D-Tek Bilgi Merkezi

Tıbbi Laboratuvarlarda Yapay Zekâ Sistemlerinin Performansının Değerlendirilmesi: Bileşenler ve Ölçütler

Prof. Dr. Diler Aslan

D-Tek Teknoloji Geliştirme, Üretim, Eğitim ve Danışmanlık Hizmetleri Ltd. Şti.

Yayın tarihi	30.08.2026
Son güncelleme	30.08.2026
Doküman sürüm	00

Tıbbi Laboratuvarlarda Yapay Zekâ Sistemlerinin Performansının Değerlendirilmesi: Bileşenler ve Ölçütler

Prof. Dr. Diler Aslan
29.08.2026

İçindekiler

“Yapay zekâ doğruluğu” çoklu bileşeni kapsar. Amaçlanan kullanıma göre birden fazla performans bileşeniyle değerlendirilir. **Yapay zekânın doğruluğu; ayırt etme ve kalibrasyon performansı, hata büyüklüğü, tekrarlanabilirlik, genellenebilirlik,**

sağlamlık, hasta gruplarına göre performans ve insan–sistem ortak performansı birlikte değerlendirilerek gösterilir.

Bunların tamamı her yapay zekâ sistemi için kullanılmaz. Seçilecek ölçütler, sistemin **amaçlanan kullanımına ve ürettiği çıktının türüne** göre belirlenir.

ISO 24051-1 – Tıbbi Laboratuvarlar – Part 1: Yapay Zekânın Tıbbi Laboratuvarlarda Uygulanmasına İlişkin Genel İlkeler Standardı bağlamında doğruluk değerlendirmesi; modelin yerel verilerde doğru çıktı üretmesinin yanı sıra tekrarlanabilirliğini, farklı hasta gruplarındaki performansını ve insanla birlikte kullanımını da kapsar.

ISO/DIS 24051-1 Açısından Temel Performans Bileşenleri

ISO 24051-1'de yerel doğrulama sırasında açıkça öne çıkan bileşenler şunlardır:

1. Yerel verilerde doğruluk
2. Hakkaniyet ve yanlılık
3. Farklı cihazlarda tekrarlanabilirlik
4. Farklı laboratuvar birimlerinde tekrarlanabilirlik
5. Yerel hasta alt gruplarında performans
6. İnsan katılımının bulunduğu sistemlerde insan–sistem ortak performansı
7. Uygun referans yöntemlerle karşılaştırma
8. Yapay zekâ kullanılarak ve kullanılmadan elde edilen çıktıların karşılaştırılması
9. Genellenebilirlik
10. Veri kayması ve veri sürüklenmesi
11. Sürekli performans izleme

Yapay Zeka Doğruluğu Bileşenleri çeşitli kaynaklardan derlenerek izleyen paragraflarda sunulmaktadır. Kaynaklara göre önerilen bileşenler ayrıntılı olarak bölümün sonundadır.

Yapay Zekâ Sistemlerinde Performans Değerlendirme Bileşenleri

1. Ayırt etme performansı

Sistemin pozitif ve negatif durumları birbirinden ayırabilme yeteneğidir.

- **Duyarlılık (sensitivity):** Gerçek pozitiflerin ne kadarını doğru bulduğu
- **Özgüllük (specificity):** Gerçek negatiflerin ne kadarını doğru tanıdığı
- **Pozitif kestirim değeri (PPV):** Pozitif çıktının gerçekten pozitif olma olasılığı
- **Negatif kestirim değeri (NPV):** Negatif çıktının gerçekten negatif olma olasılığı
- **Yanlış pozitif oranı**
- **Yanlış negatif oranı**

- **ROC eğrisi altında kalan alan (AUC):** Sistemin farklı eşiklerde ayırt etme gücü

Genel doğruluk (accuracy) ise:

$$\text{Doğruluk} = \frac{\text{Doğru pozitif} + \text{Doğru negatif}}{\text{Bütün olgular}}$$

şeklinde hesaplanır. Ancak sınıflar dengesizse yanıltıcı olabilir. Örneğin hastalık prevalansı %2 olan bir grupta herkese "negatif" diyen sistem %98 genel doğruluk gösterebilir, fakat hiçbir hastayı yakalayamaz.

2. Kalibrasyon

Kalibrasyon, modelin tahmin ettiği değerlerle gözlenen sonuçlar arasındaki uyumdur. Modelin verdiği olasılıkların gerçek olay sıklığıyla uyumudur.

Örneğin sistem, 100 hastanın her biri için hastalık olasılığını %20 olarak bildiriyorsa bu grubun yaklaşık 20'sinde hastalık bulunması beklenir.

Bir model:

- Hastaları iyi ayırt edebilir,
- Fakat verdiği risk olasılıkları gerçeği yansıtmayabilir.

Bu nedenle özellikle risk tahmini yapan sistemlerde **ayırt etme ve kalibrasyon birlikte** değerlendirilmelidir.

3. Sürekli sayısal çıktılarda doğruluk

Yapay zekâ sınıflandırma yerine sayısal bir sonuç üretiyorsa duyarlılık ve özgüllük yeterli olmaz.

Değerlendirmede:

- **Bias:** Yapay zekâ sonucu ile karşılaştırma sonucu arasındaki sistematik fark
- **Ortalama mutlak hata (MAE)**
- **Kök ortalama kare hata (RMSE)**
- **Uyum sınırları**
- **Klinik olarak kabul edilebilir hata oranı**
- **Korelasyonla birlikte yöntemler arası uyum**

incelenir.

Burada yüksek korelasyon, tek başına doğruluğu göstermez. İki yöntem arasında güçlü korelasyon bulunmasına rağmen klinik açıdan önemli sistematik fark olabilir.

4. Görüntü analizinde doğruluk

Nesne belirleme veya görüntü bölütleme sistemlerinde farklı ölçütler gerekir:

- Lezyon veya hücreyi doğru belirleme oranı
- Kaçırılan ve yanlış belirlenen nesne sayısı
- **Dice katsayısı** (Yapay zekânın belirlediği alan ile referans alan arasındaki benzerliği ölçer.)
- **Intersection over Union (IoU)** (Dice, iki alanın benzerliğini; IoU ise ortak alanın toplam kapsanan alana oranını gösterir.)
- Konumlandırma doğruluğu
- Sınırların uzman işaretlemesiyle uyumu
- Olgu veya görüntü başına yanlış pozitif sayısı

Bu ölçütlerin klinik anlamı da değerlendirilmelidir. Yüksek görüntü örtüşmesi her zaman doğru klinik sınıflandırma anlamına gelmez.

5. Tekrarlanabilirlik

Aynı veya benzer koşullarda sistemin tutarlı sonuç üretmesidir.
ISO 24051-1 özellikle şu karşılaştırmaları belirtir:

- Farklı cihaz veya analizörlerde
- Farklı laboratuvar birimlerinde
- Yerel toplumun farklı hasta alt gruplarında
- İnsan değerlendirmesinin bulunduğu sistemlerde farklı kullanıcılarla

elde edilen sonuçların tekrarlanabilirliği.

Model bir kez doğru sonuç verse bile farklı cihaz, kullanıcı veya merkezlerde aynı performansı göstermiyorsa güvenilir kabul edilemez.

6. Genellenebilirlik

Modelin geliştirildiği ortam dışında da kabul edilebilir performans göstermesidir.
Şunlara göre değerlendirilir:

- Farklı yaş grupları
- Cinsiyet
- Hastalık şiddeti ve alt türleri
- Farklı prevalanslar
- Farklı laboratuvarlar
- Farklı analizör, tarayıcı veya görüntüleme sistemleri
- Farklı örnek ve veri özellikleri

Genel performans iyi olsa bile belirli bir hasta grubunda performans düşük olabilir. Bu nedenle sonuçlar alt gruplara ayrılarak incelenmelidir.

7. Sağlamlık

Sistemin normal uygulamada karşılaşılabilecek küçük değişikliklere dayanabilmesidir.

Örnekler:

- Görüntü kalitesindeki değişiklikler
- Eksik veya gürültülü veri
- Ölçüm cihazındaki küçük farklılıklar
- Veri formatındaki değişiklikler
- Beklenen aralığın sınırındaki sonuçlar
- Daha önce görülmemiş girdi örüntüleri

Sistem bu koşullarda beklenmeyen veya aşırı güvenli çıktılar üretiyorsa genel doğruluk oranı yüksek olsa da güvenli değildir.

8. Yanlılık ve hasta gruplarına göre adil performans

Doğruluk, bütün hasta grubu için hesaplanan ortalamayla sınırlandırılmamalıdır.

Her önemli alt grupta:

- Duyarlılık
- Özgüllük
- Yanlış pozitif ve yanlış negatif oranları
- Kalibrasyon
- Hata büyüklüğü

karşılaştırılmalıdır.

Gruplar arasındaki her fark yanlılık anlamına gelmez; farkın veri yapısından mı, hastalık özelliklerinden mi, modelden mi veya sistematik bir adaletsizlikten mi kaynaklandığı araştırılmalıdır.

9. İnsan–yapay zekâ ortak doğruluğu

İnsan gözetimi bulunan uygulamalarda değerlendirilmesi gereken asıl performans, modelin tek başına doğruluğu değil, **insan–sistem bütününün doğruluğudur**.

Şunlar karşılaştırılabilir:

- Uzman tek başına
- Yapay zekâ tek başına
- Yapay zekâ destekli uzman

Aynı zamanda:

- Uzmanın yapay zekâ hatasını düzeltme oranı
- Yapay zekânın uzmanın doğru kararını yanlış yönde değiştirme oranı
- İnsan ile sistem arasındaki uyuşmazlıklar
- Otomasyon yanlılığı
- Sonuç verme süresi
- Kullanıcılar arası değişkenlik

değerlendirilmelidir.

10. Klinik doğruluk ve klinik yarar

Teknik olarak doğru bir model, klinik açıdan yararlı olmayabilir. Bu nedenle şu sorular da yanıtlanmalıdır:

- Çıktı klinik kararı doğru yönde değiştiriyor mu?
- Yanlış çıktı hastaya ne ölçüde zarar verebilir?
- Gereksiz test veya girişim oluşturuyor mu?
- Önemli bir tanının atlanmasını azaltıyor mu?
- Mevcut yöntemden daha iyi veya en azından eşdeğer mi?
- Laboratuvar ve klinik iş akışına uygun mu?

Bunların tamamı her yapay zekâ sistemi için kullanılmaz. Seçilecek ölçütler, sistemin **amaçlanan kullanımına ve ürettiği çıktının türüne** göre belirlenir.

Kaynaklara Göre Performans Bileşenleri ve Ölçütleri

1. STARD-AI

Kaynak: Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine.* 2025;31:3283–3289. doi:10.1038/s41591-025-03953-8.

STARD-AI, yapay zekâ merkezli tanısal doğruluk çalışmalarının raporlanmasına yöneliktir. Şunları destekler:

- Referans standardın tanımlanması
- Yapay zekânın indeks test olarak değerlendirilmesi
- Veri setlerinin ve veri bölünmesinin açıklanması
- Tanısal doğruluk ölçütlerinin verilmesi
- Algoritmik yanlılık ve adillik değerlendirmesi
- Bulguların uygulanabilirliği ve genellenebilirliği

- Hata analizinin raporlanması

Duyarlılık, özgüllük, pozitif ve negatif kestirim değerleri gibi ölçütleri tanısal doğruluk çerçevesi kapsamındadır.

2. TRIPOD+AI

Kaynak: Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378. TRIPOD+AI tam metni

Tanısal veya prognostik tahmin modellerinin geliştirilmesi ve değerlendirilmesine yöneliktir. Şunları destekler:

- Modelin ayırt etme performansı
- Kalibrasyon
- Model performans ölçütleri ve bunların güven aralıkları
- İç ve dış değerlendirme
- Veri kaynakları ve hasta toplumunun tanımlanması
- Alt grup değerlendirmeleri
- Eksik verilerin yönetimi
- Modelin hedef toplumdaki uygulanabilirliği

İlk Başlıktaki **ayırt etme ve kalibrasyon ayrımı**, esas olarak bu çerçeveye dayanmaktadır. TRIPOD+AI da ISO 24051-1 DIS'ta yer alır.

3. CONSORT-AI

Kaynak: Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*. 2020;26:1364–1374. doi:10.1038/s41591-020-1034-x.

İnsan–yapay zekâ etkileşimini ve klinik uygulamanın değerlendirilmesini içerir. Özellikle:

- Sistemin kullanıldığı klinik ortam
- Kullanıcının gerekli yetkinlikleri
- Yapay zekâ girdilerinin ve çıktılarının yönetimi
- İnsan–yapay zekâ etkileşimi
- Hatalı veya başarısız olguların analizi

konularının raporlanmasını ister.

4. DECIDE-AI

Kaynak: Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine.* 2022;28:924–933. doi:10.1038/s41591-022-01772-9. DECIDE-AI

Yapay zekânın gerçek klinik ortamda değerlendirilmesine yöneliktir. Derleme başlığındaki şu konularla ilişkilidir:

- Yapay zekânın gerçek iş akışı içindeki performansı
- İnsan faktörleri
- Kullanıcı–sistem etkileşimi
- Teknik performans ile gerçek klinik performansın ayrılması
- Beklenmeyen ve başarısız olgular
- Sistemin güvenliği
- Klinik yarar

Bu kaynak, “model tek başına” ile “insan–yapay zekâ sistemi” performansının ayrılması gerektiği görüşünü destekler.

5. PROBAST+AI

Kaynak: Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ.* 2025;388:e082505. doi:10.1136/bmj-2024-082505.

Şu değerlendirmeleri destekler:

- Katılımcılar ve veri kaynakları
- Kestiriciler
- Sonuç ölçütü
- Analiz yöntemi
- Yanlılık riski
- Modelin hedef topluma uygulanabilirliği
- Geliştirme ve değerlendirme verilerinin uygunluğu

Hasta alt grupları, temsil gücü, yanlılık ve genellenebilirlik açıklamalarında yararlanılabilecek temel kaynaklardan birisidir.

6. Tıbbi görüntü bölütleme ölçütleri

Kaynak: Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging.* 2015;15:29. doi:10.1186/s12880-015-0068-x.

İlk başlıkta kullanılan:

- Dice katsayısı
- Örtüşme ölçütleri
- Bölütleme sınırlarının karşılaştırılması
- Nesne ve yüzey temelli görüntü değerlendirmeleri

bu alana özgü literatüre dayanmaktadır. Bunlar **ISO 24051-1'de belirtilmez**.

7. FDA yapay zekâ destekli tıbbi cihaz yazılımı rehberi

Kaynak: FDA. *Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations*. Draft Guidance, 2025.

Şu konuları destekler:

- Amaçlanan kullanıma uygun performans ölçütlerinin seçilmesi
- Bağımsız test verileri
- Veri kaynaklarının ve alt grupların değerlendirilmesi
- Yanlılık
- İnsan faktörleri
- Yerel ve gerçek yaşam performansı
- Değişiklik yönetimi
- Yaşam döngüsü boyunca performans izleme

Performans Ölçütleri–Kaynak Eşleştirmesi

Önceki yanıttaki konu	**Başlıca kaynak**
Duyarlılık, özgüllük, PPV, NPV	STARD-AI ve tanısal doğruluk metodolojisi
AUC ve ayırt etme	TRIPOD+AI
Kalibrasyon	TRIPOD+AI
Yanlılık ve uygulanabilirlik	STARD-AI, PROBAST+AI
Alt grup performansı	STARD-AI, PROBAST+AI, FDA
İnsan–yapay zekâ etkileşimi	CONSORT-AI, DECIDE-AI
Klinik ortamda performans	DECIDE-AI
Dice ve görüntü bölütleme ölçütleri	Taha ve Hanbury
Sürekli izleme ve veri değişiklikleri	ISO 24051-1 ve FDA
Yerel doğrulama	ISO 24051-1

Yapay zekâ kullanımına ilişkin açıklama: Bu çalışmanın yapılandırılmasında ve ilk taslağının hazırlanmasında OpenAI ChatGPT'den yararlanılmıştır. İçerik, özgün kaynaklar

esas alınarak yazar tarafından doğrulanmış ve düzeltilmiş; çalışmanın son biçimi yazar tarafından verilmiştir.

Yazar hakkında

Prof. Dr. Diler Aslan, tıbbi biyokimya, tıbbi laboratuvar yönetimi, laboratuvar verilerinin yönetimi ve kalite sistemleri alanlarında uzun yıllar akademik ve uygulamalı çalışmalar yürütmüş; çeşitli kurumsal yöneticilik görevlerinde bulunmuştur. FORTRAN II Programlama Dili (1972), ANKUZEM E-Eğitimci (2013), UKAS ISO/IEC 42001 Yapay Zekâ Yönetim Sistemleri Farkındalık (2025) ve ADLM Tıbbi Laboratuvarlarda Veri Bilimi (2026) eğitimlerini tamamlamıştır. Dijital sağlık, veri kalitesi, yapay zekâ yönetişimi ve tıbbi laboratuvarlarda kalite yönetimi alanlarında eğitim ve danışmanlık çalışmalarını sürdürmektedir. ORCID: 0000-0003-4907-9445