Sunum 1: Osmanlıca Optik Karakter Tanıma (OCR) – Derin Sinir Ağlarıyla

1. Giriş ve Amaç

- **Kapsam:** Osmanlıca dokümanların dijitalleştirilmesi ve metne dönüştürülmesi.
- **Neden:** Osmanlı arşivlerindeki milyonlarca belgeye hızlı erişim sağlamak ve kültürel mirası korumak.
- **Hedef:** Nakş (matbu nesih hattı) fontu ile basılmış belgelerde karakter tanımada yüksek doğruluk oranına ulaşan bir OCR sistemi geliştirmek.

2. Veri Setleri ve Ön İşleme

- **Veri Setleri:**
  - Orijinal veri seti: Yaklaşık 1000 sayfa.
  - Sentetik veri seti: Yaklaşık 23.000 sayfa.
  - Hibrit veri seti: Her iki veri setinin birleşimi.
- **Ön İşleme:**
  - Normalizasyon, hata düzeltme algoritmaları ve harf, katar, kelime sıklık analizleri.

3. Derin Öğrenme Mimarisi: CNN + LSTM

- **CNN (Convolutional Neural Network):**
  - Görüntüden yerel özellikleri (kenarlar, köşeler, dokular) çıkartır.
  - Farklı boyutlardaki filtreler kullanılarak harflerin karakteristiklerini yakalar.
- **LSTM (Long Short-Term Memory) – RNN Yapısı:**
  - Çıkarılan özelliklerin zaman/uzaysal sıralılığını modelleyerek harf dizilerini tanır.
  - Bidirectional LSTM kullanımı, hem ileri hem geri yönlü bağlam bilgisini değerlendirmeye olanak tanır.
- **Birleştirme:**
  - CNN katmanlarından elde edilen öznitelikler, LSTM katmanına aktarılır.
  - Çıktı, karakter, bağlı karakter katarı (ligature) ve kelime tanıma oranları üzerinden değerlendirilir.

4. Deneysel Sonuçlar ve Karşılaştırma

- **Performans:**

    - Hibrit modelde karakter tanımada %88.86 (ham), %96.12 (normalize) ve %97.37 (bitişik) doğruluk oranları.
- **Karşılaştırılan Sistemler:**
    - Google Docs, Abby FineReader, Tesseract (Arapça ve Farsça modeller) ve Miletos OCR.
- **Hata Analizi:**
    - Karakter bazında ekleme, silme ve yer değiştirme hatalarının detaylı incelenmesi.

5. Sonuç ve Değerlendirme

- **Avantajlar:**
    - CNN'in güçlü özellik çıkarımı ve LSTM'nin dizisel modelleme yeteneklerinin birleşimi, tanıma doğruluğunu artırmıştır.
- **Uygulama:**
    - Geliştirilen model, osmanlica.com üzerinden kullanıma sunulmuştur.