedir (Şekil 4.14A).

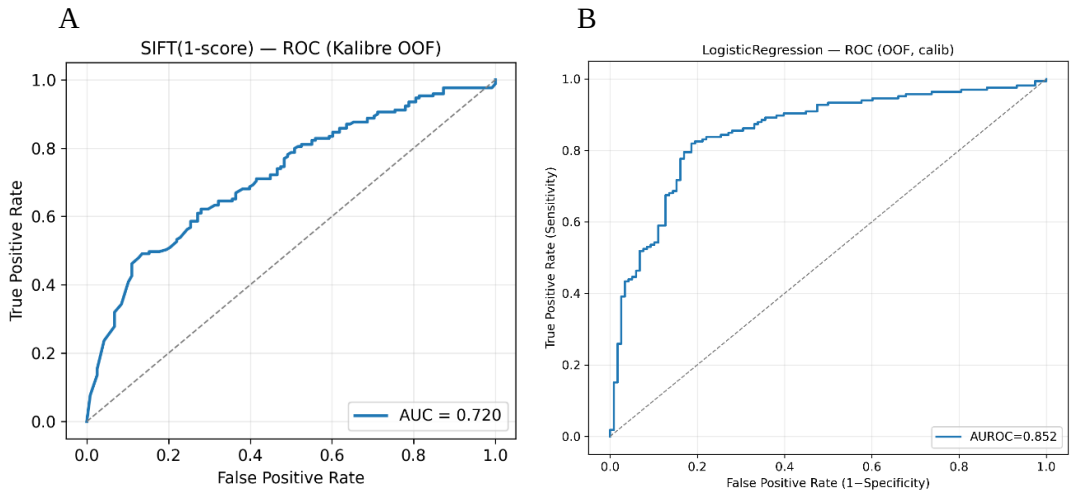
PolyPhen-2'ye bakıldığında ROC eğrisi, düşük FPR bölgesinde daha ılımlı bir yükselişle diyagonale görece yakın seyretmektedir, kalibre edilmiş ayrım gücü AUROC=0.744 olarak ölçülmüştür. Düşük yanlış pozitif oranlarında duyarlılığın sınırlı artışı, modelin sınıflar arası skor sıralamasında AlphaMissense ve Logistic Regression'a kıyasla daha zayıf bir ayrım yaptığına işaret eder. Bununla birlikte kalibrasyon adımı sayesinde eğri, orta FPR bandında (≈0.2–0.5) daha iyi bir konveksite kazanmış, bu durum, Tablo 4.2'deki AUPRC artışı ve ECE düşüşüyle birlikte değerlendirildiğinde, olasılıkların güvenilirliğinin iyileştiğini fakat sıralama gücünün sınırlı kaldığını göstermektedir (Şekil 4.14B).



Şekil 4.14 (A) Kalibrasyon sonrası AlphaMissense'e ait ROC eğrisi, (B) PolyPhen-2'ye ait ROC eğrisi

SIFT'in kalibrasyon sonrası ROC profili diyagonale en yakın çizgilerden biridir ve AUROC değeri 0.720'dir. Düşük FPR bölgesinde eğri nispeten dik yükselmekte, orta bantta ise kademeli bir artış sergilemektedir. Bu davranış, SIFT'in kalibre edilmiş durumda dahi pozitifleri, özellikle özgüllükten fazla ödün vermeden yakalamakta zorlandığını düşündürür. Tablo 4.2'de görülen yüksek Recall ile düşük Specificity dengesizliğinin, ROC eğrisinin şekline yansıdığı görülmektedir: model, daha çok pozitifleri toplama yönünde çalışırken yanlış-pozitifleri de artırmaya eğilimlidir (Şekil 4.15A).

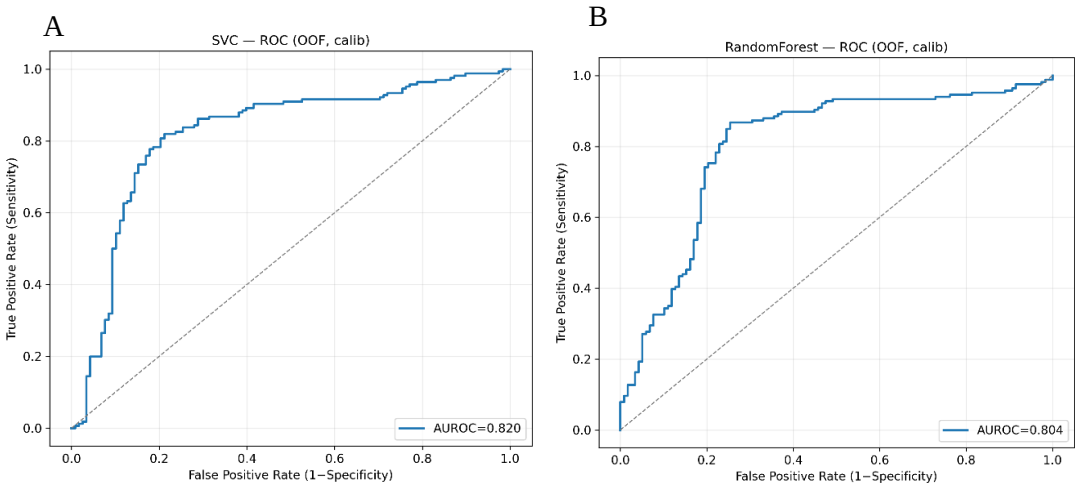
Kalibrasyon sonrasında Logistic Regression'ın ROC eğrisi, düşük yanlış-pozitif oranlarında hızlı bir yükseliş ve köşe yakınında belirgin konveks bir profil göstermiş, AUROC değeri 0.852 ile yüksek düzeyde doğrulanmıştır. Eğrinin erken bölgede dikleşmesi, özgüllükten fazla ödün vermeden duyarlılığın hızla kazanıldığını, orta bantta yani FPR yaklaşık 0,10–0,40 bandında çizginin diyagonalin belirgin üstünde seyretmesi, farklı klinik tercih noktalarında da modelin pozitif ve negatif sınıfları ayırma üstünlüğünü koruduğunu göstermektedir (Şekil 4.15B).



Şekil 4.15 (A) Kalibrasyon sonrası SIFT'e ait ROC eğrisi, (B) Logistic Regression'a ait ROC eğrisi

SVM'nin kalibre ROC profili, düşük FPR bölümünde yaklaşık 0–0.15 düzeyinde belirgin bir early lift sergilemiş, genel ayrım performansı AUROC=0.820 olarak hesaplanmıştır. Eğrinin yaklaşık 0.15–0.50 FPR aralığında konveksitesini koruyarak diyagonalin üzerinde seyretmesi, modelin yanlış pozitif artışı altında dahi duyarlılıktaki artışı sürdürebildiğini göstermektedir. Bu yapı, kalibrasyon sonrasında elde edilen görece düşük Brier ve ECE değerleriyle birlikte değerlendirildiğinde, SVM'in hem ayrıştırma hem de olasılık güvenilirliği bakımından dengeli bir profil sunduğunu düşündürür (Şekil 4.16A).

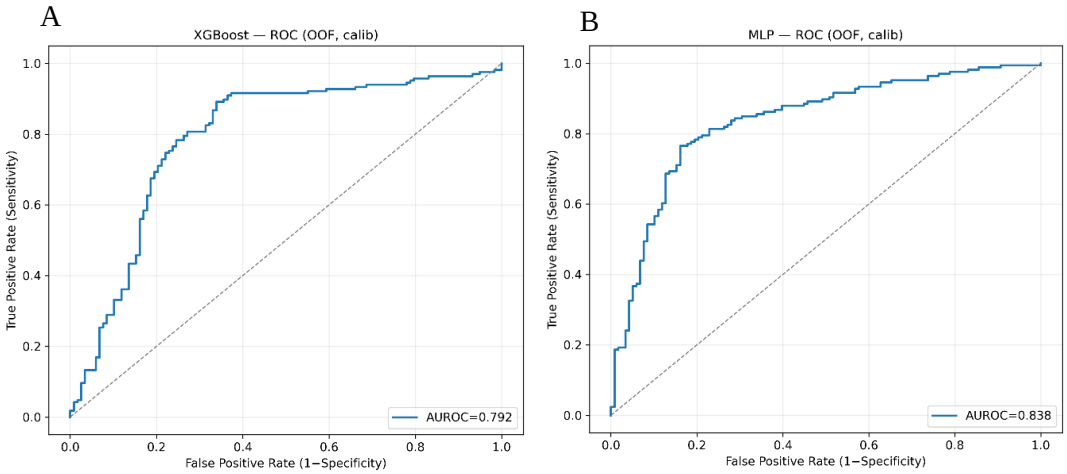
Random Forest'ın kalibre ROC eğrisi, düşük FPR bölgelerinde kademeli bir yükseliş göstermiş, ardından 0,1-0,2 bandı arasında kısa bir düşüşün ardından yeniden belirgin bir tırmanış göstermiş, bütünleşik ayrım gücü AUROC=0.804 olarak bulunmuştur. Bu desen, modelin düşük FPR bölgesinde temkinli fakat etkili bir ayrım sağladığını, daha yüksek duyarlılık düzeylerine ise kademeli olarak artan yanlış pozitif sergileyerek ulaştığı gözlemlenmektedir. Buna karşın genel olarak diyagonalin üzerinde kalarak anlamlı bir ayırım sağladığını göstermektedir. Tablo 4.2'deki F1/MCC üstünlükleriyle birlikte okunduğunda, RF'nin eşik=0,5 koşulunda dengeli bir karar davranışı sergilediği anlaşılmaktadır (Şekil 4.16B).



Şekil 4.16 (A) Kalibrasyon sonrası SVM'ye ait ROC eğrisi, (B) Random Forest'e ait ROC eğrisi

Kalibre edilmiş XGBoost'un ROC eğrisi, düşük FPR oranlarında duyarlılığı hızlı bir biçimde kaldırsada diğer modellere kıyasla diyagonale daha yakın bir eğri çizmektedir. AUROC değeri 0,792 olarak hesaplanmıştır. Eğrinin yaklaşık 0.15–0.40 FPR bandında kademeli ama daha yatay bir ilerleyiş sergilemesi, modelin bu aralıkta duyarlılık artışını nispeten özgüllük pahasına elde ettiğini düşündürür. Bu profil, ortak eşikte (0.5) raporlanan Accuracy=0.768, MCC=0.524 ve özellikle de Specificity=0.737 değerleriyle uyumludur. Bu durum dikkate alındığında XGBoost kalibrasyon sonrası makul bir ayırt ediciliği korurken, yanlış pozitifleri sınırlamada Random Forest, Logistic Regression ve Alphamissense kadar güçlü olmadığı görülmektedir (Şekil 4.17a).

MLP'ye bakıldığında, ROC eğrisi diyagonalden uzakta daha belirgin bir konveksite sunmakta ve düşük FPR bölgesinde duyarlılığı 0.75'lere hızlıca taşıyarak istikrarlı bir early lift sağlamaktadır. 0,838 AUROC değeri de bu görsel üstünlüğü sayısal olarak doğrular niteliktedir. Eğrinin orta bantta (FPR≈0.2–0.5) köşeye yakın seyrini sürdürmesi, farklı klinik tercih profillerinde de avantajın korunduğunu gösterir. Bu bulgular, ortak eşikte elde edilen yüksek Accuracy=0.792, F1=0.820 ve görece dengeli Specificity=0.771 ile tutarlıdır. Ayrıca MLP'nin Brier=0.158 ve ECE=0.049 değerleri, olasılık çıktılarının kalibrasyon sonrası güvenilir yorumlanabildiğini destekler (Şekil 4.17B)



Şekil 4.17 (A) Kalibrasyon sonrası XGBoost'a ait ROC eğrisi, (B) MLP'ye ait ROC eğrisi

4.3. İstatistiksel Analiz Sonuçları

Raporlanan kalibrasyon katsayılarının ve Hosmer–Lemeshow (HL) uyum testinin birlikte değerlendirilmesi, modellerin olasılık çıktılarının yalnızca ayırt edici gücünü değil, aynı zamanda bu çıktılara ne ölçüde güvenilebileceğini de nicel olarak ortaya koyar. Tabloya yansıyan iki temel katsayıdan calibration slope (ideal değer 1), modelin tahmin ettiği olasılıkları ne kadar gereğinden fazla abarttığını verirken; calibration intercept (ideal değer 0) ise bu olasılıkların tüm aralık boyunca sistematik olarak yukarı ya da aşağı kayıp kaymadığını, yani genel yanlılık düzeyini gösterir. Slope değerinin 1'in altında olması, modelin 0'a çok yakın ve 1'e çok yakın olasılıkları gerçek veriye kıyasla fazla üretme eğiliminde olduğunu, başka bir deyişle gerçek veriye oranla "aşırı güvenli" (over-confident) tahminler yaptığını düşündürür. Intercept'in pozitif olması ise modelin genel olarak olasılıkları olduğundan düşük tahmin ettiğini, yani veri setindeki pozitiflik oranının modelin verdiği ortalama olasılıklardan daha yüksek kaldığını (under-prediction) ifade eder. HL testi, tahmin edilen olasılık aralığını birkaç dilime bölerek her dilimde beklenen ve gözlenen pozitif sayıları karşılaştıran bir uyum-iyiliği testidir; burada $p<0,05$ elde edilmesi, beklenen ve gözlenen değerler arasındaki farkların rastlantı ile açıklanamayacak düzeyde büyük olduğunu ve modelin tahmin ettiği olasılıkların veriyle istatistiksel olarak yeterli uyum göstermediğini düşündürür. Buna karşılık $p\geq0,05$ olması, dilimlenmiş olasılık aralığında anlamlı bir uyumsuzluk saptanamadığı şeklinde

yorumlanır. Ancak testin hem örneklem büyüklüğüne hem de seçilen dilim sayısına duyarlı olduğu göz önünde bulundurulmalıdır. Bu nedenle HL sonuçları, kalibrasyon eğrilerinin görsel incelemesi, ECE ve Brier skoru gibi diğer kalibrasyon göstergeleri ile birlikte ele alınarak yorumlanmalıdır.

Genel olarak bakıldığında, modellerin tamamında calibration slope değerlerinin 0,21–0,28 aralığında yer aldığı görülmektedir. Bu durum, modellerin tahmin ettikleri olasılıkları gerçek veri oranlarına göre fazla "keskin" kullandığını, yani yüksek riskli gruplara gereğinden yüksek, düşük riskli gruplara ise gereğinden düşük olasılıklar atama eğiliminde olduğunu ve bu açıdan aşırı güvenli over confident bir profil sergilediklerini düşündürmektedir. Calibration intercept değerleri yaklaşık 0,47–0,48 civarında ve tüm modellerde sistematik olarak pozitiftir. Bu da veri setindeki gerçek pozitif oranının, modellerin verdiği ortalama olasılıklardan daha yüksek olduğunu, başka bir ifadeyle genel düzeyde pozitif sınıf için "eksik tahmin" (under-prediction) eğilimi bulunduğunu göstermektedir. Bu iki bulgu birlikte ele alındığında, modellerin ürettikleri olasılıkların hem genel biçimi (intercept>0) hem de risk grupları arasındaki farkları yansıtma biçimi (slope<1) düzeyinde düzeltmeye ihtiyaç olduğunu, buna karşın HL testi, ECE, Brier skoru ve ilgili güven aralıkları birlikte değerlendirildiğinde bazı modellerde pratik uyumun kabul edilebilir sınırlar içinde kaldığını söylemek mümkündür.

AlphaMissense için (eğim=0,223; kesişim=0,462; HL $p=1{,}64\times10^{-5}$) slope değerinin düşük, intercept'in pozitif olması, modelin risk farklarını olduğundan daha keskin gösterdiğini ve genel olarak pozitif olasılıklarını gerçek verilere göre düşük tahmin ettiğini işaret etmektedir. HL testinin anlamlı çıkması, özellikle orta ve yüksek olasılık dilimlerinde beklenen ve gözlenen pozitif oranları arasındaki farkların rastlantı ile açıklanamayacak düzeyde biriktiğini ve kalibrasyon eğrisinde şekilsel bir uyumsuzluk bulunduğunu düşündürür. Bu patern, AlphaMissense'in 0,5 ortak eşiğinde güçlü ayrım metrikleri (AUROC/AUPRC) ve yüksek Precision üretirken, verdiği olasılıkların mutlak değerlerinin gerçek risk düzeyine göre abartılı olduğunu gösteren eğim<1 bulgusuyla uyumludur (Tablo 4.3).

PolyPhen-2 (eğim=0,252; kesişim=0,486; HL p=0,344) ve SIFT (eğim=0,282; kesişim=0,482; HL p=0,501) için HL testinde p>0,05 elde edilmesi tahmin ve gerçek

veri farklarının istatistiksel olarak anlamlı bir uyumsuzluk göstermediğini, dolayısıyla kalibrasyon eğrilerinin şekil olarak kabul edilebilir düzeyde hizalandığını düşündürmektedir. Bununla birlikte slope değerlerinin 1'den belirgin düşük olması, modelin düşük riskli varyantlara gerçekte olması gerekenden daha düşük, yüksek riskli varyantlara ise daha yüksek olasılık atama eğiliminde olduğunu göstermektedir. Başka bir deyişle, risk grupları arasındaki olasılık farkı gerçek veriye göre olduğundan fazla büyütülmektedir. 0,5 ortak eşiğinde bu iki araçta, özellikle SIFT için, özgüllüğün (Specificity) görece düşük kalması da bu ölçeklenme biçiminin sınıflandırma düzeyine yansıması olarak okunabilir (Tablo 4.3).

Logistic Regression (eğim=0,215; kesişim=0,471; HL p=0,0246), Random Forest (0,211; 0,471; p=0,00884) ve XGBoost (0,235; 0,477; p=0,00440) için HL p-değerlerinin 0,05'in altında olması, bu modellerde tahmin edilen olasılık dilimlerinde beklenen ve gözlenen uyumunun istatistiksel olarak yeterli bulunmadığını göstermektedir. Bu uyumsuzluk, özellikle yaklaşık 0,40–0,70 tahmin aralığında kalibrasyon eğrisinde izlenen yukarı aşağı salınım ile ilişkilidir ve ilgili reliability diagramlarında görsel olarak da görülebilmektedir. Logistic Regression örneğinde olduğu gibi ayrım gücü (AUROC) yüksek olsa dahi, slope değerlerinin 1'den belirgin biçimde küçük olması, olasılıkların ölçekleniş biçiminde benzer bir aşırı güvenli yapı olduğunu ortaya koymaktadır (Tablo 4.3).

SVM (eğim=0,229; kesişim=0,470; HL p=0,369) ve MLP (eğim=0,230; kesişim=0,480; HL p=0,481) için HL testinde p>0,05 bulunması, bu modellerde tahmin edilen olasılıklarla gözlenen pozitiflik oranları arasında, testin kullandığı grup bazlı karşılaştırma çerçevesinde istatistiksel olarak anlamlı bir uyumsuzluk saptanmadığı şeklinde yorumlanabilir. Bu iki modelde ECE ve Brier skorlarının görece düşük olması da kalibrasyon eğrilerinden elde edilen görsel izlenim ile ortalama hata ölçütlerinin aynı yönde olduğunu desteklemektedir. Bununla birlikte, burada da slope değerlerinin 1'den uzak kalması, model olasılıklarının gerçek riske göre hâlâ fazla keskin kullanıldığını ve olasılık aralığının ideal kalibrasyon durumuna göre yeterince dengelenmediğini ima eder (Tablo 4.3).

Özetle Tablo 4.3, her modelin kalibre edilmiş olasılıklarının güvenilirliğini özetleyen üç tamamlayıcı göstergenin birlikte sunulduğu bir çerçeve sağlamaktadır. doğrusal

kalibrasyon eğimi (calibration slope; ideal 1), kalibrasyon kesişimi ve yanlılığı (calibration intercept; ideal 0) ve Hosmer–Lemeshow iyi-uyum testinin sonuçları (HL χ^2, df, p). Eğimin 1'e yaklaşması, modelin düşük olasılık tahmin ettiği gruplarda gerçekten düşük, yüksek olasılık tahmin ettiği gruplarda gerçekten yüksek pozitiflik oranları ürettiğini ve risk grupları arasındaki farkları gerçek dağılımla uyumlu bir ölçekle yansıttığını gösterir. Kesişimin (intercept) 0'a yakın olması, tüm olasılık aralığı boyunca sistematik bir yukarı ya da aşağı kayma olmadığını; HL testinde $p>0,05$ bulunması ise, olasılık aralığında beklenen ve gözlenen pozitif sayıları arasındaki farkların rastlantı ile açıklanamayacak kadar büyük olmadığını, yani kalibrasyon eğrisinin seçilen çözünürlükte yeterli uyum sergilediğini ifade eder. Bu çerçevede AlphaMissense (slope=0,223; intercept=0,462; HL $p=1,64\times10^{-5}$), Logistic Regression (slope=0,215; intercept=0,471; HL p=0,0246), Random Forest (slope=0,211; intercept=0,471; HL p=0,00884) ve XGBoost (slope=0,235; intercept=0,477; HL p=0,00440) için HL testinin anlamlı çıkması, bu modellerin olasılık dilimlerinde tahmin edilen ve gerçek verinin uyumunun istatistiksel olarak yetersiz bulunduğunu göstermektedir. Buna karşılık PolyPhen-2 (slope=0,252; intercept=0,486; p=0,344), SIFT (slope=0,282; intercept=0,482; p=0,501), SVM (slope=0,229; intercept=0,470; p=0,369) ve MLP (slope=0,230; intercept=0,480; p=0,481) için HL p-değerlerinin 0,05'in üzerinde olması, bu modellerin kalibrasyon eğrilerinde dilim bazında anlamlı bir uyumsuzluk saptanmadığı şeklinde yorumlanır. Tabloda tüm modellerin eğimlerinin 1'in belirgin biçimde altında (≈0,21–0,28) ve kesişimlerinin birbirine çok yakın pozitif değerlerde (≈0,46–0,48) kümelendiği görülmektedir; bu patern, kalibrasyonun sayısal özetinin modeller arasında tutarlı bir yapı sergilediğini ve olasılık ölçeğinin genel hizasının (intercept) ile risk farklarının büyüklüğünün (slope) sistematik biçimde raporlandığını göstermektedir (Tablo 4.3).

Tablo 4.3 Hosmer–Lemeshow uyum testi sonuçları

Model	Calibslope	CalibslopeCI lo	Calib slopeCI hi	Calib intercept	Calib intercept CI lo	Calib intercept CI hi	HL chi2	HL df	HL p	HL sig_0.05
Alphamissense	0,2230	0,1915	0,2511	0,4623	0,4085	0,5183	29,7605	5,0000	1,64E-05	significant
Polyphen-2	0,2522	0,1909	0,3126	0,4863	0,4292	0,5449	4,4898	4,0000	3,44E-01	ns
SIFT	0,2816	0,1789	0,3856	0,4822	0,4152	0,5521	2,3606	3,0000	5,01E-01	ns
Logistic R	0,2147	0,1836	0,2429	0,4707	0,4192	0,5251	14,4927	6,0000	2,46E-02	significant
SVM	0,2287	0,1919	0,2622	0,4700	0,4191	0,5248	5,3982	5,0000	3,69E-01	ns
Random F	0,2106	0,1739	0,2451	0,4708	0,4194	0,5252	18,7986	7,0000	8,84E-03	significant
Xgboost	0,2352	0,1911	0,2779	0,4770	0,4239	0,5336	17,0537	5,0000	4,40E-03	significant
MLP	0,2300	0,1942	0,2620	0,4800	0,4302	0,5334	5,5077	6,0000	4,81E-01	ns

Kalibrasyon sonrasında modellerin ayrıştırma gücü arasındaki farklar, eşleştirilmiş bootstrap yaklaşımıyla nicel olarak değerlendirildi. Her model çifti için, ΔAUC = AUC(Model A) − AUC(Model B) hesaplanmıştır. Bunun için %95 güven aralıkları ve iki yönlü p-değerleri elde edildi ve çoklu karşılaştırmalar Benjamini–Hochberg (FDR) yöntemiyle düzeltilerek q-değerleri raporlanmıştır (Tablo 4.4). Güven aralığının sıfırı içermemesi ve q<0,05 koşulunun birlikte sağlanması, AUROC açısından istatistiksel olarak anlamlı bir üstünlüğe işaret etmektedir. ΔAUC'nin işareti pozitif olduğunda Model A'nın üstün olduğu, eğer değer negatif ise Model B'nin üstün olduğu anlamına gelmektedir.

AlphaMissense-PolyPhen-2 ikilisinin ΔAUC=0,119, %95 GA: 0,072–0,169 ve q=0 değerlerine bakıldığında ve AlphaMissense-SIFT'nin ΔAUC=0,146, %95 GA: 0,090–0,202, q=0 değerlerine bakıldığında AlphaMissense'in iki araca karşı belirgin bir AUROC üstünlüğü olduğu görülmektedir. Makine öğrenmesi modellerine karşı da küçük ila orta derecede anlamlı farklar saptanmıştır (AM–SVM: ΔAUC=0,042; 0,020–0,064; q=0; AM–Random Forest: ΔAUC=0,057; 0,030–0,084; q=0; AM–XGBoost: ΔAUC=0,069; 0,035–0,105; q=0; AM–MLP: ΔAUC=0,024; 0,0037–

0,045; q=0,032). Buna karşılık AlphaMissense ile Logistic Regression arasındaki fark anlamlı değildir (ΔAUC=0,009; −0,0015–0,0196; q=0,109); bu iki model AUROC bakımından istatistiksel olarak ayırt edilememektedir. Bu durum, kalibre edilmiş 0,5 eşiğinde elde edilen Accuracy ve MCC değerlerinin birbirine çok yakın olmasıyla uyumludur (Tablo 4.4).

PolyPhen02 ise makine öğrenmesi modellerine karşı çoğunlukla daha düşük AUROC sergilemiştir. Logistic Regression (ΔAUC=−0,110; q=0), SVM (ΔAUC=−0,078; q=0,0067), Random Forest (ΔAUC=−0,062; q=0,0238) ve MLP (ΔAUC=−0,095; q=0) lehine farklar istatistiksel olarak anlamlı bulunmuştur. PolyPhen-2 ile XGBoost arasındaki karşılaştırma ise sınırda kalmış ve anlamlı bulunmamıştır (ΔAUC=−0,050; %95 GA: −0,110–0,009; q=0,121). PolyPhen-2 ile SIFT arasındaki farkın da anlamlı olmaması (ΔAUC=0,027; q=0,419), bu iki aracın AUROC düzeyinde benzer bir performans bandında yer aldığını düşündürmektedir. PolyPhen-2 ile XGBoost arasındaki karşılaştırma ise sınırda kalmış ve anlamlı bulunmamıştır (ΔAUC=−0,050; %95 GA: −0,110–0,009; q=0,121). PolyPhen-2 ile SIFT arasındaki farkın da anlamlı olmaması (ΔAUC=0,027; q=0,419), bu iki aracın AUROC düzeyinde benzer bir performans bandında yer aldığını göstermektedir (Tablo 4.4).

Sıft'e bakıldığında bütün makine öğrenmesi algoritmalarına karşı sistematik olarak daha düşük AUROC değeri sergilemiştir: Logistic Regression (ΔAUC=−0,137; q=0), SVM (ΔAUC=−0,104; $q=9{,}33\times10^{-4}$), Random Forest (ΔAUC=−0,089; q=0,0048), XGBoost (ΔAUC=−0,077; q=0,032) ve MLP (ΔAUC=−0,122; q=0). Bu bulgu, SIFT'in 0,50 eşiğinde gözlenen yüksek Sensitivity/Recall ve düşük Specificity dengesinin, bütünleşik ayrım alanına da olumsuz yansıdığını göstermektedir (Tablo 4.4).

Makine öğrenmesi algoritmalarının kendi aralarındaki karşılaştırmalara bakıldığında ise, SVM, Random Forest ve XGBoost modellerine karşı Logistic Regression'un anlamlı bir AUROC üstünlüğü sağladığı görülmektedir: (LogR–SVM: ΔAUC=0,033; 0,011–0,056; q=0,0026), (LogR–RF: ΔAUC=0,048; 0,023–0,074; q=0), (LogR–XGB: ΔAUC=0,060; 0,026–0,097; q=0). Logistic Regression ile MLP arasındaki fark ise istatistiksel olarak anlamlı değildir (ΔAUC=0,0148; %95 GA:

−0,0006–0,0301; q=0,077); bu iki modelin AUROC performansları denk kabul edilebilir. SVM–Random Forest ve SVM–XGBoost çiftlerinde de fark anlamlı bulunmamıştır ve bu üç model ayrıştırma gücü bakımından birbirine yakın bir bantta yer almaktadır. Buna karşılık MLP, Random Forest ve XGBoost'a karşı anlamlı üstünlük göstermiştir (RF–MLP: ΔAUC=−0,033; −0,061–−0,0036; q=0,0375; XGB–MLP: ΔAUC=−0,0456; −0,0847–−0,0065; q=0,032). ancak Logistic Regression'a karşı belirgin bir üstünlük sergilememiştir (Tablo 4.4).

Genel olarak AUROC sonuçları şu hiyerarşiyi işaret etmektedir: AlphaMissense ≃ Logistic Regression ≥ MLP > SVM ≃ Random Forest ≃ XGBoost > PolyPhen-2 ≃ SIFT. AlphaMissense'in makine öğrenmesi modellerine göre küçük-orta büyüklükte bir AUROC avantajı vardır. AlphaMissense'e en yakın makine öğrenmesi algoritması Logistic Regression olmuştur ve hemen arkasından MLP gelmektedir. PolyPhen-2 ve SIFT ise araç sınıfı içinde benzer düzeyde kalmış ve makine öğrenmesi modellerinin belirgin biçimde gerisinde kalmışlardır. Özetle bu tablo, ROC eğrilerinden elde edilen görsel izlenimleri nicel ve istatistiksel olarak desteklemektedir.

Tablo 4.4 Eşleştirilmiş bootstrap ΔAUC ikili model karşılaştırmaları

Model A	Model B	DeltaAUC mean	CI95_low	CI95 high	p value	q value
Alphamissense	Polyphen-2	0,119204265	0,071793204	0,1687517	0	0
Alphamissense	SIFT	0,146106114	0,090000276	0,2019345	0	0
Alphamissense	Logistic R	0,008989207	-0,001514333	0,01958543	0,0856	0,108945
Alphamissense	SVM	0,041616821	0,0199468	0,06416277	0	0
Alphamissense	Random F	0,056756365	0,030377804	0,08395071	0	0
Alphamissense	Xgboost	0,069402819	0,035080047	0,10519617	0	0
Alphamissense	MLP	0,023834001	0,003667166	0,04538615	0,0204	0,032421
Polyphen-2	SIFT	0,026901849	-0,038662855	0,0904115	0,4188	0,4188
Polyphen-2	Logistic R	-0,110215058	-0,157198241	-0,06711615	0	0
Polyphen-2	SVM	-0,077587444	-0,12861344	-0,02811399	0,0036	0,00672
Polyphen-2	Random F	-0,0624479	-0,114163435	-0,01356314	0,0136	0,0238

Tablo 4.4 Eşleştirilmiş bootstrap ΔAUC ikili model karşılaştırmaları (devamı)

Polyphen-2	Xgboost	-0,049801446	-0,109725863	0,00934556	0,0996	0,121252
Polyphen-2	MLP	-0,095370264	-0,140632755	-0,05355572	0	0
SIFT	Logistic R	-0,137116907	-0,192016887	-0,08130611	0	0
SIFT	SVM	-0,104489293	-0,163319837	-0,04458783	0,0004	0,000933
SIFT	Random F	-0,089349749	-0,147388908	-0,03133395	0,0024	0,0048
SIFT	Xgboost	-0,076703295	-0,143641564	-0,01149345	0,022	0,032421
SIFT	MLP	-0,122272113	-0,175294869	-0,07005272	0	0
Logistic R	SVM	0,032627614	0,0112998	0,05604481	0,0012	0,002585
Logistic R	Random F	0,047767158	0,022915396	0,07419436	0	0
Logistic R	Xgboost	0,060413612	0,02601372	0,0967488	0	0
Logistic R	MLP	0,014844794	-0,000621796	0,03009319	0,058	0,077333
SVM	Random F	0,015139544	-0,013481493	0,04320795	0,278	0,288296
SVM	Xgboost	0,027785998	-0,011324838	0,06596247	0,1596	0,178752
SVM	MLP	-0,01778282	-0,040412128	0,00394852	0,1156	0,134867
Random F	Xgboost	0,012646454	-0,010596765	0,03535389	0,2692	0,288296
Random F	MLP	-0,032922364	-0,061250058	-0,00357169	0,0268	0,03752
Xgboost	MLP	-0,045568818	-0,08468347	-0,00651253	0,0212	0,032421

Paired Bootstrap değer ısı haritasında (Şekil 4.18), kalibrasyon sonrası modellerin ayrıştırma gücünü (AUROC) ikili olarak karşılaştırmak için uygulanan eşleştirilmiş bootstrap testinin sonuçlarını göstermektedir. Her bootstrap yinelemesinde veri kümesindeki örnekler birlikte yeniden örneklenmiş, karşılaştırılan iki model için AUROC farkı (ΔAUC = AUC_A − AUC_B) hesaplanmış ve bu farkın sıfır etrafındaki dağılımı kullanılarak iki yönlü p-değeri elde edilmiştir. Böylece gözlenen AUC farkının yalnızca örneklem değişimine bağlılığı, yani istatistiksel olarak anlamlı bir performans ayrımı olup olmadığı test edilmiştir. Diyagonal hücrelerde p-değerinin 1.000 olması, aynı modelin kendiyle karşılaştırılması durumunda farkın

tanım gereği sıfır olduğunu yansıtmaktadır; yorumlama diagonal dışındaki hücreler üzerinden yapılır.

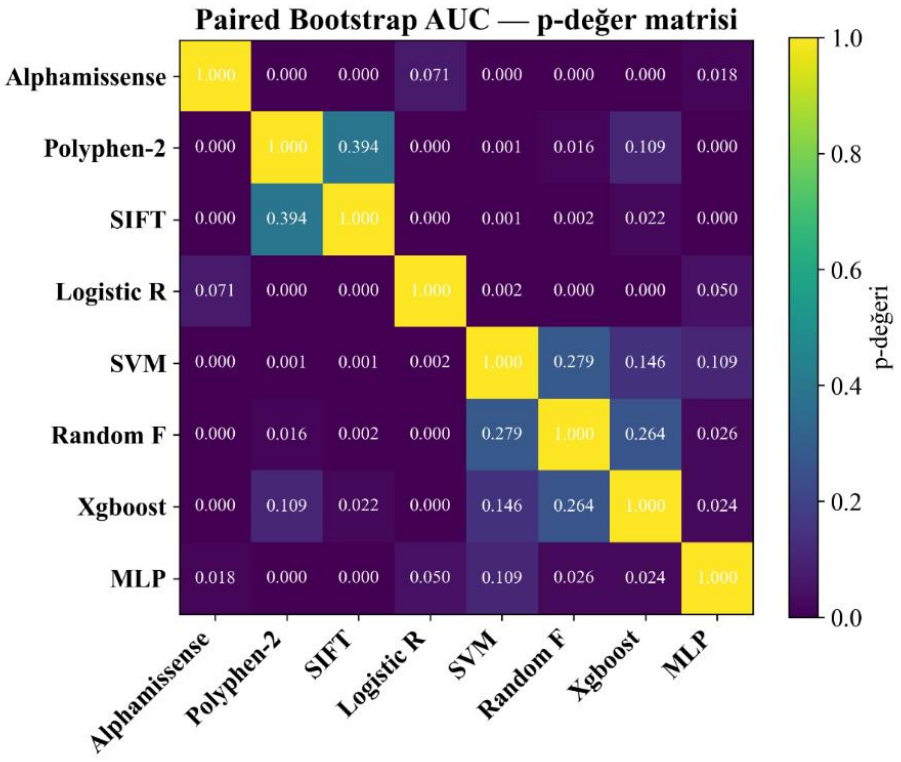
Isı haritası genel olarak üç bantta kümelenme olduğunu göstermektedir. AlphaMissense, Logistic Regression ve MLP, AUROC açısından üst düzeyde yer almakta; SVM, Random Forest ve XGBoost orta-üst grubu, PolyPhen-2 ve SIFT ise alt düzeyi oluşturmaktadır. AlphaMissense ile pek çok model arasındaki çok küçük p-değerleri (örneğin AM–SVM, AM–Random Forest, AM–XGBoost, AM–SIFT, AM–PolyPhen-2 hücrelerinde p'nin 0.001'in çok altında olması) AlphaMissense'in bu modellerden AUC bakımından istatistiksel olarak anlamlı biçimde üstün olduğunu göstermektedir. AlphaMissense ile MLP arasındaki karşılaştırmada p=0.018 bulunmuş, bu da %5 anlamlılık düzeyinde üstünlüğün korunduğuna işaret etmiştir. Buna karşılık AlphaMissense ile Logistic Regression arasındaki p=0.071 değeri, bu iki model arasında AUC farkını kanıtlamak için yeterli görülmemektedir; bu durum, kalibre OOF ROC alanlarının birbirine çok yakın olmasına ilişkin bulgularla uyumludur (Şekil 4.18).

Orta bantta yer alan SVM, Random Forest ve XGBoost arasındaki ikili karşılaştırmalarda p-değerlerinin görece yüksek kalması (SVM–Random Forest p=0.279, SVM–XGBoost p=0.146, Random Forest–XGBoost p=0.264), bu üç modelin AUROC açısından istatistiksel olarak ayırt edilemediğini düşündürmektedir. Dolayısıyla bu grupta tercih yaparken, kalibrasyon kalitesi (Brier skoru, ECE) ya da karar eşiğine bağlı ölçütler (F1, MCC, Precision/Recall) gibi ek kriterlerin dikkate alınması daha yerinde olacaktır (Şekil 4.18).

Alt bantta yer alan PolyPhen-2 ve SIFT için, aralarındaki karşılaştırmada p=0.394 elde edilmesi, bu iki aracın ayrıştırma gücünün istatistiksel olarak birbirinden ayırt edilemediğini göstermektedir. Buna karşılık, hem PolyPhen-2'nin hem de SIFT'in Logistic Regression, SVM, Random Forest, XGBoost ve MLP gibi öğrenen modellerle yapılan karşılaştırmalarında p-değerlerinin çoğunlukla 0.001–0.050 aralığında veya 0.000'a yuvarlanacak kadar küçük olması, bu araçların AUROC bakımından üst gruba karşı belirgin biçimde dezavantajlı olduğunu ortaya koymaktadır. Örneğin PP2–Logistic Regression, PP2–MLP, SIFT–Logistic Regression ve SIFT–MLP hücrelerinde p'nin 0.001'in oldukça altında olması,

öğrenen modellerin PolyPhen-2 ve SIFT'e göre anlamlı üstünlük sağladığını göstermektedir. PolyPhen-2 ile XGBoost (p=0.109) ve Logistic Regression ile MLP (p=0.050) arasındaki karşılaştırmalar ise sınırda p-değerleri vermekte; çoklu karşılaştırmalar için Benjamini–Hochberg yanlış keşif oranı düzeltmesi uygulandığında, bu ikililerin çoğu senaryoda anlamlılık eşiğinin dışında kalması beklenmektedir (Şekil 4.18).

Özet olarak bu matris, AlphaMissense'in geniş bir model yelpazesinde AUROC açısından anlamlı üstünlük sergilediğini, farkın Logistic Regression ve MLP ile karşılaştırmalarda en dar düzeye indiğini, Logistic Regression, MLP, SVM ve Random Forest ve XGBoost modellerinin kendi aralarında tam ayrışmasalar da alt banttaki araçlara karşı tutarlı üstünlük gösterdiğini, PolyPhen-2 ve SIFT'in ise birbirine benzer düzeyde, ancak üst gruba göre belirgin şekilde düşük ayrıştırma gücüne sahip olduğunu göstermektedir (Şekil 4.18).



Şekil 4.18 Kalibre modeller arası eşleştirilmiş bootstrap analizi ısı haritası

0,5 eşik değerinde yapılan McNemar testi p-değer matrisi ısı haritası, tüm yöntemler aynı karar eşiğini kullanırken hangi model çiftlerinin benzer, hangilerinin ise farklı örneklerde hata yaptığını gösteren çiftler-arası McNemar testlerinin p-değerlerini özetlemektedir. McNemar testi, iki modelin aynı örnekler üzerindeki ikili kararlarını karşılaştırır ve yalnızca uyuşmayan durumları dikkate alır: bir modelin doğru, diğerinin yanlış olduğu gözlemler ile bunun tersi durumlar sayılır. Bu iki sayı arasındaki dengenin bozulup bozulmadığına bakılarak p-değeri hesaplanır. P-değerinin küçük olması, iki modelin farklı örneklerde hata yapma eğiliminde olduğunu; p-değerinin büyük olması ise hataların büyük ölçüde aynı örneklerde kümelendiğini ve karar düzeyinde modellerin birbirine daha çok benzediğini düşündürür (Şekil 4.19).

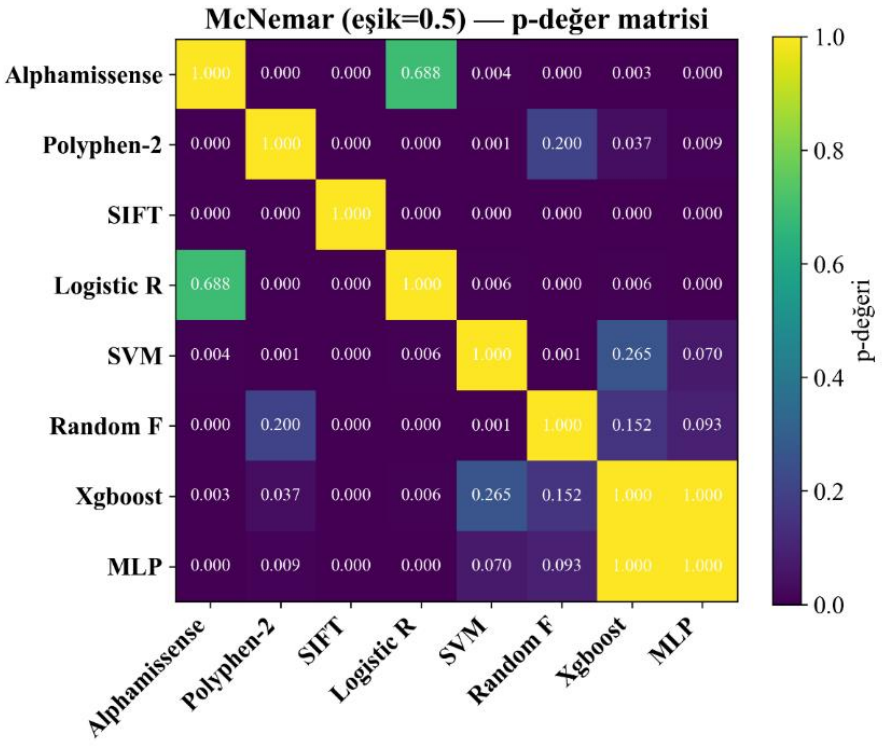
Isı haritasındaki yaygın olarak görünen çok küçük yaklaşık 0.000–0.006 bandındaki p-değerleri, özellikle AlphaMissense'in hem tahmin araçlarına hem de makine öğrenmesi modellerine kıyasla 0,5 eşiğinde belirgin şekilde farklı bir hata örüntüsüne sahip olduğunu göstermektedir. Bu, AlphaMissense'in yüksek özgüllük ve early lift (düşük FPR'de hızlı duyarlılık kazanımı) davranışıyla tutarlıdır. Yani kimi varyantları diğer modellerin yanlış sınıflandırdığı yerlerde doğru, kimi varyantları da tam tersi şekilde ele almaktadır (Şekil 4.19).

Benzer biçimde Logistic Regression'ın SVM, XGBoost ve MLP ile karşılaştırmalarında görülen çok küçük p-değerleri, bu modellerin aynı AUC bandında yer alsa bile 0,5 eşiğinde farklı pozitif varyantları kaçırıp farklı varyantlara da yanlış alarm üretme eğiliminde olduğunu göstermektedir. Buna karşılık AlphaMissense–Logistic Regression çifti için $p \approx 0.688$, PolyPhen-2–Random Forest için $p \approx 0.200$, SVM–XGBoost için $p \approx 0.265$ ve Random Forest–MLP için $p \approx 0.093$ gibi daha yüksek değerler, bu çiftlerde hataların daha çok aynı örneklerde toplandığını ve karar düzeyinde modellerin davranışlarının birbirine daha yakın olduğunu düşündürür (Şekil 4.19).

Dikkat çekici bir uç örnek $p \approx 1.000$ değeri ile XGBoost–MLP çiftidir. Bu durumda iki modelin 0,5 eşiğindeki doğru ve yanlış karar desenleri neredeyse aynıdır. Bu, performanslarının mutlaka mükemmel olduğu anlamına gelmez; yalnızca, aynı

örneklerde doğru, aynı örneklerde yanlış karara gittiklerini ve pratikte birbirinin yerine kullanılabilir olduklarını gösterir (Şekil 4.19).

Özetle, bu ısı haritasında ikili model karşılaştırmalarına ait ham p-değerleri gösterilmektedir. Çok sayıda ikili test yapıldığı için, bazı küçük p-değerlerin yalnızca rastlantıdan kaynaklanıp kaynaklanmadığını değerlendirmek amacıyla, Tablo 4.5'te raporlanan Benjamini–Hochberg FDR ile düzeltilmiş q-değerleri de dikkate alınmalıdır. Genel olarak bakıldığında AUROC için elde edilen bootstrap p-matrisinde görülen ayrıştırma farklarıyla uyumlu olarak, AUROC bakımından öne çıkan AlphaMissense ve Logistic Regression, McNemar analizinde de birçok modelle karşılaştırıldığında farklı hata örüntüleri sergilemektedir. AUC değerleri birbirine yakın olan SVM, Random Forest, XGBoost ve MLP çiftlerinde ise McNemar p-değerleri çoğu zaman yüksek kalmakta, yani 0,5 eşiğinde bu modellerin büyük ölçüde aynı örneklerde hata yaptığı görülmektedir. Dolayısıyla bu matris, hangi modelin AUROC değeri daha yüksek? sorusundan çok, aynı klinik karar eşiğinde hangi modellerin birbirinin yerine kullanılabilir olduğu ve hangilerinin farklı hastaları kaçırdığı ve yanlış alarm verdiği sorusuna yanıt sağlar. AUROC için bootstrap p-matrisi ile birlikte raporlanması, eşik-bağımsız ayrım gücü ile eşik bağımlı hata desenlerinin birlikte ve tutarlı biçimde değerlendirilmesine imkân verir.



Şekil 4.19 Ortak eşikte kalibre edilmiş modellerin McNemar testi ısı haritası

Şekil 4.19'da McNemar p-değer ısı haritasında özetlenen ilişkiler, Tablo 4.5'te sayısal olarak ayrıntılandırılmıştır. Bu tabloda yer alan b değeri "Model A'nın doğru, Model B'nin yanlış" karar verdiği örnek sayısını; c değeri ise "Model A'nın yanlış, Model B'nin doğru" karar verdiği örnek sayısını göstermektedir. Her model çifti için iki yönlü exact binom p-değeri hesaplanmış, çoklu karşılaştırma etkisini kontrol etmek üzere Benjamini–Hochberg (FDR) düzeltmesi uygulanarak q-değerleri elde edilmiştir. Böylece yalnızca özet metrik ortalamalarına bakmak yerine, modellerin aynı eşikte verdikleri kararların gerçekten birbirinin yerine geçip geçemeyeceği daha doğrudan biçimde değerlendirilmiştir.

Bulgular, 0,5 eşiğinde AlphaMissense ile Logistic Regression çıktılarının birbirine oldukça yakın olduğunu göstermektedir; uyumsuzluk sayıları düşüktür ve fark istatistiksel olarak anlamsızdır (p=0.6875, q=0.713). Bu durum, Tablo 4.2'deki benzer Accuracy ve MCC değerleriyle uyumludur. Buna karşılık AlphaMissense ile SVM, Random Forest, XGBoost ve MLP modelleri kıyaslandığında fark karşı model lehine anlamlıdır. Örneğin AlphaMissense–Random Forest çiftinde, AlphaMissense'in yanlış sınıflandırdığı fakat Random Forest'ın doğru sınıflandırdığı

örnekler belirgin biçimde fazladır ($p=1.49\times10^{-8}$, $q=5.96\times10^{-8}$). Bu sonuç, öğrenen modellerin ortak eşikte AlphaMissense'in kaçırdığı belirli alt örnek kümelerinde ek duyarlılık sağladığını düşündürmektedir. Diğer yandan AlphaMissense PolyPhen-2 ve özellikle AlphaMissense SIFT karşılaştırmalarında da fark karşı model lehine geniş bir asimetri göstermektedir. Örneğin AM ve SIFT için değerler $p=7.97\times10^{-17}$ ve $q=1.39\times10^{-15}$ şeklindedir. Bu tablo, SIFT'in 0,5 eşiğinde AlphaMissense'in yanlış sınıflandırdığı çok sayıda örneği doğru sınıfladığını yani farkın SIFT lehine belirgin asimetrik olduğunu göstermektedir. Tablo 4.2'de 0,5 eşiğinde SIFT için gözlenen yüksek Recall ve düşük Specificity birlikte değerlendirildiğinde, bu desen SIFT'in daha fazla gerçek pozitif yakalarken, aynı zamanda daha fazla yanlış pozitif ürettiği anlamına gelir (Tablo 4.5).

PolyPhen-2 ile Random Forest ikilisine bakıldığında aralarında anlamlı bir fark izlenmemiştir ($p=0.200$, $q=0.224$); iki model 0.5 eşiğinde benzer bir karar haritası üretmektedir. XGBoost ve MLP çifti neredeyse simetriktir ($p=1.000$, $q=1.000$) ve bu iki modelin karar düzeyinde büyük ölçüde ikame edilebilir çıktılar verdiğini göstermektedir. Logistic Regression ile SVM ($p=0.00635$, $q=0.00935$) ve Logistic Regression ile MLP ($p=1.45\times10^{-4}$, $q=3.12\times10^{-4}$) karşılaştırmalarında ise fark SVM ve MLP lehinedir. Bu bulgular, ortak eşikte SVM ve MLP'nin Logistic Regression'ın kaçırdığı belirli örnekleri kazandığını ve karar düzeyinde daha tutarlı bir performans sergilediğini düşündürmektedir (Tablo 4.5).

Genel olarak McNemar analizi, ROC/AUROC farklarının ötesine geçerek karar düzeyindeki hata örüntüsünü görünür kılmıştır. 0,5 ortak eşiğinde AlphaMissense ile Logistic Regression'ın benzer kararlar ürettiği; SIFT'in yüksek duyarlılık stratejisi nedeniyle birçok model karşısında farklı ve daha agresif kararlar verdiği; SVM, Random Forest, XGBoost ve MLP'nin ise AlphaMissense'e göre belirli örnek alt kümelerinde ek doğru kazandırdığı gösterilmiştir. Böylece Tablo 4.5, modellerin yalnızca ayrıştırma gücü açısından değil, aynı klinik eşiği paylaştıklarında ne ölçüde birbirlerinin yerine geçebildikleri konusunda da somut kanıt sağlamaktadır.

Tablo 4.5 Ortak eşikte McNemar testi ikili model karşılaştırma sonuçları

Model A	Model B	B(A dogru B yanlis)	C(A yanlis B dogru)	Test	p_value	q_value(BH)
Alphamissense	Polyphen-2	15	53	exact	4,116E-06	1,15247E-05
Alphamissense	SIFT	11	91	exact	7,967E-17	1,39169E-15
Alphamissense	Logistic R	2	4	exact	0,6875	0,712962963
Alphamissense	SVM	2	14	exact	0,0041809	0,006886202
Alphamissense	Random F	0	27	exact	1,49E-08	5,96046E-08
Alphamissense	Xgboost	9	28	exact	0,0025632	0,004485614
Alphamissense	MLP	2	20	exact	0,0001211	0,000282605
Polyphen-2	SIFT	18	60	exact	1,973E-06	6,13829E-06
Polyphen-2	Logistic R	49	13	exact	4,818E-06	1,22628E-05
Polyphen-2	SVM	43	17	exact	0,0010658	0,001989429
Polyphen-2	Random F	36	25	exact	0,2000314	0,224035133
Polyphen-2	Xgboost	47	28	exact	0,0369549	0,049273205
Polyphen-2	MLP	37	17	exact	0,0090733	0,012702676
SIFT	Logistic R	88	10	exact	9,941E-17	1,39169E-15
SIFT	SVM	77	9	exact	1,341E-14	1,25166E-13
SIFT	Random F	68	15	exact	3,196E-09	1,49153E-08
SIFT	Xgboost	76	15	exact	5,477E-11	3,0672E-10
SIFT	MLP	73	11	exact	2,249E-12	1,57437E-11
Logistic R	SVM	1	11	exact	0,0063477	0,009354441
Logistic R	Random F	0	25	exact	5,96E-08	2,08616E-07
Logistic R	Xgboost	9	26	exact	0,0059881	0,009314854
Logistic R	MLP	1	17	exact	0,000145	0,000312218
SVM	Random F	2	17	exact	0,0007286	0,001457214
SVM	Xgboost	11	18	exact	0,2649309	0,285310196
SVM	MLP	1	7	exact	0,0703125	0,089488636
Random F	Xgboost	16	8	exact	0,1515896	0,176854571
Random F	MLP	16	7	exact	0,0931396	0,113387398
Xgboost	MLP	17	16	exact	1	1

4.4. Genler Özelinde Kalibre Edilmiş Modellerin Metrik Performansları

Bu çalışmada kullanılan tahmin araçları ve makine öğrenmesi modellerinin genler özelindeki performansları incelendiğinde, panel genelinde homojen olmayan, belirgin derecede heterojen bir metrik dağılımı gözlenmiştir. Özellikle MVK, NLRP3, NOD2, TNFRSF1A, PSMB8, IL36RN ve OTULIN genleri için hem tekil araçların (AlphaMissense, PolyPhen-2, SIFT) hem de kalibre edilmiş makine öğrenmesi modellerinin AUROC, AUPRC, Accuracy, F1 ve MCC değerlerinin panel genel ortalamasına yakın ya da çoğu durumda bunun üzerinde seyrettiği; buna karşın MEFV ve NLRP12 gibi genlerde modellerin ayırt edici gücünün görece sınırlı kaldığı ve metriklerin panel düzeyindeki başarıya kıyasla daha düşük düzeyde gerçekleştiği görülmüştür. LPIN2 ve PSTPIP1 gibi bazı genlerde ise modeller, panel genelindeki performansa büyük ölçüde benzer veya hafifçe altında kalan bir başarı profili sergilemiş; bu genler, panel bütününün performansını belirgin biçimde yukarı veya aşağı çekmeyen, daha ara bir konumda yer almıştır. Bu heterojenliğin, genlere ait toplam varyant sayısındaki farklılıklar, patojen/benign sınıflar arasındaki dengesizlik, ilgili proteinin fonksiyonel bölgelerinin ve yapısal özelliklerinin literatürde ne ölçüde tanımlanmış ve çalışılmış olduğu ile bu sayede modellerin öğrenebildiği biyolojik sinyalin gücündeki değişkenlikten kaynaklandığı düşünülmektedir (Tablo 4.6.-Tablo 4.17.).

Öte yandan IL1RN geni için panel analizine dahil edilen varyant sayısının her iki sınıf (patojen / benign) açısından da son derece kısıtlı olması nedeniyle, ROC ve PR eğrilerinin istatistiksel olarak güvenilir biçimde oluşturulması mümkün olmamış; bu nedenle söz konusu gen için gen düzeyindeki tabloda AUROC ve AUPRC değerleri hesaplanamamış ve sunulamamıştır (Tablo 4.16.).

Tablo 4.6 MEFV geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,5	0,763131	0,270395	0,244751	0,454545	0,4	0,038576	0,666667	0,285714	0,75
PolyPhen-2	0,392857	0,609447	0,310042	0,22228	0,454545	0,4	0,038576	0,666667	0,285714	0,75
SIFT	0,214286	0,523732	0,348539	0,457868	0,363636	0,461538	-0,31053	0,5	0,428571	0,25
LogisticRegression	0,571429	0,782576	0,301799	0,300893	0,454545	0,4	0,038576	0,666667	0,285714	0,75
SVC	0,571429	0,744048	0,291926	0,307743	0,454545	0,4	0,038576	0,666667	0,285714	0,75
RandomForest	0,464286	0,647712	0,312842	0,302721	0,545455	0,545455	0,178571	0,75	0,428571	0,75
XGBoost	0,428571	0,696465	0,307869	0,390095	0,454545	0,5	-0,06901	0,6	0,428571	0,5
MLP	0,535714	0,734957	0,308731	0,280929	0,454545	0,4	0,038576	0,666667	0,285714	0,75

Tablo 4.7 MVK geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,973684	0,999325	0,144294	0,304078	0,794872	0,882353	0,296174	1	0,789474	1
PolyPhen-2	1	1	0,086667	0,283135	0,948718	0,972973	0,561951	1	0,947368	1
SIFT	1	1	0,128716	0,350306	0,948718	0,972973	0,561951	1	0,947368	1
LogisticRegression	1	1	0,127301	0,292288	0,820513	0,898551	0,319333	1	0,815789	1
SVC	0,921053	0,997921	0,120094	0,263916	0,794872	0,882353	0,296174	1	0,789474	1
RandomForest	0,921053	0,997921	0,117383	0,264531	0,871795	0,929577	0,380443	1	0,868421	1
XGBoost	0,763158	0,993196	0,142552	0,300925	0,769231	0,865672	0,276253	1	0,763158	1
MLP	1	1	0,109919	0,276039	0,794872	0,882353	0,296174	1	0,789474	1

Tablo 4.8 NLRP3 geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,936441	0,990731	0,147953	0,245595	0,731343	0,823529	0,396889	0,976744	0,711864	0,875
PolyPhen-2	0,880297	0,980405	0,199008	0,321986	0,641791	0,75	0,316376	0,972973	0,610169	0,875
SIFT	0,811441	0,973537	0,180172	0,292356	0,820896	0,892857	0,376818	0,943396	0,847458	0,625
LogisticRegression	0,936441	0,990681	0,14825	0,243758	0,716418	0,811881	0,382108	0,97619	0,694915	0,875
SVC	0,936441	0,990718	0,142194	0,23491	0,776119	0,857143	0,445766	0,978261	0,762712	0,875
RandomForest	0,864407	0,976012	0,135019	0,231572	0,835821	0,899083	0,525731	0,98	0,830508	0,875
XGBoost	0,883475	0,980962	0,140413	0,237831	0,791045	0,867925	0,463899	0,978723	0,779661	0,875
MLP	0,913136	0,987476	0,145576	0,246016	0,791045	0,867925	0,463899	0,978723	0,779661	0,875

Tablo 4.9 NLRP12 geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,954545	0,5	0,303084	0,48873	0,695652	0,222222	0,291937	0,125	1	0,681818
PolyPhen-2	0,545455	0,090909	0,348363	0,527796	0,434783	0,133333	0,170941	0,071429	1	0,409091
SIFT	1	1	0,325639	0,52825	0,304348	0,111111	0,12666	0,058824	1	0,272727
LogisticRegression	0,954545	0,5	0,313589	0,487174	0,652174	0,2	0,265908	0,111111	1	0,636364
SVC	0,818182	0,2	0,314788	0,487779	0,608696	0,181818	0,243086	0,1	1	0,590909
RandomForest	1	1	0,364521	0,537819	0,565217	0,166667	0,222681	0,090909	1	0,545455
XGBoost	0,772727	0,166667	0,339249	0,511616	0,565217	0,166667	0,222681	0,090909	1	0,545455
MLP	0,954545	0,5	0,308547	0,483441	0,608696	0,181818	0,243086	0,1	1	0,590909

Tablo 4.10 NOD2 geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,9125	0,834821	0,15908	0,265078	0,857143	0,733333	0,63901	0,785714	0,6875	0,925
PolyPhen-2	0,739844	0,493177	0,241425	0,257548	0,642857	0,565217	0,351006	0,433333	0,8125	0,575
SIFT	0,705469	0,514415	0,271387	0,277535	0,446429	0,45614	0,114764	0,317073	0,8125	0,3
LogisticRegression	0,903125	0,829858	0,149331	0,2612	0,875	0,774194	0,688583	0,8	0,75	0,925
SVC	0,867188	0,734592	0,153404	0,233129	0,857143	0,764706	0,66502	0,722222	0,8125	0,875
RandomForest	0,842188	0,663962	0,169552	0,230414	0,821429	0,75	0,650421	0,625	0,9375	0,775
XGBoost	0,83125	0,658647	0,18097	0,207699	0,785714	0,684211	0,543434	0,590909	0,8125	0,775
MLP	0,840625	0,749617	0,165887	0,228777	0,821429	0,722222	0,601041	0,65	0,8125	0,825

Tablo 4.11 TNFRSF1A geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,964646	0,989279	0,072468	0,167636	0,921053	0,941176	0,835711	1	0,888889	1
PolyPhen-2	0,939394	0,979069	0,143619	0,168211	0,842105	0,896552	0,594797	0,83871	0,962963	0,545455
SIFT	0,909091	0,968304	0,171582	0,183631	0,789474	0,866667	0,438179	0,787879	0,962963	0,363636
LogisticRegression	0,952862	0,984378	0,07252	0,106295	0,921053	0,941176	0,835711	1	0,888889	1
SVC	0,952862	0,985052	0,076335	0,133169	0,921053	0,943396	0,814697	0,961538	0,925926	0,909091
RandomForest	0,969697	0,989435	0,065794	0,142748	0,947368	0,961538	0,885142	1	0,925926	1
XGBoost	0,976431	0,990907	0,087278	0,191553	0,894737	0,923077	0,762831	0,96	0,888889	0,909091
MLP	0,952862	0,984518	0,079506	0,178496	0,921053	0,943396	0,814697	0,961538	0,925926	0,909091

Tablo 4.12 PSTPIP1 geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,8	0,616667	0,195826	0,304742	0,733333	0,6	0,4	0,6	0,6	0,8
PolyPhen-2	0,76	0,717619	0,264873	0,284307	0,466667	0,5	0,1066	0,363636	0,8	0,3
SIFT	0,28	0,275873	0,307353	0,332237	0,333333	0,444444	-0,13868	0,307692	0,8	0,1
LogisticRegression	0,76	0,6	0,193392	0,218794	0,733333	0,6	0,4	0,6	0,6	0,8
SVC	0,72	0,683333	0,204718	0,232419	0,733333	0,6	0,4	0,6	0,6	0,8
RandomForest	0,82	0,59619	0,201101	0,287612	0,733333	0,6	0,4	0,6	0,6	0,8
XGBoost	0,78	0,545	0,216607	0,348944	0,733333	0,6	0,4	0,6	0,6	0,8
MLP	0,78	0,7	0,203576	0,22885	0,8	0,727273	0,57735	0,666667	0,8	0,8

Tablo 4.13 LPIN2 geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,9375	0,833333	0,171538	0,248676	0,8	0,5	0,375	0,5	0,5	0,875
PolyPhen-2	1	1	0,137201	0,35471	0,8	0,666667	0,612372	0,5	1	0,75
SIFT	0,875	0,583333	0,217094	0,304812	0,6	0,5	0,408248	0,333333	1	0,5
LogisticRegression	0,9375	0,833333	0,159363	0,345648	0,8	0,5	0,375	0,5	0,5	0,875
SVC	0,875	0,583333	0,179681	0,216689	0,8	0,5	0,375	0,5	0,5	0,875
RandomForest	0,9375	0,833333	0,153668	0,319859	0,8	0,5	0,375	0,5	0,5	0,875
XGBoost	0,8125	0,5	0,196055	0,254192	0,8	0,5	0,375	0,5	0,5	0,875
MLP	0,9375	0,833333	0,138497	0,293084	0,9	0,8	0,763763	0,666667	1	0,875

Tablo 4.14 PSMB8 geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	0,964286	0,95	0,166057	0,345163	0,818182	0,8	0,690066	0,666667	1	0,714286
PolyPhen-2	0,821429	0,65	0,220526	0,245832	0,727273	0,727273	0,571429	0,571429	1	0,571429
SIFT	0,892857	0,720833	0,282112	0,276828	0,454545	0,571429	0,239046	0,4	1	0,142857
LogisticRegression	0,964286	0,95	0,170474	0,341816	0,818182	0,8	0,690066	0,666667	1	0,714286
SVC	0,857143	0,679167	0,200161	0,269496	0,727273	0,727273	0,571429	0,571429	1	0,571429
RandomForest	0,857143	0,830357	0,252504	0,342149	0,636364	0,666667	0,462291	0,5	1	0,428571
XGBoost	0,714286	0,65	0,254474	0,260209	0,636364	0,666667	0,462291	0,5	1	0,428571
MLP	0,964286	0,95	0,183328	0,259978	0,727273	0,727273	0,571429	0,571429	1	0,571429

Tablo 4.15 IL36RN geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	1	1	0,056655	0,207252	1	1	1	1	1	1
PolyPhen-2	0,5	0,833333	0,231364	0,08248	0,666667	0,8	0	0,666667	1	0
SIFT	1	1	0,119719	0,344843	1	1	1	1	1	1
LogisticRegression	1	1	0,059101	0,201937	1	1	1	1	1	1
SVC	1	1	0,051858	0,198475	1	1	1	1	1	1
RandomForest	1	1	0,068064	0,22936	1	1	1	1	1	1
XGBoost	1	1	0,119219	0,315033	1	1	1	1	1	1
MLP	1	1	0,128033	0,280359	0,666667	0,8	0	0,666667	1	0

Tablo 4.16 IL1RN geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense			0,345929	0,551229	0,5	0	0	0	0	0,5
PolyPhen-2			0,091119	0,301858	1	0	0	0	0	1
SIFT			0,311922	0,548166	0,5	0	0	0	0	0,5
LogisticRegression			0,356937	0,54285	0,5	0	0	0	0	0,5
SVC			0,407061	0,565425	0,5	0	0	0	0	0,5
RandomForest			0,371584	0,519919	0,5	0	0	0	0	0,5
XGBoost			0,407238	0,553678	0,5	0	0	0	0	0,5
MLP			0,321197	0,498165	0,5	0	0	0	0	0,5

Tablo 4.17 OTULIN geninde kalibre edilmiş modellerin metrik performansları

Modeller	AUROC	AUPRC	Brier	ECE	Acc	F1	MCC	Precision	Recall	Specificity
AlphaMissense	1	1	0,07903	0,236155	1	1	1	1	1	1
PolyPhen-2	1	1	0,085212	0,287954	1	1	1	1	1	1
SIFT	0,725	0,68619	0,174552	0,138324	0,777778	0,8	0,55	0,8	0,8	0,75
LogisticRegression	1	1	0,060927	0,210688	1	1	1	1	1	1
SVC	1	1	0,0625	0,229863	1	1	1	1	1	1
RandomForest	1	1	0,046001	0,200382	1	1	1	1	1	1
XGBoost	1	1	0,054896	0,224691	1	1	1	1	1	1
MLP	1	1	0,048569	0,204003	1	1	1	1	1	1

5. TARTIŞMA

Otoinflamatuvar hastalıklar, doğuştan bağışıklık yanıtının düzenleyici moleküler mekanizmalarındaki kalıtsal bozukluklar sonucunda ortaya çıkan, tekrarlayıcı ateş atakları ve sistemik inflamasyonla seyreden heterojen bir hastalık grubunu oluşturur. Özellikle IL-1 ailesi eksenindeki sinyal yolları, NLR ailesi inflamazom proteinleri ve pyrin kompleksi bu hastalıkların patogenezinde merkezi düğümler olarak tanımlanmaktadır (Kastner ve ark., 2010). Klinik yük açısından bakıldığında, özellikle Akdeniz havzasında sık görülen FMF ve diğer monogenik sendromlar morbiditeyi artırmakta; tanının gecikmesi veya yanlış sınıflandırma hastalık seyrini ve prognozu belirgin biçimde olumsuz etkileyebilmektedir (Kastner ve ark., 2010). Moleküler tanı yöntemlerinin rutin kullanıma girmesine rağmen, gen paneli taramalarında klinik önemi belirsiz tek nükleotid varyantlarının (VUS) yüksek oranlarda saptanması, yorumlama sürecinin temel darboğazlarından biri olarak varlığını sürdürmektedir. ACMG/AMP tarafından önerilen standartlaştırılmış sınıflama çerçevesi bu belirsizliği yönetmek için yapılandırılmış kriterler sunsa da, özellikle nadir ve popülasyonlara spesifik varyantlarda klinik ve fonksiyonel kanıt üretiminin sınırlı kalması, karar sürecini çoğu zaman hesaplamalı tahminleyicilerin sunduğu ek bilgiye bağımlı hale getirmektedir (Richards ve ark., 2015). Bu nedenle, variant effect predictor (VEP) olarak adlandırılan hesaplamalı patojenite araçlarının doğruluğu ve güvenilirliği kritik bir klinik mesele haline gelmiştir.

Literatürde VEP çıktılarının doğruluğunu sistematik olarak değerlendiren çok sayıda çalışma bulunmaktadır. ilk nesil karşılaştırmalı benchmark çalışmaları missense tahmin araçlarının eğitim setleri ile değerlendirme veri kümeleri arasındaki örtüşme ve veri sızıntısı nedeniyle “dairesellik” (circularity) sorunu üretebildiğini, bunun da ayrıştırma metriklerinde yapay olarak şişirilmiş, aşırı iyimser sonuçlara yol açtığını göstermiştir (Grimm ve ark., 2015). Bunu izleyen geniş ölçekli derleme ve benchmark çalışmaları, FATHMM, LRT, MutationAssessor, PolyPhen-2 ve SIFT

gibi çok sayıda yöntemi ortak veri kümeleri üzerinde karşılaştırmış; yöntemlerin ayrıştırma gücü ve kalibrasyon özelliklerinin hedef gen grubuna, varyantın yapısal/evrimsel bağlamına ve kullanılan etiketleme stratejisine duyarlı olduğunu ortaya koymuştur (Li ve ark., 2018). Bu durumda AlphaMissense, PolyPhen-2 ve SIFT gibi araçların panel-özgü performansının tüm genoma yönelik literatürden doğrudan devralınamayabileceği; bu nedenle hedef klinik panel düzeyinde benchmark yapılmasının gerekli olduğu, hatta gen bazında da farklılıkların söz konusu olduğu gösterilmiştir (Tablo 4.6-4.17). Bu bulguların doğal uzantısı olarak, çoklu VEP skorlarını ve ek omik ya da evrimsel özellikleri bütünleştiren topluluk (ensemble) yaklaşımları ön plana çıkmıştır. REVEL gibi araçlar, tekil tahminleyicilere kıyasla daha dengeli doğruluk ve daha öngörülebilir klinik performans sunarak yaygın kullanım alanı bulmuştur (Ioannidis ve ark., 2016). Benzer şekilde CADD gibi tüm genomu kapsayan çerçeveler, işlevsel ve biyokimyasal çok katmanlı ipuçlarını gömülü modeller aracılığıyla birleştirerek ayırt ediciliği artırmayı amaçlamakta ve varyant önceliklendirme süreçlerinde referans noktası olarak kullanılmaktadır (Rentzsch ve ark., 2019).

Makine öğrenmesine dayalı yaklaşımlar, genel olarak iki ana eksende gelişmiştir. Birinci grupta, CADD, M-CAP, MetaLR/MetaSVM gibi tüm genoma yönelik modeller yer almakta; bu yöntemler geniş etiket setlerinden ve çok sayıda anotasyondan yararlanarak yüksek kapsayıcılık sağlamakta, ancak dokuya, biyolojik yolaklara veya belirli protein komplekslerine özgü biyolojiyi her zaman yeterince yansıtamamaktadır (Dong ve ark., 2015; Jagadeesh ve ark., 2016). İkinci grupta ise gen-özgül veya hastalık-özgül modeller bulunmaktadır. Özellikle BRCA1/2 için geliştirilen makine öğrenmesi çerçevelerinde olduğu gibi, bu tür modeller fonksiyonel deney çıktılarının yoğun olduğu genlerde alan bilgisi, domain/motif düzeyinde yapı bilgisi ve zengin fonksiyonel veri setlerini avantaja çevirerek anlamlı performans kazandıkları bildirmiştir (Kang ve ark., 2023).

Otoinflamatuvar hastalıklar bağlamında ise, literatürdeki hesaplamalı patojenite çalışmaları, tüm genom veya geniş kanser gen panellerine ya da çeşitli hastalıklara odaklanan genel modellerle karşılaştırıldığında oldukça sınırlı kalmaktadır. FMF ve diğer monogenik otoinflamatuvar sendromlarda MEFV ve ilişkili genlere ait varyant dağılımlarını tanımlayan, NGS ile klasik yöntemleri karşılaştıran ya da tekil gen

düzeyinde sınıflandırma denemeleri sunan çalışmalar bulunmaktadır, ancak bu çalışmalar çoğunlukla ya yalnızca tanımlayıcı niteliktedir ya da belirli bir gen ya da alt popülasyona odaklıdır. Yaptığımız taramada, otoinflamatuvar gen paneli için tasarlanmış, aynı etiketli klinik veri matrisi üzerinde birden fazla VEP skorunun sistematik olarak benchmark edildiği ve bu skorların birlikte makine öğrenmesi sınıflandırıcıları için özellik uzayı olarak kullanıldığı, uçtan uca tasarlanmış bir çerçeveye rastlanmamıştır. Bu durum, klinik karar desteği açısından, otoinflamatuvar panel düzeyinde hem klasik araçların hem de bu araçlardan türeyen ML modellerinin aynı veri, aynı eşik ve aynı metrikler altında karşılaştırılmasına yönelik bir boşluğa işaret etmektedir.

Son dönemde geliştirilen AlphaMissense, protein yapı bilgisini ve evrimsel ipuçlarını derin öğrenme modelleriyle birleştirerek yaklaşık 70 milyon missense varyant için olasılık tabanlı patojenite skorları raporlamış ve çeşitli fonksiyonel veri setlerinde güçlü ayrım gücü sergilemiştir (Cheng ve ark., 2023). Bununla birlikte, mevcut çalışmalar çoğunlukla tüm proteom veya geniş gen panelleri düzeyinde, sıklıkla kanserle ilişkili genler üzerine odaklanmış; AlphaMissense'in otoinflamatuvar gen paneli bağlamında klasik patojenite tahmin araçları (VEP) ve bu araçlardan elde edilen skorları girdi olarak kullanan ML sınıflandırıcıları ile birlikte, aynı etiketli klinik veri seti üzerinde eşzamanlı olarak değerlendirilmesine yönelik bir çalışmaya rastlanmamıştır (Li ve ark., 2018; Ioannidis ve ark., 2016; Kang ve ark., 2023).

Bu tez, söz konusu boşluğu hedefleyerek; aynı otoinflamatuvar panel, aynı örneklem ve ortak metodoloji üzerinde klasik araçlar ve AlphaMissense'in çıktıları ile eğitilmiş makine öğrenmesi modellerinin ayrım gücü, kalibrasyon davranışı ve istatistiksel olarak test edilmiş farklarını bir arada raporlamayı amaçlamaktadır. Böylece panel düzeyinde klinik yorumlama açısından "hangi skorun hangi bağlamda daha güçlü olduğu ve hangi ayarlarda güvenilir kabul edilebileceği" sorusuna, metodolojik olarak sağlam, doğrudan karşılaştırılabilir ve istatistiksel olarak temellendirilmiş yanıtlar sunulmaktadır. Buna yönelik, on iki monogenik otoinflamatuvar hastalığın çekirdek genlerinden oluşturulan bir panel üzerinde, varyant etkisi tahminleyicileri (Variant Effect Predictors, VEP) ile bu araçlardan türetilen skorlar ile eğitilen makine öğrenmesi sınıflandırıcılarının, ortak bir klinik etiketli veri kümesi üzerinde sistematik ve yöntemsel olarak karşılaştırılması amaçlanmaktadır. Bu çerçevede üç

temel hedef belirlenmiştir. Birincisi, klasik VEP'ler (SIFT, PolyPhen-2) ile derin öğrenme tabanlı AlphaMissense'in ayrıştırma gücünün ROC eğrisi ve AUPRC üzerinden, olasılık kalitesinin ise Brier skoru, Expected Calibration Error (ECE), calibration slope/intercept ve Hosmer–Lemeshow istatistiği ile nicel olarak değerlendirilmesidir. İkincisi, VEP çıktılarından elde edilen kalibre olasılıkları öznitelik olarak kullanarak Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost ve Multi Layer Perceptron (MLP) gibi farklı model ailelerinin eğitilmesi ve bu modellerin panel bağlamında sağladığı ek ayrım ve kalibrasyon katkısının ortaya konulmasıdır. Üçüncüsü ise, klinik karar düzeyindeki farkların ortak bir karar eşiği altında sınanmasıdır. Bu amaçla eşik bağımlı farkların McNemar exact testiyle, eşik bağımsız ayrım üstünlüklerinin ise eşleştirilmiş bootstrap ΔAUC yaklaşımıyla istatistiksel olarak test edilmesi ve çoklu karşılaştırmalarda Benjamini–Hochberg False Discovery Rate (FDR) düzeltmesinin uygulanmasıdır.

Değerlendirme stratejisi iki katmanlı bir yapı üzerinde kurgulanmıştır. İlk katmanda, VEP'lerin ham skorları kendi içlerinde optimum Youden's J eşiğinde ve eşik bağımsız metriklerle incelenmiş, aynı zamanda bu skorlar Platt scaling ile kalibre edilerek ML sınıflandırıcılarının eğitimi için ortak bir olasılık uzayı oluşturulmuştur. İkinci katmanda ise, karşılaştırılabilirliği artırmak üzere hem VEP'ler hem de bu VEP tabanlı ML modelleri kalibrasyon sonrası sabit bir karar eşiğinde (t=0.50) değerlendirilmiş, böylece hem her yöntem için kendi doğasında (native) en iyi performans noktasına karşılık gelen model özel işletim eşiği hem de klinik uygulamada sık kullanılan sabit eşik koşulundaki göreli üstünlük ve güvenilirlik aynı veri kümesi üzerinde görünür hâle getirilmiştir.

Bu tasarım, tek bir otoinflamatuvar gen paneli ve ortak etiket seti üzerinde, ayrım (ROC/AUPRC) ve kalibrasyon (Brier, ECE, eğim/kesişim, HL) boyutlarını birlikte ele alması, ayrıca eşik düzeyindeki hata simetrisini McNemar testiyle, eşik bağımsız ayrım farklarını ise eşleştirilmiş bootstrap ΔAUC ile istatistiksel olarak temellendirmesi bakımından bütüncül bir ölçüm çerçevesi sunmaktadır. Bu bütüncül yaklaşım, panel bağlamında hem doğrudan VEP skorlarının hem de bu skorlardan öğrenen ML modellerinin hangi koşullarda daha avantajlı olduğunu, hangi ayarlarda

klinik olarak daha güvenilir kabul edilebileceğini sayısal ve karşılaştırılabilir bir zeminde tartışmaya imkân tanımaktadır.

5.1. Ana Bulgular

Monogenik otoinflamatuvar hastalıklardan oluşan hedef gen paneli ile oluşturulan varyant kümesi üzerinde, AlphaMissense, PolyPhen-2 ve SIFT varyant etki tahmin araçlarının ve bu araçlardan elde edilen skor setleri ile eğitilen makine öğrenmesi araçlarının hem ayrıştırma gücü hem de olasılık kalitesi yönlerinden karşılaştırılmıştır.

Değerlendirme, önce her aracın ve modelin kendi "doğal" davranışını (ham skorlar ve model-özel optimum eşik) ortaya koyan bir katmanda, ardından kalibrasyon sonrasında tüm yöntemlerin ortak bir karar eşiğine (t = 0,50) taşındığı ikinci katmanda gerçekleştirilmiştir. Böylece hem tek tek yöntemlerin kendi içinde ulaşabildiği en dengeli hata profili hem de panel düzeyinde sabit bir klinik eşiğe göre görece üstünlükleri görünür hâle gelmiştir.

Ham skorlar üzerinden bakıldığında, ROC eğrileri ve AUROC değerleri AlphaMissense'in otoinflamatuvar panelde belirgin bir ayrıştırma üstünlüğü sergilediğini göstermektedir. Özellikle ROC eğrisinin düşük yanlış pozitif bölgede hızla yükselmesi ve köşeye yakın seyretmesi, erken duyarlılık kazanımı açısından güçlü bir profil ortaya koymaktadır. Bu durum, Tablo 4.1'de raporlanan yüksek AUROC ve AUPRC değerleriyle de nicel olarak doğrulanmaktadır. Bu sonuç, AlphaMissense'in protein yapı bilgisi ile çok ölçekli evrimsel ipuçlarını derin öğrenme çerçevesinde birleştiren mimarisinin, otoinflamatuvar panel bağlamında da etkili olduğunu desteklemektedir. Logistic Regression, yalnızca üç kalibre edilmiş VEP olasılığından beslenmesine rağmen, ROC geometrisi ve özet metrikler itibarıyla AlphaMissense'i yakından izlemektedir. Bu desen, birden fazla VEP türevli özelliğin iyi düzenlenmiş bir lojistik regresyon modeli altında birleştirilmesinin, tekil kestiricilere kıyasla çoğu zaman güçlü ve istikrarlı performans sağlayabildiğini bildiren önceki çalışmalarla uyumludur (Jagadeesh ve ark., 2016). Buna karşılık PolyPhen-2 ve SIFT ham skorları, diyagonale daha yakın ROC profilleri ve daha kırılgan özgüllük duyarlılık dengeleri ile, bu panelde ayrım gücü açısından daha geride kalmıştır.

Olasılık kalitesi değerlendirildiğinde, kalibrasyon eğrileri ve özet istatistikler (Brier skoru, Expected Calibration Error, kalibrasyon eğimi ve kesişimi, Hosmer–Lemeshow testi) ayrım ile kalibrasyonun birbirinden bağımsız boyutlar olduğunu açık biçimde göstermektedir. Kalibrasyon sonrasında ortak 0,50 eşiğinde hesaplanan Brier ve ECE değerleri, özellikle Logistic Regression, MLP ve SVM modellerinde belirgin bir iyileşmeye işaret etmiş, bu modellerin ürettiği olasılıkların ortak eşik etrafında gerçek sınıf frekanslarıyla daha iyi örtüştüğü görülmüştür. Buna karşın AlphaMissense, Random Forest ve XGBoost için bazı senaryolarda Hosmer–Lemeshow testinin anlamlı kalması, bu araçların güçlü ayrım gücüne rağmen belirli olasılık bantlarında sistematik sapmalar gösterebildiğini düşündürmektedir. Bu desen, VEP benchmark çalışmalarında sık vurgulanan "yüksek ayrıştırma – değişken kalibrasyon" ikilisini otoinflamatuvar panel bağlamında yeniden üretmekte; genel amaçlı veya genome-wide kestiricilerin panel düzeyinde kullanılmadan önce yeniden kalibre edilmesi gerektiğine işaret etmektedir. Buna karşın AlphaMissense, Random Forest ve XGBoost için bazı senaryolarda Hosmer–Lemeshow uyum testinin anlamlı çıkması, bu yöntemlerin yüksek ayrıştırma gücüne rağmen belirli olasılık aralıklarında sistematik kalibrasyon hataları sergilediğini düşündürmektedir. Kalibrasyon eğiminin (slope) 1'in altında olması, tahmin olasılıklarının gerçekte gözlenen oranlara göre olduğundan daha uç değerlere itildiğine ve modelin kısmen "aşırı güvenli" davrandığına işaret etmektedir. Kalibrasyon sabitinin (intercept) ideal değer olan sıfırdan saparak yaklaşık 0,47–0,49 bandında kümelenmesi ise paneldeki gerçek sınıf oranına kıyasla risk düzeyinin genel olarak yukarıya kaydığını düşündürmektedir (Tablo 4.3). Başka bir ifadeyle, bu modeller bazı olasılık dilimlerinde gerçekte daha az patojen olan bir küme için fazla yüksek olasılık önermekte, bazı dilimlerde ise tersi yönde sistematik sapma üretebilmektedir. Bu desen, missense varyant kestiricilerinin çoğunda bildirilen "yüksek ayrıştırma gücü, ancak değişken kalibrasyon" örüntüsüyle uyumludur (Grimm ve ark., 2015; Li ve ark., 2018) ve ham VEP olasılıklarının klinikte doğrudan kullanılmasından ziyade, panel özgü bir yeniden kalibrasyon adımının gerekliliğini desteklemektedir.

Eşik bağımsız istatistiksel karşılaştırmalar, paired-bootstrap ΔAUC yaklaşımı ile yürütülmüş ve hem güven aralıkları hem de çoklu karşılaştırma düzeltmesi dikkate alınarak yorumlanmıştır. Bu analizler, AlphaMissense'in PolyPhen-2 ve SIFT'e

kıyasla otoinflamatuvar panel üzerinde daha yüksek AUROC değerleri elde ettiğini ve bu üstünlüğün tesadüfî dalgalanmalardan ziyade istatistiksel olarak anlamlı bir farkı yansıttığını göstermiştir (Tablo 4.4). Aynı yöntemle yapılan kıyaslamalar, Logistic Regression modelinin de bazı ağaç tabanlı Random Forest ve XGBoost modellerine karşı ΔAUC açısından avantaj sağlayabildiğini ortaya koymuştur. Bu durum, birden fazla VEP kökenli özelliğin iyi düzenlenmiş bir lojistik regresyon çerçevesinde birleştirilmesinin, missense patojenite kestiriminde çoğu zaman güçlü ve istikrarlı bir performans sunabildiğine dair önceki gözlemlerle uyumludur (Ioannidis ve ark., 2016; Li ve ark., 2018).

Karar düzeyindeki farklılıklar, 0,5 ortak eşiğinde McNemar exact testi kullanılarak değerlendirilmiştir. Her model çifti için oluşturulan 2×2 tablolarda "Model A doğru–Model B yanlış" ve "Model A yanlış–Model B doğru" hücreleri karşılaştırıldığında, bazı çiftlerde anlamlı hata asimetrileri saptanmıştır (Tablo 4.5). Bu bulgu, aynı veri üzerinde benzer AUROC değerleri üreten iki modelin, hangi örneklerde hata yaptıkları açısından klinik olarak önemli ölçüde farklı davranabileceğini göstermektedir. Başka bir deyişle, AUROC denkliği tek başına modellerin klinik pratikte birbirinin yerine sorunsuz biçimde kullanılabileceği anlamına gelmemektedir. McNemar testi, özellikle "hangi hastaların yanlış sınıflandırıldığı" sorusuna odaklanarak, ayrım metriklerinin tek başına ortaya koyamadığı farklılıkları görünür kılmıştır.

Genel resim şu şekilde özetlenebilir. Ayrıştırma açısından AlphaMissense, protein yapı–işlev temsillerini ve çok ölçekli evrimsel bilgiyi derin öğrenme ile birleştirme kapasitesi sayesinde otoinflamatuvar panelde belirgin bir early lift ve yüksek AUROC değerleri üretmiştir. Logistic Regression modeli ise yalnızca üç kalibre VEP olasılığından beslenmesine rağmen bu performansı yakından izlemiştir. Kalibrasyon açısından bakıldığında, Logistic Regression, MLP ve SVM gibi modellerin ortak 0,5 eşiğine taşındığında daha düşük Brier skoru ve daha düşük Expected Calibration Error (ECE) değerleriyle klinik açıdan daha güvenilir olasılıklar sunduğu görülmüştür. Buna karşılık, Hosmer–Lemeshow testinin anlamlı kaldığı modellerde, sapmanın hangi olasılık aralıklarında yoğunlaştığı hem kalibrasyon eğrileri hem de slope/intercept tahminleri ile ayrıntılı biçimde gösterilmiştir. İstatistiksel eşleştirilmiş kıyaslar (ΔAUC ve McNemar) ise, ayrım üstünlüğünün rastlantısal olmadığını ve

karar düzeyi hata örüntülerinin modelden modele anlamlı biçimde değişebildiğini ortaya koymuştur. Bu bulgular, daha önce tüm genom veya çeşitli özgül veri kümelerinde rapor edilen eğilimlerle uyumlu olmakla birlikte (Ioannidis ve ark., 2016; Li ve ark., 2018), otoinflamatuvar gen paneli üzerinde aynı örneklem ve ortak metodoloji altında VEP skorları ve bu skorlarla eğitilen ML modellerini, ayrıştırma, kalibrasyon ve karar düzeyi testlerini eş zamanlı olarak inceleyen ilk kapsamlı ve istatistiksel açıdan titiz karşılaştırma çerçevesini sunması bakımından literatüre özgün bir katkı sağlamaktadır.

Klinik açıdan bakıldığında, bu bulgular otoinflamatuvar gen panellerinde iki aşamalı bir karar stratejisini desteklemektedir. İlk aşamada, yüksek ayrıştırma gücüne sahip örneğin AlphaMissense veya panelde iyi performans gösteren Logistic Regression modellerinin skorlarının tarama ve önceliklendirme amacıyla kullanılması, VUS havuzunun daha yönetilebilir bir alt kümeye indirgenmesine yardımcı olabilir. İkinci aşamada ise, panel özgü yeniden kalibrasyon uygulanmış olasılıkların, hedeflenen duyarlılık/özgüllük veya Positive Predictive Value (PPV)/ Negative Predictive Value (NPV) dengesine uygun sabit bir eşikte yorumlanması önerilmektedir. Bu iki adımlı yaklaşım, VUS yorumunun iyileştirilmesine yönelik güncel önerilerle uyumludur (Grimm ve ark., 2015; Richards ve ark., 2015) ve otoinflamatuvar genetiğinde daha dengeli, öngörülebilir ve şeffaf bir klinik karar desteği sunma potansiyeline sahiptir.

5.2. Ayrıştırma Performansı Sonuçları

Ham skorların incelendiği aşamada ROC eğrileri, AlphaMissense'in düşük yanlış-pozitif oranlarında (FPR ≈ 0–0,15) belirgin bir "early-lift" ürettiğini ve duyarlılığı hızla artırdığını göstermiştir. Bu görsel desen, AUROC=0,870 ve AUPRC=0,906 değerleriyle nicel olarak da doğrulanmaktadır (Şekil 4.5A; Tablo 4.1). Aynı aşamada Logistic Regression (uncalibrated) için elde edilen ROC eğrisi, köşe bölgesine yakın konveks yapısıyla dikkat çekmiş; AUROC=0.851 ve AUPRC=0.873 değerleri, bu modelin de panel düzeyinde güçlü bir ayrıştırma performansı sunduğunu göstermiştir (Şekil 4.6B; Tablo 4.1). Buna karşılık PolyPhen-2 ve SIFT'in ROC eğrileri diyagonale daha yakın seyretmiş ve AUROC değerleri orta düzeyde kalmıştır (Tablo 4.1). Eşik bağımsız istatistiksel karşılaştırma da bu tabloyu desteklemektedir. paired-bootstrap ΔAUC analizleri, AlphaMissense ile PolyPhen-2 ve SIFT arasındaki

ΔAUC dağılımlarının sıfırın anlamlı biçimde üzerinde ve dar güven aralıklarıyla seyrettiğini göstermiş (q≈0) ve böylece AlphaMissense'in üstün ayrım gücünün rastlantısal örnekleme dalgalanmalarından bağımsız olarak sürdüğünü ortaya koymuştur (Tablo 4.4). Bu bulgular, geniş ve heterojen veri üzerinde eğitilen kestiricilerin yüksek ayrım metrikleri üretebildiğini, ancak performanslarının gen ya da yolak bağlamına göre değişebildiğini bildiren derleme çalışmalarla uyumludur (Li ve ark., 2018). Ayrıca AlphaMissense'in protein yapı temsilleri ile çok ölçekli evrimsel bilgiyi derin öğrenme modelleri aracılığıyla bütünleştirerek proteom genelinde raporlanan yüksek ayrıştırma performansı, bu tezde otoinflamatuvar panel üzerinde elde edilen ROC/AUPRC sonuçlarıyla paralellik göstermektedir (Cheng ve ark., 2023). Makine öğrenmesi sınıflandırıcıları arasında Logistic Regression'ın ROC/AUPRC düzeyindeki istikrarlı performansı, gen-özgül veri bağlamlarında, alan bilgisi ve fonksiyonel deney çıktılarıyla desteklenen modellerde gözlenen dengeli kazanımlarla benzer bir örüntü sergilemektedir (Kang ve ark., 2023). Bununla birlikte, burada gözlenen üstünlüklerin temel belirleyicisi, giriş özelliklerinin VEP skorlarından türetilmiş olması ve tüm yöntemlerin aynı otoinflamatuvar gen paneli ile aynı ClinVar etiket yapısı üzerinde karşılaştırılmasıdır. Bu yönüyle çalışma, özgün gen paneli ya da tüm düzeyindeki önceki yaklaşımlarla aynı literatür ekseninde yer almakla birlikte, panel düzeyinde ortak bir veri matrisi ve sistematik istatistiksel eşleştirme (paired-bootstrap ΔAUC) kullanması bakımından yöntemsel olarak ayrışmaktadır. Kalibrasyon sonrasında ortak eşikte (t=0.50) yapılan değerlendirmeler, yöntemler arasındaki ayrım sıralamasının genel hatlarıyla korunduğunu göstermiştir (Tablo 4.2). Bu gözlem, uygun şekilde uygulanan olasılık kalibrasyonunun ROC/AUPRC gibi eşik bağımsız ayrım ölçütlerini bozmadığını; tersine, olasılıkların doğru ölçeklenmesi sayesinde uygulama düzeyinde yorumlanan karar metriklerinin (Accuracy, F1, MCC gibi) daha tutarlı biçimde değerlendirilebildiğini düşündürmektedir. Nitekim paired-bootstrap ΔAUC sonuçları, Logistic Regression'ın ağaç tabanlı Random Forest ve XGBoost modellerine karşı anlamlı ΔAUC üstünlüğü sağlayabildiğini (q≈0) ve AlphaMissense'in PolyPhen-2 ile SIFT'e karşı ayrım üstünlüğünü koruduğunu doğrulamıştır (Tablo 4.4). Bu genel desen, VEP benchmark çalışmalarında sık vurgulanan "ayrıştırma bakımından güçlü, ancak olasılık kalitesi (kalibrasyon) açısından daha heterojen" ikilisinin

otoinflamatuvar panel bağlamında, aynı veri matrisi üzerinde ve eşleştirilmiş istatistiksel yöntemlerle yeniden gösterildiğini ortaya koymaktadır (Grimm ve ark., 2015; Li ve ark., 2018). Ayrıca literatürde tüm genom yaklaşımların örneğin REVEL, CADD gibi geniş kapsama alanı karşılığında bağlama özgü ayrıntıları her zaman tam olarak yansıtamayabildiği, bu nedenle panel düzeyinde yeniden değerlendirme ve kalibrasyonun gerekli olduğu yönündeki bulgular da bu çalışmada gözlenen örüntülerle uyumludur (Ioannidis ve ark., 2016; Rentzsch ve ark., 2019; Cheng ve ark., 2023). Bu bağlamda AlphaMissense, ROC ve AUPRC açısından panel üzerinde güçlü bir ayrıştırıcı olarak kalmakta; ancak klinik karar eşiğine yakın bölgelerde hangi modelin tercih edileceği sorusunda, kalibrasyon kalitesi ve karar düzeyindeki hata örüntüleri belirleyici hale gelmektedir.

5.3. Olasılık Skorları ve Kalibrasyon Kalitesi

Bununla birlikte, ayrıştırma tek başına klinik karar kalitesini garanti etmez. Bu yüzden ayrıştırma (ROC/AUPRC) metrikleri yanında, modellerin ürettiği olasılıkların gerçek sınıf frekanslarıyla ne kadar uyumlu olduğunu incelemek için kalibrasyon analizi yapılmıştır. Kalibrasyon diyagramları ve özet metrikler birlikte değerlendirildiğinde, iyi ayrıştıran bir modelin mutlaka iyi kalibre edilmiş olasılıklar üretmediği açıkça görülmektedir (de Hond ve ark., 2023; Huang ve ark., 2020; Steyerberg ve ark., 2010). Bu durum, klinik karar verme açısından ayrıştırma ile kalibrasyonun birbirinden bağımsız iki boyut olarak ele alınması gerektiğini desteklemektedir.

Ham skor evresinde bazı modellerin Youden noktasının sıra dışı derecede yüksek eşiklere kayması ve reliability eğrilerinde sistematik sapma göstermesi, düşük ve orta olasılık bantlarında "aşırı güven ve "yetersiz güven" örüntülerinin birlikte bulunabildiğine işaret etmektedir (Grimm ve ark., 2015; Li ve ark., 2018). Bu gözlemler Tablo 4.1'deki Brier skoru ve ECE değerleriyle uyumludur.

Kalibrasyon uygulandıktan ve tüm yöntemler 0,5 ortak karar eşiğine taşındıktan sonra, Brier skoru ve ECE'nin özellikle Logistic Regression, MLP ve SVM hattında belirgin biçimde düştüğü izlenmiştir (Tablo 4.2). Bu düşüş, söz konusu modellerin ürettiği olasılık çıktılarının klinik yorumlamaya, örneğin risk iletişimi, olasılığa bağlı karar kuralları ve eşik optimizasyonu gibi uygulamalar için daha elverişli hale

geldiğini göstermektedir (Huang ve ark., 2020; Van Calster ve Vickers, 2015). Brier skoru (tahmin olasılıkları ile gerçekleşen sonuçlar arasındaki kare farkların ortalaması) olasılık kalitesini tek bir özet değerle yakalarken, ECE (Expected Calibration Error) olasılık aralıkları bazında beklenen–gözlenen farkını nicelleştirir; her iki metrikteki eşzamanlı iyileşme, kalibrasyonun hem küresel (global) hem de yerel düzeyde işe yaradığını düşündürmektedir (Guo ve ark., 2017; Huang ve ark., 2020).

Kalibrasyonun daha ayrıntılı yorumlanmasında, doğrusal kalibrasyon katsayıları (calibration slope ve intercept) ile Hosmer–Lemeshow (HL) uyum testinin birlikte değerlendirilmesi önemlidir. Çalışmamızda Tablo 4.3'te AlphaMissense, Random Forest ve XGBoost için HL χ^2 istatistiğinin anlamlı olduğu ($p<0.05$) görülmüştür. Bu bulgu, bu modellerin "kötü" olduğu anlamına gelmemektedir; HL testi, tahmin edilen olasılık ve gerçekte gözlenen pozitiflik oranı uyumu belirli olasılık grupları içinde sınar ve güçlü ayrıştırmaya sahip ancak belirli olasılık bantlarında sistematik sapma gösteren modellerde test istatistiğinin anlamlı çıkması mümkündür (Huang ve ark., 2020; Steyerberg ve ark., 2010). Elde ettiğimiz tabloda kalibrasyon eğiminin (slope) 1'in belirgin biçimde altında kalması, tahmin olasılıklarının uçlara fazla itildiğini ve doğrusal bir "büzülme" (shrinkage) ile yeniden ölçeklendirmenin teorik olarak yerinde olacağını düşündürmektedir. Literatürde, slope<1 deseninin aşırı güven (overconfidence) ile sık ilişkili olduğu daha önce de vurgulanmıştır (Niculescu-Mizil ve Caruana, 2005; Van Calster ve ark., 2019). Kesişimin (intercept) 0'dan anlamlı sapma göstermesi ve bizim panelimizde ~0.47–0.49 bandında kümelenmesi ise, temel olay sıklığı (base rate), etiketleme stratejisi ve eğitim–değerlendirme dağılım farklarının olasılık ölçeğini sistematik biçimde kaydırabildiğine işaret etmektedir ve normalde idealde intercept ≈ 0 beklenir ve anlamlı sapma çoğunlukla tahmin edilen temel risk düzeyinin gerçeğe göre kaymış olduğunu gösterir (Steyerberg ve ark., 2010; Van Calster ve ark., 2019).

Kalibrasyon yöntemleri ile elde edilen desenler de bu yorumları desteklemektedir. Bu tezde model çıktıları için kullanılan Platt Scaling stratejisinin ardından özellikle Logistic Regression, MLP ve SVM modellerinde ECE ve Brier skorlarında gözlenen belirgin iyileşme, reliability diyagramlarında da bant-spesifik sapmaların azaldığı görsel olarak doğrulanmıştır. Buna karşın HL testinin anlamlı kalmaya devam ettiği

modellerde, sapmanın hangi olasılık aralıklarında yoğunlaştığı kalibrasyon eğrileri üzerinden ayrıntılı biçimde izlenebilmiştir.

Varyant etki kestiricileri için daha önce yapılmış geniş ölçekli karşılaştırmalarda da benzer bir örüntü rapor edilmiştir: bazı araçlar yüksek ROC/AUPRC değerleriyle güçlü ayrıştırma sağlarken, mutlak olasılıkların kalibrasyonsuz kullanımı özellikle klinik karar eşiği düzeyinde yanlı sonuçlara yol açabilmektedir (Grimm ve ark., 2015; Li ve ark., 2018). Bu tezde otoinflamatuvar panel üzerinde elde ettiğimiz bulgular, söz konusu genel eğilimi desteklemektedir. Tüm genoma yönelik ya da geniş panel kökenli kestiricilerin, belirli bir klinik panelde kullanılmadan önce panel- ve laboratuvar-özgü yeniden kalibrasyon gerektirdiğini göstermektedir. Klinik açıdan çıkarılabilecek temel sonuç şudur: kararı doğrudan tahmin edilen olasılıklara dayandıracaksak, bu olasılıkların panel bağlamında yeniden kalibre edilmesi ve kullanılacak eşiklerin (örneğin belirli bir PPV/NPV, Sensitivity/Specificity hedefi için) bu kalibre ölçek üzerinde belirlenmesi gerekir (Huang ve ark., 2020; Van Calster ve ark., 2019). Aksi takdirde, ayrıştırma gücü yüksek bir model bile belirli olasılık aralıklarında sistematik sapma nedeniyle HL testi anlamlı çıkabilir ve bu sapma, klinik düzeyde gereksiz pozitiflik artışı ya da gözden kaçan olgular şeklinde somutlaşabilir.

Bu çerçevede genel tablo birkaç temel noktada özetlenebilir. Öncelikle, yüksek ROC ve AUPRC değerleri tek başına iyi kalibre edilmiş olasılıklar anlamına gelmemektedir; aynı modele ait Brier skoru ve ECE değerleri hâlâ yüksek kaldığında, ayrım gücü güçlü olsa bile mutlak olasılıkların klinik karar kurallarında doğrudan kullanılmasının hatalı sınıflandırmalara zemin hazırlayabileceği görülmektedir. Ayrıca kalibrasyon eğiminin 1'in belirgin biçimde altında, kesişimin ise 0'dan uzak olduğu durumlarda model tahminleri gereğinden fazla uçta ve sistematik olarak kaymış hale gelmekte, modern tahmin modeli literatüründe bu tür desenler, aşırı güvenli (overconfident) olasılık tahminleriyle ilişkilendirilmekte ve en azından Platt scaling gibi bir yeniden kalibrasyon adımı uygulanmasının gerekli olduğu vurgulanmaktadır (Niculescu-Mizil ve Caruana, 2005; Van Calster ve ark., 2019). Son olarak, Hosmer–Lemeshow testinin anlamlı çıkması modelin tümüyle elenmesi gerektiği anlamına gelmemekte, daha çok olasılık ekseninin hangi bölümlerinde tahmin edilen ve gözlenen olasılıkların uyumu dengesinin bozulduğunu

gösteren istatistiksel bir uyarı işlevi görmektedir (Huang ve ark., 2020). Bu çerçeve, otoinflamatuvar panel düzeyinde klinik karar kurallarının belirlenmesinde yalnızca ayrım üstünlüğüne odaklanmanın yeterli olmadığını; model seçimi ve eşik belirleme süreçlerinde olasılık kalitesinin de aynı ölçüde dikkate alınması gerektiğini ortaya koymaktadır.

5.4. Eşik Bağımsız ve Eşik Bağımlı İstatistiksel Karşılaştırmalar

Bu çalışmada, modeller arası üstünlüğü yalnızca tek bir metrik üzerinden değil, iki tamamlayıcı istatistiksel çerçeve üzerinden değerlendirdik. Amaç, aynı otoinflamatuvar panel üzerinde hem ayrıştırma gücünü hem de karar düzeyindeki davranışı, örneklemin iç bağımlılığını dikkate alarak nesnel biçimde sınamaktır. Eşik bağımsız düzlemde, her model çifti için AUROC farkını değerlendirmek üzere eşleştirilmiş bootstrap ΔAUC prosedürü kullanıldı. Her bootstrap yinelemesinde, veri kümesindeki tüm örnekler sınıf etiketleriyle birlikte yeniden örneklenerek her iki model için AUROC hesaplandı ve fark ΔAUC = AUROC_A − AUROC_B biçiminde elde edildi. Bu farkların bootstrap dağılımından % 95 güven aralıkları üretildi ve çoklu ikili karşılaştırmalar Benjamini–Hochberg düzeltmesiyle kontrol edildi (Tablo 4.4). Sonuçlar, AlphaMissense'in PolyPhen-2 ve SIFT'e karşı ΔAUC dağılımlarının sıfırın anlamlı biçimde üzerinde konumlandığını ve q≈0 düzeyinde üstünlük sağladığını gösterdi. Logistic Regression için de, özellikle Random Forest ve XGBoost ile yapılan kıyaslamalarda pozitif ve dar güven aralıklarına sahip ΔAUC değerleri elde edildi. Bu bulgu, geniş etiketli veri üzerinde eğitilen, protein yapı ve evrim bilgisini derin temsillerle birleştiren kestiricilerin ve düzenlileştirilmiş lineer sınıflandırıcıların ayrıştırma metriklerinde tutarlı avantaj sağlayabildiğine dair literatürde bildirilen sonuçlarla uyumludur (Ioannidis ve ark., 2016; Li ve ark., 2018). ROC eğrileri ve AUC farklarının bootstrap temelli, eşleştirilmiş bir çerçevede incelenmesi, dağılım varsayımlarına görece az bağımlı oluşu ve aynı örnekler üzerindeki korelasyonu doğrudan hesaba katabilmesi nedeniyle özellikle panel büyüklüğünün sınırlı ve sınıf dengesizliğinin belirgin olduğu klinik veri setlerinde yöntemsel açıdan sağlam bir yaklaşım olarak önerilmektedir (Robin ve ark., 2011).

Eşik bağımlı düzlemde ise, modellerin sabit 0,50 karar eşiğinde aynı örneklerde hata yapma eğilimlerinin simetrik olup olmadığı McNemar testi ile değerlendirildi (Tablo

4.5). Her model çifti için 2×2'lik bir tablo oluşturuldu ve "Model A doğru / Model B yanlış" ile "Model A yanlış / Model B doğru" hücreleri üzerinden kesin binom McNemar testi uygulandı. Elde edilen sonuçlar, SIFT'in Logistic Regression, SVM ve MLP'ye karşı anlamlı düzeyde dezavantajlı olduğunu; AlphaMissense'in ise bazı ML modelleriyle kıyaslandığında, yanlış sınıflandırılan örnekler bakımından simetriden uzak hata örüntüleri sergilediğini gösterdi. q-değerleri, Benjamini–Hochberg düzeltmesi sonrasında da anlamlılığını korudu. Bu tablo, benzer AUROC değerlerine sahip iki modelin klinik düzeyde birbirinin tam ikamesi sayılamayacağını ortaya koymaktadır; çünkü AUROC yalnızca ayrıştırmanın bütünleşik bir özetini verir, hangi hastaların yanlış sınıflandırıldığına ilişkin ayrıntıyı yansıtmaz. McNemar testi ise, aynı bireylerdeki uyumsuz kararları doğrudan test ettiğinden, ayrıştırma metriklerini tamamlayan ayrı bir bilgi katmanı sağlar (Fagerland ve ark., 2013; Grimm ve ark., 2015).

Bu iki çerçeve birlikte yorumlandığında şu resim ortaya çıkmaktadır: Paired-bootstrap ΔAUC sonuçları, AlphaMissense'in PolyPhen-2 ve SIFT araçlarına karşı ayrıştırma üstünlüğünün rastlantısal bir örnekleme dalgalanması olmadığını ve Logistic Regression'ın da bazı ağaç tabanlı modellere karşı anlamlı bir performans farkı üretebildiğini göstermektedir. McNemar testleri ise, aynı 0,5 eşik altında hata deseni bakımından hangi model çiftlerinde klinik açıdan önemli asimetriler bulunduğunu ortaya koymaktadır. Örneğin, iki model benzer AUROC veya AUPRC değerleri sergilese dahi, birinin daha fazla yanlış pozitif, diğerinin ise daha fazla yanlış negatif üretmesi mümkündür; bu durum, klinik kullanımdaki tercihleri doğrudan etkileyebilir. Bizim verimizde, bu tür asimetri desenleri özellikle SIFT ile ML modelleri arasındaki karşılaştırmalarda belirginleşmiştir (Tablo 4.5).

Klinik karar desteği açısından bakıldığında, bu yaklaşım şu sonucu ima etmektedir: Laboratuvar düzeyinde bir kestiriciyi başka bir modelle değiştirme kararı yalnızca AUROC sıralamasına bakılarak verilmemelidir. Aynı panel, aynı örneklem ve aynı sabit eşik altında, hem ayrıştırma farklarının eşik-bağımsız bootstrap testleriyle istatistiksel olarak sınanması hem de McNemar testleriyle hasta bazında hata örüntülerinin karşılaştırılması gerekir. Ancak bu koşullarda, bir modelin gerçekten "daha iyi" olup olmadığı ve hangi tür hataları ne ölçüde azalttığı güvenilir biçimde söylenebilir. Bu tezde kullanılan eşik-bağımsız (paired-bootstrap ΔAUC) ve eşik-

bağımlı (McNemar exact) çerçevelerin birlikte kullanılması, otoinflamatuvar panel bağlamında modeller arası üstünlüğün rastlantısal olmadığını göstermekle kalmamış; aynı zamanda klinik karar eşiğinde hata simetrisini görünür kılarak model ikamesi ve olası ensembling stratejilerine doğrudan kanıt üretmiştir. VUS yorumunun yüksek risk taşıdığı klinik senaryolarda bu tür bir değerlendirme hattı, yalnızca ayrıştırma metriklerine dayalı sade sıralamadan daha güvenilir ve pratiğe dönük bir çerçeve sunmaktadır.

6. SONUÇ VE ÖNERİLER

Bu tezde elde edilen sonuçlar, otoinflamatuvar gen panellerinde hesaplamalı patojenite kestiricilerinin ve bu skor setlerinden eğitilmiş makine öğrenmesi (ML) modellerinin klinik iş akışına iki aşamalı bir şema ile entegre edilmesini önermektedir. İlk aşama, panel düzeyinde tarama ve önceliklendirme basamağıdır. Ham skor evresinde AlphaMissense'in düşük yanlış-pozitif oranlarında belirgin "early-lift" sağlaması ve yüksek ROC/AUPRC değerleri üretmesi, varyantların ilk sıralamasında güçlü bir avantaj sunmaktadır. Bu aşamada amaç, sınırlı doğrulama kaynaklarını (ikinci bir biyoinformatik inceleme, aile segregasyon analizi, hedefli fonksiyonel testler gibi) en fazla bilgi taşıyan adaylara yönlendirerek iş yükünü azaltmak ve tanı sürecini hızlandırmaktır. Triage aşamasında hiçbir model tek başına "nihai otorite" olarak kullanılmamalı; yüksek ayrıştırma gücüne sahip kestiriciler, yalnızca ilk filtre işlevi görmeli ve ardından fenotip uyumu, populasyon frekansı, literatür/ClinVar kanıtları birlikte değerlendirilmelidir

İkinci aşama, klinik rapora girecek olasılıkların yorumlanması ve karar eşiğinin tanımlanmasıdır. Kalibrasyon analizleri, ham olasılıkların doğrudan klinik eşiğe taşınmasının hata desenini ağırlaştırabildiğini; buna karşılık panel-özgü Platt ölçeklemesi sonrası özellikle Logistic Regression, MLP ve SVM modellerinde Brier skoru ve ECE'nin belirgin şekilde iyileştiğini göstermiştir. Bu bulgu, klinik kararın mümkün olduğunca kalibre edilmiş olasılıklar üzerinden verilmesi gerektiğini düşündürmektedir. Uygulamada izlenebilecek yol şudur: önce panel ve laboratuvarın kendi verisi üzerinde Platt temelli bir kalibrasyon kurulmalı, ardından kurumun öncelikleri doğrultusunda hedeflenen PPV/NPV veya Sensitivity/Specificity profiline uygun bir karar eşiği veri-türevi olarak belirlenmeli ve bu eşik zaman içinde periyodik olarak yeniden gözden geçirilmelidir. Böylelikle karar eşiği, bağlamdan

bağımsız sabit bir sayı yerine, panel bileşimi ve hasta popülasyonuna duyarlı, açıkça tanımlanmış bir operasyonel parametre haline gelir.

Eşik bağımsız ve eşik bağımlı testlerin birlikte kullanımı, model seçimi ve model değişimi kararları açısından önemli pratik sonuçlar doğurmaktadır. Bootstrap temelli ΔAUC analizleri, AlphaMissense ve Logistic Regression'ın ayrıştırma gücü bakımından bazı alternatiflere göre istatistiksel olarak anlamlı üstünlük sağladığını göstermiş, böylece gözlenen farkların rastlantısal örnekleme gürültüsüyle açıklanamayacağını ortaya koymuştur. McNemar sınamaları ise, 0,5 ortak eşiğinde bazı model çiftlerinin aynı örneklerde farklı türde hatalar yaptığını, yani benzer AUROC değerlerine rağmen hasta düzeyinde tamamen örtüşen karar desenleri üretmediklerini göstermiştir. Bu tablo, yalnızca AUROC/AUPRC kıyasına bakarak bir modelin diğeriyle "sorunsuz biçimde değiştirilebileceği" varsayımının güvenli olmadığını, model ikamesi kararı alınırken aynı zamanda karar düzeyi hata simetrisinin de test edilmesi gerektiğini göstermektedir. Laboratuvar uygulamasında, yeni bir model devreye alınmadan önce en azından üç boyutun birlikte incelenmesi yerinde olacaktır: ayrıştırma (ΔAUC ve güven aralıkları), kalibrasyon (Brier, ECE, slope/intercept, HL) ve karar düzeyi simetrisi (McNemar ve düzeltilmiş q-değerleri).

Bu bulgular, raporlama pratiğine de yansıtılmalıdır. Klinik raporda kullanılan modelin adı ve sürümü, uygulanan kalibrasyon yöntemi (örneğin "Logistic Regression + Platt kalibrasyonu"), karar eşiği ve bu eşiğin hangi hedefe göre seçildiği açıkça belirtilmelidir. Mümkün olduğunda, bildirilen olasılığa eşlik eden belirsizlik bilgisi (örneğin internal bootstrap'e dayalı dar bir olasılık bandı) verilmesi, farklı merkezler ya da sürümler arasındaki küçük farkların yorumlanmasını kolaylaştıracaktır. ACMG/AMP çerçevesinde PP3/BP4 kodlarının kullanımında da tek bir kestiricinin sonucu yerine, kalibre edilmiş çoklu araçların uzlaşısı veya iyi kalibre edilmiş bir ML modelinin güçlü olasılık çıktıları dikkate alınmalıdır.

Panelin gen bileşimi ve hedeflenen klinik kullanım göz önüne alındığında, eşik seçiminin gen veya kompleks alt gruplarına göre uyarlanması da mantıklı olabilir. Bu tezde bazı modellerde HL testinin anlamlı kalması ve kalibrasyon eğrilerinde belirli olasılık bantlarında sapmanın sürmesi, gen/kompleks alt-kümeleri için mikro-kalibrasyon ve koşullu eşik stratejilerinin yararlı olabileceğini düşündürmektedir.

Örneğin inflammazom çekirdek genlerinde yanlış pozitif maliyetinin yüksek olduğu senaryolarda özgüllük lehine, düzenleyici genlerde ise olası patojen varyantları kaçırmamak amacıyla duyarlılık lehine eşik ayarı tercih edilebilir. Bu tür alt-katmanlandırmaların güvenilir olabilmesi için, her alt grup için yeterli örneklem büyüklüğü ve raporda açıkça gösterilen güven aralıkları gereklidir.

Tüm bu çerçeve, aynı zamanda bir kalite güvencesi iskeleti sunmaktadır. Model ve kalibrasyon sürümlerinin kurum içinde sistematik olarak kaydedilmesi, belirli aralıklarla prospektif performans izlemi (Brier/ECE/HL ve PPV/NPV eğilimleri), veri dağılımındaki olası kaymaları işaret edecek uyarı eşikleri ve gerektiğinde yeniden kalibrasyon planlarının önceden tanımlanması, klinik karar desteğinin sürekliliğini ve şeffaflığını güçlendirecektir. Bu tezde önerilen iki aşamalı iş akışı "önce ayrıştırma gücü yüksek bir modelle triage, ardından kalibre edilmiş ML olasılıkları üzerinde özel gen panelinde bir eşik ile karar" şeklinde, otoinflamatuvar gen panelinde varyant yorumunu daha öngörülebilir, denetlenebilir ve klinik açıdan savunulabilir hale getirmeyi amaçlayan pratik bir yol haritası sunmaktadır.

7. KAYNAKLAR

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics, 76*, 7.20.1–7.20.41. https://doi.org/10.1002/0471142905.hg0720s76

Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J., & Hocking, T. D. (2018). ClinPred: Prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *American Journal of Human Genetics, 103*(4), 474–483. https://doi.org/10.1016/j.ajhg.2018.08.005

An, J., Marwaha, A., & Laxer, R. M. (2024). Autoinflammatory diseases: A review. *The Journal of Rheumatology, 51*(9), 848–861. https://doi.org/10.3899/jrheum.2023-1209

Anaconda, Inc. (n.d.). *Anaconda Distribution*. Retrieved July 16, 2025, from https://www.anaconda.com/

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

Attwood, K., & Tian, L. (2022). Confidence interval estimation of the Youden index and corresponding cut-point for a combination of biomarkers under normality. *Communications in Statistics - Theory and Methods, 51*(2), 501–518. https://doi.org/10.1080/03610926.2020.1751852

Barbosa, P., Ribeiro, M., Carmo-Fonseca, M., & Fonseca, A. (2022). Clinical significance of genetic variation in hypertrophic cardiomyopathy: Comparison of computational tools to prioritize missense variants. *Frontiers in Cardiovascular Medicine, 9*, 975478. https://doi.org/10.3389/fcvm.2022.975478

Benevenuta, S., Capriotti, E., & Fariselli, P. (2021). Calibrating variant-scoring methods for clinical decision making. *Bioinformatics (Oxford, England), 36*(24), 5709–5711. https://doi.org/10.1093/bioinformatics/btaa943

Berrar, D. (2019). Performance measures for binary classification. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (Vol. 1, pp. 546–560). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20351-8

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Boldini, D., Rodríguez-Pérez, R., Olier, I., & Vogt, M. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics, 15*(1), 73. https://doi.org/10.1186/s13321-023-00743-7

Borisov, V., Leemann, T., Sebler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 35*(6), 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors, 23*(2), 634. https://doi.org/10.3390/s23020634

Chen, W., Li, Y., Wang, S., & Pan, Z. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*. Advance online publication. https://doi.org/10.1007/s10462-024-10759-6

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science, 381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining, 14*(1), 13. https://doi.org/10.1186/s13040-021-00244-z

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Dawood, T., Chen, C., Sidhu, B. S., Ruijsink, B., Gould, J., Porter, B., Elliott, M. K., Mehta, V., Rinaldi, C. A., Puyol-Antón, E., Razavi, R., & King, A. P. (2023). Uncertainty aware training to improve deep learning model calibration for classification of cardiac MR images. *Medical Image Analysis, 88*, 102861. https://doi.org/10.1016/j.media.2023.102861

De Hond, A. A. H., Shah, V. B., Kant, I. M. J., Van Calster, B., Steyerberg, E. W., & Hernandez-Boussard, T. (2023). Perspectives on validation of clinical predictive algorithms. *NPJ Digital Medicine, 6*(1), 86. https://doi.org/10.1038/s41746-023-00832-9

Debeer, D., & Strobl, C. (2020). Conditional permutation importance revisited. *BMC Bioinformatics, 21*, 307. https://doi.org/10.1186/s12859-020-03622-2

Dey, D., Haque, M. S., Islam, M. M., Hossain, M. E., Bandyopadhyay, S., Choi, J. S., Huang, T., Yao, X., & Sihn, C. R. (2025). The proper application of logistic regression model in complex survey data: A systematic review. *BMC Medical Research Methodology, 25*, 15. https://doi.org/10.1186/s12874-024-02454-5

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics, 24*(8), 2125–2137. https://doi.org/10.1093/hmg/ddu733

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology, 13*, 91. https://doi.org/10.1186/1471-2288-13-91

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Flevaris, K., Davies, J., Nakai, S., Vučković, F., Lauc, G., Dunlop, M. G., & Kontoravdi, C. (2024). Machine learning framework to extract the biomarker potential of plasma IgG N-glycans towards disease risk stratification. *Computational and Structural Biotechnology Journal, 23*, 1234–1243. https://doi.org/10.1016/j.csbj.2024.03.008

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal, 47*(4), 458–472. https://doi.org/10.1002/bimj.200410135

Fridgeirsson, E. A., Williams, R., Rijnbeek, P., Suchard, M. A., & Reps, J. M. (2024). Comparing penalization methods for linear models on large observational health data. *Journal of the American Medical Informatics Association, 31*(7), 1514–1521. https://doi.org/10.1093/jamia/ocae109

Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., & Friede, T. (2023). Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research, 32*(2), 425–440. https://doi.org/10.1177/09622802221133557

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M.-F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)* (pp. 1050–1059). PMLR. https://proceedings.mlr.press/v48/gal16.html

Ghosh, R., Oak, N., & Plon, S. E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biology, 18*(1), 225. https://doi.org/10.1186/s13059-017-1353-5

Google DeepMind. (2023, September 19). *A catalogue of genetic mutations to help pinpoint the cause of diseases*. Retrieved October 26, 2025, from https://deepmind.google/blog/a-catalogue-of-genetic-mutations-to-help-pinpoint-the-cause-of-diseases/

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology, 23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., & Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation, 36*(5), 513–523. https://doi.org/10.1002/humu.22768

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS 2022 Datasets and Benchmarks Track* (pp. 1–15). https://papers.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)* (Vol. 70, pp.

1321–1330). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v70/guo17a.html

Gutierrez, M. J., & Lapidus, S. K. (2022). Systemic autoinflammatory diseases: A growing family of disorders of overlapping immune dysfunction. *Rheumatic Disease Clinics of North America, 48*(1), 371–395. https://doi.org/10.1016/j.rdc.2021.07.011

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7*, 94. https://doi.org/10.1186/s40537-020-00369-8

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to AUC. *Machine Learning, 77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Harris, J. K. (2021). Primer on binary logistic regression. *Family Medicine and Community Health, 9*(Suppl 1), e001290. https://doi.org/10.1136/fmch-2021-001290

Hassanzad, M., & Hajian-Tilaki, K. (2024). Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: An update review. *BMC Medical Research Methodology, 24*(1), 84. https://doi.org/10.1186/s12874-024-02198-2

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1026–1034). IEEE. https://doi.org/10.1109/ICCV.2015.123

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association, 27*(4), 621–633. https://doi.org/10.1093/jamia/ocz228

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., Powell, I. J., Cussenot, O., Cancel-Tassin, G., Giles, G. G., MacInnis, R. J., Maier, C., Hsieh, C.-L., Wiklund, F., Catalona, W. J., Foulkes, W. D., Mandal, D., Eeles, R. A., Kote-Jarai, Z., Bustamante, C. D., Schaid, D. J., Hastie, T., Ostrander, E. A., Bailey-Wilson, J. E., Radivojac, P., Thibodeau, S. N., Whittemore, A. S., & Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics, 99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.),

Proceedings of the 32nd International Conference on Machine Learning (ICML 2015) (pp. 448–456). PMLR. https://proceedings.mlr.press/v37/ioffe15.html

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics, 48*(12), 1581–1586. https://doi.org/10.1038/ng.3703

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jovel, J., & Greiner, R. (2021). An introduction to machine learning approaches for biomedical research. *Frontiers in Medicine, 8*, 771607. https://doi.org/10.3389/fmed.2021.771607

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kang, M., Kim, S., Lee, D.-B., Hong, C., & Hwang, K.-B. (2023). Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants. *Scientific Reports, 13*, 10478. https://doi.org/10.1038/s41598-023-37698-6

Kastner, D. L., Aksentijevich, I., & Goldbach-Mansky, R. (2010). Autoinflammatory disease reloaded: A clinical perspective. *Cell, 140*(6), 784–790. https://doi.org/10.1016/j.cell.2010.03.002

Khandakji, M. N., & Mifsud, B. (2022). Gene-specific machine learning model to predict the pathogenicity of BRCA2 variants. *Frontiers in Genetics, 13*, 982930. https://doi.org/10.3389/fgene.2022.982930

Kia, A., Timsina, P., Joshi, H. N., Klang, E., Gupta, R. R., Freeman, R. M., Reich, D. L., Tomlinson, M. S., Dudley, J. T., Kohli-Seth, R., Mazumdar, M., & Levin, M. A. (2020). MEWS++: Enhancing the prediction of clinical deterioration in admitted patients through a machine learning model. *Journal of Clinical Medicine, 9*(2), 343. https://doi.org/10.3390/jcm9020343

Kim, Y. D. (2023). Systemic autoinflammatory disorders: Autoinflammatory and autoimmune disorders. *Clinical and Experimental Pediatrics, 66*(10), 439–440. https://doi.org/10.3345/cep.2023.00605

Krainer, J., Siebenhandl, S., & Weinhäusel, A. (2020). Systemic autoinflammatory diseases. *Journal of Autoimmunity, 109*, 102421. https://doi.org/10.1016/j.jaut.2020.102421

Landeros, A., & Lange, K. (2022). Algorithms for sparse support vector machines. *Journal of Computational and Graphical Statistics, 32*(3), 1097–1108. https://doi.org/10.1080/10618600.2022.2146697

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., Maglott, D., Holmes, J. B., & Kattman, B. L. (2020). ClinVar: Improvements to accessing data. *Nucleic Acids Research, 48*(D1), D835–D844. https://doi.org/10.1093/nar/gkz972

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L., & Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research, 46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, C., Zhi, D., Wang, K., & Liu, X. (2022). MetaRNN: Differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Medicine, 14*(1), 115. https://doi.org/10.1186/s13073-022-01120-z

Li, J., Zhao, T., Zhang, Y., Zhang, K., Shi, L., Chen, Y., Wang, X., & Sun, Z. (2018). Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Research, 46*(15), 7793–7804. https://doi.org/10.1093/nar/gky678

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

McDermott, M. F., & Aksentijevich, I. (2002). The autoinflammatory syndromes. *Current Opinion in Allergy and Clinical Immunology, 2*(6), 511–516. https://doi.org/10.1097/00130832-200212000-00006

Meertens, B., Hoste, L., Tavernier, S. J., & Haerynck, F. (2024). The riddle of recurrent fever: A clinical approach to pediatric autoinflammatory diseases.

Frontiers in Pediatrics, 12, 1448176. https://doi.org/10.3389/fped.2024.1448176

Megahed, F. M., Chen, Y. J., Megahed, A., Ong, Y., Altman, N., & Krzywinski, M. (2021). The class imbalance problem. *Nature Methods, 18*(11), 1270–1272. https://doi.org/10.1038/s41592-021-01302-4

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging, 39*(12), 3868–3878. https://doi.org/10.1109/TMI.2020.3006437

Minton, K. (2023). Predicting variant pathogenicity with AlphaMissense. *Nature Reviews Genetics, 24*(12), 804. https://doi.org/10.1038/s41576-023-00668-9

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Christoph Molnar. https://christophm.github.io/interpretable-ml-book/

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine, 162*(1), W1–W73. https://doi.org/10.7326/M14-0698

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32, NeurIPS 2019, pp. 4694–4703). Curran Associates. https://papers.neurips.cc/paper/8717-when-does-label-smoothing-help.pdf

Nag, S., Baidya, A. T. K., Mandal, A., Mathew, A. T., Das, B., Devi, B., & Kumar, R. (2022). Deep learning tools for advancing drug discovery and development. *3 Biotech, 12*(5), 110. https://doi.org/10.1007/s13205-022-03165-8

National Center for Biotechnology Information. (2024, April 18). *Review status in ClinVar*. Retrieved June 26, 2025, from https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/

National Center for Biotechnology Information. (2024, April 4). *Identifiers in ClinVar*. Retrieved June 26, 2025, from https://www.ncbi.nlm.nih.gov/clinvar/docs/identifiers/

National Center for Biotechnology Information. (2024, August 7). *Guidelines for applying for review status in ClinVar*. Retrieved June 26, 2025, from https://www.ncbi.nlm.nih.gov/clinvar/docs/review_guidelines/

National Center for Biotechnology Information. (2024, December 14). *What is ClinVar?* Retrieved June 26, 2025, from https://www.ncbi.nlm.nih.gov/clinvar/intro/

National Center for Biotechnology Information. (2024, January 27). *The ClinVar RCV page*. Retrieved June 26, 2025, from https://www.ncbi.nlm.nih.gov/clinvar/docs/details/

National Center for Biotechnology Information. (2025, August 15). *Representation of classifications in ClinVar*. Retrieved October 22, 2025, from https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics, 34*(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research, 11*(5), 863–874. https://doi.org/10.1101/gr.176601

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research, 31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM. https://doi.org/10.1145/1102351.1102430

Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., & Ziegler, A. (2023). Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine, 42*(29), 5451–5478. https://doi.org/10.1002/sim.9921

OpenAI. (n.d.). *ChatGPT*. Retrieved October 22, 2025, from https://chat.openai.com/

Park, Y. H., Wood, G., Kastner, D. L., & Chae, J. J. (2016). Pyrin inflammasome activation and RhoA signaling in the autoinflammatory diseases FMF and HIDS. *Nature Immunology, 17*(8), 914–921. https://doi.org/10.1038/ni.3457

Pejaver, V., Byrne, A. B., Feng, B. J., Pagel, K. A., Mooney, S. D., Karchin, R., O'Donnell-Luria, A., Harrison, S. M., Tavtigian, S. V., Greenblatt, M. S., Biesecker, L. G., Radivojac, P., Brenner, S. E., & ClinGen Sequence Variant Interpretation Working Group. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *American Journal of Human Genetics, 109*(12), 2163–2177. https://doi.org/10.1016/j.ajhg.2022.10.013

Peng, Y., Zhao, S., Zeng, Z., Hu, X., & Yin, Z. (2023). LGBMDF: A cascade forest framework with LightGBM for predicting drug–target interactions. *Frontiers in Microbiology, 13*, 1092467. https://doi.org/10.3389/fmicb.2022.1092467

Politiek, F. A., & Waterham, H. R. (2021). Compromised protein prenylation as pathogenic mechanism in mevalonate kinase deficiency. *Frontiers in Immunology, 12,* 724991. https://doi.org/10.3389/fimmu.2021.724991

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3), e1301. https://doi.org/10.1002/widm.1301

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics in Medicine, 36*(14), 2302–2317. https://doi.org/10.1002/sim.7273

Python Software Foundation. (n.d.). *Python*. Retrieved July 16, 2025, from https://www.python.org/

Ramalingam, K., Yadalam, P. K., Ramani, P., Krishna, M., Hafedh, S., Badnjević, A., Cervino, G., & Minervini, G. (2024). Light gradient boosting-based prediction of quality of life among oral cancer-treated patients. *BMC Oral Health, 24*(1), 349. https://doi.org/10.1186/s12903-024-04050-x

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

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., & ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine, 17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. https://doi.org/10.1186/1471-2105-12-77

Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schnappauf, O., Chae, J. J., Kastner, D. L., & Aksentijevich, I. (2019). The pyrin inflammasome in health and disease. *Frontiers in Immunology, 10*, 1745. https://doi.org/10.3389/fimmu.2019.01745

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics, 43*(4), 1716–1741. https://doi.org/10.1214/15-AOS1321

Shan, G. (2015). Improved confidence intervals for the Youden index. *PLOS ONE, 10*(7), e0127272. https://doi.org/10.1371/journal.pone.0127272

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Steyerberg, E. W., & Harrell, F. E., Jr. (2016). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology, 69*, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Takeshima, H. (2022). Deep learning and its application to function approximation for MR in medicine: An overview. *Magnetic Resonance in Medical Sciences, 21*(4), 553–568. https://doi.org/10.2463/mrms.rev.2021-0040

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Thiele, C., & Hirschfeld, G. (2021). cutpointr: Improved estimation and validation of optimal cutpoints in R. *Journal of Statistical Software, 98*(11), 1–27. https://doi.org/10.18637/jss.v098.i11

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation, 32*(4), 358–368. https://doi.org/10.1002/humu.21445

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Van Calster, B., & Vickers, A. J. (2015). Calibration of risk prediction models: Impact on decision-analytic performance. *Medical Decision Making, 35*(2), 162–169. https://doi.org/10.1177/0272989X14547233

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association, 29*(9), 1525–1534. https://doi.org/10.1093/jamia/ocac093

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91. https://doi.org/10.1186/1471-2105-7-91

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols, 11*(1), 1–9. https://doi.org/10.1038/nprot.2015.123

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Welzel, T., & Kuemmerle-Deschner, J. B. (2021). Diagnosis and management of the cryopyrin-associated periodic syndromes (CAPS): What do we know today? *Journal of Clinical Medicine, 10*(1), 128. https://doi.org/10.3390/jcm10010128

Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., & Burton, J. (2025). A tutorial and use case example of the eXtreme Gradient Boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clinical and Translational Science, 18*(3), e70172. https://doi.org/10.1111/cts.70172

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks, 94*, 103–114. https://doi.org/10.1016/j.neunet.2017.07.002

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 694–699). ACM. https://doi.org/10.1145/775047.775151

Zhang, T., Hao, Z., Jiang, Q., Zhu, L., & Ye, L. (2025). Machine learning-based prediction model for post-stroke cerebral–cardiac syndrome: A risk stratification study. *Scientific Reports, 15*(1), 30657. https://doi.org/10.1038/s41598-025-10104-z

GEN KÖK
Genom ve Kök Hücre Merkezi
Genome and Stem Cell Center

Sayı: 68742022- 604.01.043
Konu: Değerlendirme, Onay

11/09/2024

Sayın ; Doç.Dr.Özkan Ufuk NALBANTOĞLU

"Otoinflamatuvar Hastalıklarda Klinik Önemi Bilinmeyen Patojenik Tek Nükleotid Varyantlarının Yapay Zeka Temelli Araçlarla Tespit Edilmesi" başlıklı proje çalışmasının Merkezimiz Biyoinformatik Birim'inde yapılması ve bu belgeyi tez önerisinde kullanmanız tarafımızca uygun bulunmuştur.

Gereğini bilgilerinize rica ederim.

Not: Bu projeden çıkan bütün yayınlarınızda projenin Betül Ziya Eren Genom ve Kök Hücre Merkezinde yapıldığına dair ibarenin kullanılması zorunludur.

Kullanım şekli: "Genom ve Kök Hücre Merkezi (GENKÖK)/ Genome and Stem Cell Center (GENKÖK)

Erciyes Üniversitesi Genom ve Kök Hücre Merkezi
Erciyes Üniversitesi Merkez Kampüsü / 38039 Kayseri / TÜRKİYE Telefon: +90 352 207 66 66 / 13 600 Telefon: +90 352 437 93 15
Faks: +90 352 437 93 16 E-mail: info@genkok.com E-mail: genkok@erciyes.edu.tr Web: www.genkok.com

OTOİNFLAMATUVAR HASTALIKLARDA KLİNİK ÖNEMİ BİLİNMEYEN PATOJENİK TEK NÜKLEOTİD VARYANTLARININ YAPAY ZEKA TEMELLİ ARAÇLARLA TESPİT EDİLMESİ

ORİJİNALLİK RAPORU

%1	%1	%0	%0
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	<%1
2	www.mdpi.com İnternet Kaynağı	<%1
3	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Öğrenci Ödevi	<%1
4	breast-cancer-research.biomedcentral.com İnternet Kaynağı	<%1
5	acikbilim.yok.gov.tr İnternet Kaynağı	<%1
6	dergipark.org.tr İnternet Kaynağı	<%1
7	ijfans.org İnternet Kaynağı	<%1
8	export.arxiv.org İnternet Kaynağı	<%1

ÖZGEÇMİŞ

**Sadece sorumlu araştırmacı tarafından doldurulacaktır.*

1. KİŞİSEL BİLGİLER

Adı Soyadı:	Tugay TOPAK
Doğum yeri ve yılı:	
Görev yeri:	
Görev unvanı:	
Yazışma adresi:	
Dahili telefon:	
GSM no:	
E-posta adresi:	

2. EĞİTİM BİLGİLERİ

Mezun olduğu üniversite/fakülte:	Nevşehir Hacı Bektaş Veli Üniversitesi, Fen Edebiyat Fakültesi, Moleküler Biyoloji ve Genetik
Mezuniyet tarihi (yıl):	2022

Varsa, Akademik Unvanlar		
Unvan	Tarih (yıl)	Alındığı Üniversite-Fakülte-Bölüm/Anabilim Dalı
Tıpta uzmanlık:		
Yan dal Uzmanlık:		
Doktora:		
Doçent:		
Profesör:		

3. İŞ TECRÜBESİNE AİT BİLGİLER

Tarih (Yıl):	Bugüne kadar çalıştığı kurum/kuruluşlar

4. GÖREV ALDIĞI ARAŞTIRMA PROJELERİ

Tarih (Yıl)	Görev*	Proje Türü**	Proje Adı

***:** *Yürütücü, araştırmacı/uzman, danışman, diğer*

**: *Kurumsal, ulusal, uluslar arası*

5. ARAŞTIRMA KONUSU İLE İLGİLİ BİLİMSEL YAYIN VE BİLDİRİLERİ

1.	
2.	
3.	
4.	

6. ÖZGEÇMİŞ SAHİBİNİN İMZASI

Özgeçmiş Sahibinin (*El yazısıyla*)	
Adı ve Soyadı:	Tugay TOPAK
Tarih:	
İmza:	