

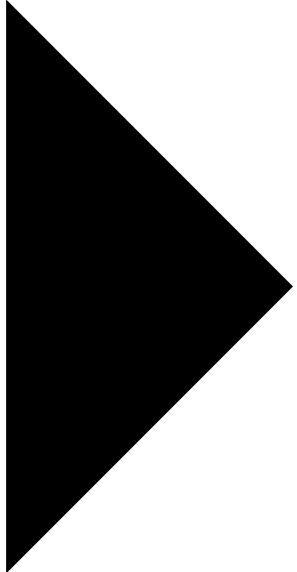
**Osmanlıca Optik  
Karakter Tanıma  
(OCR) ve Derin Sinir  
Ağları  
Hazırlayan : Semih  
Temur**

# 1. Giriş

---

- Osmanlıca, 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu'nda kullanılan ve Arap alfabesiyle yazılan bir yazı dilidir.
- Osmanlıca metinler, Arapça ve Farsça kelimeler içerdiğinden ve Latin alfabesine geçiş sonrası kullanılmadığından okunması günümüzde zor hale gelmiştir.
- Osmanlıca eserlerin dijitalleştirilmesi, tarihi ve kültürel mirasın korunması açısından önemlidir.
- OCR (Optical Character Recognition - Optik Karakter Tanıma) teknolojisi, basılı metinleri dijital metne dönüştürerek bu süreci hızlandırabilir.
- Osmanlıca için mevcut OCR araçları sınırlı olup, tanıma doğruluk oranları düşüktür. Bu çalışmada, Osmanlıca matbu nesih hattıyla yazılmış metinleri otomatik olarak tanıyan bir OCR modeli geliştirilmiştir.

## 2. Çalışmanın Amacı



Derin öğrenme teknikleri kullanarak Osmanlıca OCR doğruluğunu artırmak.  
Mevcut OCR araçlarıyla kıyaslandığında daha yüksek doğruluk oranları elde eden bir model geliştirmek.

Osmanlıca harf, kelime ve katar frekanslarını analiz ederek OCR başarımını artırmak.

Modelin doğruluk seviyesini artırmak için harflerin karakteristik özelliklerini belirlemek ve hataları minimize etmek.

Geliştirilen modelin web tabanlı olarak herkesin erişimine açılmasını sağlamak.

# 3. Kullanılan Yöntemler ve Modeller

---

## **Derin Öğrenme Modeli:**

Evrişimsel Sinir Ağları (CNN) ile karakter ve desen tanıma.

Tekrarlayan Sinir Ağları (RNN) ve LSTM (Uzun Kısa Süreli Bellek) ile dizisel karakter tanıma.

CTC (Connectionist Temporal Classification) ile karakterlerin sıralamasını öğrenme.

## **Veri Setleri:**

Orijinal Veri Seti: 1.000 sayfa Osmanlıca metinden oluşmaktadır.

Sentetik Veri Seti: Algoritmalarla üretilmiş 23.000 sayfa sentetik Osmanlıca metin.

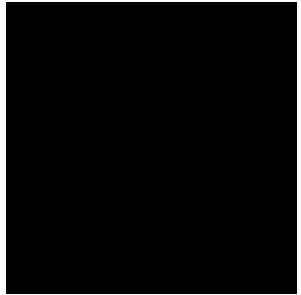
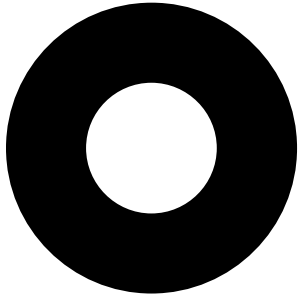
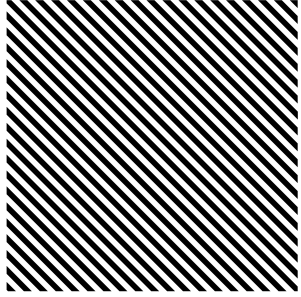
Hibrit Veri Seti: Orijinal ve sentetik veri setlerinin birleşimi.

## **Karşılaştırılan OCR Modelleri:**

Google Docs  
Abby FineReader  
Miletos

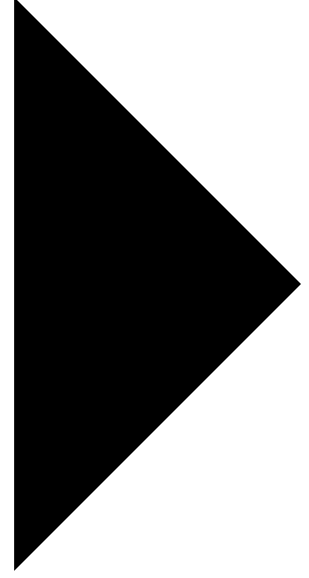
Tesseract (Arapça ve Farsça modelleri)  
OCR sonuçları harf, katar ve kelime bazında analiz edilerek modelin başarımı test edilmiştir.

# 4. Sonular ve Karşılaştırma



- Hibrit modelin tanıma doğruluk oranları:
  - Ham Metin Tanıma: %88,86
  - Normalize Metin Tanıma: %96,12
  - Bitişik Metin Tanıma: %97,37
- Google Docs, Abby FineReader ve Tesseract gibi yaygın OCR modellerine kıyasla daha başarılı sonuçlar elde edilmiştir.
- Karakter seviyesinde hata analizi yapılarak modelin eksiklikleri belirlenmiş ve iyileştirme sağlanmıştır.
- Osmanlı alfabesi harfleri, noktalı/noktasız oluşları, bitişkenlikleri gibi özellikler açısından incelenmiş ve OCR süreçlerine etkileri analiz edilmiştir.
- Osmanlıca metinlerde en sık kullanılan harfler, kelimeler ve kelime yapıları belirlenerek dil modellemesi geliştirilmiştir.

# 5. Sonuç ve Değerlendirme



- Osmanlıca OCR'da derin sinir ağları kullanımı ile tanıma başarımında önemli artış sağlanmıştır.
- Modelin eğitimi için orijinal ve sentetik veri kümeleri kullanılmış, bu sayede geniş çapta veri ile öğrenme sağlanmıştır.
- Geliştirilen model, harf ve kelime bazında karşılaştırmalı analizlerle mevcut araçlardan daha yüksek doğruluk oranlarına ulaşmıştır.
- Osmanlıca karakter tanıma üzerine detaylı hata analizleri yapılmış, hangi harflerin yanlış tanındığı ve nedenleri incelenmiştir.
- Osmanlica.com platformunda bu OCR modeli halka açık şekilde kullanıma sunulmuştur.

# 6. Gelecek Çalışmalar



El yazısı ile yazılmış Osmanlıca belgeler için OCR modelinin geliştirilmesi.

Modelin farklı yazı tipleriyle eğitilerek daha geniş kullanım alanına uygun hale getirilmesi.

Kelime ve cümle bazlı OCR modellemelerinin eklenmesiyle daha yüksek doğruluk oranlarına ulaşılması.

Osmanlıca-Türkçe otomatik çeviri modelleriyle entegrasyon sağlanarak OCR sonrası dil işleme çalışmalarının yapılması.