JOURNAL of
MATERIALS and MECHATRONICS:A

e-ISSN 2717-8811
JournalMM, 2026, 7(1), 34-55
https://doi.org/10.55546/jmm.1797083

Araştırma Makalesi / Research Article

Zor Örnek Farkındalıklı Kalp Hastalığı Tahmini: SHAP Tabanlı Etkileşim Mühendisliği ve Hibrit Topluluk Öğrenme Yaklaşımı

Emin DEMİR[1*], Yusuf Ziya AYIK[2]

[1] Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, Yönetim ve Bilişim Sistemleri Anabilim Dalı, Erzurum, Türkiye,
ORCID ID: https://orcid.org/0009-0007-9657-2079, emin.demir22@ogr.atauni.edu.tr
[2*] Atatürk Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Bilgisayar Teknolojileri Programı, Erzurum, Türkiye,
ORCID ID: https://orcid.org/0000-0002-7857-9417, ziyaayik@atauni.edu.tr

Geliş/ Received: 04.10.2025**;** **Revize/Revised:** 05.12.2025 **Kabul / Accepted:** 19.01.2026

ÖZET: Bu çalışma, kalp hastalığı teşhisinde karşılaşılan sınıflandırma problemlerine yönelik, güvenilirliği ve doğruluğu artırmayı hedefleyen aşamalı bir model geliştirme süreci (T0–T3) sunmaktadır. Bu süreç, varsayılan modellerden (T0) başlayarak, SHapley Additive exPlanations (SHAP) tabanlı özellik mühendisliği ve hiperparametre optimizasyonu (T1), olasılık kalibrasyonu (T2) ve zor örnekler için özel bir hibrit yaklaşımı (T3) içerir. Veriler %70 eğitim/%30 test şeklinde ayrılmış ve en iyi model çok ölçütlü seçim yöntemleri (tekil metrikler, bileşik puanlama) kullanılarak belirlenmiştir. Bulgular, kalibrasyonun olasılık güvenilirliğini artırarak triyaj eşiği belirlemede kolaylık sağladığını; hibrit mimarinin ise zor örneklerde duyarlılığı artırırken genel performansta rekabetçi kaldığını göstermektedir. Bu çalışma, sadece doğruluk metriklerine değil, aynı zamanda güvenilirlik ölçütlerine de odaklanarak klinik karar desteği için şeffaf ve hesap verebilir bir çerçeve sunmaktadır.

Anahtar Kelimeler: Makine öğrenmesi, Aşamalı model geliştirme, SHAP, Klinik karar desteği

*Sorumlu yazar / Corresponding author: emin.demir22@ogr.atauni.edu.tr
Bu makaleye atıf yapmak için /To cite this article

Demir, E., & Ayık, Y. Z. (2026). Zor örnek farkındalıklı kalp hastalığı tahmini: shap tabanlı etkileşim mühendisliği ve hibrit topluluk öğrenme yaklaşımı. *Journal of Materials and Mechatronics: A, 7*(1), 34-55.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Hard-Aware Heart Disease Prediction: SHAP-Based Interaction Engineering and Hybrid Ensemble Approach

ABSTRACT: This study presents a stepwise model development process (T0–T3) aimed at improving reliability and accuracy in addressing classification problems encountered in the diagnosis of heart disease. The process begins with baseline models (T0) and includes SHapley Additive exPlanations (SHAP) based feature engineering and hyperparameter optimization (T1), probability calibration (T2), and a specialized hybrid approach for hard examples (T3). Data was split into a 70% training and 30% test set, and the best model was identified using a multi-criteria selection approach (single metrics, composite scoring). Our findings show that calibration improves the reliability of probability scores, which in turn facilitates the determination of triage thresholds. The hybrid architecture, while boosting sensitivity for challenging cases, remains competitive with the overall performance of a single best model. This work offers a transparent and accountable framework for clinical decision support by focusing not only on accuracy metrics but also on reliability measures.

Keywords: Machine learning, Tiered model development, SHAP, Clinical decision support

1. GİRİŞ

Kalp hastalıklarının erken teşhisi, mortalite ve morbiditenin azaltılmasında kritik bir rol oynamaktadır. Bu amaçla makine öğrenmesi tabanlı sınıflandırıcılar giderek daha fazla kullanılmakta; ancak mevcut çalışmaların çoğu, performansı yalnızca genel doğruluk, F1 veya AUC gibi klasik metrikler üzerinden değerlendirmektedir. Bu yaklaşım, modellerin çoğunluk örneklerde başarılı olurken nadir ve zor vakalarda sistematik hatalar yapabileceğini göz ardı etmektedir. Oysa klinik açıdan en riskli grubu bu zor vakalar oluşturmaktadır; yanlış sınıflandırıldıklarında en yüksek hata maliyetiyle sonuçlanabilmektedir. Dolayısıyla yalnızca yüksek genel doğruluk yeterli değildir; modellerin hem zor örneklerde duyarlılığı artırması (Hard-Recall, Kept-F1) hem de ürettikleri olasılık çıktılarının güvenilirliği (kalibrasyon; Brier, ECE, Güvenilirlik Diyagramları) kritik önem taşımaktadır. Klinik bağlamda bir modelin "%80 risk" ifadesi, gerçek dünyada da yaklaşık %80 isabetle karşılık gelmeli; böylece hem karar desteği güvenilir hem de hasta güvenliği açısından anlamlı hale gelmelidir.

Mevcut çalışmaların önemli bir kısmı ya yalnızca hiperparametre optimizasyonu (Demir vd., 2023; Küçükmanisa & Kilimci, 2024; Sungur & Bakır, 2024; Takcı, 2022) ve klasik topluluk yöntemlerine odaklanmakta (Erdem ve ark., 2024; Yapıcı ve ark., 2024) ya da açıklanabilirlik (ör. SHAP) (Kaya & Görmez, 2025; Öznacar & Sertkaya, 2024) üretse de bunu doğrudan çıkarım (inference) stratejisine entegre etmemektedir (Lundberg & Lee, 2017 & Lee, 2017). Ayrıca, kalibrasyon çoğu çalışmada ihmal edilmekte (Niculescu-Mizil & Caruana, 2005) veya sadece değerlendirme adımı olarak görülmektedir (Platt, 1999). Zor örnek farkındalıklı bir iş akışının, triyaj ve hibrit tahmin mekanizmasıyla birleştirilmesi ise görece sınırlı incelenmiştir (Ganie ve ark., 2025; P. Shah ve ark., 2025).

Bu çalışmanın amacı, kalp hastalığı tahmininde zor örnek farkındalığını merkezine alan yeni bir yaklaşım önermektir. Çalışmanın özgün katkıları şu şekilde özetlenebilir:

- Çok Aşamalı Makine Öğrenmesi İş Akışı
- SHAP-Güdümlü Özellik Mühendisliği
- Kalibrasyonun Operasyonelleştirilmesi

- Gelişmiş Zor Örnek Tespiti
- Zor Örnek Farkındalıklı Hibrit Çıkarım
- Çok Ölçütlü Model Seçimi
- Çift Boyutlu Değerlendirme

Böylece hem genel doğruluk hem de zor vakalarda hata payının azaltılması hedeflenmektedir. Literatürde çoğunlukla tek değişkenli önem analizleri öne çıkarken, bu çalışma SHAP etkileşimlerinden yararlanarak özellik kombinasyonlarının rolünü de ortaya koymaktadır (Lundberg & Lee, 2017). Ayrıca, olasılık kalibrasyonu sayesinde modellerin risk tahminleri güvenilir hale getirilmiş; zor örneklerin belirlenerek ayrı bir modda değerlendirilmesiyle klinik karar destek sistemleri için pratik, şeffaf ve güvenli bir yol haritası geliştirilmiştir.

2. MATERYAL VE YÖNTEM

2.1 Veri Seti

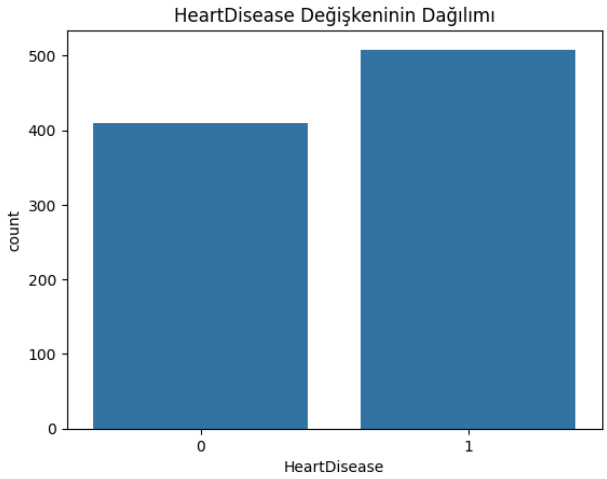
Bu çalışmada kullanılan veri seti, literatürdeki beş farklı kalp hastalığı veri tabanının birleştirilmesiyle oluşturulmuştur (Fedesoriano, 2021). Bu derleme; Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel ve V.A. Medical Center Long Beach tarafından sağlanan veriler (Detrano ve ark., 1989) ile Statlog projesine ait verileri (Feng ve ark., 1992) kapsamaktadır. Farklı kaynaklardan gelen bu verilerin birleştirilmesi, modelin genelleme yeteneğini artırarak tek merkezli çalışmalardaki yanlılık riskini azaltmaktadır. Veri seti toplamda 918 bireye ait klinik ve demografik değişkenleri içermektedir. Katılımcılar; yaş, cinsiyet, göğüs ağrısı tipi, kan basıncı, kolesterol düzeyi, maksimum kalp atım hızı, egzersize bağlı anjina, ST segmenti eğimi, eskiye dönük iskemi ölçümü (Oldpeak) gibi çok sayıda kardiyovasküler risk faktörü açısından değerlendirilmiştir.

Hedef değişken, bireylerde kalp hastalığı bulunup bulunmadığını gösteren HeartDisease değişkenidir (ikili sınıf: 0 = sağlıklı, 1 = kalp hastalığı). Bu değişken, klinik tanı kriterlerine dayalı olarak oluşturulmuş olup çalışmada bağımlı değişken olarak kullanılmıştır.

Veri seti, model geliştirme ve değerlendirme süreçlerinde eğitim ve test olarak ayrılmıştır. Toplam örneklemin yaklaşık %70'i (n=642) eğitim, %30'si (n=276) test amacıyla kullanılmıştır. Veri seti %70 eğitim ve %30 test olarak ayrılmıştır. Tek seferlik ayrımın yaratabileceği varyans riskini minimize etmek amacıyla, eğitim seti üzerindeki model seçimi ve hiperparametre optimizasyonu süreçleri 5-Katlı Tekrarlı Tabakalı Çapraz Doğrulama (Repeated Stratified 5-Fold CV) yöntemiyle gerçekleştirilmiştir. %30'luk test seti ise model geliştirme sürecine dahil edilmemiş, yalnızca nihai modelin genelleme yeteneğini ölçmek için kullanılmıştır.

2.2 Veri Ön İşleme

Ön işleme süreci, modelleme adımlarının güvenilirliği açısından kritik bir aşama olmuştur. Veri setinde eksik gözlem bulunmadığından, doğrudan temizleme işlemi yapılmamıştır. Şekil 1'de Çalışmada hedef değişken olan HeartDisease sınıfı dengeli dağılmış olup; toplam örneklemin yaklaşık %55'i kalp hastalığı tanısı almış, %45'i ise sağlıklı bireylerden oluşmaktadır.



Şekil 1. Veri setinde bulunan hasta sayısı (HeartDisease) dağılımı

Çizelge 1. Veri setinde bulunan değişkenler ve açıklamaları

Değişken	Açıklama
Age	Hastanın yaşı (yıl)
Sex	Cinsiyet (0=Kadın, 1=Erkek)
ChestPainType	Göğüs ağrısı tipi (ATA, NAP, ASY, TA)
RestingBP	İstirahat kan basıncı (mmHg)
Cholesterol	Serum kolesterol (mg/dl)
FastingBS	Açlık kan şekeri (0=<=120 mg/dl, 1=>120 mg/dl)
RestingECG	İstirahat EKG sonuçları (Normal, LVH, ST)
MaxHR	Maksimum kalp atım hızı
ExerciseAngina	Egzersiz kaynaklı anjina (0=Hayır, 1=Evet)
Oldpeak	Egzersizle indüklenen ST depresyonu
ST_Slope	Egzersiz sonrası ST segment eğimi (Up, Flat, Down)
HeartDisease	Hedef değişken (0=Sağlıklı, 1=Kalp hastalığı)

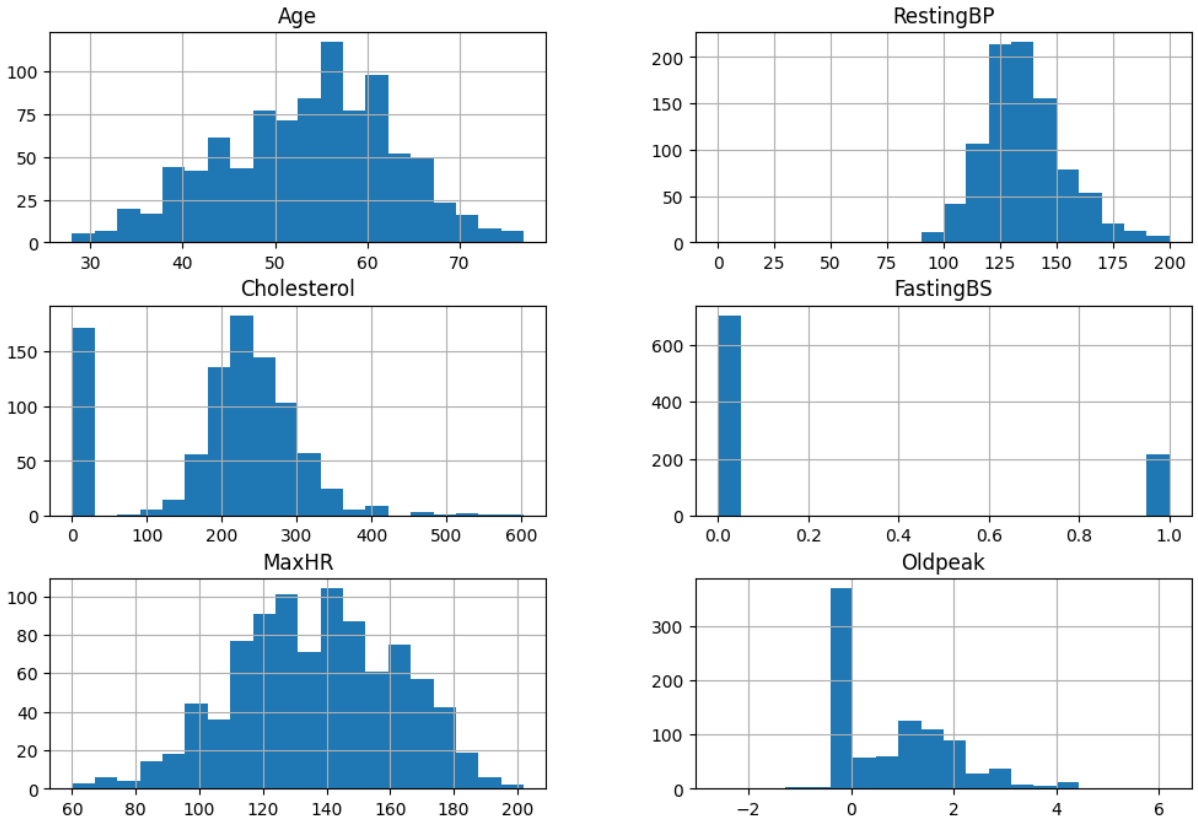
Değişkenlerin tanımları Çizelge 1'de sunulmuştur. Bu değişkenler arasında yaş, kolesterol, istirahat kan basıncı, maksimum kalp atım hızı ve Oldpeak gibi sürekli değişkenler yer almaktadır.

Söz konusu değişkenlerin temel istatistiksel değerleri (ortalama, standart sapma, minimum, maksimum ve çeyrekler) Çizelge 2'de verilmiştir.

Çizelge 2. Veri setinde bulunan sayısal değişkenlerin temel istatistik bilgileri

Count	mean	std	min	25%	50%	75%	max
918	53,51089	9,432617	28	47	54	60	77
918	132,3965	18,51415	0	120	130	140	200
918	198,7996	109,3841	0	173,25	223	267	603
918	0,233115	0,423046	0	0	0	0	1
918	136,8094	25,46033	60	120	138	156	202
918	0,887364	1,06657	-2,6	0	0,6	1,5	6,2

Histogram grafiklerinde (Şekil 2) kolesterol ve Oldpeak değişkenlerinde sağa çarpıklık, istirahat kan basıncında ise uç değerlerin varlığı dikkat çekmektedir. Ancak bu değerler klinik açıdan olası senaryoları temsil ettiğinden, çalışmada aykırı değer temizleme işlemi uygulanmamıştır. Böylece nadir fakat klinik olarak anlamlı örneklerin korunması amaçlanmıştır.



Şekil 2. Veri setindeki sayısal değişkenlerin histogram dağılım grafikleri

Modelleme sürecinde sayısal değişkenler ortalama=0 ve standart sapma=1 olacak şekilde z-skor standardizasyonuna tabi tutulmuştur. Bu işlem, özellikle mesafe tabanlı algoritmaların (ör. KNN) dengeli performans göstermesi açısından önemlidir. Kategorik değişkenler ise One-Hot Encoding (OHE) yöntemiyle ikili göstergelere dönüştürülmüştür. Ön işleme adımları yalnızca eğitim verisi üzerinde gerçekleştirilmiş, test verisine ise aynı dönüşümler eğitimden öğrenilen parametrelerle yansıtılmıştır. Böylece veri sızıntısının önüne geçilmiştir.

2.3 Sınıflandırma Algoritmaları

Bu çalışmada, literatürde yaygın olarak kullanılan ve farklı öğrenme paradigmalarını temsil eden on farklı sınıflandırma algoritması kullanılmıştır. İstatistiksel tabanlı yöntemlerden Lojistik Regresyon (LR) ve Naive Bayes (NB); mesafe tabanlı yöntemlerden K-En Yakın Komşu (KNN) ve Destek Vektör Makineleri (SVM); ağaç tabanlı yöntemlerden Karar Ağaçları (DT), Rastgele Ormanlar (RF), Ekstra Ağaçlar (ET), Gradyan Artırma (GB) ve XGBoost (XGB); ayrıca sinir ağı tabanlı Çok Katmanlı Algılayıcı (MLP) modelleri değerlendirilmiştir. Modellerin teorik altyapıları standart literatürle uyumlu olup (Chen & Guestrin, 2016; Hastie ve ark., 2009), bu çalışmadaki odak noktası algoritmaların temel matematiksel yapıları değil, önerilen aşamalı geliştirme sürecindeki performanslarıdır.

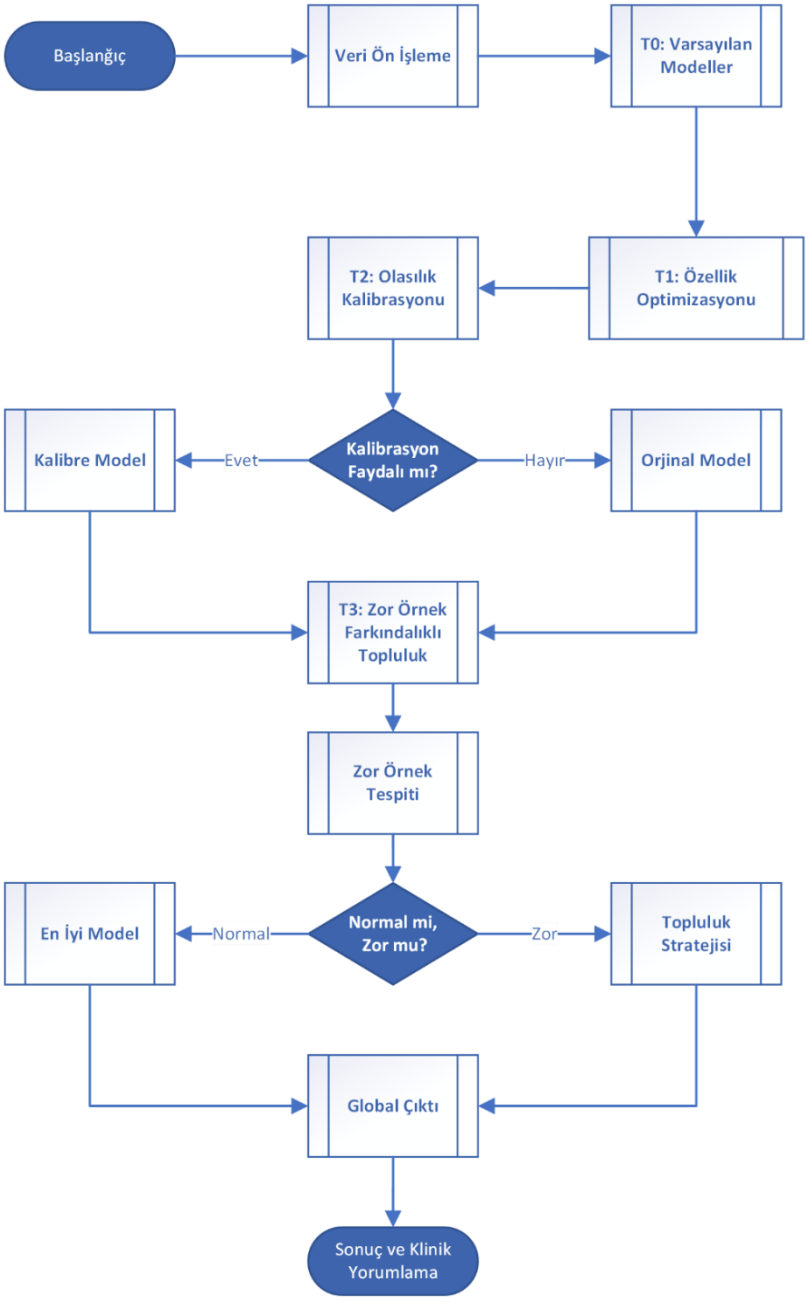
2.4 Açıklanabilir Yapay Zekâ SHAP Yöntemi

Model kararlarının şeffaflığını artırmak ve özellikler arasındaki karmaşık ilişkileri ortaya çıkarmak amacıyla SHAP yöntemi kullanılmıştır (Lundberg & Lee, 2017). Oyun teorisine dayanan bu yaklaşım, her bir özelliğin model tahminine olan marjinal katkısını hesaplayarak hem yerel (tekil örnek bazında) hem de global (tüm veri seti bazında) açıklamalar sunar. Klasik özellik önem analizlerinin aksine SHAP, özelliklerin birbirleriyle olan etkileşimlerini de nicel olarak

ölçebilmektedir. Bu çalışmada SHAP analizi, özellik mühendisliği sürecini yönlendiren bir araç olarak kullanılmıştır. Özellikle yüksek etkileşim skoruna sahip değişken çiftleri (örneğin; Göğüs Ağrısı Tipi ve Maksimum Kalp Atım Hızı) belirlenerek yeni hibrit özellikler türetilmiş ve model performansına katkıları test edilmiştir.

2.5 Aşamalı/Yinelemeli Model Geliştirme Süreci (T0–T3)

Bu çalışma, literatürde "Çok Aşamalı Makine Öğrenmesi İş Akışı" olarak adlandırılan yaklaşımdan esinlenerek dört aşamalı bir yapı üzerinde kurgulanmıştır (Alférez ve ark., 2022; Aragonés Lozano ve ark., 2023; Haihong ve ark., 2019; Filippou ve ark., 2023). Her aşama, bir öncekinin çıktısı üzerine inşa edilmekte ve gerekirse geri besleme mekanizmasıyla yeniden değerlendirmeye olanak tanımaktadır. Böylece modelleme süreci yalnızca tek seferlik bir optimizasyon değil, yinelemeli ve sistematik bir geliştirme hattı olarak tasarlanmıştır. (Şekil 3)



Şekil 3. Varsayılan modellerden (T0) başlayarak; SHAP tabanlı özellik optimizasyonu (T1), olasılık kalibrasyonu (T2) ve zor örnek farkındalıklı hibrit mimariye (T3) uzanan çok aşamalı makine öğrenmesi iş akışı

Şekil 3'te T0–T3 aşamalarını gösteren çok aşamalı makine öğrenmesi iş akışı yapısı çıkarılmıştır. Global tahmin, normal örnekler için en iyi tahmin yapan model ve zor örnekler için en iyi topluluk öğrenme yönteminin birlikte kullanılmasıyla elde edilmiştir

T0 – Varsayılan Modeller: İlk aşamada tüm sınıflandırıcılar ilgili kütüphanelerin varsayılan hiperparametreleri ile eğitilmiş, böylece herhangi bir optimizasyon yapılmadan elde edilen taban performans (baseline) belirlenmiştir. Bu adım, sonraki aşamalarda sağlanan katkıların taban çizgiye kıyasla ne kadar gelişim sağladığını göstermek için referans noktası işlevi görmektedir.

T1 – SHAP Tabanlı Özellik Mühendisliği ve Hiperparametre Optimizasyonu: İkinci aşamada klasik korelasyon veya Variance Inflation Facto (VIF) tabanlı değişken seçiminden farklı olarak, SHAP yöntemi kullanılmıştır. SHAP analizi, yalnızca lineer değil doğrusal olmayan (nonlineer) ilişkileri ve özellik etkileşimlerini de dikkate alarak daha kapsamlı bir değişken önemlendirmesi sağlamaktadır. Bu doğrultuda önce özelliklerin önem sıralaması ve etkileşimleri incelenmiş, ardından yeni değişkenler türetilmiştir (ör. ChestPainType × MaxHR etkileşimi). Böylece tekil değişkenlerin yakalayamadığı kombinasyonel etkiler öğrenme sürecine kazandırılmıştır. Elde edilen genişletilmiş veri seti üzerinde her bir algoritmanın parametre alanı literatür taraması ve önceki çalışmaların deneyimleri ışığında tanımlanmış, RandomizedSearchCV ile optimize edilmiştir. Ayrıca eğitim sürecinde out-of-fold zorluk skorları kullanılarak zor örneklerin ağırlıkları artırılmış, böylece modelin daha dengeli öğrenmesi sağlanmıştır.

Hiperparametre optimizasyonu aşamasında (T1), hesaplama maliyeti ve başarım dengesi gözetilerek RandomizedSearchCV algoritması tercih edilmiştir. Model genelleme yeteneğini artırmak ve varyansı düşürmek amacıyla Tekrarlı Tabakalı 5-Katlı Çapraz Doğrulama (Repeated Stratified 5-Fold Cross-Validation, n_repeats=2) uygulanmıştır. Her bir algoritma için belirlenen geniş parametre uzayında 60 farklı kombinasyon (n_iter=60) denenmiş ve validasyon setleri üzerinde en yüksek F1 skorunu veren parametre seti nihai model için seçilmiştir. İncelenen parametre aralıkları ve optimizasyon sonucunda elde edilen en iyi değerler Çizelge 3'de sunulmuştur.

Çizelge 3. Hiperparametre optimizasyonunda kullanılan arama uzayları ve belirlenen en iyi parametreler

Model	Arama Uzayı	En İyi Parametreler
LR	C: [0.001, 0.0052, 0.0268, 0.1389, 0.5, 0.7197, 1.0, 2.0, 3.728, 19.31, 100] Penalty: ['l2'] Solver: ['lbfgs', 'newton-cg', 'liblinear']	C: 1.0 Solver: 'lbfgs' Penalty: 'l2'
SVM	C: [0.01, 0.0278, 0.0774, 0.2154, 0.5, 0.5995, 1.0, 1.668, 2.0, 4.642, 12.92, 35.94, 100] Gamma: ['scale', 'auto'] Kernel: 'rbf'	C: 1.0 Gamma: 'scale' Kernel: 'rbf'
RF	n_estimators: [100, 300, 500, 800] max_depth: [None, 4, 6, 8, 10] min_samples_leaf: [1, 2, 4] max_features: ['sqrt', 'log2', None]	n_estimators: 800 max_depth: 8 min_samples_leaf: 1 max_features: 'sqrt'
ET	n_estimators: [100, 300, 600, 800]max_depth: [None, 4, 6, 8, 10] min_samples_leaf: [1, 2, 4]max_features: ['sqrt', 'log2', None]	n_estimators: 300 min_samples_leaf: 4max_features: 'log2'
GB	n_estimators: [100, 200, 400, 600] learning_rate: [0.1, 0.05, 0.02] max_depth: [2, 3, 4] subsample: [1.0, 0.8, 0.6]	n_estimators: 400 learning_rate: 0.02 max_depth: 2 subsample: 1.0
XGB	n_estimators: [100, 300, 600, 900] learning_rate: [0.1, 0.05, 0.02] max_depth: [3, 4, 5] reg_lambda: [0.0, 1.0, 2.0] colsample_bytree: [1.0, 0.8, 0.6]	n_estimators: 100 learning_rate: 0.05 max_depth: 3 reg_lambda: 2.0
KNN	n_neighbors: [5, 7, 9, 11, 13, 15] weights: ['uniform', 'distance'] metric: ['minkowski', 'manhattan']	n_neighbors: 15 weights: 'distance' metric: 'manhattan'
DT	max_depth: [None, 3, 5, 7, 9] min_samples_leaf: [1, 2, 4] criterion: ['gini', 'entropy', 'log_loss']	max_depth: 3 min_samples_leaf: 4 criterion: 'gini'
NB	var_smoothing: [1e-9, 1e-8, 1e-10, 1e-11]	var_smoothing: 1e-09
MLP	hidden_layer_sizes: [(100,), (64,32), (80,40), (60,)] activation: ['relu', 'tanh'] alpha: [0.0001, 0.001, 0.0005] learning_rate_init: [0.001, 0.0005, 0.002] max_iter: [300, 400, 600]	hidden_layer_sizes: (60,) activation: 'relu' alpha: 0.0001 learning_rate_init: 0.0005 max_iter: 400

T2 – Olasılık Kalibrasyonu: Üçüncü aşamada, farklı algoritmaların ürettikleri olasılık tahminleri sigmoid (Platt scaling) yöntemiyle kalibre edilmiştir. Varsayılan durumda bazı algoritmalar (örn. SVM, Naive Bayes) aşırı uç olasılıklara eğilimliyken, bazıları (örn. RF, GB) daha temkinli değerler üretmektedir. Bu farklılık topluluk öğrenme stratejileri ve triyaj mekanizması için sorun oluşturabilmektedir. Kalibrasyon ile olasılık çıktıları tutarlı bir ölçeğe çekilmiştir. Önceki çalışmalardan farklı olarak, bu aşamada kalibrasyon öncesi ve sonrası performans metrikleri karşılaştırılmış; yalnızca fayda sağladığı durumlarda kalibre edilmiş model tercih edilmiştir. Böylece kalibrasyon, yalnızca teorik bir adım olmaktan çıkarılmış, operasyonel bir seçim kriteri haline getirilmiştir. Kalibrasyon sürecinde parametrik olmayan Isotonic Regression yerine parametrik Platt Scaling (Sigmoid) yöntemi tercih edilmiştir. Isotonic Regression, büyük ve gürültüsüz veri setlerinde yüksek esneklik sağlasa da, örneklem sayısı sınırlı olduğunda ($N < 1000$) eğitim verisine aşırı uyum (overfitting) riski taşımakta ve çapraz doğrulama sırasında kararsız sonuçlar üretmektedir (Kull & Flach, 2014; Kumar ve ark., 2019; Niculescu-Mizil & Caruana, 2005) Bu çalışmada kullanılan veri setinin toplam 918 örnek içermesi nedeniyle, daha az parametreye sahip ve genelleme performansı daha kararlı olan Platt Scaling yöntemi seçilmiştir. Uygulamada her model için hem Platt Scaling hem de Isotonic Regression denenmiş; Platt Scaling'in Brier Score ve ECE metriklerinde sistematik olarak daha düşük veya eşdeğer değerler verdiği, aynı zamanda daha kısa eğitim süresi sunduğu gözlenmiştir. Bu nedenle tüm modellerde sigmoid tabanlı kalibrasyon kullanılmıştır.

T3 – Zor Örnek Farkındalıklı Topluluk Öğrenmesi: Son aşamada, modellerin zor örneklerdeki performansı ayrı bir stratejiyle ele alınmıştır. Bunun için önce gelişmiş zor örnek tespiti yöntemleri uygulanmıştır: (i) modeller arası varyans, (ii) olasılık dağılımındaki belirsizlik, (iii) her iki ölçütün birleşimi ve (iv) olasılık aralığına dayalı klasik yöntem. Bu yöntemlerden en dengeli dağılım sağlayan yaklaşım seçilmiş ve zor örnekler (kept set) belirlenmiştir. Normal örnekler doğrudan en iyi tek modelle tahmin edilirken, zor örnekler üç farklı topluluk stratejisiyle değerlendirilmiştir:

- Eşit Topluluk Yaklaşımı: Modellerin olasılık çıktıları eşit ağırlıkla ortalanmıştır.
- Ağırlıklı Topluluk Yaklaşımı: Zor örnekler üzerindeki F1 skorlarına göre her modele ağırlık atanmış ve ağırlıklı ortalama alınmıştır.
- Yığınlama Topluluğu Yaklaşımı: Zor örneklerde modellerin çıktıları, meta-seviyede lojistik regresyon kullanılarak birleştirilmiştir.

Bu dört aşamalı yapı sayesinde, Global set tüm örnekleri (normal + zor) kapsamaktadır. Bu aşamada normal örnekler en iyi tek model ile, zor örnekler ise en iyi topluluk öğrenme yöntemi ile tahmin edilmiştir. Böylece hem genel doğruluk korunmuş hem de zor vakalarda duyarlılık artırılmıştır.

2.6 Zor Örnek Tanımı

Sınıflandırma problemlerinde bazı örnekler, modeller tarafından yüksek güvenle sınıflandırılırken, bazıları belirsiz veya çelişkili tahminlerle karşılanmaktadır. Bu tür örnekler literatürde zor örnekler olarak adlandırılmakta ve genellikle modelin hata yapma ihtimalinin en yüksek olduğu, klinik açıdan ise yanlış sınıflandırıldığında en yüksek risk taşıyan vakaları temsil etmektedir. Bu çalışmada zor örneklerin belirlenmesi, klasik tek yöntemli yaklaşımların ötesinde, çoklu ölçütlere dayalı bir stratejiyle gerçekleştirilmiştir.

Klasik Yaklaşım (Olasılık Aralığı); En basit yöntem, modelin tahmin olasılıklarının belirli bir aralıkta kalmasıdır. Örneğin, tahmin olasılığı %45–%55 aralığında olan örnekler "emin değilim bölgesi" olarak kabul edilerek zor örnek kümesine dâhil edilebilir. Bu yöntem basit olmakla birlikte, yalnızca tek bir modelin çıktısına dayandığı için sınırlıdır.

Modeller Arası Varyans (Uyuşmazlık); Bu çalışmada ek olarak farklı algoritmaların aynı örnek üzerindeki tahmin varyansı incelenmiştir. Modeller arasında yüksek varyans bulunan örnekler, yapısal olarak ayrıştırılması güç ve belirsizliği yüksek vakalar olarak değerlendirilmiş ve zor örnek sınıfına dâhil edilmiştir.

Belirsizlik; Bir diğer ölçüt olarak, tahmin edilen olasılık dağılımının entropisi kullanılmıştır. Entropi yüksek olduğunda, modelin 0 veya 1 sınıfı arasında belirgin bir tercih yapamadığı anlaşılmaktadır. Bu tür yüksek entropili örnekler de zor örnekler kategorisine eklenmiştir.

Kombine Yöntem; Daha dengeli bir tanımlama için, model varyans (%60 ağırlık) ve belirsizlik (%40 ağırlık) metrikleri birleştirilmiş ve kombine yöntem skoru hesaplanmıştır (Denklem 1). Bu yöntem, hem modeller arası görüş ayrılığını hem de tek modelin içsel belirsizliğini birlikte dikkate alarak daha güvenilir bir zor örnek tespiti sağlamaktadır.

$$Skor\_e_{birleşik\ yöntem} = 0.6 \times Skor\_e_{varyans} + 0.4 \times Skor\_e_{belirsizlik} \quad (1)$$

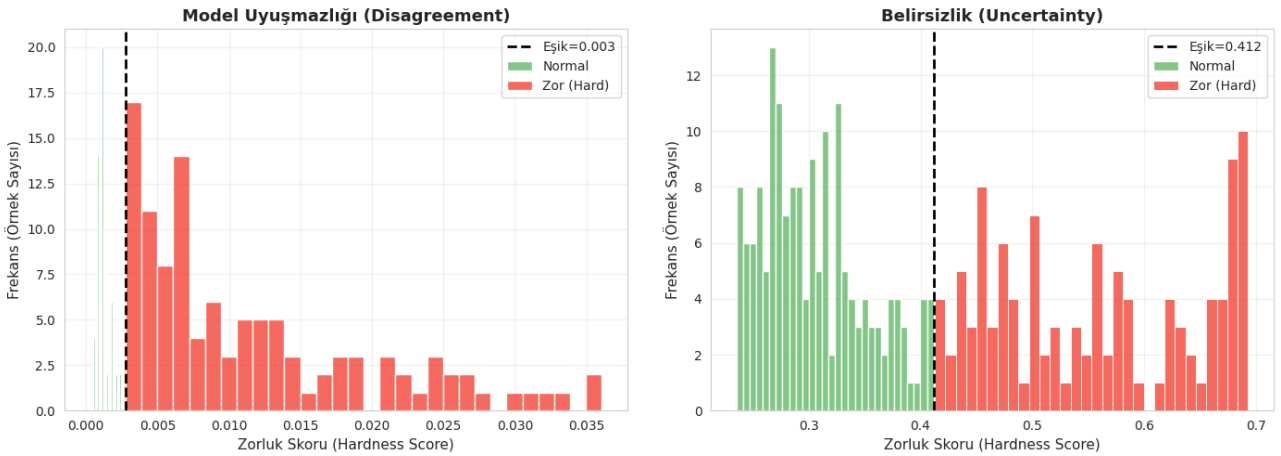
Çalışmada, zor örnek havuzunun sınıf dengesizliğine yol açmayacak ancak istatistiksel olarak anlamlı büyüklüğe sahip bir oranda olması hedeflenmiştir. Uygulanan yöntemler sonucunda bu oran yaklaşık %40 olarak gerçekleşmiş ve analizler bu küme üzerinden yürütülmüştür.

Zor örneklerin ayrılmasında hedeflenen oran, literatürdeki müfredat öğrenme stratejilerine benzer bir denge gözetilerek belirlenmiştir. Bu oran çok düşük olduğunda (%5-10) topluluk modelleri için yeterli veri oluşmamakta; çok yüksek olduğunda (%50+) ise zorluk kavramı ayırt ediciliğini yitirmektedir. Bu nedenle, çalışmada %30-40 bandı, istatistiksel olarak anlamlı bir alt küme oluşturmak için optimum aralık olarak kabul edilmiştir.

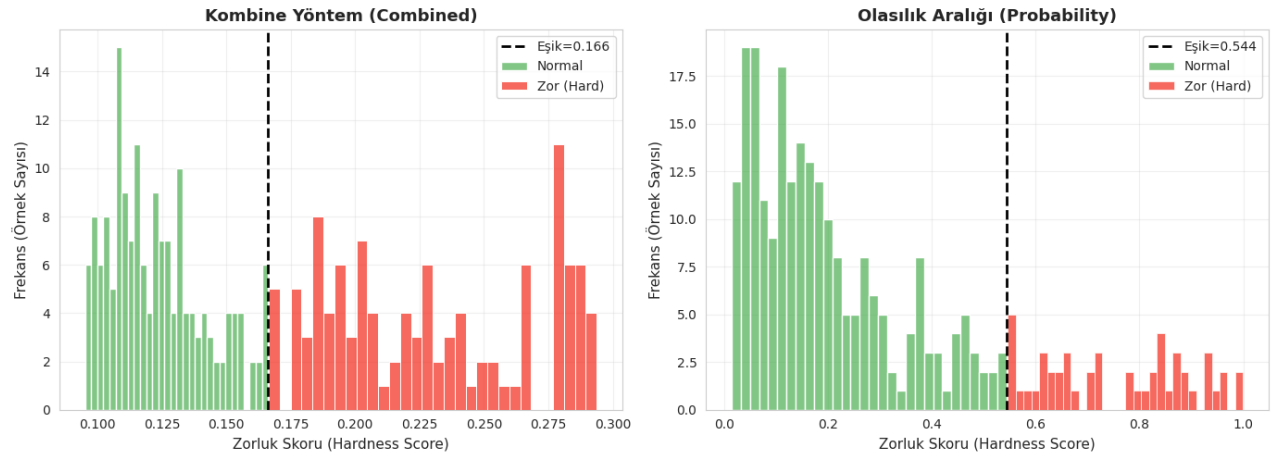
Çizelge 4. Yöntemlere göre zor örnek sayısı ve oranları

Method	Zor Örnek Sayısı	Zor Örnek Oranı (%)
Varyans	110	39.9
Belirsizlik	110	39.9
Kombine	110	39.9
Olasılık Aralığı	49	17.8

Şekil 4 ise zorluk skor dağılımlarını sunarak normal ve zor örneklerin ayrışma biçimini görselleştirmektedir. Söz konusu şekil incelendiğinde adet ve oran aynı olsa bile farklı metodların zorluk dağılımlarının farklı olduğu net bir şekilde anlaşılmaktadır.

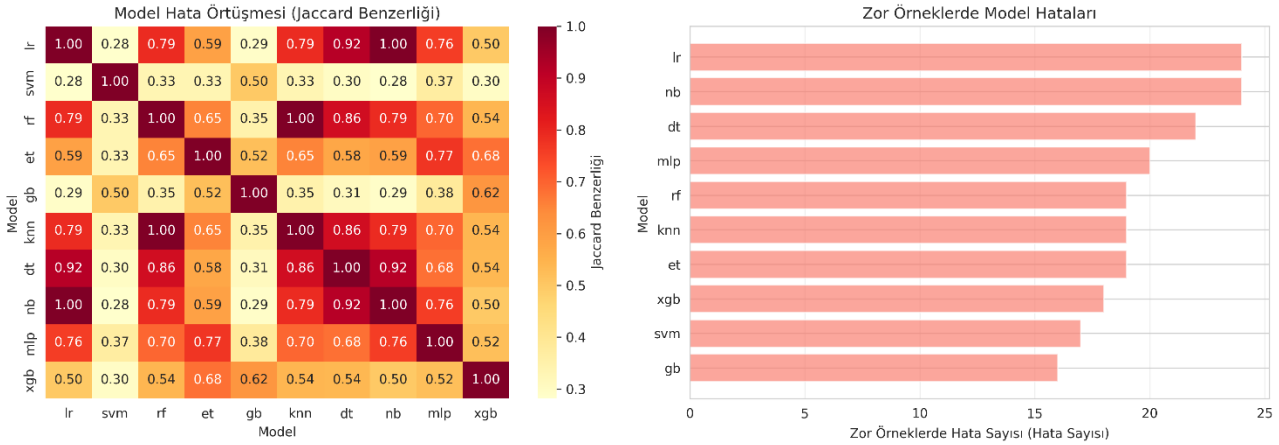


Şekil 4. Farklı yöntemlerin zorluk skor dağılımları



Şekil 4. Farklı yöntemlerin zorluk skor dağılımları (devamı)

Nihai olarak seçilen yöntemle elde edilen zor örnek kümesi (kept set), sonraki aşamada hibrit modellerin özel olarak ele aldığı alt grup olarak kullanılmıştır.



Şekil 5. Zor örnekler üzerinde modellerin hata analizi

Şekil 5 zor örnekler üzerinde modellerin hata analizini göstermektedir. Sol panelde yer alan Jaccard benzerliği matrisi, farklı algoritmaların hata kümeleri arasındaki örtüşme düzeyini ortaya koymaktadır. Örneğin, RF ve XGB benzer hataları paylaşırken, NB ve DT modellerinin hataları diğerlerinden daha farklı bir dağılım göstermektedir. Sağ panel ise her modelin zor örnekler üzerinde yaptığı hata sayısını göstermektedir. Bu sonuçlar, bazı modellerin (örneğin Naive Bayes ve Decision Tree) zor örneklerde sistematik olarak daha fazla hata yaptığını; buna karşın KNN'in nispeten daha az hata ürettiğini göstermektedir. Bu gözlem, zor örnekler için topluluk öğrenme yaklaşımlarının neden gerekli olduğunu görsel olarak desteklemektedir.

2.7 Topluluk ve Hibrid Stratejileri

Makine öğrenmesi literatüründe topluluk öğrenme yöntemleri, farklı modellerin güçlü yönlerini bir araya getirerek tekil modellerin sınırlılıklarını aşmayı amaçlamaktadır (Pearly & Karthik, 2025; M. Shah ve ark., 2023). Bu çalışmada hem global performansı hem de zor örneklerdeki duyarlılığı artırmak amacıyla üç farklı topluluk öğrenme stratejisi uygulanmıştır.

Eşit Topluluk Yaklaşımı (Soft Voting): İlk yaklaşımda, tüm sınıflandırıcıların olasılık tahminleri eşit ağırlıkla ortalanmış ve karar bu ortalama olasılıklar üzerinden verilmiştir. Bu yöntem,

basitliği ve yorumlanabilirliği nedeniyle literatürde sıkça tercih edilmektedir (Behera ve ark., 2021; Dogan & Birant, 2019; Kusuma ve ark., 2023).

Ağırlıklı Topluluk Yaklaşımı (Weighted Ensemble): İkinci yaklaşımda, modellerin özellikle zor örnekler üzerindeki performansları dikkate alınmıştır. Her bir modelin zor örnek kümesindeki F1 skoru, ilgili modele ağırlık olarak atanmış ve tahminler bu ağırlıklar üzerinden birleştirilmiştir. Böylece zor örneklerde daha başarılı olan modellerin katkısı artırılarak, hata payının azaltılması hedeflenmiştir.

Topluluk Yığınlama Yaklaşımı (Stacking Ensemble): Üçüncü stratejide, farklı algoritmaların çıktıları meta-öğrenici olarak tanımlanan Lojistik Regresyon (LR) modeline girdi olarak verilmiştir. Bu yaklaşım, modeller arasındaki doğrusal olmayan ilişkileri de öğrenerek daha esnek ve güçlü bir tahminleme yapabilmektedir (Saihood & Sonuç, 2023).

Hibrit Global Model (zor örnek farkındalıklı): Bu üç stratejinin ötesinde, çalışmada ek olarak hibrit bir yapı da uygulanmıştır. Bu yaklaşımda normal örnekler en iyi model ile, zor örnekler ise en iyi performans gösteren topluluk öğrenme stratejisi ile tahmin edilmiştir. Böylece hem genel doğruluk korunmuş hem de zor vakalarda duyarlılık artırılmıştır. Bu hibrit yapı, global tahminleme sürecini Denklem 2. ile ifade edilebilir:

$$\hat{y}(x) = \begin{cases} f_{best}(x), & x \in Normal \\ f_{ensemble}(x), & x \in Hard \end{cases} \quad (2)$$

Burada f_{best} en iyi tek modeli, $f_{ensemble}$ ise Equal, Weighted ve Yığınlama yöntemleri arasından zor örnekler için en başarılı olanını temsil etmektedir. Uygulanan bu stratejilerin performans karşılaştırmaları bulgular bölümünde Çizelge 10'da ayrıntılı olarak sunulmuştur.

2.8 Triyaj Stratejisi

Çalışmanın 2.5 Zor Örnek Tanımı bölümde ayrıntılı biçimde açıklanan zor örnek tespiti, bu çalışmada triyaj stratejisinin temelini oluşturmaktadır. Triyaj yaklaşımı sayesinde, modelin yüksek güvenle sınıflandırdığı normal örnekler doğrudan en iyi tek modelle tahmin edilirken; belirsizlik içeren zor örnekler farklı topluluk öğrenme yöntemleriyle yeniden değerlendirilmiştir (Şekil 6). Klinik bağlamda bu yaklaşım, modelin "emin değilim" dediği vakaların güvenli bir şekilde alternatif yöntemlere yönlendirilmesini sağlamaktadır. Böylece hem genel doğruluk korunmuş hem de riskli vakalarda hata payı azaltılmıştır.

Şekil 6. Normal örneklerin en iyi tekil modelle, zor örneklerin ise topluluk stratejileriyle değerlendirildiği Zor Örnek Farkındalıklı hibrit triyaj akış şeması

2.9 Değerlendirme Metrikleri

Bu çalışmada, geliştirilen modellerin performansı yalnızca genel doğruluk üzerinden değil, farklı boyutları dikkate alan çoklu metriklerle değerlendirilmiştir. Böylece hem ayırma gücü hem olasılık güvenilirliği hem de zor örneklerdeki duyarlılık birlikte incelenmiştir.

Klasik sınıflandırma metrikleri arasında F1 skoru, AUC (Area Under the ROC Curve), Precision (Pozitif Öngörü Değeri) ve Recall (Duyarlılık) öne çıkmaktadır. F1 skoru, precision ve recall arasındaki dengeyi yansıtarak özellikle dengesiz sınıf dağılımlarında daha anlamlı bir gösterge sunmaktadır (Aydemir, 2022). AUC, modelin pozitif ve negatif sınıfları ayırt etme kapasitesini ölçmekte (Tomak & Bek, 2011); precision yanlış pozitifleri, recall ise yanlış negatifleri doğrudan yakalayarak klinik risklerin değerlendirilmesine katkı sağlamaktadır (Alvarez, 2002; Fränti & Mariescu-Istodor, 2023; Solanki ve ark., 2020). Ayrıca sınıf dengesizliği durumunda klasik accuracy yanıltıcı olabileceği için, her iki sınıfın doğruluğunu eşit önemle alan Denklem 3. Dengeli Doğruluk metriği de hesaplanmıştır:

$$Dengeli\ Doğruluk = \frac{\mathrm{TPR} + \mathrm{TNR}}{2} \tag{3}$$

Burada TPR (Doğru Pozitif Oranı) ve TNR (Gerçek Negatif Oran) değerlerinin ortalaması alınarak, dengesiz veri setlerinde daha adil bir ölçüm sağlanmıştır (García ve ark., 2009).

Modelin yalnızca sınıfları ayırt etmesi yeterli değildir; üretilen olasılıkların da güvenilir olması gerekir. Klinik karar destek sistemlerinde, örneğin bir modelin "%80 risk" dediği durumda, gerçek vakaların da yaklaşık %80 oranında pozitif çıkması beklenir. Bu güvenilirliği ölçmek amacıyla Brier Skoru ve Güvenilirlik Diyagramı kullanılmıştır. Denklem 4. Brier Skoru, tahmin edilen olasılıklarla gerçek sonuçlar arasındaki farkın karesini alarak genel hata düzeyini ölçer:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_{i-}y_i)^2 \tag{4}$$

Burada p_i modelin pozitif sınıfa verdiği olasılık, $y_{i\,\in\{0,1\}}$ gerçek etikettir. Düşük Brier Skoru, modelin olasılıklarının gerçeğe daha yakın olduğunu gösterir. Güvenilirlik Diyagramı ise modelin tahmin olasılıklarını gerçek pozitiflik oranlarıyla karşılaştırarak kalibrasyon düzeyini görsel olarak ortaya koymaktadır. Kalibrasyona özgü bir diğer ölçüt olan Beklenen Kalibrasyon Hatası - Expected Calibration Error (ECE), tahmin edilen olasılıkların segmentler halinde gruplanmasıyla beklenen sapmayı ölçer; daha düşük ECE, daha güvenilir olasılık çıktıları anlamına gelir (Sanders, 1967; Stehouwer ve ark., 2025).

Son olarak, bu çalışmanın özgün yönlerinden biri belirsiz veya sınırda vakaların (zor örnekler) ayrıca değerlendirilmesidir. Bu kapsamda Hard-Recall ve kept-F1 metrikleri tanımlanmıştır. Hard-Recall yalnızca zor örnekler üzerinde modelin duyarlılığını ölçmekte; kept-F1 ise zor örnek seti (kept set) üzerinde precision ve recall arasındaki dengeyi göstermektedir Denklem 5.

$$Hard - Recall = \frac{TP_{hard}}{TP_{hard} + FN_{hard}} \tag{5}$$

Bu metrikler, global F1 veya AUC'nin gölgelediği zor örnek performansını görünür kılmakta ve klinik riskli vakalar açısından daha hassas bir değerlendirme sunmaktadır. Bu çok boyutlu değerlendirme çerçevesi sayesinde, modellerin yalnızca yüksek doğruluk üretip üretmediği değil, aynı zamanda klinik açıdan kritik olan güvenilirlik ve zor örnek duyarlılığı da analiz edilmiştir.

3. BULGULAR VE TARTIŞMA

Bu çalışmada elde edilen bulgular bu bölümde mikroyapı oluşumu ve mekanik özellikler olmak üzere iki alt başlık altında tartışılacaktır.

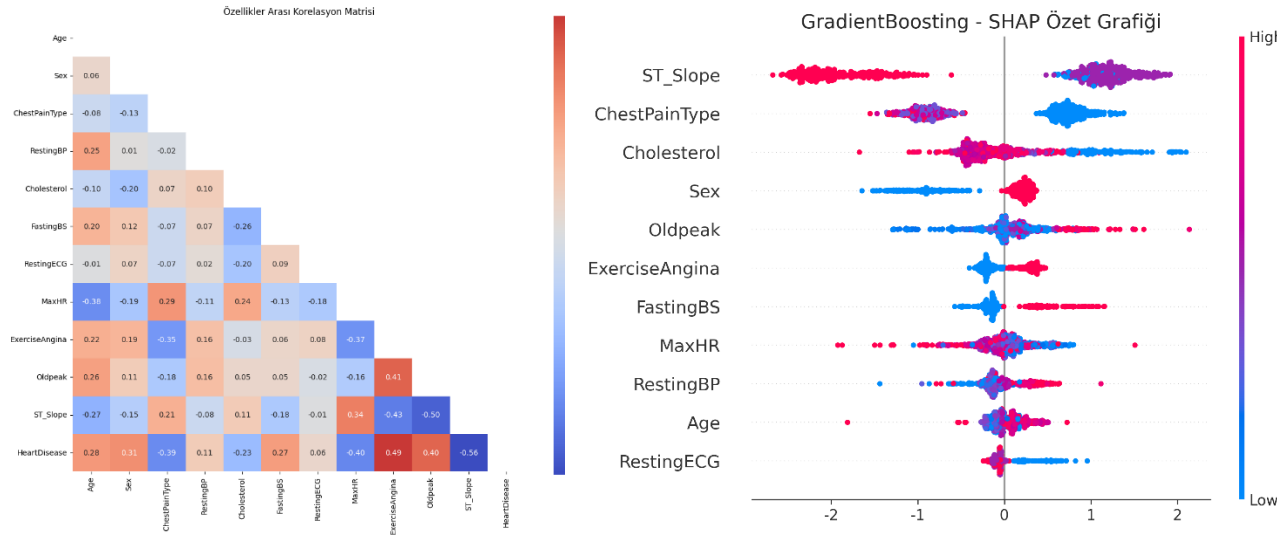
3.1 T0 ve T1 Karşılaştırması

Çizelge 5'de T0 (varsayılan parametreler) ve T1 (SHAP + tuning) modellerinin performans karşılaştırmasını sunmaktadır. Görüldüğü üzere tüm metriklerde (özellikle F1 ve AUC) T1 aşamasında anlamlı bir iyileşme sağlanmıştır. En yüksek başarı KNN ile elde edilmiştir.

Çizelge 5. T0 ve T1 aşamaları performans karşılaştırması

Model	AUC	F1	ACC	Precision	Recall	Brier	Bal_Acc	Thr	Aşama
DT	0,824	0,835	0,8224	0,8611	0,81	0,178	0,8239	0,05	T0_default
DT	0,864	0,858	0,837	0,8292	0,889	0,145	0,8306	0,09	T1_FE_tuned
ET	0,942	0,909	0,8949	0,878	0,941	0,087	0,8892	0,4	T0_default
ET	0,943	0,914	0,9021	0,8937	0,935	0,091	0,8982	0,5	T1_FE_tuned
GB	0,934	0,926	0,9166	0,9166	0,935	0,09	0,9144	0,37	T0_default
GB	0,934	0,926	0,9166	0,9166	0,935	0,09	0,9144	0,37	T1_FE_tuned
KNN	0,934	0,895	0,8731	0,8277	0,974	0,092	0,8609	0,21	T0_default
KNN	**0,951**	**0,926**	**0,9166**	**0,9113**	**0,941**	**0,08**	**0,9136**	**0,46**	**T1_FE_tuned**
LR	0,932	0,901	0,8876	0,8812	0,922	0,097	0,8835	0,46	T0_default
LR	0,932	0,901	0,8876	0,8812	0,922	0,097	0,8835	0,46	T1_FE_tuned
MLP	0,931	0,889	0,8804	0,9166	0,863	0,1	0,8825	0,68	T0_default
MLP	0,938	0,898	0,8876	0,9066	0,889	0,093	0,8875	0,57	T1_FE_tuned
NB	0,918	0,894	0,884	0,906	0,882	0,114	0,8842	0,57	T0_default
NB	0,918	0,894	0,8841	0,906	0,882	0,114	0,8842	0,57	T1_FE_tuned
RF	0,949	0,916	0,9057	0,9096	0,922	0,086	0,9038	0,54	T0_default
RF	0,941	0,916	0,9057	0,9044	0,928	0,097	0,903	0,54	T1_FE_tuned
SVM	0,951	0,921	0,9094	0,8902	0,954	0,079	0,9039	0,44	T0_default
SVM	0,951	0,919	0,9094	0,9102	0,928	0,081	0,9071	0,49	T1_FE_tuned
XGB	0,932	0,919	0,9094	0,9102	0,928	0,091	0,9071	0,38	T0_default
XGB	0,942	0,922	0,913	0,9215	0,922	0,085	0,912	0,52	T1_FE_tuned

Bu iyileşmenin iki temel nedeni bulunmaktadır. İlk olarak, SHAP tabanlı özellik mühendisliği ile klasik korelasyon analizinde göz ardı edilen değişkenler ve etkileşimler modele kazandırılmıştır. Şekil 7'de bağımsız değişkenlerin bağımlı (hedef) değişken ile korelasyonunu ve SHAP değişken önem sıralamasını göstermektedir. İkinci olarak, hiperparametre optimizasyonu sayesinde modeller, varsayılan ayarlara kıyasla daha uygun parametrelerle çalıştırılmış ve kapasitesini en üst düzeye çıkarmıştır.



a) b)

Şekil 7. Değişkenler arası doğrusal ilişkileri gösteren korelasyon matrisi (a) ve özelliklerin model tahminlerine olan katkılarını ve önem düzeylerini küresel ölçekte sunan SHAP özet grafiği (b)

Çizelge 6'da ise bağımsız değişkenlerin Korelasyon ve SHAP değişken önem sıralamasındaki değişkenlik tablo halinde sunulmuştur.

Çizelge 6. Korelasyon ve SHAP değişken önem sıralaması

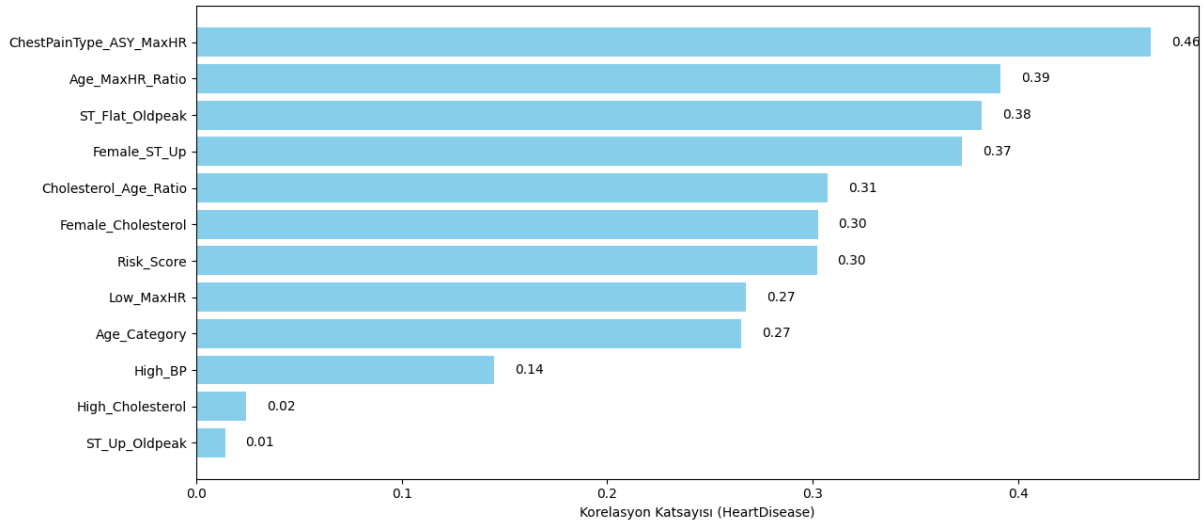
Özellik	Korelasyon Sıralaması	SHAP Sıralaması
ST_Slope	1	1
ExerciseAngina	2	6
Oldpeak	3	5
MaxHR	4	8
ChestPainType	5	2
Sex	6	4
Age	7	10
FastingBS	8	7
Cholesterol	10	3
RestingBP	11	9
RestingECG	12	11

SHAP analizi kullanılarak özelliklerin model tahminlerine katkıları belirlendi. En yüksek SHAP önem değerine sahip ilk 5 özellik (ST_Slope, Oldpeak, ChestPainType vb.) temel alınarak literatür bilgisi ışığında yeni türetilmiş özellikler oluşturuldu. Bu özellikler risk skorları, oran tabanlı özellikler ve kategorik etkileşimleri içermektedir.

Buna göre ST_Slope ve ChestPainType en yüksek katkıyı sağlamış; ayrıca Cholesterol değişkeni düşük korelasyon değerine rağmen yüksek SHAP skoru ile öne çıkmıştır. Böylece SHAP yalnızca açıklayıcı bir araç olarak değil, aynı zamanda özellik mühendisliğine yön veren bir araç olarak kullanılmıştır. Bu durum, SHAP'ın klasik yöntemlere kıyasla daha kapsamlı bir değişken değerlendirmesi sunduğunu ortaya koymaktadır.

Şekil 8'de ise SHAP etkileşim analizlerinden elde edilen dikkat çekici bir sonucu sunmaktadır. ChestPainType_ASY × MaxHR kombinasyonu, asemptomatik göğüs ağrısı ile egzersiz sırasındaki kalp atım hızının klinik açıdan önemli bir ilişkiyi yansıttığını göstermiştir. Benzer olarak Yaş ile MaxHR oranı, ST_Flat ile Oldpeak gibi değişkenler arasında tespit edilen korelasyon katsayıları Şekil

8'de sunulmuş olup bu etkileşim özelliği modele eklenmiş ve T1 performansının yükselmesine katkı sağlamıştır.



Şekil 8. SHAP ile yeni eklenen değişkenlerin hedef değişken ile korelasyonu

3.2 T2 – Kalibrasyon Sonuçları

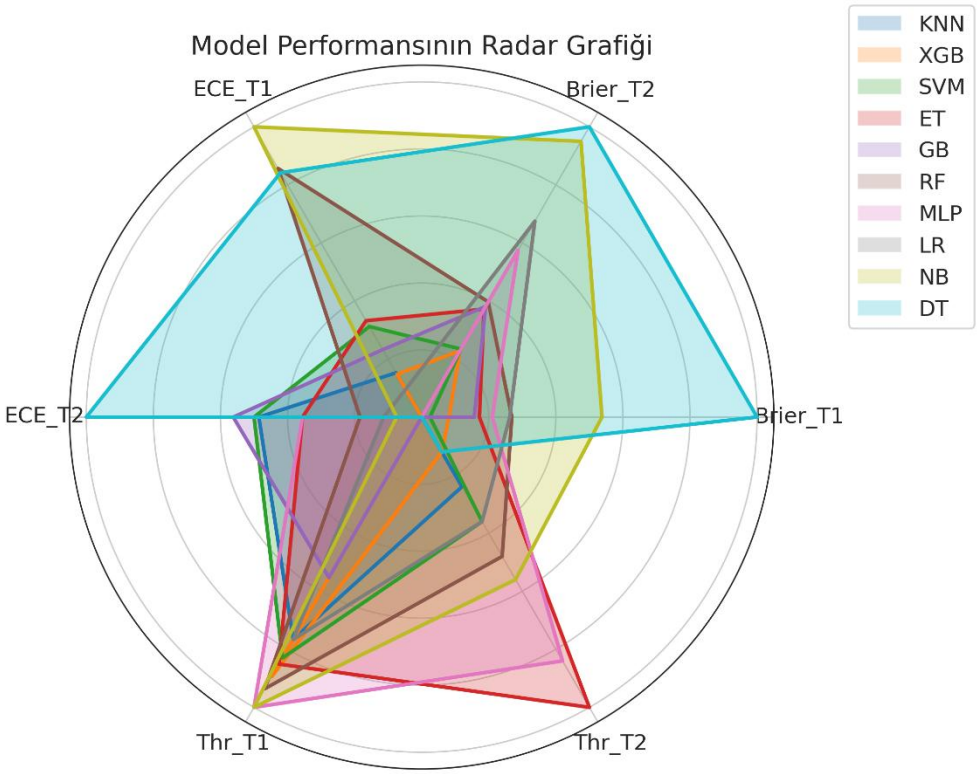
Kalibrasyon aşamasının temel amacı, modellerin ayırma gücünü (F1, AUC, doğruluk gibi metrikler) doğrudan artırmak değildir. Bunun yerine, üretilen olasılık tahminlerinin güvenilirliğini iyileştirmek hedeflenmiştir. Bu aşamada temel amaç olasılık güvenilirliğini artırarak topluluk öğrenme stratejileri ve Triyaj mekanizması için daha sağlam bir zemin hazırlayabilemtir. Klinik bağlamda bir modelin "%80 risk" demesi, gerçek pozitiflik oranının da yaklaşık %80 olması gerektiği anlamına gelir; aksi halde modelin çıktıları doğru yorumlanamaz. Bu nedenle kalibrasyonun başarısı, özellikle Brier Skoru, beklenen kalibrasyon hatası (ECE) ve eşik değeri (threshold, thr) üzerinden değerlendirilmiştir.

Çizelge 7. T1 (kalibrasyon öncesi) ve T2 (kalibrasyon sonrası) metrik karşılaştırma

Model	Brier_T1	Brier_T2	ECE_T1	ECE_T2	Thr_T1	Thr_T2
KNN	0,0795	0,0755	0,0524	0,0686	0,46	0,45
XGB	0,0848	0,0828	0,052	0,0451	0,52	0,42
SVM	0,0811	0,083	0,0644	0,0693	0,49	0,48
ET	0,0907	0,0874	0,0659	0,0623	0,5	0,64
GB	0,0897	0,0877	0,0584	0,0723	0,36	0,39
RF	0,097	0,0882	0,1054	0,054	0,54	0,51
MLP	0,0932	0,0939	0,0409	0,0623	0,57	0,6
LR	0,0966	0,0971	0,0481	0,0506	0,46	0,48
NB	0,1145	0,1059	0,1162	0,0487	0,57	0,53
DT	0,1446	0,1075	0,1043	0,0935	0,1	0,42

Çizelge 7'de T1 (kalibrasyon öncesi) ve T2 (kalibrasyon sonrası) aşamalarında bu üç metriği karşılaştırmaktadır. Sonuçlara göre kalibrasyon sonrasında Brier Skoru değerleri düşmüş, bu da tahmin edilen olasılıkların gerçeğe daha yakın hale geldiğini göstermiştir. ECE değerlerinde genel bir azalma gözlenmiş, yani olasılık çıktılarının güvenilirliği artmıştır. Threshold (thr) değerleri ise daha

tutarlı bir aralığa oturmuştur; bu durum karar verme sürecinde klinik uygulamaları kolaylaştıracak bir denge sağlamaktadır.



Şekil 9. T1 ve T2 aşamalarında Brier Skoru, ECE ve Threshold metriklerinin değişimi

Şekil 9'da radar grafiği, Brier Skoru, ECE ve threshold metriklerinin T1 ve T2 aşamalarındaki değişimini bütüncül olarak sunmaktadır. Grafik, kalibrasyonun özellikle olasılık güvenilirliği boyutunda sağladığı katkıyı görsel olarak ortaya koymaktadır. Özellikle Decision Tree (DT) ve Naive Bayes (NB) gibi uç olasılık üreten modellerde kalibrasyonun belirgin fayda sağladığı görülmektedir. Bununla birlikte, KNN ve SVM gibi zaten daha istikrarlı olasılık dağılımları üreten modellerde iyileşme sınırlı kalmıştır.

3.3 T3 - Zor Örnek Farkındalıklı Topluluk Öğrenmesi Sonuçları

Zor örnek farkındalıklı topluluk öğrenmesi sonuçları normal, zor ve global (normal+zor) örnekler olarak ayrı ayrı hesaplanarak farklı örnek tiplerinde farklı yaklaşımların başarı metrikleri hesaplanmıştır.

Çizelge 8. T3 aşamasında Normal örnekler için modellerin hesaplanan başarı metrikleri

AUC	F1	ACC	Precision	Recall	Brier	Bal_Acc	Set	Model
0,9726	0,9524	0,9458	0,9375	0,9677	0,0509	0,9428	Normal	En İyi Model (KNN)
0,9632	0,9524	0,9458	0,9375	0,9677	0,0521	0,9428	Normal	Eşit Topluluk Yaklaşımı
0,9638	0,9524	0,9458	0,9375	0,9677	0,052	0,9428	Normal	Ağırlıklı Topluluk
0,9676	0,9524	0,9458	0,9375	0,9677	0,0523	0,9428	Normal	Yığınlama Yaklaşımı

Çizelge 8 incelendiğinde Normal örneklerde en iyi model performansı: Normal veri kümesinde en iyi model (KNN) en yüksek AUC skoru (0.9726) ile öne çıkmakta olup, F1 skoru (0.9524), doğruluk (0.9458) ve dengeli doğruluk (0.9428) gibi metriklerde güçlü bir performans

sergilemektedir. Bu sonuçlar, normal örneklerin daha öngörülebilir yapıda olduğunu ve temel modelin bu alanda yeterli olduğunu göstermektedir; topluluk öğrenme yöntemleri (ör. Topluluk Yığınlama Yaklaşımı ile AUC 0.9676) benzer F1 ve doğruluk değerleri sunsa da en iyi modelin üstünlüğü AUC ve Brier skoru (0.0509) gibi kalibrasyon metriklerinde belirgindir.

Çizelge 9. T3 aşamasında zor örnekler için modellerin hesaplanan başarı metrikleri

AUC	F1	ACC	Precision	Recall	Brier	Bal_Acc	Set	Model
0,9317	0,8852	0,8727	0,871	0,9	0,1127	0,87	Zor	En İyi Model (KNN)
0,905	0,8852	0,8727	0,871	0,9	0,1326	0,87	Zor	Eşit Topluluk Yaklaşımı
0,9047	0,8852	0,8727	0,871	0,9	0,1322	0,87	Zor	Ağırlıklı Topluluk
0,9377	0,9153	0,9091	0,931	0,9	0,0949	0,91	Zor	Topluluk Yığınlama

Çizelge 9'da, Zor örneklerde en iyi model, eşit topluluk yaklaşımı ve ağırlıklı topluluk yöntemleri benzer F1 skorları (0.8852) ve doğruluk (0.8727) ile sınırlı performans gösterirken, Topluluk Yığınlama belirgin bir iyileşme sağlamaktadır (F1: 0.9153, AUC: 0.9377, doğruluk: 0.9091). Bu, Yığınlama'nın zor örneklerdeki gürültü ve karmaşıklığı daha iyi yönettiğini göstermektedir; ayrıca Brier skoru (0.0949) ve dengeli doğruluk (0.91) gibi metriklerde de üstünlük sağlamıştır. Eşit Topluluk ve Ağırlıklı Topluluk yöntemleri ise en iyi modele benzer şekilde daha düşük AUC (≈0.905) ile kalmaktadır.

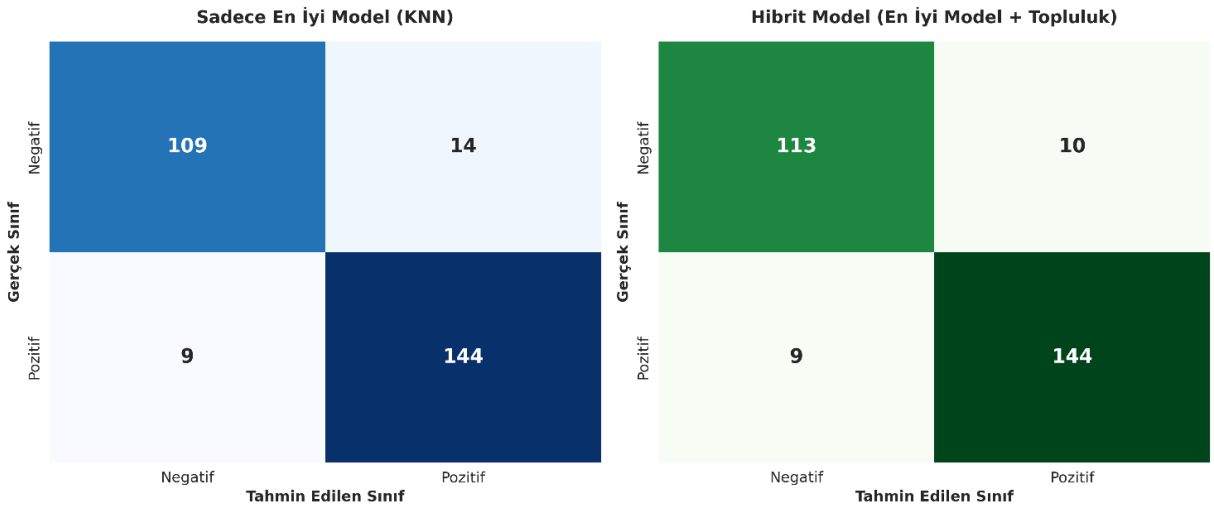
Çizelge 10. T3 aşamasında Global (Normal+Zor) örnekler için modellerin hesaplanan başarı metrikleri

AUC	F1	ACC	Precision	Recall	Brier	Bal_Acc	Set	Model
0,9574	0,926	0,9167	0,9114	0,9412	0,0755	0,9137	Global	En İyi Model (KNN)
0,9455	0,926	0,9167	0,9114	0,9412	0,0842	0,9137	Global	Eşit Topluluk Yaklaşımı
0,9457	0,926	0,9167	0,9114	0,9412	0,084	0,9137	Global	Ağırlıklı Topluluk
0,9535	0,9381	0,9312	0,9351	0,9412	0,0693	0,9299	Global	Topluluk Yığınlama
0,9574	0,9381	0,9312	0,9351	0,9412	0,0684	0,9299	Global	Hibrid (En iyi + Yığınlama)

Çizelge 10'da Global veri kümesinde (Normal + Zor), Hibrid (En İyi Model + Topluluk Yığınlama) yaklaşımı F1 skoru (0.9381) ve doğruluk (0.9312) ile Topluluk Yığınlama+ya yakın performans gösterirken, AUC skoru (0.9574) ile en iyi modelin seviyesine ulaşmaktadır. Hibrit model bireysel yöntemlerin güçlü yönlerini birleştirerek genel performansı optimize etmiş ve Brier skoru (0.0684) gibi kalibrasyon metriklerinde de iyileşme sunmuştur. Karşılaştırmada Eşit Topluluk ve Ağırlıklı Topluluk F1'de (0.926) geride kalırken, Yığınlama ve Hibrid global dengeli doğrulukta (0.9299) öne çıkmaktadır.

Zor örnek farkındalıklı topluluk öğrenmesi iyileştirme: Zor örnek farkındalıklı topluluk öğrenmesi stratejisi, zor örneklerde Topluluk Yığınlama yöntemi ile +0.0301 F1 skoru iyileştirmesi sağlamış; global hibrit model ise en iyi modele göre +0.0121 F1 skoru artışı sunmuştur. Bu kazanımlar AUC'de (Zor set için +0.0060; globalde eşit seviye) ve Brier skorlarında (Zor set için −0.0178) da yansımıştır; sonuç olarak zor örnek farkındalıklı topluluk öğrenmesi yaklaşım, zor örnek performansını yükseltirken genel model güvenilirliğini artırmıştır.

Elde edilen performans artışları (örneğin F1 skorunda +0.0121) sayısal olarak mütevazı görünmekle birlikte, yüksek performanslı modellerde (AUC > 0.95) marjinal kazanımların zorlaştığı "doygunluk bölgesi" dikkate alındığında değerlidir. Bu çalışmada istatistiksel anlamlılık testleri yapılmamış olmakla birlikte, odak noktası genel istatistiksel farktan ziyade, klinik açıdan kritik olan "zor vakalardaki" (Yanlış Negatif) azalmadır.



a) b)

Şekil 10. Global (normal + zor) örneklerde en iyi model ile hibrid model karşılaştırılması

Şekil 10'da görüldüğü üzere, global (normal + zor) test kümesi üzerinde, KNN tabanlı en iyi model pozitif sınıfı yüksek doğrulukla tahmin ederken (144 TP, 9 FN), negatif sınıfta 14 yanlış pozitif üretmiştir. Hibrit yaklaşım ise pozitif sınıf performansını koruyarak (144 TP, 9 FN), negatif sınıfın doğru tahminini artırmış (TN 109'dan 113'e yükselmiş) ve yanlış pozitif sayısını azaltmıştır (FP 14'ten 10'a düşmüştür). Bu sonuç, hibrit modelin özellikle negatif sınıf tahminlerini optimize ederek daha dengeli bir performans sağladığını ve genel doğruluk ile hassasiyet metriklerinde avantaj sunduğunu göstermektedir.

3.4 Tartışma

Bu çalışmada kalp hastalığı sınıflandırması için uygulanan aşamalı model geliştirme sürecinde, farklı aşamaların (T0–T3) performans katkıları sistematik olarak incelenmiştir. Bulgular, hem tekil modellerin güçlü yönlerini hem de hibrit stratejilerin sunduğu ek katkıları ortaya koymaktadır.

Öncelikle tekil modeller arasında KNN ve SVM öne çıkmıştır. Özellikle KNN, hem normal örneklerde yüksek ayırma gücü (AUC = 0.9726) hem de güçlü F1 skorları ile dikkat çekmiştir. Bu sonuç, literatürde de raporlandığı üzere, benzer tıbbi sınıflandırma problemlerinde KNN'in güvenilir bir alternatif olduğunu göstermektedir. SVM ise marjine dayalı karar yapısı sayesinde dengeli performans sağlamış, özellikle doğruluk ve recall metriklerinde rekabetçi olmuştur.

Zor örnekler üzerinde yapılan analiz, topluluk öğrenme stratejilerinin önemini ortaya koymuştur. Eşit Topluluk ve Ağırlıklı Topluluk yöntemleri beklenen katkıyı sağlayamazken, topluluk yığınlama yöntemi zor örneklerde belirgin bir iyileşme sunmuştur (F1: 0.9153). Bu bulgu, yığınlama'nın modeller arası heterojenliği kullanarak sınırda örneklerin daha iyi öğrenilmesini sağladığını göstermektedir.

Çalışmanın özgün katkısı olan Hibrid (Zor Örnek Farkındalıklı) Global Model, normal örneklerde en iyi model (KNN), zor örneklerde ise en iyi ensemble yöntemi (yığınlama) kullanarak global düzeyde performans artışı sağlamıştır. Her ne kadar F1 iyileştirmesi sınırlı (+0.0121) olsa da, bu artışın modelin zaten yüksek performans seviyesine ulaşmış olduğu (doygunluk aşaması) göz önüne alındığında önemli olduğu söylenebilir. Doygun modele yapılan bu küçük ek katkılar, aslında yüksek riskli zor örneklerdeki hataların azaltılmasında klinik açıdan ciddi bir fark yaratabilmektedir.

Buna ek olarak, SHAP analizleri yalnızca özelliklerin açıklanması için değil, aynı zamanda aktif bir özellik mühendisliği aracı olarak da kullanılmıştır. Özellikler arası etkileşimlerin SHAP değerleri üzerinden çıkarılması, klasik korelasyon temelli yöntemlerin ötesine geçilmesini sağlamış

ve modele yeni anlamlı değişkenlerin kazandırılmasına yol açmıştır. Bu yaklaşım, SHAP'ın yalnızca açıklayıcı değil, aynı zamanda yapıcı (model geliştirme sürecine entegre edilebilen) bir rol üstlenebileceğini göstermektedir.

Son olarak, bu çalışmada izlenen aşamalı model geliştirme süreci (Çok Aşamalı İş Akışı) de dikkat çekicidir. Varsayılan modellerden başlayarak hiperparametre optimizasyonu, SHAP tabanlı özellik mühendisliği, olasılık kalibrasyonu ve zor örnek odaklı hibrit stratejiler adım adım uygulanmıştır. Bu iteratif süreç, her aşamada modelin farklı bir boyutunu (ayırma gücü, açıklanabilirlik, kalibrasyon, zor örnek duyarlılığı) geliştirmiş ve sonuçta daha güvenilir bir klinik karar destek sistemi elde edilmiştir.

4. SONUÇ

Bu çalışma, kalp hastalığı sınıflandırmasında aşamalı model geliştirme sürecini temel alarak, model performansının yalnızca genel doğruluk değil, aynı zamanda olasılık güvenilirliği ve zor örnek duyarlılığı boyutlarıyla da değerlendirilebileceğini göstermiştir. T0'dan T3'e kadar ilerleyen adımlar, her aşamada farklı bir katkı sunmuştur: varsayılan modeller temel performansı sağlamış, hiperparametre optimizasyonu ve SHAP tabanlı özellik mühendisliği ayırma gücünü artırmış, kalibrasyon olasılık çıktılarının güvenilirliğini iyileştirmiş, hibrit zor örnek farkındalıklı stratejiler ise zor örneklerdeki hata oranlarını azaltarak genel performansı güçlendirmiştir.

Elde edilen bulgular, özellikle K-En Yakın Komşu (KNN) ve Destek Vektör Makineleri (SVM) gibi modellerin normal örneklerde yüksek performans gösterdiğini, ancak zor örneklerin üstesinden gelmek için topluluk öğrenme ve hibrit yaklaşımların gerekli olduğunu ortaya koymuştur. Hibrid (en iyi model + Topluluk Yığınlama) modelin +0.0121'lik F1 artışı, doygun performans seviyesinde klinik bağlamda kritik bir kazanım sunmaktadır. Bu iyileşme, örneğin yanlış negatiflerin azalmasıyla daha fazla hastanın doğru tanı almasını sağlayabilir. Ayrıca, SHAP analizlerinin yalnızca açıklama amacıyla değil, aktif bir özellik mühendisliği aracı olarak da kullanılabileceği gösterilmiştir.

Gelecek çalışmalar için üç yön öne çıkmaktadır. İlk olarak, derin öğrenme tabanlı belirsizlik ölçütleri veya Bayesian yaklaşımlar gibi farklı zor örnek tespit yöntemlerinin test edilmesi, bu örneklerin daha güvenilir ayrıştırılmasını sağlayabilir. İkinci olarak, daha büyük ve çok merkezli veri setlerinin kullanılması, bulguların genellenebilirliğini artıracaktır. Üçüncü olarak, boosting tabanlı hibritler veya attention mekanizmalarıyla desteklenen yığınlama yapıları gibi yeni topluluk stratejilerinin geliştirilmesi, zor örneklerdeki iyileştirmeleri daha da güçlendirebilir. Özellikle çok merkezli veri setleriyle genellenebilirlik üzerine odaklanmak, klinik uygulamalarda daha geniş bir etki yaratabilir.

Sonuç olarak, bu çalışma yalnızca yüksek doğruluk elde etmeyi değil, aynı zamanda model güvenilirliğini ve zor vakaların yönetimini önceliklendiren çok boyutlu bir yaklaşım sunmaktadır. Bu yaklaşım, klinik karar destek sistemlerinin güvenli ve etkin şekilde kullanılabilmesi için yol gösterici niteliktedir.

5. ÇIKAR ÇATIŞMASI

Yazarlar, bilinen herhangi bir çıkar çatışması veya herhangi bir kurum/kuruluş ya da kişi ile ortak çıkar bulunmadığını onaylamaktadırlar.

6. YAZAR KATKISI

Emin DEMİR çalışmanın kavramsal ve tasarım süreçlerinin belirlenmesi, veri toplama, veri analizi ve yorumlama, makale taslağının oluşturulması, son onay ve tam sorumluluk kısımlarında katkıda bulunmuştur. Yusuf Ziya AYIK çalışmanın kavramsal ve tasarım süreçlerinin yönetimi, fikirsel içeriğin eleştirel incelemesi, son onay ve tam sorumluluk kısımlarında katkıda bulunmuştur.

7. KAYNAKLAR

Alférez, G. H., Esteban, O. A., Clausen, B. L., & Ardila, A. M. M. (2022). Automated machine learning pipeline for geochemical analysis. *Earth Science Informatics*, *15*(3), 1683-1698.

Alvarez, S. A. (2002). An exact analytical relation among recall, precision, and classification accuracy in information retrieval. *Boston College, Boston, Technical Report BCCS-02, 1*, 1-22.

Aragonés Lozano, M., Pérez Llopis, I., & Esteve Domingo, M. (2023). Threat hunting system for protecting critical ınfrastructures using a machine learning approach. *Mathematics*, *11*(16), 3448.

Aydemir, Ö. (2022). Dengesiz veri kümelerinin sınıflandırılmasında poligon alan metriğinin sınıflandırıcı performans değerlendirilmesi için kullanılması. *Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *27*(2), 194-205.

Behera, D. K., Dash, S., Behera, A. K., & Dash, C. S. K. (2021, December). Extreme gradient boosting and soft voting ensemble classifier for diabetes prediction. In *2021 19th OITS International Conference on Information Technology (OCIT)* (pp. 191-195). IEEE.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794).

Demir, E., Bozkurt, F., & Ayık, Y. Z. (2023). Early-stage heart failure disease prediction with deep learning approach. *Journal of Scientific Reports-A*, *055*, 34-49.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304-310.

Dogan, A., & Birant, D. (2019). A Weighted Majority Voting Ensemble Approach for Classification. *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 1-6. 2019 4th International Conference on Computer Science and Engineering (UBMK).

Erdem, K., Yasin, E., Yıldız, M. B., & Koklu, M. (2024). Classification of heart diseases with ensemble learning algorithms. *Sinop Üniversitesi Fen Bilimleri Dergisi*, *9*(2), 369-387.

Fedesoriano. (2021). *Heart Failure Prediction Dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Feng, C., Sutherland, A., King, S., Muggleton, S., & Henery. (1992). *Statlog (Heart)* [Dataset]. UCI Machine Learning Repository.

Filippou, K., Aifantis, G., Papakostas, G. A., & Tsekouras, G. E. (2023). Structure learning and hyperparameter optimization using an automated machine learning (AutoML) pipeline. *Information*, *14*(4), 232.

Fränti, P., & Mariescu-Istodor, R. (2023). Soft precision and recall. *Pattern Recognition Letters*, *167*, 115-121.

Ganie, S. M., Pramanik, P. K. D., & Zhao, Z. (2025). Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. *Scientific Reports*, *15*(1), 13912.

García, V., Mollineda, R. A., & Sánchez, J. S. (2009, June). Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis* (pp. 441-448). Berlin, Heidelberg: Springer Berlin Heidelberg.

Haihong, E., Zhou, K., & Song, M. (2019, December). Spark-based machine learning pipeline construction method. In *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)* (pp. 1-6). IEEE.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). Model assessment and selection. In *The elements of statistical learning: data mining, inference, and prediction* (pp. 219-259). New York, NY: Springer New York.

Kaya, E. N., & Görmez, Y. (2025). Proje efor tahmini için makine öğrenmesi modellerinin geliştirilmesi ve shap yöntemi kullanılarak açıklanması. *Mühendislik Bilimleri ve Tasarım Dergisi*, *13*(2), 528-544.

Kull, M., & Flach, P. A. (2014, September). Reliability maps: a tool to enhance probability estimates and improve classification accuracy. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 18-33). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kumar, A., Liang, P. S., & Ma, T. (2019). Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, *32*.

Kusuma, C. C., Chandra, N., Utomo, G. N., Parimartha, P. A., Hidayat, J., Anggreainy, M. S., & Parmonangan, I. H. (2023, September). Diagnosing Diabetes Using Soft Voting Ensemble Machine Learning Model. In *2023 10th International Conference on ICT for Smart Society (ICISS)* (pp. 1-6). IEEE.

Küçükmanisa, A., & Kilimci, Z. H. (2024). Heart Disease Prediction with Machine Learning-Based Approaches. *Sakarya University Journal of Science*, *28*(1), 101-107.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning* (pp. 625-632).

Öznacar, T., & Sertkaya, Z. T. (2024). Heart Failure Prediction: A Comparative Study of SHAP, LIME, and ICE in Machine Learning Models. *International Journal of Computational and Experimental Science and Engineering*, *10*(4).

Pearly, A. A., & Karthik, B. (2024, October). A Hybrid Deep Learning Framework for Lung Disorder Detection Using Multi-Scale Feature Extraction and Ensemble Classification. In *International Conference on Advancements in Smart Computing and Information Security* (pp. 118-132). Cham: Springer Nature Switzerland.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifier*, *10*(3), 61-74.

Saihood, Q., & Sonuç, E. (2023). A practical framework for early detection of diabetes using ensemble machine learning models. *Turkish Journal of Electrical Engineering and Computer Sciences*, *31*(4), 722-738.

Sanders, F. (1967). The Verification of Probability Forecasts. *Journal of Applied Meteorology*, *6*(5), 756-761.

Shah, M., Gandhi, K., Patel, K. A., Kantawala, H., Patel, R., & Kothari, A. (2023). Theoretical Evaluation of Ensemble Machine Learning Techniques. *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 829-837.

Shah, P., Shukla, M., Dholakia, N. H., & Gupta, H. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. *Scientific Reports*, *15*(1), 17927.

Solanki, S., Verma, S., & Chahar, K. (2020). A comparative study of information retrieval using machine learning. *Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019*, 35-42.

Stehouwer, N., Rowland-Seymour, A., Gruppen, L., Albert, J. M., & Qua, K. (2025). Validity and reliability of Brier scoring for assessment of probabilistic diagnostic reasoning. *Diagnosis*, *12*(1), 53-60.

Sungur, F., & Bakır, H. (2024). Hiperparametre ayarlama ve veri dengelemenin kalp hastalığı tahmini için kullanılan makine öğrenimi algoritmaları üzerindeki etkilerinin incelenmesi. *Bilişim Teknolojileri Dergisi*, *17*(1), 45-58.

Takcı, H. (2022). Optimum parametreler yardımıyla performansı artırılmış KNN algoritması tabanlı kalp hastalığı tahmini. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, *38*(1), 451-460.

Tomak, L., & Bek, Y. (2011). İşlem Karakteristik Eğrisi Analizi ve Eğri Altında Kalan Alanların Karşılaştırılması. *Deneysel ve Klinik Tıp Dergisi*, *27*(2).

Yapıcı, İ. Ş., Arslan, R. U., & Erkaymaz, O. (2024). Kalp Yetmezliği Tanılı Hastaların Hayatta Kalma Tahmininde Topluluk Makine Öğrenme Yöntemlerinin Performans Analizi. *Karaelmas Fen ve Mühendislik Dergisi*, *14*(1), 59-69.