AÇIKLANABİLİR YAPAY ZEKÂ: YÖNTEMLER, UYGULAMALAR VE ETİK KONULAR

Osman Sefa Coşar
Gazi Üniversitesi Yapay Zeka Uygulama ve Araştırma Merkezi
osmansefac18@gmail.com
DOI: 10.5281/zenodo.16890925

1. GİRİŞ

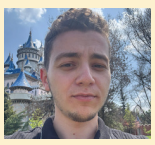
Yapay zekâ (YZ) sistemlerinin karmaşıklığı arttıkça bu sistemlerin karar alma süreçlerini şeffaf ve anlaşılır hâle getirme ihtiyacı da artmıştır. Derin öğrenme gibi güçlü ancak "kara kutu" niteliğindeki modellerin kullanımı, özellikle zaman serisi verileri gibi karmaşık yapılarda yorumlanması zor kararlarla sonuçlanmaktadır. Bu nedenle kararların neden alındığını açıklamak ve kullanıcı güvenini sağlamak amacıyla açıklanabilir yapay zekâ (explainable artificial intelligence-XAI) yöntemlerine olan ilgi giderek artmaktadır (Schlegel vd., 2019).

Açıklanabilirlik, bir yapay zekâ modelinin aldığı kararların mantıksal temellerinin anlaşılabilir olmasını sağlarken; şeffaflık, modelin iç işleyişine dair bilgi sunabilmeyi ifade eder. Hesap verebilirlik ise, bu kararların sonuçlarından sorumlu tutulabilecek mekanizmaların oluşturulmasını gerektirir. Bu üç unsur bir araya geldiğinde, yalnızca teknik başarı değil aynı zamanda etik sorumluluk da sağlanmış olur.

Avrupa Birliği'nin Genel Veri Koruma Tüzüğü (General Data Protection Regulation-GDPR) gibi düzenlemeler, algoritmik kararların açıklanabilir olmasını yasal bir zorunluluk hâline getirmiştir. Ayrıca, adalet, mahremiyet ve güvenilirlik gibi kriterler de yeni nesil yapay zekâ modellerinin seçiminde belirleyici olmaya başlamıştır. Bu kapsamda, XAI yöntemleri ya doğrudan yorumlanabilir modellerin tercih edilmesiyle ya da karmaşık modellerin üzerine eklenen açıklama katmanlarıyla uygulamaya konmaktadır.

XAI yöntemleri görselleştirme, özellik önemi (feature importance) ve etki analizleri gibi tekniklerle modellerin karar mantığını ortaya koyar. Bu açıklamalar kullanıcıya güven vermek, modelin hatalarını tespit etmek ve karar süreçlerini doğrulamak açısından büyük önem taşımaktadır. Ayrıca bazı yöntemler, zaman serisi gibi geleneksel görüntü ve metin dışı verilerde de uygulanabilir hâle getirilmiştir. Bu bağlamda LIME, SHAP ve Saliency Maps gibi açıklayıcı tekniklerin farklı veri türlerinde kullanımı ve etkililiği çeşitli çalışmalarda incelenmiştir (Schlegel vd., 2019).

Uygulamalı XAI sistemlerinin geliştirilmesi, yalnızca model açıklanabilirliğini artırmakla kalmamakta, aynı zamanda güvenlik ve karar destek sistemlerinin daha etkili çalışmasını da sağlamaktadır. Örneğin, yol güvenliği alanında geliştirilen bir XAI tabanlı çukur tespit sistemi sayesinde, modelin hangi görüntü özelliklerine göre karar verdiği açıkça gözlemlenebilmekte ve bu sayede kararların doğruluğu kullanıcı tarafından değerlendirilebilmektedir (Senbagavalli vd., 2025). Benzer şekilde, yangın sonrası yanık alanların tespiti için geliştirilen başka bir sistemde XAI kullanımı, uzaktan algılama verilerinin analizinde daha doğru ve anlaşılır sonuçlar sunmaktadır (DASA, 2022).

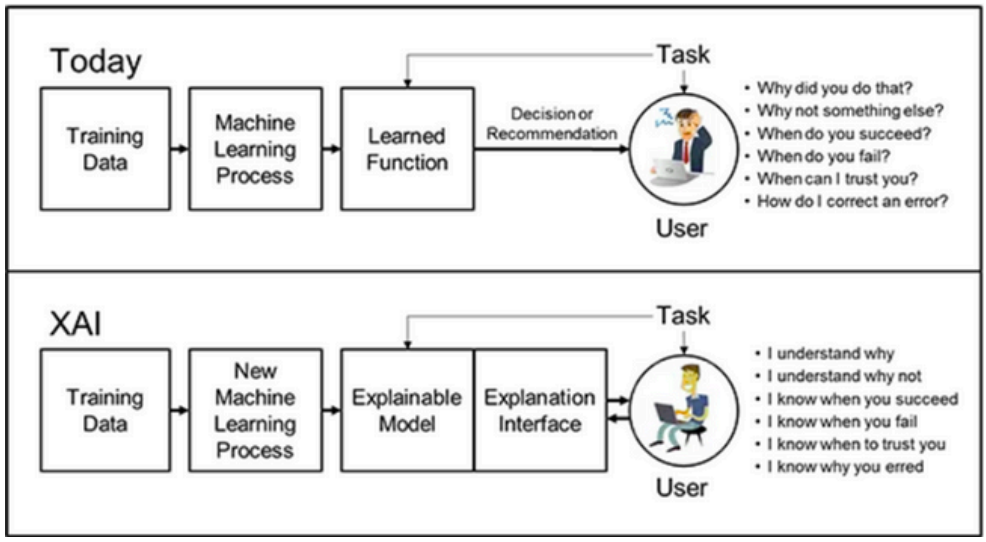


Osman Sefa Coşar
Gazi Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü 3. sınıf öğrencisiyim. Gazi Üniversitesi Yapay Zeka Uygulama ve Araştırma Merkezi'nde kısmi zamanlı olarak çalışmaktayım. Bilgisayarlı görü, derin öğrenme, görüntü işleme ve veri bilimi alanları ile ilgilenmekteyim.

2. XAI NEDİR?

Yapay zekâ sistemlerinin özellikle sağlık alanında teşhis, risk sınıflandırması ve hasta yönlendirme gibi kritik süreçlerde yaygınlaşması, bu teknolojilerin güvenilirliğini ve açıklanabilirliğini ön plana çıkarmaktadır. Derin öğrenme temelli modeller yüksek doğruluk oranlarına ulaşsalar da, karar alma süreçlerinin karmaşıklığı nedeniyle genellikle "kara kutu" olarak değerlendirilmekte ve bu durum kullanıcıların güvenini ciddi biçimde zedelemektedir.

Arnaud vd. (2024) tarafından vurgulandığı üzere, özellikle acil servislerde kullanılan 3P-U sistemi gibi klinik karar destek modellerinin pratikte kabul görmemesinin temel nedenlerinden biri, bu sistemlerin neden ve nasıl karar verdiklerine dair yeterince anlaşılır açıklamalar sunamamasıdır. Bu bağlamda XAI, sadece teknik bir kolaylık değil, aynı zamanda kullanıcı güvenini tesis etmek, sistemin hatalı kararlarını tespit edebilmek ve etik-hukuki düzenlemelere uyum sağlamak açısından da kritik bir ihtiyaç hâline gelmektedir. Şekil 1'de geleneksel makine öğrenimi ile açıklanabilir yapay zekâ arasındaki temel ayrımlar açıkça ortaya koyulmaktadır



Şekil 1. Geleneksel Makine Öğrenimi ile Açıklanabilir Yapay Zekanın Karşılaştırılması

Avrupa Birliği'nin GDPR kapsamında bireylerin otomatik karar alma süreçlerine dair açık ve anlaşılır açıklama talep etme hakkı, XAI sistemlerinin geliştirilmesini yasal bir yükümlülük haline getirmiştir. Arnaud ve arkadaşları, bu ihtiyaca yanıt olarak "tutarlılık" (consistency) metriğini önermiş ve XAI açıklamalarının klinik uzmanların bilgi temelli değerlendirmeleriyle ne ölçüde örtüştüğünü istatistiksel testler yoluyla analiz etmişlerdir. Bu değerlendirme sadece modelin teknik doğruluğunu değil, aynı zamanda uzman görüşleriyle ne kadar uyumlu açıklamalar sunduğunu da ölçerek XAI sistemlerinin gerçek dünyadaki kullanılabilirliğini artırmayı hedeflemiştir. Bu sayede XAI, yalnızca bir modelin karar mantığını anlamayı kolaylaştırmakla kalmaz; aynı zamanda bu kararların klinik bağlamda makul, güvenilir ve insan uzmanlığı ile uyumlu olup olmadığını da değerlendirme imkânı sunar. Dolayısıyla XAI'nin gerekliliği yalnızca modelin iç işleyişini açıklamaktan ibaret değil; aynı zamanda insan-makine iş birliğini güçlendirmek, hataları azaltmak, karar süreçlerine güven tesis etmek ve kullanıcıların sistemle etkileşimini kolaylaştırmak gibi çok katmanlı bir işlevsellik sunmaktadır (Arnaud vd., 2024).

3. XAI YAKLAŞIMLARI

XAI alanında geliştirilen yöntemler, genel olarak modele bağımlı (model-specific) ve modele bağımsız (model-agnostic) olmak üzere ikiye ayrılır. Bu ayrım, kullanılan yöntemin bir makine öğrenmesi modelinin iç yapısına ne ölçüde erişim gerektirdiğine göre yapılır.

3.1. Modele Bağımlı ve Modele Bağımlı Olmayan Yöntemler

Modele bağımlı açıklama teknikleri, modelin yapısına özel olarak geliştirilmiştir ve genellikle içsel ağırlıklar, katmanlar ya da aktivasyon haritaları gibi öğelere doğrudan erişimi gerektirir. Örneğin, gradient-weighted class activation mapping (Grad-CAM) ve Guided Backpropagation gibi yöntemler, özellikle konvolüsyonel sinir ağları (convolutional neural networks-CNN) gibi belirli derin öğrenme modellerine uygulanabilir ve bu modellerin karar alma süreçlerini görselleştirerek anlamlandırmayı hedefler. Bu açıklamalar sınıfla ilişkili bölgeleri vurgulayarak kararın nasıl verildiğini görsel olarak sunar (Devireddy, 2025).

Diğer yandan modele bağımsız yöntemler, modelin iç yapısına doğrudan erişmeden sadece giriş ve çıkış bilgileri üzerinden çalışır. Bu yaklaşım açıklamaların daha geniş bir model ailesine uygulanabilir olmasını sağlar. Local Interpretable Model Agnostic Explanation (LIME) ve SHapley Additive exPlanations (SHAP) gibi yöntemler, özellikle farklı model türleriyle birlikte çalışabilmeleri ve kullanıcıya sezgisel açıklamalar sunmaları açısından yaygın olarak tercih edilmektedir. LIME, belirli bir girdiye karşılık verilen tahmini açıklamak için girişin küçük varyasyonlarını üretir ve bu varyasyonlar üzerinden basit bir model öğrenerek açıklama sunar. SHAP ise shapley değerleri kullanarak her özelliğin modele katkısını nicel olarak ifade eder. Her iki yöntem de modele bağımlı yöntemlerin aksine daha fazla işlem gücü gerektirse de, esneklik açısından önemli avantajlar sunar (Devireddy, 2025).

İlgili çalışmada yapılan karşılaştırmalı analizde, modele bağımlı yöntemlerin mimariye özgü detaylı açıklamalar sunduğu, ancak genellenebilirliklerinin sınırlı olduğu ifade edilmektedir. Öte yandan modele bağımsız yöntemlerin daha geniş çaplı uygulamalarda kullanılabileceği, ancak bazı durumlarda açıklamaların daha yüzeysel kalabileceği belirtilmektedir. İlgili çalışmada, bu yöntemlerin birlikte kullanılmasının daha tutarlı ve güvenilir açıklamalar üretebileceği de vurgulanmaktadır (Devireddy, 2025).

Son olarak, modele bağımlı ve modelden bağımsız açıklama yöntemlerinin avantaj ve sınırlılıklarını daha net kavrayabilmek adına iki önemli materyale yer verilmiştir. İlk olarak Çizelge 1 bu iki yöntem türünü uygulanabilirlik, esneklik, hesaplama maliyeti ve yorumlanabilirlik açısından karşılaştırmaktadır. Buna ek olarak, Şekil 2'de ise LIME, SHAP, Grad-CAM ve Guided Backpropagation gibi açıklama yöntemlerinin farklı hayvan türleri üzerindeki görsel çıktıları bir araya getirilmiştir. Her bir XAI yönteminin farklı veri türlerinde nasıl çalıştığını sezgisel olarak göstermesi açısından bu karşılaştırmalı görsel, açıklanabilirlik yöntemleri arasındaki farkları görselleştirerek sunma açısından oldukça önemlidir.

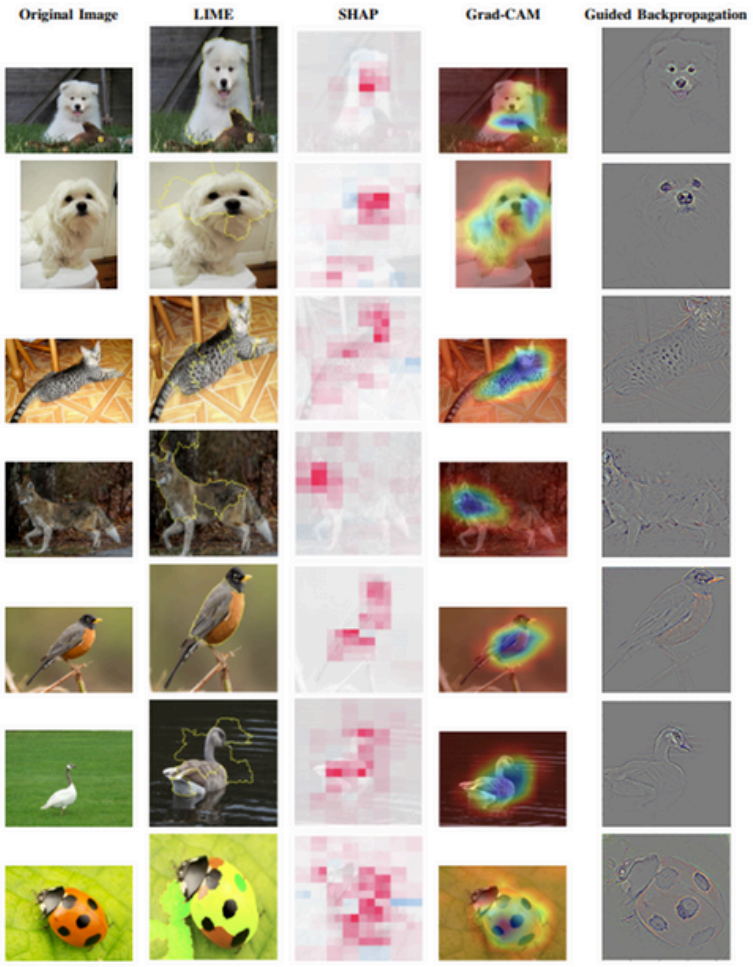
Çizelge 1. Modele bağımlı ve bağımsız açıklanabilirlik yöntemlerinin karşılaştırılması

Özellik	Model Bağımsız Yöntemler	Modele Bağımlı Yöntemler
Uygulanabilirlik	Herhangi bir ML modeline uygulanabilir (CNN, Karar Ağaçları vb.)	Sadece belirli modellere uygulanabilir (örneğin CNN'ler)
Esneklik	Yüksek	Düşük
Hesaplama Maliyeti	Post-hoc analiz nedeniyle daha yüksek	Modelin içine gömülü olduğundan daha düşük
Yorumlanabilirlik	Geniş kapsamlı açıklamalar sunar ancak daha az hassas olabilir.	Belirli mimarilere yönelik daha hassas açıklamalar üretir.

3.2. Post-hoc ve İçsel Açıklanabilirlik

XAI alanında geliştirilen yöntemler genel olarak iki temel yaklaşıma dayanmaktadır: açıklanabilirliğin modelin doğasına yerleştirildiği içsel (ante-hoc) yöntemler ve eğitilen modellerin kararlarını sonradan yorumlamayı amaçlayan post-hoc açıklama teknikleri. Bu iki yaklaşım, yapay zekâ modellerinin kullanıcılar tarafından anlaşılabilirliğini sağlama biçimlerine göre farklılaşır. İçsel açıklanabilirlik, modelin başından itibaren şeffaf ve anlaşılır olacak şekilde tasarlanmasını ifade eder. Karar ağaçları, doğrusal regresyonlar veya generalized additive models (GAMs) gibi yöntemler bu kapsama girer.

Bu modellerin yapısı, modelin nasıl çalıştığının kullanıcılar tarafından doğrudan izlenebilmesini mümkün kılar. Örneğin GAM'ler, değişkenler arasındaki doğrusal olmayan ilişkileri açık şekilde ifade edebildiği için özellikle sağlık gibi hem doğruluğun hem de yorumlanabilirliğin kritik olduğu alanlarda kullanılmaktadır (Retzlaff vd., 2024). Benzer şekilde, Bayesian Rule List (BRL) gibi yapılar, if-then kurallarıyla çalışan ve karar sürecini açık biçimde ortaya koyan modellere örnek olarak verilebilir. Bu sayede yüksek boyutlu özellik uzayı, insan tarafından yorumlanabilir kurallara indirgenerek karar mantığı daha anlaşılır hâle getirilebilmektedir.



Şekil 2. Farklı türler için XAI yöntemlerinin karşılaştırılması (her satır bir tür, her sütun farklı bir XAI yöntemidir.)

Buna karşılık post-hoc açıklama teknikleri, özellikle "kara kutu" olarak adlandırılan derin öğrenme ve topluluk modellerinin çıktılarının neden ve nasıl üretildiğini açıklamaya çalışır. Bu yöntemler modelin iç yapısını değiştirmez, ancak modelin çıktısına dair çeşitli analizler sunarak kullanıcıların karar mekanizmasını daha iyi anlamasını hedefler. SHAP, LIME, Anchors ve karşı olgusal (counterfactual) açıklamalar bu kapsama girer (Retzlaff vd., 2024; Madathil vd., 2024). Özellikle SHAP yöntemi, her bir özelliğin tahmine ne kadar katkı sağladığını oyun teorisi temelli olarak hesaplayarak, kullanıcıya güven veren detaylı bir açıklama sunar. SHAP'ın sezgisel açıklamalar sağlaması, onu tıp gibi kritik kararların alındığı alanlarda önemli bir araç hâline getirmiştir (Retzlaff vd., 2024; Madathil vd., 2024). LIME ise modelin çevresinde oluşturulan yerel vekil modeller aracılığıyla, belirli bir tahminin neden yapıldığını açıklamaya çalışır. Bu yönüyle, özellikle görselleştirmeyle desteklendiğinde kullanıcılar tarafından kolayca anlaşılabilir ve etkileşimli bir yorumlama imkânı sunar (Bhati vd., 2024).

Tıbbi görüntüleme gibi alanlarda, post-hoc tekniklerin yaygın olarak kullanıldığı gözlemlenmektedir. Grad-CAM, Integrated Gradients, Occlusion ve LIME gibi yöntemlerle, derin öğrenme modellerinin hangi görüntü bölgelerine odaklandığı görselleştirilebilmekte ve böylece karar mekanizması daha güvenilir hâle getirilebilmektedir. Özellikle Grad-CAM gibi teknikler, medikal görüntüler üzerinde ısı haritaları oluşturarak, modelin hangi bölgelere dayanarak karar verdiğini göstermekte etkili olmaktadır (Bhati vd., 2024). Bu durum, özellikle sağlık profesyonellerinin modeli daha kolay doğrulamasını ve yorumlamasını sağlamakta önemli bir rol oynamaktadır.

İçsel ve post-hoc yöntemler arasında belirgin bir fark, açıklanabilirliğin ne zaman ve nasıl sağlandığıyla ilgilidir. İçsel açıklamalar modelin karar sürecini doğrudan açığa çıkarırken, post-hoc açıklamalar yalnızca gerekçelendirme sunar. Bu bağlamda post-hoc teknikler kullanıcıya sonradan açıklama sunmakla birlikte, modelin karar alma sürecinin doğrudan şeffaflaştırılması açısından sınırlıdır (Madathil vd., 2024). Öte yandan, içsel açıklama sunan modellerin geliştirilmesi genellikle daha düşük karmaşıklık ve sınırlı doğrulukla sonuçlanabilir. Özellikle yüksek riskli alanlarda, modelin kararlarının hem doğru hem de kullanıcıya anlaşılır olması beklendiğinde bu iki yaklaşımın birlikte kullanımı daha yaygındır (Retzlaff vd., 2024; Bhati vd., 2024; Madathil vd., 2024).

Bu bağlamda, açıklama yöntemlerinin özelliklerini karşılaştırmalı olarak değerlendirmek, hangi yöntemin hangi bağlamda tercih edilmesi gerektiğini belirlemek açısından önemlidir. Çizelge 2'de, yaygın olarak kullanılan açıklama tekniklerinin destekledikleri veri türleri, açıklama düzeyleri (yerel/küresel), doğruluk (fidelity) ve kapsam gibi teknik kriterler bakımından karşılaştırmasını sunmakta; ayrıca her bir yöntemin özel zayıf yönlerine dair uyarılara yer vermektedir (Retzlaff vd., 2024).

Çizelge 2. Açıklanabilir yapay zeka yöntemlerinin karşılaştırılması

Yöntem	Desteklenen Veri Türü	Yerel/Küresel	Doğruluk (Fidelity)	Kapsam	Açıklama
Özellik Önemi	Metin, Tablo, Görüntü	Yerel ve Küresel	Yüksek	Yüksek	Saldırılara açık olabilir
Scoped Kurallar	Metin, Tablo	Yerel	Mükemmel	Düşük	Açıklamalar için ayarlama gerekir
Karşı-olgular	Metin, Tablo, Görüntü	Yerel	Mükemmel	Düşük	Aşırı genelleme ve gerçek dışı örnek üretme riski
Karar Ağaçları	Tablo (Sayısal, Kategorik)	N.A.	Mükemmel	Mükemmel	Aşırı öğrenmeye yatkın,Küçük varyasyonlara duyarlı
Bayesyen Kural Listeleri	Tablo (Kategorik)	N.A.	Mükemmel	Mükemmel	Hesaplama açısından maliyetli olabilir
Genelleştirilmiş Eklemlenmiş Modeller	Tablo (Sayısal, Kategorik)	N.A.	Mükemmel	Mükemmel	Gürültüye duyarlı, yüksek boyutlu uzaylarda temel fonksiyon seçimi zor

3.3. Görselleştirme Temelli Yöntemler

Derin öğrenme modellerinin karar verme süreçlerini daha anlaşılır kılmak amacıyla geliştirilen açıklanabilir yapay zekâ teknikleri arasında görselleştirme temelli yöntemler önemli bir yer tutar. Bu yöntemler, özellikle sağlık gibi yüksek riskli alanlarda, modellerin "neden bu kararı verdiği" sorusuna görsel destekle yanıt vererek güven oluşturmayı hedefler (Alicioglu & Sun, 2021).

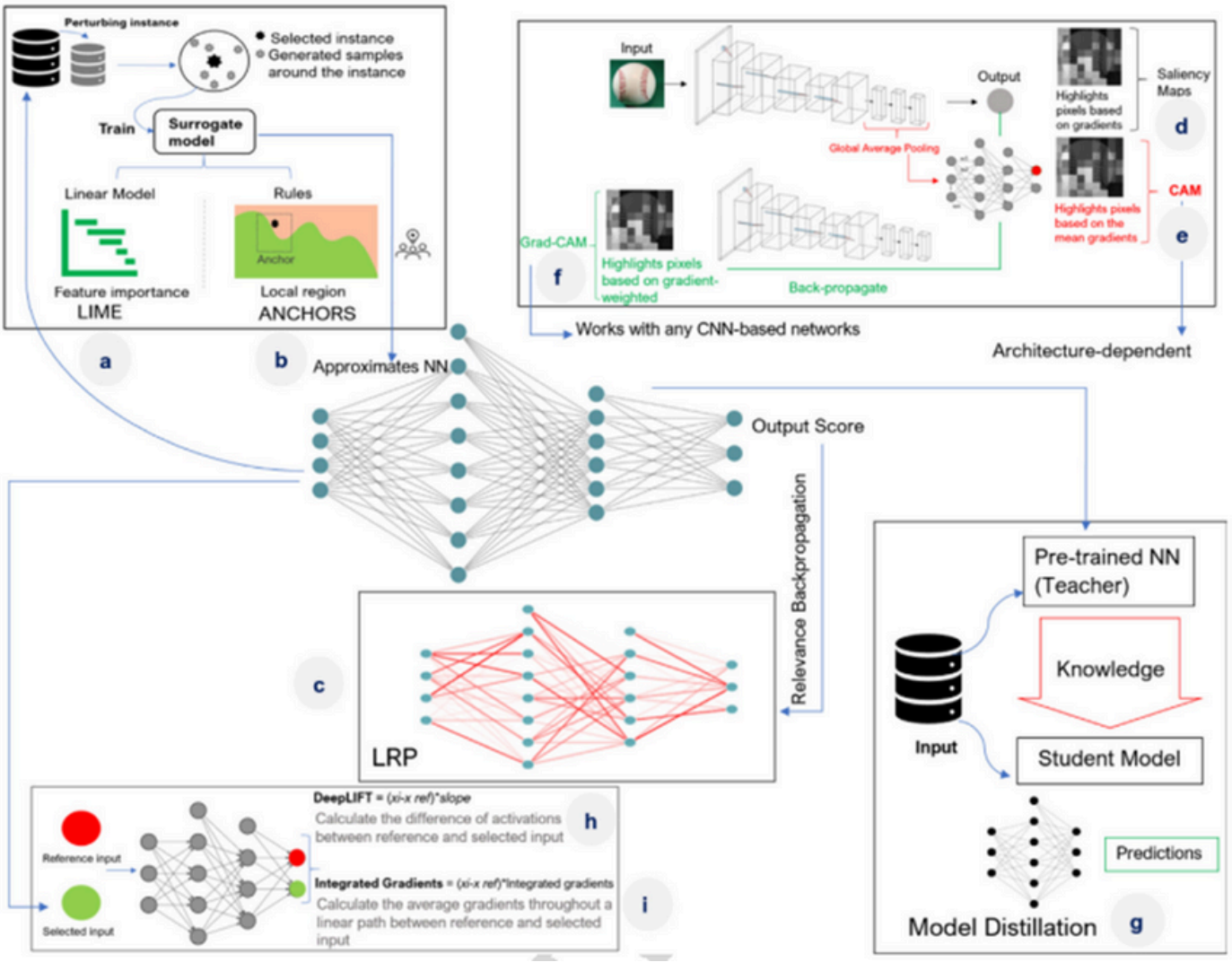
Bhati ve arkadaşlarının (2024) çalışmasında görselleştirme temelli XAI yöntemleri dört ana gruba ayrılmıştır: bozulmaya (perturbation) dayalı, gradyan tabanlı, ayrıştırma (decomposition) tabanlı ve dikkat (attention) mekanizmaları. Bu teknikler, derin öğrenme modellerinin iç işleyişini daha şeffaf hale getirerek, özellikle tıbbi görüntüleme alanında klinik karar destek sistemlerine katkı sunmaktadır. Bozulma tabanlı yöntemler, modelin giriş verisi üzerinde kasıtlı değişiklikler yaparak hangi bölgelerin karar üzerinde etkili olduğunu görselleştirir. Örneğin, LIME yöntemi, süperpiksel seviyesinde girişleri değiştirerek modelin tepkisini analiz eder (Bhati vd., 2024). Bu yöntem özellikle lokal açıklamalar üretmede etkilidir ve görüntülerin anlam bütünlüğünü koruyarak gerçek zamanlı uygulamalara uygundur (Alicioglu & Sun, 2021).

Gradyan tabanlı yöntemler ise geri yayılım (backpropagation) sırasında elde edilen gradyan bilgilerini kullanarak modelin hangi bölgelere duyarlı olduğunu görselleştirir. Grad-CAM, bu alandaki en yaygın tekniklerden biridir. Bhati ve arkadaşları (2024), Grad-CAM'in mimari bağımsızlığı sayesinde geniş bir model yelpazesinde kullanılabildiğini ve özellikle tıbbi görüntülerde insan uzmanlarla örtüşen bölgeleri başarılı şekilde vurguladığını belirtmiştir. Ayrıca saliency map ve guided backpropagation gibi yöntemler de bu gruba dahildir ve özellikle sınıf ayrımı yapan bölgeleri görsel olarak ortaya koyar (Alicioglu & Sun, 2021).

Ayrıştırma tabanlı yöntemler arasında en dikkat çekeni layer-wise relevance propagation (LRP) yöntemidir. Bu teknik, modelin kararını hangi nöronların ne ölçüde etkilediğini katman katman analiz ederek bir önem haritası üretir. Bhati ve arkadaşları (2024), LRP'nin Alzheimer ve Multipl Skleroz gibi hastalıkların teşhisinde belirleyici beyin bölgelerini öne çıkardığını ve klasik gradyan tabanlı yöntemlere kıyasla daha açıklayıcı sonuçlar verdiğini ifade etmiştir.

Dikkat mekanizmaları ise modelin karar sürecinde hangi bölgelere daha çok odaklandığını doğrudan öğrenir ve bu bilgiyi görsel olarak sunar. Özellikle trainable attention modelleri, sinir ağlarının farklı bölgelerinin önem derecesini öğrenerek hem performansı hem de açıklanabilirliği artırır (Bhati vd., 2024). Alicioglu ve Sun (2021), dikkat mekanizmalarının görselleştirme tabanlı XAI uygulamalarında güçlü ve sezgisel açıklamalar sunduğunu ve son yıllarda giderek yaygınlaştığını vurgulamıştır.

Tüm bu yöntemler, görselleştirmenin sağladığı sezgisel açıklamalar sayesinde hem uzman olmayan kullanıcıların modeli anlamasını kolaylaştırmakta hem de model geliştiricilerine hata ayıklama ve güvenilirlik artırımı açısından değerli bilgiler sunmaktadır (Alicioglu & Sun, 2021; Bhati vd., 2024). Bu yöntemlerin genel yapılarını ve birbirleriyle olan ilişkilerini bütüncül şekilde gösteren bir özet Şekil 3'te sunulmuştur.



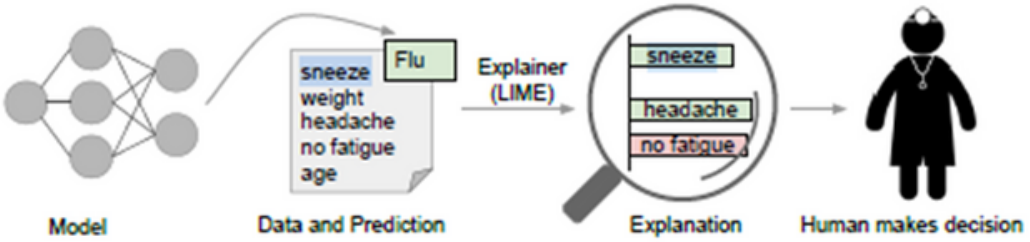
Şekil 3. Görselleştirme temelli XAI yöntemleri ve mimari bağlamda yerleşimleri

3.4. Yerel ve Global Açıklamalar

Açıklanabilir yapay zekâ alanında bir modelin karar verme sürecine dair açıklamalar sunarken, bu açıklamaların kapsamı önemli bir rol oynar. Açıklamalar genellikle iki temel yaklaşıma sahiptir; yerel (local) ve global (global) açıklamalar. Yerel açıklamalar, modelin tek bir örnek üzerindeki kararını anlamaya odaklanırken; global açıklamalar, modelin genel davranışı hakkında fikir vermeyi hedefler (Doshi-Velez & Kim, 2017).

Yerel açıklamalarda amaç, belirli bir girdinin çıktısına neden olan faktörleri belirlemektir. LIME bu amaçla geliştirilen ve sınıflandırıcının tekil bir tahminine yakın bölgede, insan tarafından anlaşılabilir bir model oluşturarak açıklama üretmeyi sağlayan yöntemlerden biridir (Ribeiro vd., 2016). Bu sayede kullanıcı, modelin belirli bir girdiye karşı neden o sonucu verdiğini daha rahat değerlendirebilir.

LIME'ın bu yerel açıklama yaklaşımı, özellikle kritik karar süreçlerinde kullanıcıya güven sağlar. Örneğin, bir hastanın grip olup olmadığına karar veren bir modelde, LIME çıktısı "hapşırma" ve "baş ağrısı" semptomlarının bu kararı desteklediğini, "yorgunluk olmaması" durumunun ise kararın aleyhine işlediğini açıklamada vurgular. Bu süreç, doktorun karar verme mekanizmasına müdahil olup modeli sorgulamasına imkân tanır. Bu tür bir senaryo Şekil 4'de görsel olarak sunulmaktadır.



Şekil 4. LIME algoritmasının yerel açıklama çıktısı için örnek

Burada, modelin yaptığı tahmin, LIME tarafından açıklanmış ve açıklama doktorun kararına rehberlik edecek şekilde sunulmuştur (Ribeiro vd., 2016). Bir hastaya ait semptomlar üzerinden "grip" tahmini yapılmakta; pozitif katkı sağlayan özellikler (hapşırma, baş ağrısı) ve negatif katkı sağlayan unsur (yorgunluk olmaması) vurgulanmaktadır

Öte yandan global açıklamalar, modelin tüm veri üzerindeki genel karar yapısını anlamaya yöneliktir. Burada amaç, modelin hangi özelliklere genel olarak nasıl ağırlık verdiğini ve hangi tür verilerde nasıl davrandığını görmektir. LIME makalesinde bu amaca yönelik olarak SP-LIME (Submodular Pick for LIME) adında bir yöntem önerilmiş ve modelin farklı örnekler üzerindeki kararlarını temsil eden açıklamaların seçilmesi yoluyla, modelin genel davranışının kullanıcıya gösterilmesi hedeflenmiştir (Ribeiro vd., 2016).

Doshi-Velez ve Kim (2017) ise yerel ve global açıklamaların hangi durumlarda tercih edilmesi gerektiğini uygulama bağlamına göre değerlendirmiştir. Örneğin, tıbbi bir sistemde doktorun bir hastaya özel hızlı bir karar vermesi gerekiyorsa yerel açıklamalar ön planda olmalıdır. Ancak sistemin güvenilirliğini genel olarak değerlendirmek isteyen bir sağlık yöneticisi için global açıklamalar daha faydalı olabilir (Doshi-Velez & Kim, 2017).

Sonuç olarak, yerel açıklamalar kullanıcıya bireysel kararlar hakkında sezgisel bir anlayış kazandırırken; global açıklamalar, modelin güvenilirliğini ve genel eğilimlerini değerlendirmede kritik rol oynar. Bu iki yaklaşım birlikte kullanıldığında, hem anlık hem bütünsel düzeyde modelin açıklanabilirliği sağlanabilir.

4. XAI TEKNİKLERİ VE ARAÇLARI

Tabular veri yapısına sahip veri kümeleri, özellikle finans, sağlık ve sosyal bilimlerde yoğunlukla kullanıldığından, bu veriler üzerinde çalışan yapay zeka modellerinin açıklanabilirliği oldukça kritiktir. Bu bağlamda, XAI teknikleri arasında LIME, SHAP, partial dependence plots (PDP) ve feature importance gibi yöntemler öne çıkmaktadır.

4.1. Tabular Veri: LIME, SHAP, PDP ve Feature Importance Yöntemleri

LIME, modelden bağımsız çalışmasıyla öne çıkar ve belirli bir tahminin yakın çevresindeki karar mantığını açıklamaya odaklanır. LIME, orijinal veri noktasına benzer yapay örnekler oluşturarak bu örneklerdeki tahmin sonuçlarını analiz eder ve bu sonuçlara göre yorumlanabilir basit bir model (genellikle lineer) kurar (Wang, 2024). Böylece karmaşık modellerin kararları, küçük bir bölgedeki davranışları üzerinden açıklanabilir hale gelir. Ancak bu yöntem sadece yerel açıklamalar sunar ve doğrusal olmayan ilişkilerde yetersiz kalabilir (Salih vd., 2024). Ayrıca LIME, özellikler arasındaki korelasyonları dikkate almadan çalıştığı için bazı senaryolarda yanlış yorumlamalara neden olabilir (Salih vd., 2024).

SHAP, oyun teorisine dayalı bir yöntem olup her bir özelliğin modele katkısını adil bir şekilde dağıtmayı amaçlar. SHAP, modelin tüm özellik kombinasyonlarını göz önünde bulundurarak her bir özelliğin tahmine ne kadar katkı sunduğunu hesaplar (Wang, 2024). SHAP'ın hem lokal hem de global açıklama sunabilmesi büyük avantaj sağlarken, özellikle karar ağaçları gibi modellerde daha stabil sonuçlar verdiği gösterilmiştir (Hall, 2020). Ancak SHAP'da bazı sınırlamalara sahiptir. Özellikle kollinear (birbirine çok bağlı) özellikler olduğunda, SHAP bu bağımsızlığı varsayarak gerçek katkıları olduğundan düşük hesaplayabilir (Salih vd., 2024). Bu durum, aynı anda önemli olan iki değişkenin etkisinin birinin üzerine yıkılmasıyla sonuçlanabilir.

Bu iki yaygın XAI yöntemi, özellikleri açısından farklılıklar taşır. Çizelge 3, LIME ve SHAP yöntemlerinin temel yönlerden nasıl ayrıştığını açık biçimde ortaya koymaktadır (Salih vd., 2024).

PDP yöntemi ise, belirli bir veya iki özelliğin değerlerinin sistematik olarak değiştirilerek modelin çıktısındaki değişimin gözlemlenmesi esasına dayanır. Bu grafikler, ilgili özelliğin hedef değişkene olan ortalama etkisini görselleştirir (Wang, 2024). PDP'ler özellikle global düzeyde model davranışlarını incelemede etkilidir. Ancak bu yöntemin en büyük zayıflığı, değişkenler arası korelasyonu göz ardı etmesidir. PDP bir özelliğin etkisini izole etmeye çalışırken, diğer bağlı değişkenlerin etkilerini de dolaylı olarak hesaba katabilir ve bu durum yanıltıcı sonuçlara yol açabilir (Hall, 2020).

Feature importance ise modelin çıktısını en çok hangi değişkenlerin etkilediğini göstermek amacıyla kullanılır. Karar ağaçlarında kullanılan "gini importance" veya "gain" metrikleri buna örnek verilebilir. Ancak bu klasik yöntemler model içi bilgiye dayalıdır ve özellikle korelasyonlu verilerde bazı özelliklerin önemini olduğundan yüksek gösterebilir (Hussain vd., 2022). Modern uygulamalarda SHAP ve LIME gibi yöntemlerle desteklenerek daha iyi yorumlar yapılması önerilmektedir.

Sonuç olarak, tablo verilerinde açıklanabilirliği sağlamak için kullanılan LIME, SHAP, PDP ve feature importance gibi yöntemlerin her birinin güçlü yönleri olduğu kadar bazı sınırlamaları da bulunmaktadır. Bu nedenle modelin karmaşıklığı, veri kümesindeki özellikler arası ilişkiler ve açıklama ihtiyacının lokal ya da global oluşu dikkate alınarak uygun yöntemin seçilmesi büyük önem taşımaktadır (Wang, 2024; Salih vd., 2024; Hall, 2020; Hussain vd., 2022).

Çizelge 3. LIME ve SHAP yöntemlerinin karşılaştırılması

Kriter	SHAP	LIME
Temel Kavram	Mevcut modelin çıktısını doğrudan açıklar	Karmaşık modeli açıklamak için yerel bir basit model kurar
Teorik Dayanak	Oyun teorisine dayalı katkı paylaşımı	Özellik bozulumu (perturbation) temelli yöntem
Yöntem Türü	Modelden bağımsız, sonradan açıklama (post-hoc)	Modelden bağımsız, sonradan açıklama
Veri Türü	Görüntü, tablo ve sinyaller	Tablo verisi
Açıklama Kapsamı	Hem yerel hem global açıklamalar	Yalnızca yerel açıklama
Korelasyon Duyarlılığı	Orijinal yöntemde korelasyon dikkate alınmaz	Korelasyon dikkate alınmaz
Doğrusal Olmayan İlişkiler	Kullanılan modele bağlı olarak açıklanabilir	Doğrusal olmayan yapıları yakalayamaz
Hesaplama Süresi	Görece yüksek	Daha düşük
Görselleştirme	Waterfall, Beeswarm ve Özet grafikleri	Sadece tek grafik

4.2. Görüntü ve Video

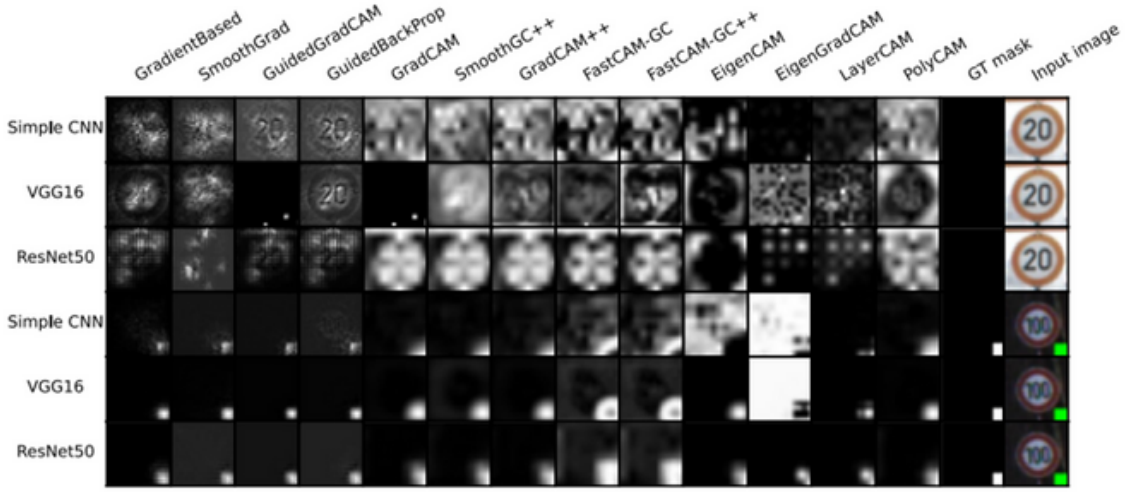
Görüntü ve video verileri üzerinde çalışan derin sinir ağlarının kararlarını anlamlandırmak, bu sistemlere duyulan güvenin artırılması açısından kritik bir gerekliliktir. Görsel açıklanabilirlik teknikleri bu karmaşık modellerin hangi bölgelere odaklandığını görselleştirerek kullanıcıya içgörü sağlar.

Bu alandaki temel yöntemlerden biri olan Grad-CAM, bir görüntüde modelin hangi alanlara odaklandığını gösteren sınıf-ayrıştırıcı görselleştirmeler üretir. Grad-CAM, hedeflenen sınıf için çıktı alınırken, son konvolüsyon katmanına doğru geri yayılım yapan gradyanları kullanarak önemli bölgeleri haritalar (Selvaraju vd., 2019). Bu yaklaşım, modelin bir "kedi"yi neden o şekilde tanıdığına dair görsel bir gerekçe sunabilir. Ayrıca, guided backpropagation ile birleştirilerek yüksek çözünürlüklü ve sınıf-özgü saliency haritaları oluşturan Guided Grad-CAM yöntemi de geliştirilmiştir.

Grad-CAM'in bir başka avantajı ise herhangi bir CNN mimarisine ek bir yeniden eğitim veya yapısal değişiklik gerekmeden uygulanabilmesidir. Bu da onu hem sınıflandırma, hem görüntü açıklama (captioning), hem de görsel soru-cevap gibi çoklu görevler için oldukça esnek kılar (Selvaraju vd., 2019).

Bununla birlikte, yapılan bazı çalışmalar, bu saliency haritalarının eğitim sırasında rastgele başlatma (random initialization) gibi faktörlerden önemli ölçüde etkilendiğini ortaya koymuştur. Özellikle, aynı veri ve aynı model mimarisi kullanılsa bile, sadece farklı rastgele başlangıçlarla elde edilen saliency haritalarının anlamlı ölçüde değiştiği görülmüştür. Bu durum, bu tür açıklamaların güvenilirliğini sorgulatmaktadır (Woerl, Disselhoff & Wand, 2023). Çözüm olarak, farklı başlangıçlarla eğitilmiş birden fazla modelin sonuçlarının ortalamasını almak gibi "marginalization" tabanlı yaklaşımlar önerilmiştir.

Saliency map yöntemlerinin güvenilirliğini sistematik olarak değerlendiren karşılaştırmalı bir çalışmada ise, Grad-CAM, Guided Backpropagation, Smooth Grad, EigenCAM gibi birçok tekniğin farklı senaryolarda ne ölçüde doğru lokalizasyon sağladığı analiz edilmiştir. Bu yöntemlerin görsel çıktılarını karşılaştırmalı şekilde sunan Şekil 5'de görüldüğü gibi, Grad-CAM tabanlı yaklaşımlar farklı ağ mimarilerinde (Simple CNN, VGG16, ResNet50) daha belirgin ve hedef odaklı bölgeler üretebilmektedir. Özellikle hedef nesnenin bulunduğu alanı başarıyla vurgulayan bu teknikler, diğer yöntemlere kıyasla daha yüksek tutarlılık sergilemektedir.



Şekil 5. GTSRB* test veri kümesi üzerinde eğitilen ResNet50, VGG16 ve SCNN mimarileri için üretilen saliency (dikkat) haritalarından örnekler. İlk üç satır, desen içermeyen giriş görüntülerini (sınıf 0), son üç satır ise tanımlı bir desen içeren giriş görüntülerini (sınıf 1) göstermektedir

Bu analizde, özel olarak hazırlanan veri kümeleri ve geliştirilmiş bir karşılaştırma metriğiyle hangi yöntemlerin güvenilir olduğu belirlenmiştir. Sonuçlara göre, Grad-CAM gibi gradyan tabanlı yöntemler hem model kararlarını açıklamada hem de eğitim verisindeki önyargıları tespit etmede etkili olmuştur (Szczepankiewicz vd., 2023).

Bununla birlikte, bazı durumlarda Grad-CAM yetersiz kalabilir; örneğin, aynı sınıfa ait birden fazla nesne içeren görüntülerde tüm hedefleri tam olarak kapsayamayabilir. Bu sorunu aşmak için geliştirilen Grad-CAM++, yüksek mertebeden türevler kullanarak çoklu hedefleri daha başarılı şekilde tespit edebilmektedir. Ancak yine de çözünürlük sınırlamaları sürmektedir (Gao vd., 2023). Bu nedenle, Augmented Grad-CAM++ gibi yeni yaklaşımlar önerilmiştir. Bu yöntemde, aynı görüntünün farklı geometrik dönüşümlerine dayalı olarak oluşturulan çoklu aktivasyon haritaları birleştirilerek daha keskin ve yüksek çözünürlüklü bir saliency haritası elde edilmektedir. Özellikle endüstriyel kusur tespiti gibi hassas uygulamalarda bu yaklaşımın doğruluk ve görsel yorumlama başarısını anlamlı şekilde artırdığı gösterilmiştir.

4.3. Metin

Doğal dil işleme (natural language processing-NLP) alanında derin öğrenme modellerinin kullanımı giderek yaygınlaşsa da, bu modellerin "kara kutu" doğası modelin nasıl karar verdiğini anlamayı oldukça güçleştirmektedir. Bu nedenle, özellikle metin verileri üzerinde çalışan modellerde açıklanabilirlik ihtiyacı ön plana çıkmaktadır. Metin sınıflandırma, duygu analizi, makine çevirisi ve okuduğunu anlama gibi görevlerde kullanılan açıklanabilir yapay zeka teknikleri, kullanıcıların modelin verdiği kararları daha iyi anlamasını ve bu kararların güvenilirliğini değerlendirmesini sağlar.

LIME bu alanda yaygın olarak kullanılan model-agnostik bir tekniktir. LIME, bir modelin tek bir tahmini üzerindeki davranışını açıklamak için çalışır. Metin verilerinde bu yaklaşım, belirli kelimelerin modelin çıktısına etkisini ölçmek amacıyla, bu kelimelerin kaldırılması ya da maskelenmesi gibi bozulmalar uygulayarak yerel bir doğrusal model oluşturur. Bu şekilde, hangi kelimelerin karar üzerinde ne kadar etkili olduğunu anlaşılır hale getirir (Mersha vd., 2024). LIME'ın en büyük avantajı, modelden bağımsız çalışması ve özellikle sınıflandırma görevlerinde kelime düzeyinde etkili açıklamalar sunabilmesidir. Ancak global düzeyde modelin genel davranışını anlamakta yetersiz kalabilir.

Şekil 6'da görüldüğü gibi, LIME yöntemi bir kara kutu modelin belirli bir tahmini üzerinde, o tahmine yakın örnekler üretip bu örnekler üzerinden daha basit bir açıklayıcı model kurarak çalışır. Bu açıklayıcı model, siyah kutunun yerel davranışını taklit ederek insan tarafından yorumlanabilir hale getirir.

Integrated Gradients ise daha model-özgü bir yaklaşımdır ve modelin çıktısının girdilere duyarlılığını, yani gradyanlarını inceleyerek çalışır. Bu teknik, bir temel girdiden (genellikle sıfır bilgi içeren bir giriş) başlayarak gerçek girdiye kadar yapılan interpolasyon boyunca gradyanları toplar ve böylece her kelimenin modele katkısını hesaplar. Bu yöntem, LIME'a kıyasla daha tutarlı sonuçlar üretme eğilimindedir ve modelin iç yapısını da dikkate aldığı için özellikle transformer tabanlı modellerde öne çıkmaktadır (Gurrapu vd., 2023).

Bir diğer önemli XAI yöntemi ise dikkat görselleştirmeleridir. Bu teknik, transformer mimarisinin doğasında bulunan dikkat (attention) ağırlıklarını kullanarak modelin hangi kelimelere ne kadar "odaklandığını" görselleştirir. Örneğin, bir BERT modeli üzerinde yapılan analizlerde, dikkat ağırlıklarının görselleştirilmesiyle modelin hangi kelime çiftleri arasındaki ilişkileri dikkate aldığı anlaşılabilir hale gelir (Mershaa vd., 2025). Bununla birlikte, yalnızca dikkat ağırlıklarının yorumlanması, her zaman modelin karar mantığını tam olarak yansıtmayabilir; zira bazı durumlarda model yüksek dikkat verdiği kelimeleri değil, daha az dikkat verdiği ancak bağlamda kritik olan kelimeleri de kullanabilir.

Son yıllarda geliştirilen değerlendirme çerçeveleri sayesinde bu XAI tekniklerinin doğruluğu, tutarlılığı ve insan akıl yürütmesiyle olan uyumu ölçülmektedir. Örneğin, ruman-reasoning agreement, robustness ve consistency gibi metriklerle yapılan değerlendirmelerde, LIME genellikle insan akıl yürütmesine en çok benzeyen açıklamaları üretirken; dikkat görselleştirme teknikleri tutarlılık açısından öne çıkmaktadır (Mershaa vd., 2025).

Ayrıca, son dönemde öne çıkan bir başka yaklaşım ise rationalization yöntemidir. Bu yaklaşım, modelin verdiği tahminleri doğal dilde açıklama üretimiyle gerekçelendirmeye çalışır. Rationalization, özellikle teknik bilgisi olmayan kullanıcılar için oldukça erişilebilir ve sezgisel bir açıklama biçimi sunar. Rasyonelleştirme, iki ana türe ayrılır: girdiden önemli cümlelerin çıkarıldığı extractive yöntemler ve modelin kendi ifadesiyle açıklama ürettiği abstractive yöntemler. Her iki yaklaşım da modelin kararlarını daha insan-merkezli ve anlaşılır hale getirmeyi amaçlamaktadır (Gurrapu vd., 2023).

Bu farklı XAI tekniklerinin her biri, kendi güçlü ve zayıf yönlerine sahiptir. Bu nedenle, metin tabanlı uygulamalarda açıklanabilirlik sağlamak için tek bir yönteme bağlı kalmak yerine, göreve, modele ve kullanıcı profiline göre uygun kombinasyonların tercih edilmesi daha etkili sonuçlar verebilir.

4.4. Açık Kaynak Kütüphaneler

Son yıllarda XAI alanındaki ilerlemeler, açıklanabilirliği sağlayan yöntemlerin yazılımsal araçlara dönüşmesini beraberinde getirmiştir. Bu doğrultuda geliştirilen açık kaynak kütüphaneler, araştırmacılar ve mühendisler için XAI uygulamalarını daha erişilebilir ve tekrarlanabilir hale getirmektedir. Ancak, bu araçların sunduğu açıklamaların kararlılığı ve doğruluğu, hala araştırılması gereken önemli bir konudur.

Özellikle LIME, SHAP, DeepLIFT ve Integrated Gradients gibi yöntemlerin yaygın olarak desteklendiği açık kaynak kütüphaneler, birçok kullanıcı tarafından güvenilir kabul edilmekte olsa da, bu tekniklerin eğitim sürecindeki rassallıklara duyarlı olduğu gözlemlenmiştir.

Woerl ve arkadaşlarının (2023) yürüttüğü çalışmada, bu yöntemlerin ürettiği saliency haritalarının, modelin rastgele ağırlık başlatmalarına (initialization) ve eğitim sırasında seçilen mini-batch'lere bağlı olarak önemli ölçüde değişkenlik gösterdiği ortaya konmuştur. Örneğin, SHAP ve LIME gibi kütüphaneler aracılığıyla üretilen açıklamalar, aynı giriş verisi için farklı model başlatmalarında birbirinden oldukça farklı sonuçlar üretebilmektedir. Bu durum, özellikle karar destek sistemleri gibi kritik uygulamalarda, kullanıcı güvenini zedeleyebilecek potansiyel bir risk oluşturmaktadır (Woerl vd., 2023).

Woerl ve arkadaşları, bu sorunu azaltmak amacıyla önerdikleri "marginalization" tekniğiyle, aynı yapıda fakat farklı başlatmalarla eğitilmiş çoklu modellerin çıktılarının ortalamasını alarak, açıklamaların daha tutarlı hale getirilebileceğini göstermiştir. Özellikle GradCAM gibi yöntemlerin üst katmanlarda kullanıldığında daha kararlı açıklamalar ürettiği, buna karşın SHAP ve LIME gibi yöntemlerin düşük SNR (signal-to-noise ratio) değerlerine sahip olduğu raporlanmıştır (Woerl vd., 2023).

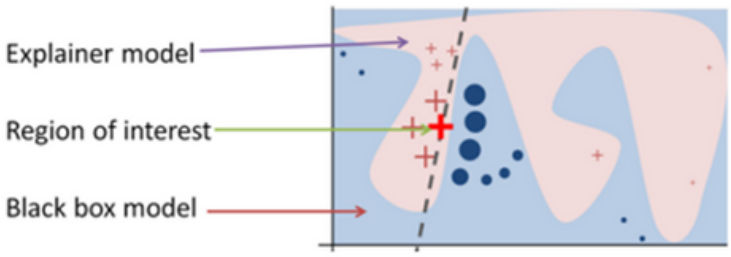
Öte yandan, Wang (2024), LIME ve SHAP'in sektörel uygulamalardaki gücüne dikkat çekmiştir. LIME'ın tıbbi teşhis gibi bireysel açıklamanın önemli olduğu alanlarda, SHAP'in ise finansal risk değerlendirmeleri gibi detaylı özellik katkısı gereken senaryolarda başarılı olduğu vurgulanmıştır. Bu kütüphaneler, kullanıcıya hem lokal hem de global açıklamalar sunabilme kapasitesine sahiptir. Ancak, bu güçlü özelliklerin yanında, açıklamaların yorumlanması için uzman bilgisine ihtiyaç duyulabileceği ve SHAP gibi yöntemlerin yüksek hesaplama maliyeti nedeniyle büyük ölçekli uygulamalarda sınırlamalarla karşılaşabileceği belirtilmiştir (Wang, 2024).

Sonuç olarak, ELI5, Captum, Alibi ve InterpretML gibi açık kaynak kütüphaneler, LIME, SHAP ve benzeri yöntemleri kapsayarak kullanıcıya esnek çözümler sunsalar da, açıklamaların güvenilirliği konusunda temkinli olunması gerektiği görülmektedir. Açıklanabilirlik araçlarının sunduğu görselleştirmelerin yalnızca tek bir model çıktısına dayandırılması yerine, çoklu model entegrasyonlarıyla desteklenmesi önerilmektedir. Bu sayede, XAI araçlarının sunduğu çıktılar daha sağlam ve güvenilir hale getirilebilir.

5. XAI UYGULAMA ALANLARI

5.1. Sağlık

Sağlık sektörü, yapay zekânın (YZ) potansiyelinden en fazla fayda sağlaması beklenen alanlardan biridir; ancak bu teknolojilerin güvenle benimsenmesi için açıklanabilirlik büyük önem taşımaktadır. Son yıllarda geliştirilen açıklanabilir yapay zekâ (XAI) yöntemleri, klinik karar destek sistemlerinin şeffaflığını artırmayı ve sağlık çalışanlarının bu sistemlere olan güvenini sağlamayı hedeflemektedir (Noor vd., 2025). Özellikle "kara kutu" olarak adlandırılan karmaşık makine öğrenmesi modelleri, neye göre karar verdiklerini açıklayamadıkları için hekimler ve hastalar açısından güven sorunu oluşturmaktadır. Bu sorunu çözmek üzere geliştirilen SHAP, LIME, Grad-CAM gibi yöntemler, algoritmaların hangi nedenlerle belirli kararlar verdiğini görünür kılarak sistemin yorumlanabilirliğini artırmaktadır (Band vd., 2023).



Şekil 6.LIME açıklama mantığı

Bu konuda yapılan somut çalışmalardan biri, Stanford Üniversitesi tarafından geliştirilen CheXplain adlı sistemdir. Bu sistem, göğüs röntgeni görüntülerinde yapay zekânın tanı koyarken dikkat ettiği bölgeleri Grad-CAM yöntemiyle görselleştirerek, hekimlerin model kararlarını daha iyi anlamalarına ve doğrulamalarına olanak tanımaktadır (Arun vd., 2021).

Yapılan çalışmalar, açıklanabilirliğin sadece teknik bir özellik değil, aynı zamanda etik ve yasal bir zorunluluk olduğunu ortaya koymaktadır. Avrupa Birliği'ndeki GDPR gibi düzenlemeler, otomatik karar alma süreçlerinde bireylerin açıklama hakkını tanımlamaktadır (Caterson vd., 2024). Bu da, XAI sistemlerinin sağlık alanında uygulanabilmesi için yalnızca doğruluk açısından değil, açıklanabilirlik açısından da güçlü olmalarını zorunlu kılmaktadır.

XAI uygulamaları özellikle elektronik sağlık kayıtları (EHR) üzerinde yapılan çalışmalarda dikkat çekmektedir. SHAP yöntemi, bu alanda en yaygın kullanılan XAI tekniği olarak öne çıkmakta ve bireysel tahminler için özniteliklerin katkı düzeyini nicel olarak açıklamaktadır. SHAP, hem bireysel kararların anlaşılmasını sağlamakta hem de modelin genel davranışı hakkında bilgi sunmaktadır (Caterson vd., 2024). Bunun yanında, LIME yöntemi ise belirli bir tahminin neden o şekilde yapıldığını lokal düzeyde açıklamakta, böylece hekimlerin örnek bazında model kararlarını sorgulamasına olanak tanımaktadır.

Görüntü verileriyle yapılan çalışmalarda da XAI tekniklerinin sağlık alanındaki etkisi oldukça belirgindir. Özellikle Grad-CAM gibi görselleştirme tabanlı yöntemler, medikal görüntülerin analizinde hangi bölgelerin karar üzerinde etkili olduğunu ortaya koyarak, radyologlara karar süreçlerinde destek olmaktadır (Band vd., 2023; Noor vd., 2025). Covid-19 tanısından kronik böbrek hastalıklarına, beyin tümörlerinden mantar enfeksiyonlarına kadar pek çok klinik alanda açıklanabilir modeller, sadece doğru karar vermekle kalmayıp bu kararların ardındaki nedenleri de gösterebilmişlerdir.

Bununla birlikte, XAI sistemlerinin tasarımı kadar, bu sistemlerin kullanıcıya sunduğu açıklamaların kalitesi de sağlık alanında güven inşa etmede belirleyici olmaktadır. Bazı çalışmalar, açıklamaların fazla teknik ya da karmaşık olmasının hekimlerin güvenini azaltabileceğini, bu nedenle açıklamaların klinik bağlamla uyumlu ve sade olması gerektiğini vurgulamaktadır (Rosenbacke vd., 2024). Çünkü açıklanabilirliğin varlığı tek başına güveni artırmak için yeterli değildir; önemli olan, bu açıklamaların sağlık profesyonelleri tarafından gerçekten nlaşılabilir ve değerlendirilebilir olmasıdır.

Sonuç olarak, XAI sistemleri, sağlık hizmetlerinde sadece teknolojik değil aynı zamanda etik, yasal ve kullanıcı odaklı bir gereklilik olarak öne çıkmaktadır. Yapay zekâ uygulamalarının açıklanabilirlik boyutunu güçlendirmek, sadece güveni artırmakla kalmayacak, aynı zamanda hasta güvenliği ve klinik etkinlik açısından da önemli katkılar sağlayacaktır.

5.2. Eğitim

Yapay zeka (YZ) sistemlerinin eğitim alanında artan rolü, bu sistemlerin nasıl karar verdiğini açıklayabilme ihtiyacını beraberinde getirmiştir. Özellikle öğrenci performans tahmini, bireyselleştirilmiş öğretim sistemleri ve öğrenme analizleri gibi uygulamalarda kullanılan karmaşık modellerin, şeffaf biçimde anlaşılması gerekmektedir. Bu nedenle XAI (Explainable Artificial Intelligence), eğitim teknolojileri içinde giderek daha merkezi bir yere oturmaktadır (Khosravi vd., 2022).

Eğitimde XAI kullanımını ele alan literatürde en dikkat çeken katkılardan biri XAI-ED çerçevesidir. Bu çerçeve; öğrenciler, öğretmenler, yöneticiler ve politika yapıcılar gibi paydaşlara nasıl açıklamalar sunulması gerektiği, hangi modellerin kullanıldığı, insan merkezli arayüzlerin nasıl tasarlandığı ve açıklamaların potansiyel tuzaklarını kapsayan altı boyuttan oluşmaktadır. Bu yaklaşım sayesinde, açıklanabilirliğin yalnızca teknik değil aynı zamanda pedagojik bir mesele olduğu vurgulanır. Örneğin, açıklamalar yalnızca kararın nedenini öğretmene sunmakla kalmaz, aynı zamanda öğrencinin öğrenme sürecine dair farkındalığını da artırabilir (Khosravi vd., 2022).

Bu çerçevede geliştirilen somut uygulamalardan biri, IBM Research tarafından oluşturulan XAITutor adlı sistemdir. Bu sistem, öğrenciye verilen önerilerin neden o şekilde sunulduğunu şeffaf biçimde açıklarken, aynı zamanda öğretmenin pedagojik kararlarına da destek olacak şekilde ayrıntılı geri bildirim sağlar. Böylece hem öğrenen hem de öğreten taraf için açıklanabilirlik çift yönlü olarak işletilmektedir.

Makineden gelen geri bildirimlerin öğrencilere ne şekilde sunulacağı da önemli bir konudur. Özellikle “Open Learner Models” (OLM) adı verilen açık öğrenci modelleri, öğrencinin kendi öğrenme sürecini izlemesini ve bu süreç hakkında düşünmesini sağlayarak metabilişsel farkındalık geliştirmesine katkı sunar. Bu sistemler aracılığıyla, öğrenci hangi konularda yeterli olduğunu veya nerelerde zorlandığını görerek hedefli çalışma planları oluşturabilir (Khosravi vd., 2022).

Bununla birlikte, öğretmenler için de açıklanabilirlik kritik önemdedir. Bir sınıfın genel başarı durumu, öğrencilerin hangi konularda zorluk yaşadığı ya da hangi öğretim yöntemlerinin etkili olduğu gibi çıkarımlar, öğretmenin öğretim tasarımı üzerinde yeniden düşünmesini sağlar. Bu geri bildirim süreci, öğretmenin pedagojik bilgi birikimini geliştirerek öğretim kalitesini artırabilir (Khosravi vd., 2022).

İkinci çalışmada ise, XAI uygulamalarının son yıllarda eğitim dâhil birçok alanda yaygınlaştığı vurgulanmaktadır. Özellikle SHAP ve LIME gibi modelden bağımsız açıklama yöntemlerinin eğitim verilerinde de sıkça kullanıldığı belirtilmiştir (Saarela & Podgorelec, 2024). Eğitim alanındaki uygulamalar; öğrenci performansının erken tahmini, mezuniyet sonrası başarı tahmini, öğrenci bağlılığı ve öğrenme süreçlerine ilişkin ileri düzey analizleri içermektedir. Bu sistemlerin çoğunda kararların nedenleri, model içgörüleri ve öğrenci özelindeki öngörüler SHAP/LIME gibi araçlarla görselleştirilerek sunulmaktadır (Saarela & Podgorelec, 2024)

Ancak her iki çalışmada da dikkat çekilen temel sorun, açıklamaların değerlendirilmesinde kullanılan metriklerin eksikliğidir. Eğitim alanında açıklamalar genellikle anekdotlara ya da uzman görüşlerine dayanmakta; sağlam, nicel değerlendirme çerçeveleri yeterince kullanılmamaktadır (Saarela & Podgorelec, 2024). Bu durum, geliştirilen sistemlerin güvenilirliğini sorgulatmakta ve XAI'nin eğitimde yaygınlaşmasının önünde engel oluşturmaktadır.

Sonuç olarak, eğitimde XAI kullanımı yalnızca algoritmik kararların şeffaflığını sağlamakla kalmaz, aynı zamanda öğrenci merkezli öğrenme anlayışını da destekler. Ancak bunun etkili olabilmesi için kullanıcı dostu açıklama yöntemlerinin geliştirilmesi, paydaşların ihtiyaçlarının dikkate alınması ve açıklamaların geçerliliğinin sağlam ölçütlerle değerlendirilmesi gerekmektedir.

5.3. Savunma

Savunma alanında YZ sistemlerinin kullanımı her geçen gün artarken, bu sistemlerin güvenilirliğini ve şeffaflığını sağlamak kritik bir ihtiyaç hâline gelmiştir. Özellikle otomatik silah sistemleri ve video analizine dayalı güvenlik uygulamalarında, kararların sadece doğru olması değil, aynı zamanda nasıl alındığının da anlaşılabilir olması beklenmektedir (Gohel vd., 2021).

XAI'nin savunmadaki önemi, video analizine dayalı tehdit tespit sistemleri üzerinden açıkça görülmektedir. Geleneksel güvenlik sistemlerinde insan gözüyle sürekli video takibi mümkün olmadığından, YZ tabanlı nesne ve yüz tanıma sistemleri, havalimanı gibi kritik yerlerde ziyaretçilerin tehdit oluşturup oluşturmadığını belirlemek için kullanılmaktadır. Ancak bu sistemlerin hatalı olarak masum bireyleri tehdit olarak algılaması hem yasal hem de etik sorunlara yol açabilmektedir. Bu nedenle, sistemin bir kişiyi neden şüpheli olarak işaretlediğini açıkça ortaya koyabilmesi büyük önem taşır. Aksi takdirde, yaşanabilecek kamuya açık arama ve aşağılanmalar ciddi hukuki sonuçlar doğurabilir (Gohel vd., 2021).

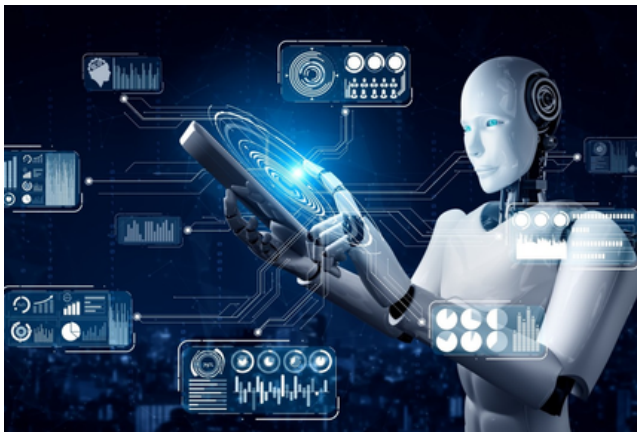
Bu bağlamda somut bir örnek, DARPA'nın XAI programı kapsamında geliştirilen, savunma operasyonlarında karar süreçlerini açıklayabilen yapay zekâ sistemleridir. Bu sistemlerde, örneğin bir insansız hava aracının neden belirli bir hedefi tehdit olarak sınıflandırdığı, kullanılan XAI teknikleriyle operatöre açıklanabilir şekilde sunulmaktadır. Böylece hem insan denetimi kolaylaşmakta hem de algoritmik kararlar daha güvenli hale gelmektedir (Gunning & Aha, 2019).

XAI'nin sunduğu açıklanabilirlik, yalnızca güvenlik güçlerinin kararları sorgulamasını kolaylaştırmakla kalmaz; aynı zamanda mahkemede sunulabilecek rasyonel gerekçelerin oluşturulmasını da sağlar. Modelin belirli bir kararı hangi özelliklere dayanarak verdiğini göstermesi, adil ve tarafsız karar verme süreçlerini destekler (Gohel vd., 2021).

Bunun yanında, savunma sektöründe gerçek zamanlı tehdit tespitine yönelik YZ modellerinde kullanılan görüntü işleme tekniklerinde de açıklanabilirlik gereklidir. Örneğin, bir modelin bir görüntüyü neden savaş uçağı olarak sınıflandırdığını anlamak için kullanılan karşı-gerçeksel (counterfactual) yaklaşımlar, kararın hangi görüntü bölgesine dayandığını gösterebilir. Böylece sistemlerin iç işleyişi daha iyi anlaşılır ve kullanıcı güveni artırılır (Gohel vd., 2021).

Sonuç olarak, savunma sektöründe XAI yalnızca teknik bir gereklilik değil, aynı zamanda hukuki ve etik sorumlulukların da bir parçasıdır. Kararların anlaşılabilir, şeffaf ve gerekçeli olması, sistemlerin güvenilirliğini artırmakla kalmaz, aynı zamanda kamuoyunun bu sistemlere olan güvenini de güçlendirir.

5.4. Hukuk



Yapay zekâ (YZ) sistemlerinin hukuk ve adalet alanında kullanımı, beraberinde önemli etik ve toplumsal tartışmaları da getirmiştir. Kararların bireylerin özgürlüklerini doğrudan etkilediği bu alanda, şeffaflık ve hesap verilebilirlik büyük önem taşır. Bu noktada, Açıklanabilir Yapay Zekâ teknikleri devreye girerek yargı sistemindeki algoritmik kararları anlamlandırmayı mümkün kılar (Ikumapayi & Muhammed, 2024).

Özellikle tahmine dayalı polislik uygulamalarında, XAI teknikleri sayesinde bir bölgenin suç oranlarının ya da bireylerin tekrar suç işleme olasılıklarının nasıl belirlendiği açıklanabilir hâle gelir. Bu sayede, sistemin potansiyel önyargılar taşıyıp taşımadığı daha rahat analiz edilebilir (Ikumapayi & Muhammed, 2024).

Bu bağlamda ABD'de kullanılan COMPAS adlı karar destek sistemi örnek olarak öne çıkar. Bu sistem, sanıkların yeniden suç işleyip işlemeyeceğini tahmin etmekte kullanılmış, ancak kararlarının açıklanamaması kamuoyunda ciddi eleştiriler doğurmuştur. Sonrasında, sistemin yaş, ırk ve sosyoekonomik durum gibi değişkenlere karşı önyargılı olup olmadığını anlamak üzere LIME ve SHAP gibi XAI araçlarıyla analizler yapılmış ve algoritmik şeffaflık çabaları başlatılmıştır (Dressel & Farid, 2018).

Yargı destek sistemlerinde ise, örneğin kefalet ya da ceza süreleri gibi kararlarda YZ'den yardım alındığında, açıklanabilirlik sağlamak hem yargıcın karar sürecine katkı sağlar hem de kararların gerekçesinin daha net sunulmasına yardımcı olur. Bu da hukukta adaletin tecellisine doğrudan katkıda bulunur (Ikumapayi & Muhammed, 2024).

Hukuki araştırma ve dava sonucu tahmini gibi alanlarda da XAI uygulamaları öne çıkar. Özellikle önceki içtihatları analiz eden modellerin hangi faktörleri dikkate alarak sonuca vardığını ortaya koymak, avukatların daha sağlam argümanlar üretmesini sağlar (Ikumapayi & Muhammed, 2024).

Ayrıca, e-keşif (e-discovery) süreçlerinde kullanılan yapay zekâ sistemleri, belgelerdeki hangi bölümlerin önemli olduğunu sadece işaretlemekle kalmayıp, bunun nedenini de açıklayarak avukatlara zaman kazandırır ve bilgiye dayalı karar alma süreçlerini destekler (Ikumapayi & Muhammed, 2024).

Son olarak, adalet sisteminde kullanılan YZ modellerinde görülebilecek önyargıların tespiti açısından da XAI önemli bir araçtır. Bu teknikler, karar süreçlerinde cinsiyet, etnik köken veya sosyoekonomik durum gibi değişkenlerin sistem üzerindeki etkisini şeffaf şekilde ortaya koyabilir. Böylece, ayrımcılığın önlenmesine yönelik somut adımlar atılabilir (Ikumapayi & Muhammed, 2024).

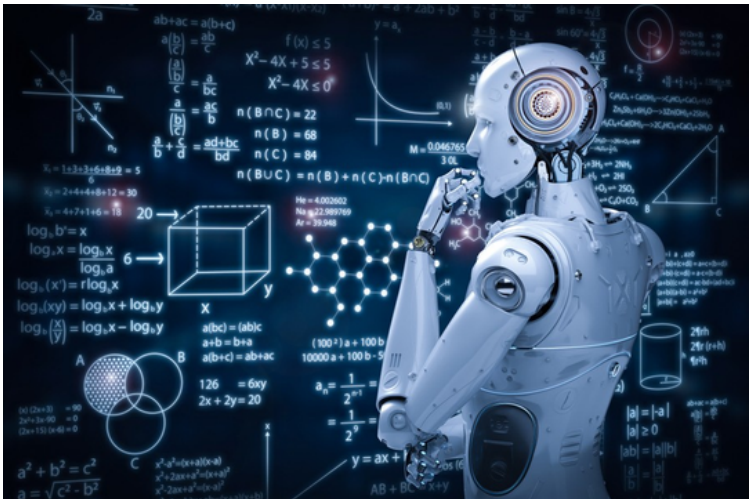
5.5. Finans

Finans sektörü, yapay zekâ modellerinin karar verme süreçlerine entegre edilmesiyle birlikte büyük bir dönüşüm geçirmektedir. Ancak bu modellerin "kara kutu" doğası, şeffaflık ve güven eksikliği gibi sorunları da beraberinde getirmiştir. Bu nedenle, açıklanabilir yapay zekâ yöntemlerinin finans alanındaki kullanımı giderek daha kritik hale gelmektedir. Özellikle kredi yönetimi, risk değerlendirme, dolandırıcılık tespiti ve yatırım kararları gibi yüksek hassasiyet gerektiren uygulamalarda, modellerin nasıl karar verdiğini anlamak büyük önem taşımaktadır (Černevičienė & Kabašinskas, 2024).

Makaledeki sistematik incelemeye göre, XAI'nin en yoğun kullanıldığı alan kredi yönetimidir. Bu alan, kredi skoru tahmininden risk değerlendirmeye kadar birçok alt görevi kapsamaktadır. Özellikle SHAP gibi özellik önem derecesi sunan yöntemler, hangi değişkenlerin modele ne kadar katkı sunduğunu belirlemede sıklıkla tercih edilmektedir (Černevičienė & Kabašinskas, 2024). Bunun yanında LIME gibi yerel açıklama yöntemleri, bireysel tahminlerin arkasındaki mantığı açığa çıkararak kullanıcıların model davranışını daha iyi kavramasına olanak tanır.

Bu çerçevede öne çıkan bir örnek, 2018 yılında düzenlenen FICO Explainable ML Challenge adlı yarışmadır. Bu yarışmada, katılımcılardan kredi skorlama sistemlerini hem yüksek doğrulukla hem de açıklanabilir şekilde tasarlamaları istenmiştir. Kazanan modeller, SHAP ve LIME gibi araçları kullanarak her bir müşteriye özel olarak hangi değişkenlerin kredi sonucunu etkilediğini açıkça gösterebilmiş ve bu sayede finansal kararların gerekçesi kullanıcıya sunulmuştur (FICO, 2018).

Ayrıca bazı çalışmalarda, modelin kararlarını sadeleştirerek anlaşılır hale getiren kural tabanlı yaklaşımlar ve karar ağaçları gibi yöntemler kullanılmıştır. Örneğin, kredi risk değerlendirme sistemlerinde recursive rule extraction algoritmaları ile karar verme süreçleri daha şeffaf hale getirilmiştir (Černevičienė & Kabašinskas, 2024).



XAI uygulamalarının yatırım yönetimi, portföy optimizasyonu, sigorta fiyatlandırması gibi daha geniş finansal alanlarda da kullanıldığı belirtilmiştir. Görselleştirme destekli açıklama araçları, özellikle zamana bağlı modellerin kararlarını kullanıcıya sezgisel şekilde sunmak adına etkili olmuştur (Černevičienė & Kabašinskas, 2024).

Sonuç olarak, finans sektöründe XAI tekniklerinin benimsenmesi sadece model performansını değil, aynı zamanda güvenilirliği, yasal uyumluluğu ve kullanıcı memnuniyetini de artırmaktadır. Ancak model karmaşıklığı ile açıklanabilirlik arasında bir denge kurmak hâlâ önemli bir zorluk olarak varlığını sürdürmektedir (Černevičienė & Kabašinskas, 2024).

5.6. Otonom Araçlar

Otonom araçlar, çevrelerini algılayabilen, karar alabilen ve insan müdahalesi olmadan hareket edebilen gelişmiş yapay zeka sistemleriyle donatılmıştır. Ancak bu sistemlerin karar alma süreçleri çoğu zaman "kara kutu" olarak nitelendirildiğinden, hem kullanıcılar hem de düzenleyici kurumlar açısından güven ve kabul sorunları ortaya çıkmaktadır. Bu nedenle, XAI yaklaşımları otonom araçlar alanında kritik bir rol üstlenmektedir (Atakishiyev vd., 2024).

Özellikle derin öğrenmeye dayalı otonom araçlar sistemleri yüksek doğruluk sunsa da şeffaflıktan yoksundur. Bu durum güvenlik ve sorumluluk gibi alanlarda soru işaretlerine neden olur. Örneğin, Kuznietsov ve arkadaşları, XAI'nin yalnızca kullanıcıların sisteme olan güvenini artırmakla kalmayıp, aynı zamanda geliştiricilerin hata ayıklamasına, denetçilerin inceleme yapmasına ve yasal mercilerin sorumluluk analizini gerçekleştirmesine de katkı sunduğunu vurgulamaktadır (Kuznietsov vd., 2024).

Bu noktada, Waymo gibi şirketler tarafından geliştirilen sistemler dikkat çekmektedir. Waymo'nun test araçlarında, aracın neden fren yaptığı, şerit değiştirdiği veya bir yayaya öncelik verdiği gibi kararlar, kullanıcıya görsel açıklamalarla sunulmaktadır. Bu açıklamalarda sensör verileri, nesne sınıflandırmaları ve tahmini niyetler görselleştirilerek hem gerçek zamanlı güvenliği artırmakta hem de kullanıcı farkındalığını desteklemektedir.

Bu alandaki XAI çalışmaları genel olarak beş ana katkı üzerinden değerlendirilmektedir: yorumlanabilir tasarım, yorumlanabilir vekil modeller, yorumlanabilir izleme sistemleri, yardımcı açıklamalar ve yorumlanabilir doğrulama mekanizmaları. Bu katkılar sayesinde, AV sistemleri daha güvenli ve şeffaf hale getirilmeye çalışılmaktadır (Kuznietsov vd., 2024).

Öte yandan Atakishiyev ve arkadaşları, açıklanabilirliğin farklı paydaşlar için farklı anlamlar taşıdığını belirtmektedir. Kullanıcılar için açıklamalar, sistemin ne zaman ve neden belli bir karar verdiğini anlamak açısından önemliyken; geliştiriciler, düzenleyici kurumlar ve sigorta şirketleri için daha teknik ve detaylı açıklamalar gereklidir. Bu nedenle, açıklama mekanizmalarının hedef kullanıcıya göre özelleştirilmesi gerektiği vurgulanmaktadır (Atakishiyev vd., 2024).

Ayrıca, açıklamaların sadece doğru içerik üretmesi değil, aynı zamanda uygun zamanlama ve uygun arabirimlerle sunulması da önemlidir. Özellikle gerçek zamanlı karar alma gerektiren AV sistemlerinde, açıklamaların zamanında ve anlaşılır şekilde sunulması güvenliğin sağlanmasında hayati rol oynamaktadır (Atakishiyev vd., 2024).

Sonuç olarak, XAI teknolojilerinin otonom araçlara entegrasyonu; güvenlik, şeffaflık, izlenebilirlik ve toplumsal kabul açısından büyük önem taşımaktadır. Bu yaklaşımlar sayesinde, hem teknik açıdan sistemlerin daha güvenilir hale gelmesi hem de toplumsal düzeyde bu sistemlere olan güvenin artması beklenmektedir (Kuznietsov vd., 2024; Atakishiyev vd., 2024).

6. ETİK VE SOSYAL BOYUTLAR

6.1. Adalet, Önyargı ve Hesap Verebilirlik

Yapay zekâ sistemlerinin karar alma süreçlerine entegre edilmesi, beraberinde adalet, önyargı ve hesap verebilirlik gibi etik soruları da gündeme taşımıştır. Özellikle XAI sistemlerinin bu alanlarda nasıl bir rol oynayabileceği, yapay zekâ sistemlerinin karar alma süreçlerine entegre edilmesi, beraberinde adalet, önyargı ve hesap verebilirlik gibi etik soruları da gündeme taşımıştır. Özellikle XAI sistemlerinin bu alanlarda nasıl bir rol oynayabileceği, son dönemde hem teknik hem de sosyal bilimler literatüründe yoğun şekilde tartışılmaktadır.

Adalet konusunda XAI çoğu zaman bir umut olarak görülmektedir; çünkü sistemlerin nasıl çalıştığını açıklamak, özellikle dezavantajlı gruplara karşı oluşabilecek önyargıları ortaya çıkarmada etkili olabilir. Nitekim Deck ve arkadaşları (2024), XAI'ın adalet ile ilişkisine dair yapılan birçok iddianın genellikle fazla basitleştirildiğini, net bir normatif temelden yoksun olduğunu ve XAI'ın gerçek teknik kapasitesiyle uyumsuz olduğunu belirtmektedirler. Bu bağlamda XAI, tüm adalet problemlerini çözebilecek bir "etik panzehir" değil; ancak adalet arayışında kullanılabilecek birçok araçtan yalnızca biridir (Deck vd., 2024).

XAI'ın önyargıları tespit etme ve azaltma konusundaki rolü ise yine karmaşıktır. Özellikle denetim ve muhasebe alanlarında yapılan sistematik bir inceleme, yapay zekâ tabanlı sistemlerin veri eksiklikleri, demografik homojenlik, anlamsız korelasyonlar ve tasarımcıların bilişsel önyargıları gibi birçok kaynaktan beslenen adaletsizlikleri yeniden üretebildiğini ortaya koymuştur (Murikah vd., 2024). Örneğin, geçmişteki ayrımcılıkların izlerini taşıyan veri setleriyle eğitilen bir sistem, bu önyargıları geleceğe taşıyabilir. Bu durum, özellikle azınlık grupların daha fazla denetlenmesine veya riskli olarak sınıflandırılmasına neden olabilir.

Diğer yandan, hesap verebilirlik meselesi, XAI ile ilgili en tartışmalı konulardan biridir. Lima ve arkadaşları (2022), açıklanabilir sistemlerin bazen sorumluluğu yanlış yerlere kaydırabileceğini, özellikle de açıklama sunan algoritmaların kendi başına "blameworthy" yani suçlanabilir olarak algılanabileceğini öne sürmektedir. Bu durum, aslında sistemin arkasında yer alan geliştiricilerin veya kurumların sorumluluktan kaçmasına yol açabilir. Ayrıca, XAI'ın sunduğu açıklamalar kullanıcıya sahte bir kontrol hissi verebilir, bu da kurbanın sistem kararından kısmen sorumlu tutulmasına neden olabilir (Lima vd., 2022).

Bu bağlamda XAI, hem adaletin sağlanmasında bir araç hem de yeni etik risklerin kaynağı olabilir. XAI'ın adalet üzerindeki etkilerini değerlendirirken, hangi adalet ilkesine (örneğin grup adaleti mi bireysel adalet mi) hizmet ettiği, hangi paydaşları etkilediği ve nasıl bir rol üstlendiği (gözlemleyici mi yoksa müdahale edici mi) gibi sorular net bir şekilde tanımlanmalıdır (Deck vd., 2024). Aksi takdirde, sistemler açıklama sunsalar dahi önyargıları yeniden üretebilir veya sorumluluğu gerçek karar alıcılardan uzaklaştırabilir.

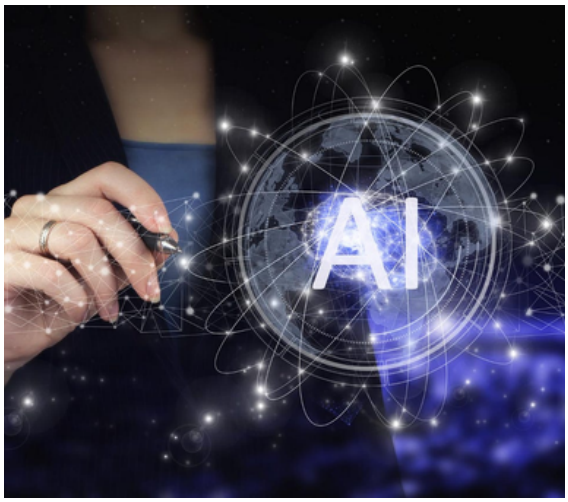
6.2. Yasal ve Hukuki Düzenlemeler

Açıklanabilir yapay zekâ sistemlerinin hukukla kesişiminde en çok tartışılan konulardan biri, otomatik kararların açıklanabilirliğine dair birey haklarıdır. Avrupa Birliği'nin GDPR düzenlemesi algoritmik karar verme süreçlerine ilişkin bazı şeffaflık ve hesap verebilirlik yükümlülükleri getirerek bu alanda önemli bir hukuki çerçeve sunmuştur (Kaminski, 2019).

Özellikle GDPR'ın 22. maddesi, yalnızca otomatik işleme dayalı ve birey üzerinde "önemli etki" yaratan kararların belirli durumlarda insan müdahalesiyle sınırlandırılabileceğini belirtir. Ancak bu düzenleme, neyin "önemli etki" sayılacağı veya hangi durumlarda "yalnızca otomatik" olduğu gibi birçok belirsizlik içerdiğinden, gerçek dünyadaki uygulamada sınırlı koruma sunmaktadır (Edwards & Veale, 2017).

Yine GDPR'ın 13, 14 ve 15. maddeleriyle getirilen bildirim ve erişim hakları, kararın "mantığına dair anlamlı bilgiler" verilmesini öngörür. Ancak bu açıklamaların kapsamı, algoritmaların teknik doğası ve ticari sır kaygıları nedeniyle çoğu zaman sınırlı kalmaktadır (Kaminski, 2019). Bu noktada, bazı araştırmacılar bireylere teknik sistemlerin iç mantığını değil, kararın sonucunu değiştirmek için hangi değişkenlerin etkili olduğunu gösteren "koşulsuz karşıolgusal açıklamalar" (unconditional counterfactuals) sunulmasını savunmuştur (Wachter, Mittelstadt & Russell, 2018). Örneğin, bir bireye "Geliriniz 30.000£ olduğu için kredi alamadınız, eğer 45.000£ olsaydı kredi alırdınız." şeklindeki açıklama, kararın iç işleyişini açmadan bile bireyin sistemle etkileşimini anlamlı şekilde artırabilir.

Bununla birlikte, bazı akademisyenler “açıklama hakkı” etrafında dönen tartışmaların, gerçek hukuki güvencelere dair yanlış bir algı oluşturduğunu savunur. Edwards ve Veale (2017), GDPR’daki bu hakkın büyük ölçüde yan açıklamalarda yer aldığını ve yasal bağlayıcılığının sınırlı olduğunu vurgulayarak, bireylere anlamlı koruma sağlamak için sistem düzeyinde denetim mekanizmalarının daha etkili olabileceğini belirtmiştir.



Sonuç olarak, XAI sistemlerinin hukuki düzenlemelere uyumlu hâle getirilmesi yalnızca birey odaklı açıklama hakları ile değil; aynı zamanda sistem tasarımında gizliliğe duyarlı mimarilerin benimsenmesi, etki analizleri yapılması ve bağımsız denetim yapılarına alan açılmasıyla mümkündür (Kaminski, 2019; Edwards & Veale, 2017). Bu yaklaşım, hem kullanıcıların algoritmalar karşısında güçlenmesini sağlayacak hem de XAI uygulamalarının toplumsal kabulünü artıracaktır.

6.3. Aşırı Basitleştirilmiş veya Yanıltıcı Açıklama Riskleri

Açıklanabilir yapay zekâ (XAI) yöntemlerinin kullanımı, çoğu zaman sistemin nasıl çalıştığına dair güven verici bir algı oluştursa da, bu yöntemlerin teknik sınırları göz önüne alındığında, aslında birçok risk barındırmaktadır (Chung vd., 2024). Özellikle siyah kutu olarak nitelendirilen derin öğrenme sistemlerinde, geliştirilen açıklama yöntemleri modelin davranışını yalnızca yüzeysel olarak yansıtmakta, çoğu durumda eksik ya da yanıltıcı bilgiler sunmaktadır (Chung vd., 2024).

Makalede vurgulandığı üzere, XAI yöntemleri genellikle insanlara güven vermek amacıyla sadeleştirilmiş açıklamalar üretir; fakat bu sadeleştirme, karmaşık karar süreçlerinin gerçek doğasını maskeleyebilir. Örneğin, bir modelin karar verirken hangi öznitelikleri dikkate aldığına dair sunulan saliency haritaları ya da dikkat haritaları, çoğu zaman modelin gerçekte hangi bilgilere odaklandığını yansıtmayabilir (Chung vd., 2024). Bu durum, kullanıcıların modelin performansına aşırı güven duymasına neden olabilir.

Bununla birlikte, makalede dikkat çekilen önemli bir diğer risk de açıklamaların hedef kitlenin ihtiyaçlarına göre yeterince özelleştirilmemiş olmasıdır. Farklı uzmanlık seviyelerine sahip kullanıcılar için farklı açıklama seviyeleri gerekmesine rağmen, günümüzdeki sistemler çoğunlukla tek tip açıklamalar üretmektedir (Chung vd., 2024). Bu durum, özellikle yüksek riskli uygulamalarda (örneğin sağlık ya da adalet sistemleri gibi) kararların doğru anlaşılmasını ve denetlenmesini zorlaştırabilir. XAI yöntemlerinin teknik açıdan karşılaştığı bir diğer sorun da açıklamaların kararlılığıdır. Ufak girdi değişiklikleri bile, açıklama sonuçlarını ciddi şekilde değiştirebilir. Bu da kullanıcıyı, modelin hangi verilere dayandığı konusunda yanlış yönlendirebilir (Chung vd., 2024).

Son olarak, insanlar açıklamaları değerlendirirken bilişsel önyargılara sahiptir. Sunulan açıklama, modelin gerçek mantığıyla örtüşmese bile, insan zihnine mantıklı geliyorsa kolaylıkla kabul görebilir. Bu da açıklamaların gerçeği yansıtma kapasitesinden çok, ikna ediciliğiyle değerlendirildiği anlamına gelir ve bu durum büyük bir güvenlik riski oluşturur (Chung vd., 2024).

Tüm bu nedenlerle, XAI sistemlerinin sunduğu açıklamaların mutlak doğruymuş gibi algılanması, hem teknik hem de etik açıdan ciddi riskler barındırmaktadır. Yapay zekâ sistemlerinin güvenli ve denetlenebilir olabilmesi için açıklamaların yalnızca sunulması değil, aynı zamanda bu açıklamaların doğruluğu, istikrarı ve hedef kullanıcıya uygunluğu da sistematik olarak değerlendirilmelidir (Chung vd., 2024).

7. ZORLUKLAR VE GELECEĞE YÖNELİK TAHMİNLER

7.1. XAI’nin Mevcut Sınırlamaları

Açıklanabilir Yapay Zekâ (XAI), karar alma süreçlerinde şeffaflık sağlamayı hedeflese de mevcut XAI yaklaşımları hâlâ çeşitli sınırlamalarla karşı karşıyadır. Bu sınırlamaların başında kavramsal netlik eksikliği gelmektedir. Literatürde "açıklanabilirlik" ve "yorumlanabilirlik" kavramlarının tanımları konusunda hâlâ fikir birliği bulunmamaktadır; aynı kavramlar farklı adlarla anılabilirken, farklı kavramlar da aynı adla ifade edilebilmektedir (Saeed & Omlin, 2023). Bu durum, yeni geliştirilen XAI yöntemlerinin ne ölçüde başarılı olduklarının değerlendirilmesini zorlaştırmaktadır.

Bir diğer önemli sınırlama ise değerlendirme kriterlerinin eksikliğidir. XAI sistemlerinin etkinliğini ölçmek için kullanılan metrikler çoğu zaman öznel temellidir ve standart bir değerlendirme çerçevesi bulunmamaktadır. Bunun temel nedenlerinden biri, çoğu zaman "doğru açıklama"ya dair bir "ground truth"un olmamasıdır (Saeed & Omlin, 2023). Mevcut değerlendirme ölçütleri güvenilirlik, anlaşılabilirlik, sadelik ve doğru temsil gibi çeşitli boyutları kapsasa da, bu metrikler arasında nasıl bir denge kurulacağı ya da hangi bağlamda hangisinin öncelikli olacağı hâlâ belirsizdir (Saeed & Omlin, 2023).

XAI sistemlerinin kullanıcı odaklılığı da ayrı bir sınırlama alanıdır. Farklı kullanıcı profilleri—örneğin alan uzmanları, karar vericiler ve sıradan kullanıcılar—farklı açıklama türlerine ihtiyaç duymaktadır. Ancak mevcut sistemler genellikle bu farklılıkları göz önünde bulundurmamaktadır (Yang vd., 2023). Açıklamaların kullanıcı deneyimine göre kişiselleştirilmemesi, açıklamaların anlaşılabilirliğini ve etkililiğini azaltmaktadır.

Bir başka sınırlayıcı unsur da XAI'nin yüksek doğruluk ve açıklanabilirlik arasında kurmaya çalıştığı denge ile ilgilidir. Derin öğrenme gibi kompleks modeller genellikle yüksek doğruluk sağlar ancak bu modellerin açıklanabilirliği oldukça düşüktür. Öte yandan, daha yorumlanabilir modeller genellikle doğruluk açısından yetersiz kalmaktadır. Bu ikilemin özellikle sağlık, finans gibi kritik alanlarda çözülmesi büyük önem taşımaktadır (Yang vd., 2023).

Son olarak, mevcut XAI yöntemlerinin pek çoğu görsel ve metinsel veri üzerinde yoğunlaşmakta, heterojen ve yapılandırılmamış veri türlerine uygulanabilirlik konusunda sınırlı kalmaktadır (Saeed & Omlin, 2023). Bu da XAI'nin çok alanlı uygulanabilirliğini azaltmakta ve yaygınlaştırılmasını zorlaştırmaktadır. Tüm bu sınırlamalar, XAI alanında daha sistematik, kullanıcı odaklı ve ölçülebilir çözümlere olan ihtiyacı açıkça ortaya koymaktadır.

7.2. Açıklanabilirlik-Doğruluk Denge İkilemi

Makine öğrenimi modelleri geliştirilirken karşılaşılan temel sorunlardan biri, açıklanabilirlik ile doğruluk arasındaki dengeyi sağlama zorunluluğudur. Daha doğrusu, yüksek performans gösteren modellerin çoğu zaman "kara kutu" yapıda olması, bu modellerin karar alma süreçlerini kullanıcıya veya uzmanlara açıklamada yetersiz kalmasına yol açmaktadır (Abusitta vd., 2024). Bu durum özellikle sağlık, savunma ve finans gibi yüksek riskli alanlarda güven sorunlarını beraberinde getirir.

Basit modeller olan doğrusal regresyon ya da Naive Bayes gibi yöntemler, genellikle anlaşılır yapıları sayesinde açıklanabilirlik açısından güçlüdür. Ancak bu modeller karmaşık ve gürültülü verilerle başa çıkmakta zorlandığından, doğrulukları sınırlı kalabilir (Abusitta vd., 2024). Öte yandan, derin sinir ağları gibi gelişmiş yaklaşımlar, çok yüksek doğruluklara ulaşabilirken, modelin iç işleyişini anlamak oldukça zordur (Arrieta vd., 2019). Bu sebeple araştırmacılar ve geliştiriciler, bir modelin şeffaflığı ile performansı arasında genellikle bir ödünleşim yapmak zorunda kalmaktadır.

XAI alanındaki çalışmalarda, bu ikilem açıkça vurgulanmaktadır. Geliştirilen tekniklerin çoğu, kara kutu modellerin iç işleyişini sonradan açıklamaya çalışan "post-hoc" yöntemlere yönelmiştir. Ancak bu yöntemlerin sunduğu açıklamaların, modelin karar süreciyle ne derece örtüştüğü hala tartışmalıdır (Arrieta vd., 2019). Diğer yandan, modelin açıklanabilirliğini artırmak adına yapısı basitleştirildiğinde, bu durum genellikle doğrulukta kayıplara neden olur (Abusitta vd., 2024).

Bu bağlamda, ideal bir makine öğrenimi modelinin hem yüksek doğruluk sağlaması, hem de kullanıcıya karar gerekçelerini açık bir şekilde sunabilmesi beklenmektedir (Abusitta vd., 2024). Ancak şu anki teknik seviyede bu iki özelliği aynı anda optimize etmek oldukça zordur. Bu nedenle, açıklanabilirlik-doğruluk dengesini kurmak, sadece teknik değil, aynı zamanda etik ve kullanım bağlamına bağlı stratejik bir karar halini almaktadır (Arrieta vd., 2019).

Sonuç olarak, XAI araştırmalarının önündeki en büyük engellerden biri bu yapısal çelişkiyi aşmaktır. Gelecekte geliştirilecek XAI çözümlerinin, bu dengeyi daha başarılı şekilde kurabilmesi; yalnızca model performansını değil, aynı zamanda insan-merkezli güven inşasını da destekleyecektir.

7.3. İnsan Merkezli, Bağlam Farkında AI Açıklamaları

XAI sistemlerinin gerçek dünyada benimsenebilmesi için yalnızca teknik doğruluk değil, aynı zamanda kullanıcı ihtiyaçlarını anlayan insan-merkezli açıklamalar üretmeleri gerekmektedir.

Warren, Keane ve Byrne (2022), kullanıcıların bir sistemin kararlarını anlamasında açıklama türlerinin ve kullanılan özniteliklerin belirleyici olduğunu göstermektedir. Özellikle, nedensel ve karşıolgusal açıklamalar karşılaştırıldığında, kullanıcıların kararları tahmin etme doğruluğunda anlamlı bir fark oluşmazken; karşıolgusal açıklamalar kullanıcılar tarafından daha tatmin edici ve güvenilir bulunmuştur (Warren vd., 2022). Bu durum, bazı açıklamaların objektif bilgi kazandırmaktan çok, öznel düzeyde güven ve memnuniyet hissi uyandırdığını göstermektedir.

Bu bulgu, Rong vd. (2024) tarafından yapılan daha geniş çaplı sistematik inceleme ile de desteklenmektedir. Çalışma, açıklamaların çoğunlukla öznel anlayışı artırdığı, ancak kullanıcıların sistemin işleyişine dair gerçek bilgilerini her zaman geliştirmediğini ortaya koymuştur. Özellikle algılanan adalet ya da memnuniyet gibi hislere yönelik kazanımlar ile objektif görev başarımı arasındaki farklar, insan-merkezli XAI sistemlerinin tasarımında dikkat edilmesi gereken önemli bir gerilim alanı oluşturur (Rong vd., 2024).

Ayrıca, Warren ve arkadaşlarının yaptığı deneysel çalışma, kullanıcıların sürekli nicel (örneğin ağırlık, içki birimi gibi) özellikler yerine kategorik özelliklere (örneğin cinsiyet, mide doluluğu gibi) dayanan açıklamaları daha kolay anladığını göstermiştir. Bu durum, açıklamaların yalnızca teknik olarak anlamlı değil, aynı zamanda insanın bilişsel yapısına uygun bir şekilde sunulması gerektiğini ortaya koyar (Warren vd., 2022). Cinsiyet gibi kategorik özelliklerin yorumlanabilirliği, kullanıcıların karar sınırlarını daha net kavramasını sağlamaktadır.

Rong ve arkadaşları da açıklamaların etkililiğini değerlendirirken kullanıcıların görev performansı, algılanan güven, kullanım kolaylığı gibi farklı boyutların dikkate alınması gerektiğini belirtmektedir. Ancak bu değerlendirmelerin çoğunun psikolojik ya da bilişsel bilim temelli değil, mühendislik odaklı olduğu vurgulanmıştır. Bu bağlamda, XAI araştırmalarının insan-bilgisayar etkileşimi ve bilişsel psikoloji alanlarından daha fazla beslenmesi gerektiği önerilmektedir (Rong vd., 2024).

Sonuç olarak, insan-merkezli ve bağlam farkında açıklamalar üretmek, sadece teknik bir gereklilik değil, aynı zamanda etik ve kullanıcı deneyimi açısından da zorunludur. Açıklamanın anlaşılabilir olması, açıklamanın doğruluğu kadar önemlidir. Açıklamaların kullanıcıya göre özelleştirilmesi, hangi açıklamanın hangi kullanıcı tipi için daha faydalı olduğunu anlamayı gerektirir. Bu da XAI'nin yalnızca algoritma performansıyla değil, insan psikolojisi ve bağlamsal farkındalıkla birlikte değerlendirilmesini gerektirir.

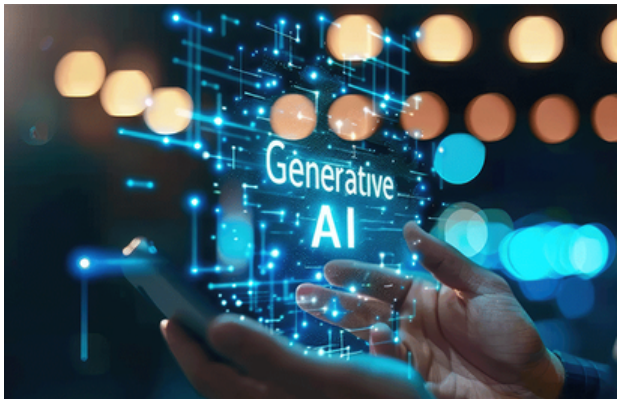
7.4. Nedensel XAI, Etkileşimli Açıklamalar, Üretken Yapay Zekada Açıklanabilirlik

XAI'nin geleceği açısından nedensel açıklamalar önemli bir araştırma yönü olarak öne çıkmaktadır. Özellikle, bir modelin çıktılarının sadece hangi girdilere bağlı olduğunu değil, bu girdilerin sonuçlar üzerindeki nedensel etkisini anlamaya odaklanan yaklaşımlar, kullanıcıların karar alma süreçlerinde daha güvenilir bilgi edinmelerine olanak tanır (Beckers, 2022). Bu doğrultuda, Beckers çalışmasında “Sufficient Explanations” ve “Counterfactual Explanations” kavramlarını nedensel modelleme çerçevesinde yeniden tanımlayarak, eylem yönlendirici açıklamaların nasıl üretilebileceğini tartışmaktadır.

Özellikle, bir eylemin belli bir çıktıyı garanti altına aldığı durumlar için yeterli açıklamalar, var olan bir çıktının alternatif bir senaryo altında nasıl değişebileceğini değerlendiren durumlar için ise karşı-olgusal açıklamalar ön plana çıkarılmaktadır (Beckers, 2022).

Ayrıca, Beckers çalışması “actual causation” kavramını da bu çerçevede yeniden ele almakta ve yalnızca korelasyonlara değil, eylemle yönlendirilebilecek nedensel yapıya odaklanan açıklamaların daha etkili olduğu vurgulanmaktadır. Bu tür açıklamaların, özellikle adalet gibi etik kaygıların ön planda olduğu uygulamalarda da yol gösterici olabileceği belirtilmektedir (Beckers, 2022).

Etkileşimli açıklamalar ise XAI'nin kullanıcıyla daha derin ilişkiler kurabilmesi adına önem arz eder. Longo ve arkadaşları, etkileşimli XAI'nin kullanıcıların modelin iç işleyişini yalnızca gözlemlemekle kalmayıp, aynı zamanda açıklamalara dair sorular sormasına ve açıklama biçimini şekillendirmesine olanak sağladığını belirtmektedir (Longo vd., 2023). Bu, kullanıcı ile sistem arasında çift yönlü bir açıklama süreci kurarak sadece açıklamaların doğruluğunu değil, kullanıcıya olan faydasını da artırmayı hedeflemektedir. Bu bağlamda "multi-faceted explanations" ve "computational argumentation" gibi yaklaşımlar, modelin karar sürecine dair farklı perspektifleri açıklamak için geliştirilmiştir (Longo vd., 2023).



Üretken yapay zeka modelleri (örneğin GPT türü modeller veya generatif görsel modeller) bağlamında açıklanabilirlik ise henüz tam anlamıyla çözülememiş karmaşık bir zorluk olarak önümüzde durmaktadır. Bu modellerin sahip olduğu milyarlarca parametre, klasik açıklama yöntemlerinin yetersiz kalmasına neden olmaktadır (Longo vd., 2023). Özellikle polisemantik nöronlar (birden fazla kavramı temsil eden nöronlar) nedeniyle açıklamalarda kavramsal tutarlılığı sağlamak zorlaşmakta, bu da kullanıcıların yorumlarını zorlaştırmaktadır. Bu sorunun çözümüne yönelik olarak önerilen "mechanistic interpretability" yaklaşımı, modelin iç yapısını tersine mühendislikle analiz ederek açıklamaları kavramsal temellere dayandırmayı hedeflemektedir (Longo vd., 2023).

Ancak bu sistemlerin güvenilirliği yalnızca teknik doğruluklarıyla sınırlı değildir. Baniecki ve Biecek'in çalışması, XAI sistemlerinin açıklamalarının kötü niyetli aktörlerce manipüle edilebildiğini ortaya koyarak önemli bir uyarıda bulunmaktadır (Baniecki & Biecek, 2024). Özellikle karşı-olgusal açıklamalar gibi güçlü araçlar bile, adversarial saldırılarla yanlış yönlendirici hale getirilebilmektedir. Bu nedenle, üretken yapay zeka sistemlerinde açıklanabilirliğin yalnızca nedensel ya da teknik temellere değil, aynı zamanda güvenlik ve manipülasyon direncine de dayanması gerektiği önerilmektedir (Baniecki & Biecek, 2024).

Sonuç olarak, nedensel açıklamalar daha güçlü eylem rehberliği sunarken, etkileşimli açıklamalar kullanıcıyı merkeze koyarak açıklamanın kalitesini artırmakta; üretken yapay zeka bağlamında açıklamalar ise çok boyutlu, güvenli ve insan-anlamlı hale getirilmeye çalışılmaktadır. Bu üç alanın kesişimi, gelecekte XAI sistemlerinin hem teknik olarak sağlam hem de insan-merkezli ve güvenli olabilmesi için kritik bir gelişim alanıdır.

8. SONUÇLAR

Bu çalışma kapsamında açıklanabilir yapay zekâ, yalnızca teknik bir yetenek olarak değil; aynı zamanda etik, sosyal ve hukuki bir zorunluluk olarak ele alınmış ve çok katmanlı bir perspektifle incelenmiştir. Derin öğrenme ve diğer gelişmiş yapay zekâ modelleri, sundukları yüksek doğruluk oranlarına rağmen, karar alma süreçlerinin kullanıcılar tarafından anlaşılabilir olmaması nedeniyle güven, denetlenebilirlik ve sorumluluk açısından ciddi sorunlar yaratmaktadır. Bu bağlamda XAI, yalnızca modellerin açıklanabilirliğini artırmakla kalmamakta; aynı zamanda yapay zekâ sistemlerinin toplumsal kabulünü ve kurumsal uygulanabilirliğini de güçlendiren temel bir yapı taşı hâline gelmektedir.

Çalışma boyunca, XAI yöntemleri modele bağımlı/bağımsız ve içsel/post-hoc olmak üzere farklı sınıflandırmalar altında detaylandırılmış; LIME, SHAP, Grad-CAM, Integrated Gradients, Attention mekanizmaları, LRP gibi çeşitli teknikler hem teorik olarak açıklanmış hem de uygulamalı örneklerle pekiştirilmiştir. Özellikle tabular veri, görüntü ve metin gibi farklı veri türleri üzerinde bu yöntemlerin nasıl kullanıldığı, hangi senaryolarda avantaj veya dezavantaj oluşturduğu kapsamlı şekilde analiz edilmiştir.

Yerel ve global açıklama düzeyleri arasında yapılan ayrım, kullanıcı ihtiyaçlarına göre açıklama stratejilerinin farklılaştırılmasının önemine işaret etmektedir. Bireysel kararlar için yerel açıklamalar güven inşa ederken; modelin genel davranışını anlamak için global açıklamalara duyulan ihtiyaç, sistemlerin bütüncül değerlendirilmesi açısından vazgeçilmezdir. Bu iki yaklaşımın birlikte kullanımı, hem mikro düzeyde bireylerin kararları sorgulamasını hem de makro düzeyde sistemin tutarlılığının izlenmesini mümkün kılmaktadır.

Makalenin uygulama alanlarına dair bölümlerinde, XAI'nin sektörel katkıları çarpıcı biçimde ortaya konmuştur. Sağlık alanında açıklanabilirlik, klinik karar destek sistemlerinin güvenilirliğini artırmakta ve hasta güvenliği açısından yaşamsal öneme sahip olmaktadır. Eğitimde ise hem öğrencilerin öğrenme sürecine dair farkındalık geliştirmesi hem de öğretmenlerin pedagojik kararlarını iyileştirmesi XAI sayesinde mümkün olmaktadır.

Finans sektöründe SHAP ve LIME gibi yöntemlerle risk değerlendirme süreçleri daha şeffaf hale getirilirken; hukuk ve savunma alanlarında açıklanabilirlik, yalnızca algoritmik kararların rasyonelleştirilmesini değil, aynı zamanda etik ve yasal sorumlulukların yerine getirilmesini de mümkün kılmaktadır. Otonom araçlarda ise gerçek zamanlı açıklamaların sistem güvenliği, kullanıcı memnuniyeti ve yasal denetim açısından kritik olduğu gösterilmiştir.

Çalışmada ayrıca, XAI tekniklerinin kararlılığı, doğruluğu ve insan akıl yürütmesiyle ne ölçüde örtüştüğü gibi önemli değerlendirme kriterleri de ele alınmış; bazı yöntemlerin eğitim sürecindeki rastlantısallıklardan etkilendiği ve yanıltıcı açıklamalar üretebildiği gösterilmiştir. Bu durum, açıklamaların körü körüne kabul edilmesinin yaratabileceği risklere işaret etmektedir. Aynı şekilde, açıklamaların hedef kitlenin uzmanlık düzeyine uygun olmaması hâlinde güvenin azalabileceği ve XAI sistemlerinin benimsenmesinin zorlaşabileceği de vurgulanmıştır.

Etik boyutta ise XAI'nin adalet, önyargı ve hesap verebilirlik gibi kavramlarla ilişkisi detaylı biçimde irdelenmiştir. Açıklamalar, geçmişteki önyargıları yeniden üretme riski taşıyan modellerin denetlenebilirliğini artırmakla birlikte, sorumluluğu birey yerine algoritmaya yükleyen yanıltıcı bir araç hâline de gelebilir. Bu nedenle XAI sistemlerinin tasarımı kadar, sunduğu açıklamaların içeriği, zamanlaması, dili ve sunum şekli de etik güvenilirlik açısından belirleyici olmaktadır.

Geleceğe dönük olarak, XAI sistemlerinin daha güvenilir hâle getirilmesi için çoklu model entegrasyonu, açıklama tutarlılığı ölçütleri, insan odaklı arayüz tasarımları ve sektöre özgü açıklama politikalarının geliştirilmesi gerektiği sonucuna varılmıştır. Açıklanabilirlik yalnızca teknik bir özellik değil; yapay zekâ sistemlerinin meşruiyeti, kabul edilebilirliği ve insanla birlikte çalışabilirliği için temel bir gerekliliktir. Bu bağlamda XAI, yapay zekânın geleceğini şekillendirecek en kritik alanlardan biri olarak öne çıkmaktadır.

9. KAYNAKÇA

- Alicioglu, G., & Sun, B. (2021). A survey of visual analytics for explainable artificial intelligence methods. Computers & Graphics, 99, 1–19. https://doi.org/10.1016/j.cag.2021.09.002.
- Arnaud, L., Moens, F., Clarysse, B., & Mertes, P. (2024). Consistency of XAI Models Against Medical Expertise: An Assessment Protocol.
- Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, 101286. https://doi.org/10.1016/j.imu.2023.101286.
- Bhati, D., Neha, F., & Amiruzzaman, M. (2024). A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. Journal of Imaging, 10(239). https://doi.org/10.3390/jimaging10100239.
- Caterson, J., Lewin, A., & Williamson, E. (2024). The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review. Digital Health, 10, 1–12. https://doi.org/10.1177/20552076241272657.
- Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57, 216. https://doi.org/10.1007/s10462-024-10854-8.
- DASA. (2022). International Conference on Decision Aid Sciences and Applications. XAI approach to improved and informed detection of burnt scar.
- Devireddy, K. (2025). A Comparative Study of Explainable AI Methods: Model-Agnostic vs. Model-Specific Approaches.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv Preprint. arXiv:1702.08608. https://arxiv.org/abs/1702.08608.
- Gao, Y., Liu, J., Li, W., Hou, M., Li, Y., & Zhao, H. (2023). Augmented Grad-CAM++: Super-Resolution Saliency Maps for Visual Interpretation of Deep Neural Network. Electronics, 12(4846). https://doi.org/10.3390/electronics12234846.
- Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions
- Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., & Batarseh, F. A. (2023). Rationalization for explainable NLP: A survey. Frontiers in Artificial Intelligence, 6, 1225093. https://doi.org/10.3389/frai.2023.1225093.
- Hall, P. (2020). On the Art and Science of Explainable Machine Learning: Techniques, Recommendations, and Responsibilities. KDD '19 XAI Workshop, Anchorage, AK. https://arxiv.org/abs/1810.02909.
- Hussain, S., Aziz, S., Bharathy, G., & Kolluru, D. M. (2022). A Systematic Literature Review on Explainable Artificial Intelligence Techniques: In reference to Banking and Financial Services. SSRN. https://ssrn.com/abstract=4910749.
- Ikumapayi, N. A., & Muhammed, B. (2024). Explainable AI: Making Machine Learning Decisions Transparent. SSRN.
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., ... & Gašević, D. (2022). Explainable Artificial Intelligence in Education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074.
- Madathil, A. P., Luo, X., Liu, Q., Walker, C., Madarkar, R., Cai, Y., Liu, Z., Chang, W., & Qin, Y. (2024). Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. Journal of Intelligent Manufacturing, 35, 4159–4180. https://doi.org/10.1007/s10845-023-02266-2.
- Mersha, M.A., Bitewa, M., Abay, T., & Kalita, J. (2024b). Explainability in Neural Networks for Natural Language Processing Tasks. arXiv preprint arXiv:2412.18036.

- Mersha, M. A., Yigezu, M. G., & Kalita, J. (2025). Evaluating the Effectiveness of XAI Techniques for Encoder-Based Language Models. arXiv preprint arXiv:2501.15374.
- Noor, A. A., Manzoor, A., Qureshi, M. D. M., Qureshi, M. A., & Rashwan, W. (2025). Unveiling explainable AI in healthcare: Current trends, challenges, and future directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(2), e70018. https://doi.org/10.1002/widm.70018.
- Saarela, M., & Podgorelec, V. (2024). Recent Applications of Explainable AI (XAI): A Systematic Literature Review. Applied Sciences, 14(8884). https://doi.org/10.3390/app14198884.
- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. arXiv. https://arxiv.org/abs/2305.02012Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., & Keim, D. A. (2019). Towards a rigorous evaluation of XAI methods on time series.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv preprint arXiv:1610.02391v4.
- Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. Cognitive Systems Research, 86, 101243. https://doi.org/10.1016/j.cogsys.2024.101243.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 1135–1144. https://doi.org/10.1145/2939672.2939778.
- Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. JMIR AI, 3(1), e53207. https://doi.org/10.2196/53207
- Senbagavalli, M., Mahapatra, S., Fathima, A. F. M. K., & Ruchita, L. (2025). Revolutionizing road safety: An innovative approach to pothole detection and prevention using XAI and deep learning.
- Szczepankiewicz, K., Popowicz, A., Charkiewicz, K., Nałęcz-Charkiewicz, K., Szczepankiewicz, M., Lasota, S., Zawistowski, P., & Radlak, K. (2023). Ground truth based comparison of saliency maps algorithms. Scientific Reports, 13, 16887. https://doi.org/10.1038/s41598-023-42946-w.
- Wang, Y. (2024). A Comparative Analysis of Model Agnostic Techniques for Explainable Artificial Intelligence. Research Reports on Computer Science, 3(2). https://doi.org/10.37256/rrcs.3220244750.
- Woerl, A.-C., Disselhoff, J., & Wand, M. (2023). Initialization Noise in Image Gradients and Saliency Maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1766–1775.