Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 473-485, 11, 2024

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 473-485, 11, 2024

Research Article

Makine Öğrenmesi ve Özellik Seçimi ile Anemi Hastalığı Sınıflandırması

Kerim Berkay BUÇAN[*], Serhat KILIÇARSLAN [2]

[1]*Yazılım Mühendisliği / Lisansüstü Eğitim Enstitüsü, Bandırma Onyedi Eylül Üniversitesi, Türkiye*
[2]*Yazılım Mühendisliği / Mühendislik ve Doğa Bilimleri Fakültesi, Bandırma Onyedi Eylül Üniversitesi, Türkiye*

[*]*(kerimbucan@ogr.bandirma.edu.tr)*

(Received: 12 December 2024, Accepted: 17 December 2024)

(4th International Conference on Frontiers in Academic Research ICFAR 2024, December 13-14, 2024)

ATIF/REFERENCE: Buçan, K. B. & Kılıçarslan, S. (2024). Makine Öğrenmesi ve Özellik Seçimi ile Anemi Hastalığı Sınıflandırması. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(11), 473-485.

Özet – Anemi, kandaki kırmızı kan hücrelerinin veya hemoglobin seviyesinin vücudun ihtiyaç duyduğu oksijeni taşıyamayacak kadar düşük olduğu yaygın bir sağlık sorunudur. Bu durum, yorgunluk, baş dönmesi ve nefes darlığı gibi belirtilerle kendini gösterirken, özellikle çocuklar, kadınlar ve yaşlılar gibi hassas gruplarda daha ciddi etkiler yaratmaktadır. Geleneksel anemi teşhis yöntemleri, zaman alıcı ve maliyetli olması nedeniyle sınırlı sağlık kaynaklarına sahip bölgelerde uygulanabilirliği kısıtlıdır. Bu çalışmada, anemi teşhisinde makine öğrenmesi yöntemlerinin ve özellik seçimi tekniklerinin uygulanabilirliği incelenmiştir. Veri seti 15.300 bireyden oluşmaktadır. Farklı yaş ve cinsiyet gruplarından bireylerin hematolojik ve biyokimyasal özelliklerini içermektedir. Bu çalışmada, özellik seçimi için Korelasyon Matrisi, Recursive Feature Elimination ve Boruta algoritmaları uygulanmış ve elde edilen özellikler, Random Forest, Gradient Boosting Machine , XGBoost, Karar Ağacı ve Lojistik Regresyon makine öğrenmesi modelleri üzerinde değerlendirilmiştir. Deneysel sonuçlar, Gradient Boosting Machine modelinin %99,97 doğruluk, %99,97 duyarlılık, %100 özgüllük ve %99,97 F1 skoru ile en yüksek performansı sergilediğini ortaya koymuştur. Özellikle Gradient Boosting Machine modeli, hem Boruta algoritmasıyla hem de Korelasyon Matrisiyle seçilen özellikler üzerinde bu üstün performansı göstermiştir. XGBoost modeli ise %99,90 doğruluk, %99,91 duyarlılık, %100 özgüllük ve %99,90 F1 skoru ile GBM'in hemen ardından gelmiştir. Bu model, Korelasyon ve Boruta algoritmalarıyla seçilen özelliklerle etkili sonuçlar sağlamıştır. Karar Ağacı modeli, %99,90 doğruluk, %99,90 duyarlılık, %100 özgüllük ve %99,90 F1 skoru ile XGBoost modeliyle benzer performans sergilemiştir ve en iyi sonuçlarını Boruta algoritmasıyla seçilen özellikler üzerinde göstermiştir.

Anahtar Kelimeler – Anemi Hastalığı, Sınıflandırma, Özellik Seçimi, Makine Öğrenmesi.

I. GİRİŞ

Anemi, kandaki kırmızı kan hücrelerinin veya hemoglobin seviyesinin vücudun ihtiyaç duyduğu oksijeni taşıyamayacak kadar düşük olduğu yaygın bir sağlık sorunudur [1]. Bu durum, vücudun dokularına yeterli oksijen taşınamamasına neden olur ve yorgunluk, halsizlik, baş dönmesi, nefes darlığı gibi belirtilerle kendini gösterebilir. Anemi, bireyin yaşı, cinsiyeti ve genel sağlık durumuna göre değişen normal hemoglobin değerlerinin altında tespit edilir [2]. Bu durum, demir, folik asit (B9 vitamini) ve B12 vitamini gibi temel besin maddelerinin eksikliklerinden kaynaklanabilir. Örneğin, demir eksikliği, kırmızı kan hücrelerinin yeterli miktarda hemoglobin üretememesine yol açarak oksijen taşıma kapasitesini

azaltır [3]. Anemi, özellikle çocuklar, üreme çağındaki kadınlar ve yaşlılar gibi savunmasız popülasyon gruplarında yaygındır ve ciddi sağlık sorunlarına yol açabilir. Demir eksikliği anemisi, çocukların zihinsel ve fiziksel gelişimini olumsuz etkileyerek bağışıklık sistemini zayıflatır ve enfeksiyonlara yatkınlığı artırır [2]. Bu durum sadece bireysel sağlık açısından değil, aynı zamanda toplum sağlığı ve ekonomik üretkenlik üzerinde de önemli etkilere sahiptir. Dünya genelinde yaklaşık 1,6 milyar insanın anemiden etkilendiği tahmin edilmektedir ve bu oran dünya nüfusunun yaklaşık %24'üne karşılık gelmektedir [4].

Anemi teşhisinde, hemoglobin konsantrasyonu (Hb), hematokrit (HCT) değeri ve kırmızı kan hücrelerinin (RBC) sayısı gibi hematolojik parametrelerin yanı sıra serum demir düzeyi, toplam demir bağlama kapasitesi (TIBC) ve ferritin seviyeleri gibi biyokimyasal göstergeler de önemli bir rol oynar [5]. Örneğin, demir eksikliği durumunda ferritin seviyesinin düşük olduğu gözlemlenirken, enflamatuvar durumlarla ilişkili anemilerde serum demir düzeyi düşük, ancak ferritin seviyesi yüksek olabilir. Geleneksel teşhis yöntemleri, bu parametrelerin laboratuvar testleri yoluyla ölçülmesini ve tıbbi uzmanlar tarafından yorumlanmasını gerektirir. Ancak bu süreç hem zaman alıcıdır hem de maliyetlidir, bu da özellikle kaynakları kısıtlı sağlık sistemlerinde önemli bir zorluk oluşturur.

Makine öğrenmesi (ML) ve veri madenciliği, sağlık sektöründe giderek artan bir şekilde kullanılmaktadır ve anemi teşhisi gibi karmaşık problemleri çözmede etkili araçlar sunmaktadır. Bu yöntemler, büyük veri kümelerindeki karmaşık örüntüleri analiz ederek otomatik ve hızlı teşhis imkanı sağlar. Anemi teşhisinde ML'nin sağladığı avantajlar arasında daha yüksek doğruluk, daha düşük maliyet ve zaman tasarrufu gibi faktörler yer alır. Örneğin, geleneksel yöntemlerde bir laboratuvar teknisyeni tarafından manuel olarak yorumlanan veriler, ML algoritmaları ile otomatik olarak analiz edilebilir ve bu da daha hızlı karar verme süreçlerine olanak tanır [6].

Literatürde, anemi teşhisi ve sınıflandırılması için birçok ML ve derin öğrenme yöntemi kullanılmıştır. Bu çalışmaların tarihsel gelişimi ve doğruluk oranlarına göre performansları incelenmiştir.

Khan vd. (2019), Bangladeş'te çocukluk çağı anemisini tahmin etmek amacıyla çeşitli ML algoritmalarını karşılaştırmış ve Rastgele Orman (RF) algoritmasının %68,53 doğruluk oranı ile en iyi sonuçları verdiğini tespit etmiştir [9]. Veri dengesizliği ve sınırlı özellik seti gibi zorluklar, bu çalışmanın odak noktalarından biri olmuştur.

Kilicarslan vd. (2021), Tokat Gaziosmanpaşa Üniversitesi'nden toplanan 15,300 hastanın kan verileri üzerinde çalışarak genetik algoritma ile optimize edilmiş GA-SAE ve GA-CNN hibrit modelleri geliştirmiş ve GA-CNN modelinin %98,50 doğruluk oranıyla en iyi performansı sağladığını göstermiştir [4]. Aynı yıl, Sen vd. (2021), mikroskobik görüntülerde orak hücrelerin tanımlanması ve sınıflandırılmasında RF algoritmasını kullanarak %92 doğruluk oranı ile başarılı sonuçlar elde etmiştir [11].

Verma vd. (2022), Hindistan'da ayakta tedavi gören hastalardan alınan kan sayımı verilerini kullanarak Naive Bayes algoritmasının diğer yöntemlere göre daha iyi bir performans gösterdiğini ve %96,09 doğruluk oranı ile öne çıktığını vurgulamıştır [8]. Benzer şekilde, Vohra vd. (2022), veri dengesizliğini gidermek için SMOTE yöntemini kullanarak Çok Katmanlı Algılayıcı (MLP) modeliyle anemi hastalığını hafif, orta ve ağır olarak sınıflandırmış ve %99,35 doğruluk oranına ulaşmıştır [7].

Hasan vd. (2023), fetal anemi tahmininde K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM) ve Light Gradient Boosting Machine (LGBM) algoritmalarının bir oylama sınıflandırıcısı ile birleştirilmesi sonucu %99,95 doğruluk oranı elde etmiştir [10]. Aynı yıl, Saputra vd. (2023), beta talasemi ve demir eksikliği anemisi gibi türlerin ayırt edilmesi için Extreme Learning Machine (ELM) algoritması kullanarak %99,21 doğruluk oranına ulaşmıştır [12]. Yağmur vd. (2023), Kaggle veri tabanından alınan kan verilerini kullanarak Öğrenmeli Vektör Kuantalama (LVQ), Rekabetçi Katman

Sinir Ağı (CLNN) ve Kendiliğinden Organize Olan Harita (SOM) gibi çeşitli yapay sinir ağı yöntemleriyle anemi sınıflandırması yapmış ve PRNN yönteminin %99,88 doğruluk oranıyla en başarılı sonuçları verdiğini belirtmiştir [6].

Özellik seçimi, büyük veri kümelerinde en anlamlı ve etkili değişkenlerin belirlenmesi amacıyla kullanılan önemli bir veri ön işleme tekniğidir. Bu yöntemler, özellikle tıbbi teşhis alanında, veri boyutunu azaltarak hesaplama maliyetini düşürmekte ve modelin genel doğruluğunu artırmaktadır. Literatürde, anemi teşhisi için geliştirilen modellerde de sıklıkla özellik seçimi tekniklerinin kullanıldığı görülmektedir.

Zhang vd. (2017) tarafından önerilen ağırlıklandırma ve sıralama tabanlı hibrit bir özellik seçimi yöntemi bulunmaktadır. Bu yöntem, özellikle felç riski tahmininde etkili olmuş ve 28 özellikten sadece 9'unun kullanılmasıyla yüksek doğruluk oranları elde edilmiştir [13].

Singh vd. (2021), Kardiyovasküler Sağlık Çalışması (CHS) veri setini kullanarak felç tahmini için karar ağacı tabanlı bir özellik seçimi yöntemi önermiş ve ardından ana bileşen analizi (PCA) ile boyut indirgeme işlemi gerçekleştirmiştir. Sonuçlar, optimal özellik seti ile %97,7 doğruluk oranına ulaşarak diğer yöntemleri geride bırakmıştır [14].

Pathan vd. (2022) genetik algoritma (GA) ve Recursive Feature Elimination (RFE) gibi yöntemlerin, optimal özellik setlerini belirleyerek sınıflandırma doğruluğunu artırdığını vurgulamaktadır. Örneğin, GA-SVM algoritması, Cleveland veri setinde tüm özelliklerle %83,34 doğruluk sağlarken, seçilen özelliklerle bu oran %88,34'e yükselmiştir [15].

Bu bulgular, özellik seçiminin hem anemi hem de diğer hastalıkların teşhisinde model performansını artırma potansiyelini ortaya koymaktadır. Özellikle büyük veri setlerinde, gereksiz özelliklerin çıkarılması, makine öğrenmesi modellerinin daha hızlı ve doğru bir şekilde çalışmasını sağlamaktadır.

Bu çalışma, anemi hastalığının teşhisini makine öğrenmesi yöntemleri ve özellik seçimi teknikleriyle daha etkili ve kolay hale getirmeyi amaçlamaktadır. Boruta, RFE ve korelasyon tabanlı özellik seçimi yöntemleri kullanılarak veri setindeki anlamlı özellikler belirlenmiş, ardından bu özellikler çeşitli makine öğrenmesi algoritmalarıyla uygulanmıştır. Çalışmanın temel hedefi, hem tanı süreçlerini hızlandırarak hem de doğruluk oranını artırarak sağlık alanına katkı sağlamaktır.

Performans değerlendirmeleri için, karmaşıklık matrisinden türetilen duyarlılık, özgüllük, doğruluk ve F1 Skoru metrikleri kullanılmıştır. Bu metrikler, modellerin hem doğru pozitif hem de doğru negatif sınıflandırmalarını ölçerek genel başarılarını ve doğruluk oranlarını kapsamlı bir şekilde değerlendirmek için tercih edilmiştir. Çalışma, bu metrikler aracılığıyla makine öğrenmesi modellerinin etkili bir şekilde kıyaslanmasına ve anemi teşhisi için en uygun yöntemlerin belirlenmesine katkı sağlanmasını hedeflemiştir.

Çalışmanın geri kalan bölümleri şu şekilde düzenlenmiştir: Bölüm 2'de çalışmada kullanılan veri seti ve yöntemler açıklanmıştır. Bölüm 3'te ise deneysel sonuçlar paylaşılmıştır. Son olarak, Bölüm 4'te çalışmanın sonuçları ele alınmıştır.

II. MATERYAL VE YÖNTEM

A. Veri Seti

Bu çalışmada kullanılan veri seti, anemi hastalığı üzerine Tokat Gaziosmanpaşa Üniversitesi Tıp Fakültesi'nden elde edilen veri setimiz 15300 bireyden ve 29 özellik oluşmaktadır [4]. Veri seti, hamile kadınlar, çocuklar ve kanser hastaları hariç olmak üzere, farklı yaş gruplarından ve cinsiyetlerden bireyleri kapsamaktadır. Veri setinde yer alan özellikler arasında Beyaz Kan Hücresi (WBC), RBC, HGB, HCT ve serum demir, B12, folat gibi biyokimyasal parametreler bulunmaktadır. Veri setindeki hedef değişken, bireylerin anemi durumunu ve türünü temsil etmektedir. Veri ön işleme aşamasında, hedef değişken dışındaki bağımsız değişkenler seçilmiş ve gereksiz sütunlar veri setinden çıkarılmıştır.

B. Özellik Seçimi

Makine öğrenmesi modellerinin performansı, kullanılan özelliklerin kalitesine ve anlamlılığına bağlıdır. Ancak, veri setinde bulunan bazı özellikler yüksek korelasyon, düşük bilgi değeri veya hedef değişkenle zayıf ilişkiler nedeniyle modellerin başarısını olumsuz etkileyebilir [16]. Gereksiz ve ilişkisiz özellikler, modelin hem doğruluk oranını hem de hesaplama maliyetini artırabilir. Bu sebeple, özellik seçimi, veri boyutunu azaltarak yalnızca anlamlı özelliklerin modele dahil edilmesini sağlar ve modelin hem doğruluğunu artırır hem de hesaplama süresini kısaltır. Bu çalışmada, özellik seçimi için üç farklı yöntem uygulanmıştır: Korelasyon Matrisi, RFE ve Boruta Algoritması.

Korelasyon Matrisi

Korelasyon matrisi, veri setindeki özelliklerin birbirleriyle olan ilişkisini ölçmek ve aşırı ilişkili (multicollinearity) özellikleri tespit etmek için kullanılmıştır. Bu matriste, özellikler arasındaki korelasyon katsayıları hesaplanmıştır. Korelasyon katsayısı rrr, bir özelliğin diğer bir özellik üzerindeki doğrusal ilişki düzeyini ifade eder [16]. Katsayının mutlak değeri 0.9'dan büyük olan özellikler yüksek derecede ilişkili kabul edilmiş ve bu durum gereksiz bilgi tekrarına yol açabileceğinden bir özelliğin çıkarılması gerektiğine karar verilmiştir. Yüksek korelasyon gösteren özelliklerden biri rastgele seçilerek veri setinden çıkarılmış ve böylece modelin gereksiz bilgiyle yüklenmesi engellenmiştir. Bu yaklaşım, veri setinin boyutunu düşürmüş ve modellerin daha genel ve etkili hale gelmesine katkı sağlamıştır.

RFE

RFE, özelliklerin model performansına katkısını iteratif bir süreçle değerlendirerek düşük önem derecesine sahip özellikleri eler [17]. Bu çalışmada, Lojistik Regresyon (LG) tabanlı bir RFE yöntemi uygulanmıştır. RFE, modelin tahmin doğruluğunu en üst düzeye çıkaran özellikleri belirlemek için başlangıçta tüm özelliklerle modeli eğitip, her iterasyonda en az katkı sağlayan özellikleri çıkararak ilerler. Bu işlem sonucunda, en yüksek önem derecesine sahip 10 özellik belirlenmiş ve bu özellikler kullanılarak daha kompakt ve hedef odaklı bir veri seti oluşturulmuştur. Böylece, modelin daha iyi genelleme yapması ve hesaplama süresinin azaltılması sağlanmıştır.

Boruta Algoritması

Boruta, Random Forest tabanlı bir özellik seçimi algoritmasıdır ve modelin başarısını artırmak için önemli özellikleri belirlerken gereksiz olanları eler [18]. Bu algoritma, orijinal veri setindeki her özelliğin önem düzeyini değerlendirir ve ardından rastgele oluşturulan "gölge özelliklerle" karşılaştırır. Boruta, hedef değişkenle anlamlı ilişkiler gösteren özellikleri seçerek, veri setini daha etkili hale getirir. Bu çalışmada Boruta algoritması, veri setindeki özelliklerin önem derecelerini ölçerek yalnızca anlamlı özelliklerin modele dahil edilmesini sağlamış ve gereksiz özelliklerin etkisini ortadan kaldırmıştır.

C. ML Algoritmaları

Lojistik Regresyon (LG)

LG, iki sınıflı ya da çok sınıflı sınıflandırma problemleri için sıklıkla kullanılan temel bir makine öğrenmesi modelidir [19]. Verilerin doğrusal olarak ayrılabildiği durumlarda oldukça etkili sonuçlar sunar. Model, sınıflandırma problemlerini çözmek için sigmoid fonksiyonu kullanır ve verilerin bir sınıfa ait olma olasılığını tahmin eder. Bu çalışmada, lojistik regresyon modelinin özellikle basitliği ve hızlı hesaplama avantajları nedeniyle tercih edildiği görülmektedir. Model performansı duyarlılık, özgüllük, doğruluk ve F1 skoru gibi metriklerle değerlendirilmiştir.

Random Forest (RF)

RF, birden fazla karar ağacını birleştirerek sınıflandırma veya regresyon yapan bir topluluk öğrenme algoritmasıdır [20]. Algoritma, her bir karar ağacını farklı veri alt kümeleri ve özellik setleriyle eğiterek genelleme performansını artırır ve aşırı öğrenme (overfitting) riskini azaltır. Özellikle karmaşık ve büyük veri setlerinde yüksek doğruluk oranları sunar. Bu çalışmada, RF algoritması, farklı özellik seçimi yöntemleriyle elde edilen veri setleri üzerinde uygulanmış ve modelin doğruluğu, özgüllüğü ve duyarlılığı gibi metriklerle performansı ölçülmüştür.

Gradient Boosting Machine (GBM)

GBM, sınıflandırma hatalarını iteratif olarak minimize eden bir topluluk öğrenme yöntemidir [21]. Her bir model, önceki modelin hatalarından öğrenerek performansını adım adım iyileştirir. Bu yaklaşım, veri setindeki karmaşık yapıları öğrenmede oldukça başarılıdır. GBM, modelin tahmin performansını artırmak için hata azaltımını optimize eder ve genellikle sınırlı veri setlerinde dahi etkili sonuçlar sunar. Performans değerlendirmesi duyarlılık, özgüllük, doğruluk ve F1 skoru metrikleri ile yapılmıştır.

XGBoost

XGBoost, GBM'nin optimize edilmiş ve hızlandırılmış bir versiyonudur [22]. Özellikle büyük veri setlerinde hızlı çalışması ve bellek kullanımını minimize etmesiyle dikkat çeker. XGBoost, veri setlerindeki eksik verilerle başa çıkabilir ve düzenlileştirme teknikleri sayesinde aşırı öğrenmeyi önler. Bu çalışmada XGBoost algoritması, farklı özellik seçimi yöntemleriyle eğitilmiş ve model performansı duyarlılık, doğruluk ve F1 skoru gibi metriklerle ölçülmüştür.

Karar Ağacı

Karar ağaçları, verileri dallara ayırarak basit ve açıklanabilir sınıflandırma modelleri oluşturur [23]. Her bir dal, veriyi belirli bir özelliğe göre ayrıştırır ve bu süreç, hedef değişkene en iyi şekilde ulaşılana kadar devam eder. Karar ağaçları, özellikle açıklanabilirliği yüksek modeller gerektiren durumlarda tercih edilir. Bu çalışmada, karar ağacı algoritması ile elde edilen sonuçlar duyarlılık, özgüllük, doğruluk ve F1 skoru metrikleriyle değerlendirilmiştir.

Bu algoritmalar, veri setine uygulanan Boruta, RFE ve korelasyon tabanlı özellik seçimi yöntemleriyle birleştirilmiş ve anemi teşhisinde en uygun yöntemi belirlemek için performans karşılaştırmaları yapılmıştır.

III. DENEYSEL SONUÇLAR

Deneysel çalışmalar, Google Colab platformu üzerinde NVIDIA L4 GPU ile gerçekleştirilmiştir. NVIDIA L4 GPU, enerji verimliliği ve yapay zeka iş yükleri için optimize edilmiş olup, yüksek performanslı hesaplamalar için 22.5 GB GPU belleği sunmaktadır [2]. Bu GPU, özellikle derin öğrenme ve makine öğrenmesi gibi büyük ölçekli veri işleme görevlerinde etkili bir şekilde kullanılmaktadır. Google Colab, bu tür zorlu işlemleri desteklemek için 53.0 GB'a kadar sistem RAM'i ve güçlü bir işlem gücü sağlamaktadır. Çalışma sırasında sistem RAM'inin 2.2 GB'ı, GPU RAM'inin ise 0.0 GB'ı kullanılmıştır. Bu platform, büyük veri kümeleri üzerinde etkili eğitim ve test süreçleri için uygun bir hesaplama gücü sağlamıştır.

A. Özellik Seçimi Sonuçları

Anemi teşhisi ve sınıflandırma işlemlerinde kullanılan veri setinde bulunan özelliklerin fazlalığı ve bazı özellikler arasındaki yüksek korelasyon nedeniyle model başarısının olumsuz etkilenme ihtimali göz önüne alınmıştır. Bu durumu engellemek ve yalnızca modelin performansına en çok katkı sağlayan özellikleri belirlemek amacıyla üç farklı özellik seçimi yöntemi uygulanmıştır: Korelasyon Matrisi, RFE ve Boruta algoritması.

- Korelasyon Matrisi ile seçilen özellikler: GENDER, WBC, NE#, LY#, MO#, EO#, BA#, RBC, HGB, MCV, MCH, MCHC, RDW, PLT, MPV, PCT, PDW, SD, TSD, SDTSD, FERRITIN, FOLATE, B12 (Toplam 23 özellik).
- RFE ile seçilen özellikler: GENDER, MO#, EO#, RBC, HGB, HCT, MCV, MCH, MCHC, FOLATE (Toplam 10 özellik).
- Boruta algoritması ile seçilen özellikler: GENDER, LY#, MO#, RBC, HGB, HCT, MCV, MCH, MCHC, RDW, PLT, PCT, PDW, SD, TSD, SDTSD, FERRITIN, FOLATE, B12 (Toplam 19 özellik).
- Üç yöntemin ortak olarak seçtiği özellikler: GENDER, MO#, RBC, HGB, MCV, MCH, MCHC, FOLATE (Toplam 8 özellik).
- Korelasyon Matrisi ve RFE'nin ortak olarak seçtiği özellikler: GENDER, MO#, EO#, RBC, HGB, MCV, MCH, MCHC, FOLATE (Toplam 9 özellik).
- Korelasyon Matrisi ve Boruta algoritmasının ortak olarak seçtiği özellikler: GENDER, LY#, MO#, RBC, HGB, MCV, MCH, MCHC, RDW, PLT, PCT, PDW, SD, TSD, SDTSD, FERRITIN, FOLATE, B12 (Toplam 18 özellik).
- RFE ve Boruta algoritmasının ortak olarak seçtiği özellikler: GENDER, MO#, RBC, HGB, HCT, MCV, MCH, MCHC, FOLATE (Toplam 9 özellik).

B. Makine Öğrenmesi Performans Sonuçları

Çalışmada model eğitiminde kullanılan parametreler, hem genel eğitim süreçlerinde tutarlılığı sağlamak hem de farklı algoritmalar için optimize edilmiş hiperparametrelerin belirlenmesi açısından dikkatlice seçilmiştir. Ortak parametreler, tüm modellerde kullanılan eğitim ayarlarını temsil ederken, spesifik hiperparametreler, her bir algoritmanın performansını maksimize etmek için özel olarak ayarlanmıştır. Tablo 1, tüm modeller için kullanılan ortak eğitim parametrelerini sunarken, Tablo 2'de her bir modelin başarı oranını artırmak amacıyla kullanılan spesifik hiperparametreler detaylı olarak verilmiştir. Bu düzenleme, algoritmaların hem karşılaştırılabilir bir temel üzerinde değerlendirilmesini hem de her bir modelin kendi içinde optimize edilmesini sağlamaktadır.

Tablo 1. Çalışmada Kullanılan Ortak Parametreler Tablosu

Parametre	Değer
Mini topluluk boyutu	32
Maksimum dönem	10
Başlangıç öğrenme oranı	1e-4

Tablo 2. Çalışmada Kullanılan Spesifik Parametreler Tablosu

Model	Parametre	Değer
RF	n_estimators	100
	criterion	gini
	min_samples_split	2
	min_samples_leaf	1
	max_features	sqrt
	bootstrap	True
	random_state	42
GBM	learning_rate	0.1
	n_estimators	100
	max_depth	3
	min_samples_split	2
	min_samples_leaf	1
	subsample	1.0
	random_state	42
XGBoost	learning_rate	0.3
	n_estimators	100
	max_depth	6
	gamma	0
	min_child_weight	1
	subsample	1.0
	colsample_bytree	1.0
	eval_metric	logloss
	use_label_encoder	False
Karar Ağacı	criterion	gini
	splitter	best
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
LG	solver	lbfgs
	max_iter	1000
	penalty	l2
	C	1.0
	fit_intercept	True

Deneysel çalışmada kullanılan modeller RF, GBM, XGBoost, Karar Ağacı ve LG'dur. Modellerin eğitimi, %80 eğitim ve %20 test oranında bir veri bölünmesi ile gerçekleştirilmiştir. Çalışmada, modellerin performanslarını değerlendirmek amacıyla karmaşıklık matrisinden türetilen doğruluk, özgüllük, duyarlılık ve F1 Skoru gibi çeşitli performans ölçütleri kullanılmıştır. Karmaşıklık matrisi, Gerçek Pozitif (GP), Yanlış Pozitif (YP), Gerçek Negatif (GN) ve Yanlış Negatif (YN) olmak üzere dört bileşenden oluşmaktadır. Performans metriklerinin matematiksel ifadeleri sırasıyla Denklem (1)-(4) olarak verilmiştir.

$$Doğruluk = \frac{(GP + GN)}{(GP + YP + GN + YN)} \quad (1)$$

$$Özgüllük = \frac{(GN)}{(GN + GP)} \quad (2)$$

$$Duyarlılık = \frac{(GP)}{(GP + YN)} \quad (3)$$

$$F - skor = 2 \times \frac{(GP)}{(2 \times GP + YP + YN)} \quad (4)$$

Tablo 3, Tablo 4, Tablo 5, Tablo 6 ve Tablo 7 sırasıyla RF, GBM, XGBoost ,Karar Ağacı ve LG modellerinin özellik seçimi yöntemleriyle performans sonuçlarını içermektedir. Tablo 8'de ise bütün modellerin özellik seçimi yöntemleriyle olan performans sonuçları doğruluk performans ölçütüne göre sıralanmıştır.

Tablo 3. RF Modelinin Özellik Seçimi Yöntemleriyle Performans Sonuçları

Özellik Seçimi	Doğruluk	Hassasiyet	Özgüllük	F-Skor
Normal	0,991176	0,991176	1	0,99076
RFE	0,91732	0,904403	1	0,905645
Korelasyon	0,992157	0,992202	1	0,991774
Boruta	0,994444	0,994645	1	0,994275

Tablo 4. GBM Modelinin Özellik Seçimi Yöntemleriyle Performans Sonuçları

Özellik Seçimi	Doğruluk	Hassasiyet	Özgüllük	F-Skor
Normal	0,999673	0,99968	1	0,999675
RFE	0,914052	0,897637	1	0,901596
Korelasyon	0,999673	0,99968	1	0,999675
Boruta	0,999673	0,99968	1	0,999675

Tablo 5. XGBoost Modelinin Özellik Seçimi Yöntemleriyle Performans Sonuçları

Özellik Seçimi	Doğruluk	Hassasiyet	Özgüllük	F-Skor
Normal	0,99902	0,999059	1	0,999023
RFE	0,915686	0,90324	1	0,906841
Korelasyon	0,99902	0,999059	1	0,999023
Boruta	0,99902	0,999059	1	0,999023

Tablo 6. Karar Ağacı Modelinin Özellik Seçimi Yöntemleriyle Performans Sonuçları

Özellik Seçimi	Doğruluk	Hassasiyet	Özgüllük	F-Skor
Normal	0,99902	0,999023	1	0,999014
RFE	0,888889	0,891074	1	0,889933
Korelasyon	0,99902	0,999023	1	0,999014
Boruta	0,99902	0,999023	1	0,999014

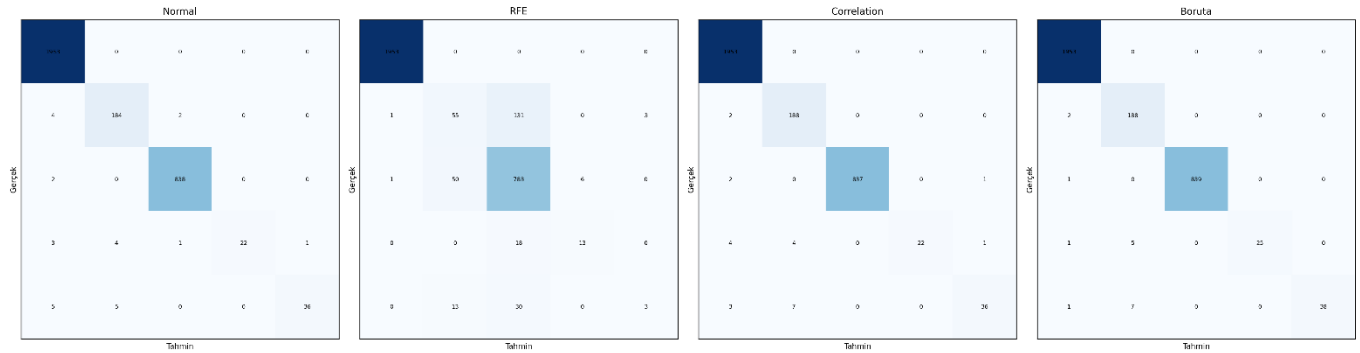
Tablo 7. LG Modelinin Özellik Seçimi Yöntemleriyle Performans Sonuçları

Özellik Seçimi	Doğruluk	Hassasiyet	Özgüllük	F-Skor
Normal	0,908497	0,90338	0,989496	0,90473
RFE	0,912745	0,88777	1	0,892499
Korelasyon	0,902288	0,894975	0,992134	0,89597
Boruta	0,910458	0,905074	0,992643	0,906435

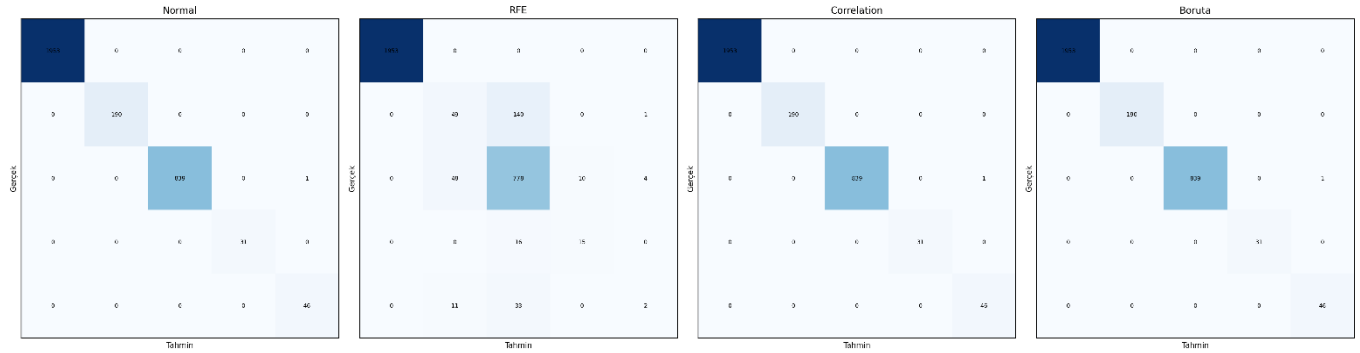
Tablo 8. Tüm Modellerin Özellik Seçimi Yöntemleriyle Performans Sonuçları

Sıra	Model	Özellik Seçimi	Doğruluk	Hassasiyet	Özgüllük	F-Skor
1	GBM	Normal	0,999673	0,99968	1	0,999675
1	GBM	Korelasyon	0,999673	0,99968	1	0,999675
1	GBM	Boruta	0,999673	0,99968	1	0,999675
2	XGBoost	Korelasyon	0,99902	0,999059	1	0,999023
2	XGBoost	Boruta	0,99902	0,999059	1	0,999023
2	Karar Ağacı	Boruta	0,99902	0,999023	1	0,999014
2	XGBoost	Normal	0,99902	0,999059	1	0,999023
2	Karar Ağacı	Korelasyon	0,99902	0,999023	1	0,999014
2	Karar Ağacı	Normal	0,99902	0,999023	1	0,999014
3	RF	Boruta	0,994444	0,994645	1	0,994275
4	RF	Korelasyon	0,992157	0,992202	1	0,991774
5	RF	Normal	0,991176	0,991176	1	0,99076
6	RF	RFE	0,91732	0,904403	1	0,905645
7	XGBoost	RFE	0,915686	0,90324	1	0,906841
8	GBM	RFE	0,914052	0,897637	1	0,901596
9	LG	RFE	0,912745	0,88777	1	0,892499
10	LG	Boruta	0,910458	0,905074	0,992643	0,906435
11	LG	Normal	0,908497	0,90338	0,989496	0,90473
12	LG	Korelasyon	0,902288	0,894975	0,992134	0,89597
13	Karar Ağacı	RFE	0,888889	0,891074	1	0,889933

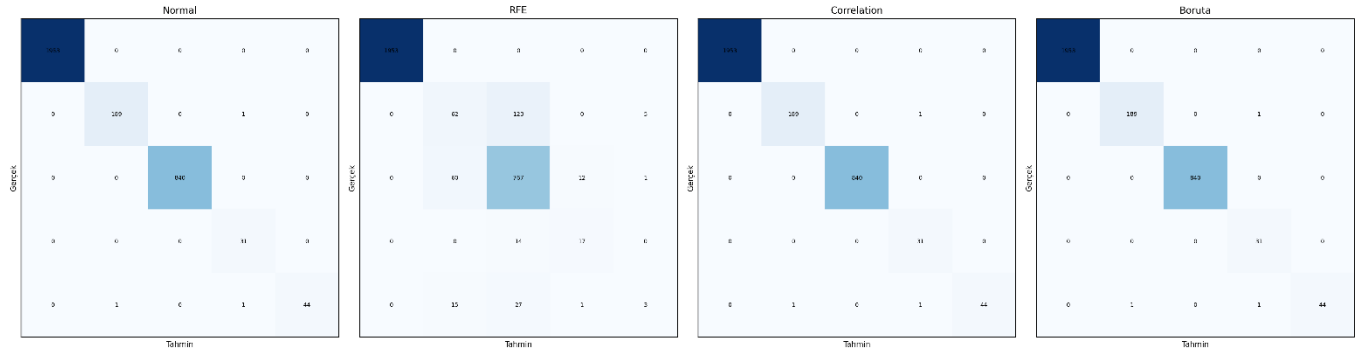
RF, GBM, XGBoost, Karar Ağacı ve LG modellerinin; normal, Korelasyon Matrisi, RFE, ve Baruta özellik seçimine göre elde edilen karışıklık matrisleri sırasıyla Şekil 1, Şekil 2, Şekil 3, Şekil 4 ve Şekil 5'te sunulmuştur.



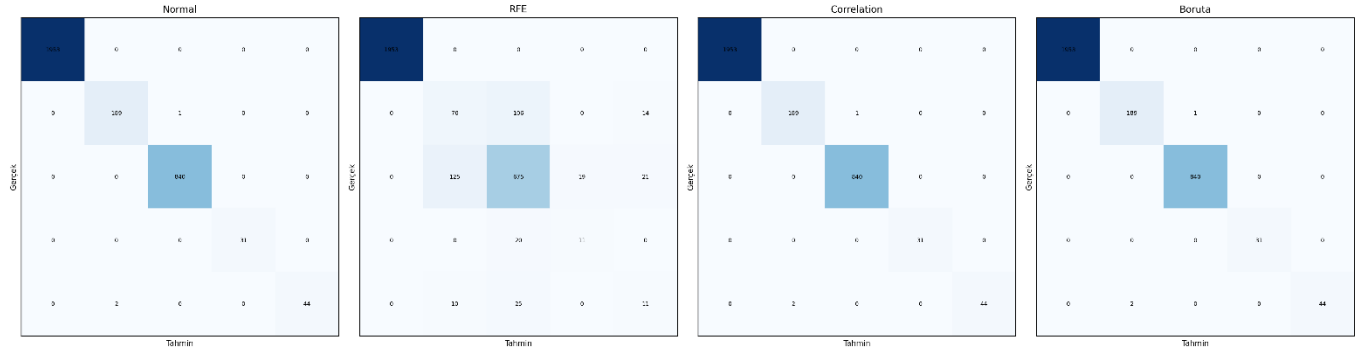
Şekil 1. RF Modelinin Normal,RFE,Korelasyon ve Baruta ile Özellik Seçimi Yapılmış Karışıklık Matrisleri



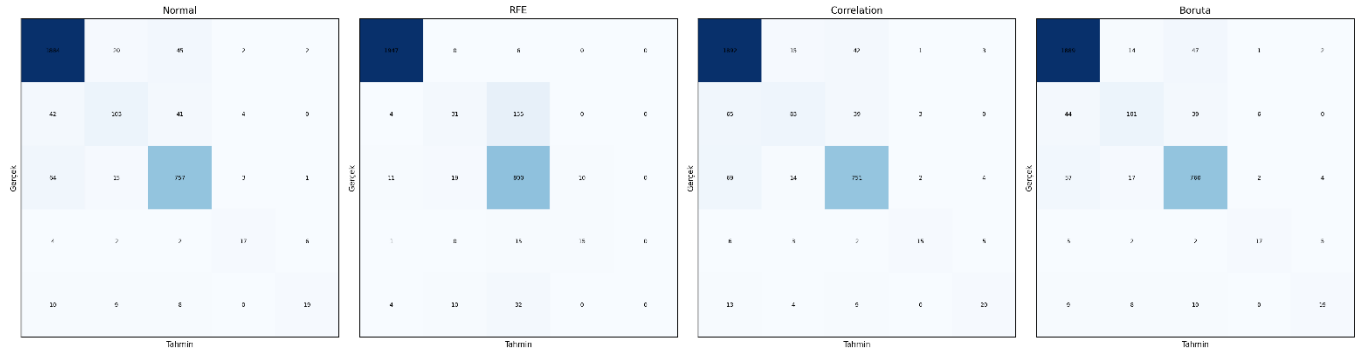
Şekil 2. GBM Modelinin Normal, RFE, Korelasyon ve Baruta ile Özellik Seçimi Yapılmış Karışıklık Matrisleri



Şekil 3. XGBoost Modelinin Normal, RFE, Korelasyon ve Baruta ile Özellik Seçimi Yapılmış Karışıklık Matrisleri



Şekil 4. Karar Ağacı Modelinin Normal, RFE, Korelasyon ve Baruta ile Özellik Seçimi Yapılmış Karışıklık Matrisleri



Şekil 5. LG Modelinin Normal, RFE, Korelasyon ve Baruta ile Özellik Seçimi Yapılmış Karışıklık Matrisleri

IV. TARTIŞMA

GBM modeli, doğruluk metriği açısından tüm modeller arasında en iyi performansı sergilemiştir. GBM modeli, normal veri setinde, Korelasyon Matrisi ve Boruta algoritmalarıyla seçilen özelliklerle %99,97 doğruluk oranına ulaşmıştır. Bu tutarlı performans, GBM'in veri setindeki karmaşık ilişkileri ve özellik etkileşimlerini etkili bir şekilde öğrenme kapasitesine sahip olduğunu göstermektedir. Karışıklık Matrisi (KM) sonuçlarına göre GBM modeli, yanlış pozitif ve yanlış negatif tahminleri minimumda tutarak dengeli bir performans göstermiştir.

XGBoost modeli, Korelasyon Matrisi ve Boruta algoritmalarıyla %99,90 doğruluk oranına ulaşarak GBM'den sonra en iyi performansı sergilemiştir. Ayrıca, normal veri setinde de aynı doğruluk oranını korumuştur. KM analizine göre XGBoost, özellikle düşük sınıflandırma hatası ile dikkat çekmiş ve sınıflar arasında yüksek hassasiyet sağlamıştır. Bu durum, XGBoost modelinin yüksek genelleme kapasitesine sahip olduğunu ve seçilen özelliklere duyarlı bir şekilde çalıştığını göstermektedir.

Karar Ağacı modeli, Boruta algoritması ile %99,90 doğruluk oranına ulaşmış ve Korelasyon Matrisi ile normal veri setinde de benzer sonuçlar elde etmiştir. KM verileri, Karar Ağacı modelinin daha az karmaşıklığa sahip olmasına rağmen, sınıfları iyi ayırabildiğini ve genellikle yanlış pozitif oranlarının düşük olduğunu göstermektedir. Bu durum, Karar Ağacı modelinin seçilen özelliklerin performans üzerindeki etkisinden faydalandığını ve kritik özellikleri etkili bir şekilde kullanabildiğini ortaya koymaktadır.

RF modeli, Boruta algoritması ile optimize edildiğinde %99,44 doğruluk oranına ulaşmış, bu da Korelasyon Matrisi ile %99,22 ve normal veri setindeki %99,11 doğruluk oranlarından daha yüksektir. KMverileri, RF modelinin nispeten daha yüksek yanlış negatif oranına sahip olduğunu, ancak genel doğruluk açısından kabul edilebilir sonuçlar verdiğini ortaya koymaktadır. RF modeli, genel olarak yüksek performans sergilese de, GBM ve XGBoost kadar etkileyici sonuçlar elde edememiştir.

LG modelinin performansı diğer modellere göre daha sınırlı kalmış ve en yüksek doğruluk oranı RFE ile %91,27 olarak elde edilmiştir. LG modelinin daha düşük doğruluk oranları, özellikle veri setindeki doğrusal olmayan ilişkilerin sınırlı bir şekilde anlaşılmasından kaynaklanabilir. KM analizine göre LG modeli, özellikle düşük özgüllük değerine sahip sınıflarda daha fazla hata yapmıştır. Bu, LG modelinin karmaşık veri setlerinde sınıf ayrımında zorlandığını göstermektedir.

Özellik seçimi yöntemlerinin etkisi incelendiğinde, Boruta algoritmasının genellikle modellerin doğruluk oranlarını artırmada en etkili yöntem olduğu görülmüştür. Özellikle GBM, XGBoost, Karar Ağacı ve RF modellerinde Boruta algoritması ile optimize edilen veri setleri, yüksek doğruluk oranlarına ulaşmıştır. KM verileri de Boruta'nın sınıflandırma hatalarını önemli ölçüde azalttığını ve modellerin genel performansını artırdığını desteklemektedir.

Korelasyon Matrisi yöntemi de birçok model için benzer şekilde etkili olmuştur. GBM, XGBoost ve Karar Ağacı modellerinde Korelasyon Matrisi ile seçilen özellikler, model performansını en üst seviyede tutmayı başarmıştır. Bu durum, Korelasyon Matrisi ile belirlenen özelliklerin, hedef değişken ile güçlü ilişkiler içerdiğini göstermektedir. KM analizine göre, bu yöntemle optimize edilen modellerde yanlış pozitif ve yanlış negatif oranlarının genellikle daha düşük olduğu gözlemlenmiştir.

RFE yöntemi, daha az sayıda özellik seçmesi nedeniyle doğrusal modellerde (örneğin LG ve Karar Ağacı) daha düşük doğruluk oranları ile sonuçlanmıştır. Bununla birlikte, bu yöntemin model karmaşıklığını azaltarak hesaplama maliyetlerini düşürdüğü gözlemlenmiştir. KM analizine göre, RFE ile seçilen özellikler bazı modellerde sınıf ayrımını güçleştirmiş ve yanlış negatif oranlarını artırmıştır.

V. SONUÇLAR

Bu çalışma, anemi teşhisi ve sınıflandırmasında farklı makine öğrenmesi modellerinin ve özellik seçimi yöntemlerinin bir arada kullanımının, model performansını optimize etmek için etkili bir strateji olduğunu ortaya koymuştur. Çalışmanın sonuçları, GBM modelinin hem doğruluk oranları hem de diğer performans metrikleri açısından en iyi sonuçları sağladığını göstermiştir. GBM modeli, %99,97 doğruluk oranıyla anemi teşhisi için güçlü bir model olarak öne çıkmıştır.

Özellik seçimi yöntemleri arasında Boruta algoritması, özellikle GBM, XGBoost ve Karar Ağacı modelleriyle birleştirildiğinde, model performansını artırmada en etkili yöntem olarak belirlenmiştir. Boruta algoritmasının gereksiz özellikleri eleyerek yalnızca hedef değişkenle anlamlı ilişkiye sahip olan özellikleri seçmesi, modellerin doğruluğunu ve genel başarımını önemli ölçüde artırmıştır. Korelasyon Matrisi yöntemi de yüksek doğruluk oranları sağlamış ve özellikle GBM ile birlikte etkili sonuçlar vermiştir. Ancak, RFE yöntemi daha az sayıda özellik seçmesi nedeniyle bazı modellerde doğruluk oranlarının düşmesine neden olmuş, buna karşın model karmaşıklığını azaltarak hesaplama maliyetlerini düşürmüştür.

Çalışmanın sonuçları, özellikle büyük ve karmaşık veri setleri üzerinde çalışan sağlık profesyonelleri ve araştırmacılar için önemli çıkarımlar sunmaktadır. Özellik seçimi yöntemlerinin kullanımı, makine öğrenmesi modellerinin daha verimli ve etkili bir şekilde kullanılmasını sağlamakta, aynı zamanda modellerin yorumlanabilirliğini artırmaktadır. Gelecekteki çalışmalarda, daha geniş veri setleri ve derin öğrenme yöntemlerinin entegre edilmesiyle model performansının daha da artırılabileceği öngörülmektedir.

KAYNAKLAR

[1] J. Fitriany and A. I. Saputri, "Anemia Defisiensi Besi," *AVERROUS: Jurnal Kedokteran dan Kesehatan Malikussaleh*, vol. 4, no. 2, pp. 1–14, Nov. 2018.

[2] S. Dogan and I. Turkoglu, "Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees," *International Journal of Science & Technology*, vol. 3, pp. 85–92, Jan. 2008.

[3] M. Dugdale, "Anemia," *Obstetrics and Gynecology Clinics of North America*, vol. 28, no. 2, pp. 363–382, Jun. 2001.

[4] S. Kilicarslan, M. Celik, and S. Sahin, "Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification," *Biomedical Signal Processing and Control*, vol. 63, Jan. 2021.

[5] E. R. Eichner, "Observations on Iron, Anemia, and Sickle Cell Trait," *Current Sports Medicine Reports*, vol. 16, no. 1, p. 2, Feb. 2017.

[6] N. Yağmur, H. Temurtaş, and İ. Dağ, "Anemi Hastalığının Yapay Sinir Ağları Yöntemleri Kullanılarak Sınıflandırılması," *Journal of Scientific Reports-B*, no. 008, pp. 20–34, Dec. 2023.

[7] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, "Multi-Class Classification Algorithms for the Diagnosis of Anemia in an Outpatient Clinical Setting," *PLOS ONE*, vol. 17, no. 7, Jul. 2022.

[8] P. Verma and V. Chopra, "A Review on Machine Learning Algorithms for Anemia Disease Prediction," *International Research Journal*, vol. 4, no. 5, 2022.

[9] J. R. Khan, S. Chowdhury, H. Islam, and E. Raheem, "Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh," *Journal of Data Science: JDS*, vol. 17, pp. 195–218, Jan. 2019.

[10] M. Hasan, M. S. Tahosin, A. Farjana, M. A. Sheakh, and M. M. Hasan, "A Harmful Disorder: Predictive and Comparative Analysis for Fetal Anemia Disease by Using Different Machine Learning Approaches," in *Proc. 2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, 2023, pp. 1–6.

[11] B. Sen, A. Ganesh, A. Bhan, S. Dixit, and A. Goyal, "Machine Learning Based Diagnosis and Classification of Sickle Cell Anemia in Human RBC," in *Proc. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021, pp. 753–758.

[12] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, "A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia," *Healthcare*, vol. 11, no. 5, Jan. 2023.

[13] Y. Zhang, Y. Zhou, D. Zhang, and W. Song, "A Stroke Risk Detection: Improving Hybrid Feature Selection Method," *Journal of Medical Internet Research*, vol. 21, no. 4, Apr. 2019.

[14] M. S. Singh and P. Choudhary, "Stroke Prediction Using Artificial Intelligence," in *Proc. 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, 2017, pp. 158–161.

[15] "Risk Detection of Stroke Using a Feature Selection and Classification Method," *IEEE Journals & Magazine*.

[16] S. Buyrukoğlu and A. Akbaş, "Machine Learning Based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach Using Correlation Matrix with Heatmap and SFS," *Balkan Journal of Electrical and Computer Engineering*, vol. 10, no. 2, pp. 110–117, Apr. 2022.

[17] K. Yan and D. Zhang, "Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, Jun. 2015.

[18] M. Hasan, P. Roy, and A. M. Nitu, "Cervical Cancer Classification Using Machine Learning with Feature Importance and Model Explainability," in *Proc. 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, 2022, pp. 1–4.

[19] A. Karuppasamy, A. Abdesselam, R. Hedjam, H. Zidoum, and M. Al-Bahri, "Feed-Forward Networks Using Logistic Regression and Support Vector Machine for Whole-Slide Breast Cancer Histopathology Image Classification," *Intelligence-Based Medicine*, vol. 9, Jan. 2024.

[20] H. Ren, T. Yang, X. Yin, L. Tong, J. Shi, J. Yang, Z. Zhu, and H. Li, "Prediction of High-Level Fear of Cancer Recurrence in Breast Cancer Survivors: An Integrative Approach Utilizing Random Forest Algorithm and Visual Nomogram," *European Journal of Oncology Nursing*, vol. 70, Jun. 2024.

[21] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, Sep. 2021.

[22] A. Ogunleye and Q.-G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020.
[23] A. K. Jakhar, A. Gupta, and M. Singh, "SELF: A Stacked-Based Ensemble Learning Framework for Breast Cancer Classification," *Evolutionary Intelligence*, vol. 17, no. 3, pp. 1341–1356, Jun. 2024.
[24] Google Colab website. [Online]. Available: https://colab.research.google.com/