# Derin Sinir Ağlarıyla Osmanlıca Optik Karakter Tanıma (OCR)

**İshak Dölek, Atakan Kurt**

**1. Giriş**

* Osmanlıca, 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu’nda kullanılan Arap harfleriyle yazılmış bir dildir.
* Osmanlı arşivlerinde milyonlarca belge bulunmasına rağmen, bunların yalnızca küçük bir kısmı dijital ortama aktarılmıştır.
* Osmanlıca OCR (Optical Character Recognition) çalışmaları sınırlıdır ve mevcut araçların doğruluk oranları düşük kalmaktadır.
* Çalışmanın amacı, Osmanlıca matbu nesih hattını tanıyabilen derin sinir ağları tabanlı bir OCR modeli geliştirmektir.
* metin, mektup, harf, el yazısı, doküman, belge içeren bir resim

  Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**2. Yöntem**

* **Model:** CNN (Evrişimli Sinir Ağı) ve LSTM (Uzun Kısa Süreli Bellek) mimarisi kullanıldı.
* diyagram, metin, taslak, plan içeren bir resim

  Yapay zeka tarafından oluşturulan içerik yanlış olabilir.
* **Veri Setleri:**
  + **Orijinal Veri Seti:** 1000 Osmanlıca sayfa içeriyor.
  + **Sentetik Veri Seti:** 23.000 sayfa algoritmik olarak üretildi.
  + **Hibrit Veri Seti:** Orijinal ve sentetik veri setlerinin birleşimi.
* **Karşılaştırmalı Analiz:**
  + Geliştirilen model, **Google Docs, Abby FineReader, Miletos, Tesseract (Arapça & Farsça)** OCR araçlarıyla karşılaştırıldı.
* **Değerlendirme Metrikleri:**
  + Ham, normalize ve birleştirilmiş metinlerde **karakter, bağlı harf ve kelime doğruluğu** hesaplandı.

metin, diyagram, ekran görüntüsü, plan içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**3. Sonuçlar**

* **Hibrit model**, diğer OCR araçlarına kıyasla **daha yüksek doğruluk oranı** sağladı:
  + **Karakter Tanıma Başarı Oranı:**
    - **Ham:** %88.86
    - **Normalize:** %96.12
    - **Bitişik:** %97.37
  + **Bağlı Harf Tanıma Başarı Oranı:**
    - **Ham:** %80.48
    - **Normalize:** %91.60
  + **Kelime Tanıma Başarı Oranı:**
    - **Ham:** %44.08
    - **Normalize:** %66.45

**4. Sonuç ve Değerlendirme**

* Çalışma, CNN ve LSTM tabanlı modelin **mevcut OCR araçlarından daha başarılı** olduğunu göstermiştir.
* metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

  Yapay zeka tarafından oluşturulan içerik yanlış olabilir.
* **Osmanlı alfabesindeki harf grupları ve karakteristikleri analiz edilerek** modelin doğruluğu artırılmıştır.
* Model, **osmanlica.com** adresinde bir OCR aracı olarak kullanıma sunulmuştur.
* Gelecekte **el yazması metinlerin tanınması** üzerine çalışmalar yapılabilir.