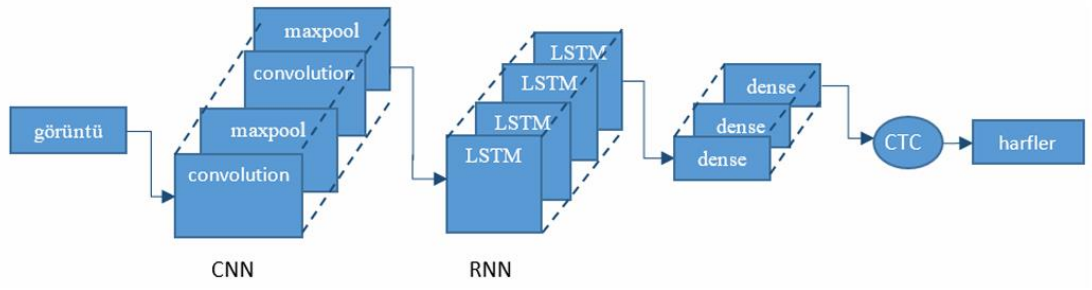
# Derin sinir ağlarıyla Osmanlıca optik karakter tanıma

Bu makalede, Nesih yazısıyla basılmış Osmanlıca belge görüntülerini CNN+RNN tabanlı derin sinir ağı modelleri kullanarak metne dönüştüren web tabanlı bir optik karakter tanıma (OCR) sistemi sunulmaktadır. Bu proje, Tesseract, Google Docs, Abby FineReader ve Miletos gibi mevcut araçların yetersiz sonuçlar sağladığı Osmanlıca OCR'nin zorluklarını ele almaktadır. Araştırma, karakter tanıma doğruluğunu önemli ölçüde iyileştiren CNN+RNN tabanlı bir derin sinir ağı modeli sunmaktadır.

## Derin Öğrenme Model Mimarisi

Önerilen derin öğrenme modeli, aşağıdakileri birleştiren bir CRNN (Konvolüsyonel Tekrarlayan Sinir Ağı) mimarisini takip eder:

- **CNN (Konvolüsyonel Sinir Ağları):** Kenarlar, şekiller ve karakter yapıları gibi metin görüntülerinden düşük ve yüksek seviyeli özellikleri çıkarır.
- **Çift yönlü LSTM (Uzun Kısa Süreli Bellek):** Sıralı metin verilerini işler, bir kelime veya bağda hem geçmiş hem de gelecekteki karakterlerden bağlamı yakalar.
- **CTC (Bağlantıcı Zamansal Sınıflandırma) kayıp işlevi:** Modelin, Osmanlıca gibi el yazısı yazılar için çok önemli olan bireysel karakter segmentasyonu gerektirmeden metin dizilerini tanımasını sağlar.




**Şekil 1:** Osmanlıca OCR için CRNN mimarisi (CRNN architecture for Ottoman OCR)

## Eğitim Süreci ve Veri Hazırlığı

Modeli geliştirmek için üç farklı veri kümesi oluşturuldu:

- **Orijinal veri kümesi:** Dikkatlice etiketlenmiş ve işlenmiş 1.000 sayfa gerçek Osmanlıca basılı metin içerir.
- **Sentetik veri kümesi:** Eğitim verilerini genişletmek için metinden görüntüye dönüştürme araçları kullanılarak oluşturulan 23.000 sayfadan oluşur.
- **Hibrit veri kümesi:** Hem gerçek dünya hem de yapay olarak oluşturulmuş metnin güçlü yönlerinden yararlanan, hem orijinal hem de sentetik veri kümelerinin bir kombinasyonu.

Veri kümesinin hazırlanması, normalleştirme, gürültü azaltma ve segmentasyon gibi görüntü ön işleme tekniklerini içeriyordu. Sentetik veri kümesi, etiketli Osmanlıca metin verilerinin sınırlı kullanılabilirliğinin üstesinden gelmede özellikle yararlıydı.

OCR modelini değerlendirmek ve doğruluğunu yaygın olarak kullanılan diğer OCR araçlarıyla karşılaştırmak için 21 sayfalık bir test seti kullanıldı. Değerlendirme üç seviyeye odaklandı:

1. Karakter tanıma doğruluğu
2. Bağlantılı bileşen (ligatür) tanıma doğruluğu
3. Kelime tanıma doğruluğu

Hibrit model en iyi sonuçları verdi:

| Model | Ham | Normalize | Bitişik | Değişen | Silinen | Eklenen |
|---|---|---|---|---|---|---|
| Osmanlica Hibrit | 88,86 | 96,12 | 97,37 | 1,60 | 1,93 | 2,50 |
| Osmanlica Orijinal | 87,73 | 94,87 | 96,16 | 2,30 | 2,50 | 2,81 |
| Osmanlica Sentetik | 73,16 | 77,64 | 78,10 | 14,92 | 5,77 | 6,15 |
| Google Docs | 83,86 | 92,02 | 91,43 | 4,24 | 3,19 | 3,50 |
| Abby FineReader | 71,98 | 80,19 | 81,05 | 13,47 | 8,23 | 3,45 |
| Tesseract Arabic | 76,92 | 82,37 | 81,27 | 12,79 | 6,15 | 2,89 |
| Tesseract Persian | 75,30 | 83,85 | 83,48 | 11,18 | 7,14 | 2,51 |
| Miletos | 75,76 | 86,46 | 86,88 | 10,94 | 6,21 | 1,57 |

**Tablo 1:** Karakter tanıma doğruluk oranı ve normalize metin hata dağılımları (%)

| Model | Ham | Normalize | Bitişik | Değişen | Silinen | Eklenen |
|---|---|---|---|---|---|---|
| Osmanlica Hibrit | 80,48 | 91,60 | 92,14 | 7,22 | 0,26 | 0,21 |
| Osmanlica Orijinal | 78,34 | 89,10 | 88,75 | 9,57 | 0,52 | 0,39 |
| Osmanlica Sentetik | 55,64 | 61,63 | 56,59 | 31,65 | 3,46 | 1,61 |
| Google Docs | 75,51 | 83,11 | 72,63 | 15,20 | 0,38 | 0,41 |
| Abby FineReader | 51,52 | 61,58 | 57,59 | 35,57 | 2,73 | 1,21 |
| Tesseract Arabic | 59,32 | 65,89 | 59,05 | 30,45 | 1,39 | 0,99 |
| Tesseract Persian | 57,90 | 66,94 | 61,47 | 31,14 | 0,87 | 0,90 |
| Miletos | 60,56 | 73,61 | 69,81 | 27,63 | 0,71 | 0,33 |

**Tablo 2:** Katar tanıma doğruluk oranı ve hata dağılımları (%)

| Model | Ham | Normalize | Değişen | Silinen | Eklenen |
|---|---|---|---|---|---|
| Osmanlica Hibrit | 44,08 | 66,45 | 31,27 | 0,56 | 0,28 |
| Osmanlica Orijinal | 40,84 | 61,13 | 35,49 | 0,56 | 0,64 |
| Osmanlica Sentetik | 15,55 | 24,53 | 70,86 | 0,60 | 2,64 |
| Google Docs | 38,64 | 50,78 | 44,88 | 0,47 | 0,94 |
| Abby FineReader | 13,28 | 24,40 | 75,01 | 0,86 | 0,81 |
| Tesseract Arabic | 20,05 | 26,43 | 66,95 | 1,67 | 6,51 |
| Tesseract Persian | 16,59 | 27,02 | 69,44 | 2,09 | 2,33 |
| Miletos | 14,92 | 31,22 | 70,80 | 0,00 | 1,70 |

**Tablo 3:** Kelime tanıma doğruluk oranı ve hata dağılımları (%)

Bu sonuçlar, derin öğrenme modelinin Tesseract (Arapça ve Farsça modeller), Google Docs OCR ve Abby FineReader gibi geleneksel OCR araçlarından daha iyi performans gösterdiğini açıkça göstermektedir. Hibrit yaklaşım, hem gerçek hem de sentetik eğitim örneklerinden öğrenerek karakter tanımayı iyileştirdi.