Journal of the Faculty of Engineering and Architecture of Gazi University 38:4 (2023) 2579-2593

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Ottoman Optical Character Recognition with deep neural networks

İshak Dölek[1], Atakan Kurt[2]*
[1]Mina R&D Informatics, Entertech office:322, Istanbul- University, Cerrahpaşa, 34320, Istanbul, Turkey
[2]Department of Computer Engineering, Faculty of Engineering, Istanbul University, Cerrahpaşa, 34320, Istanbul, Türkiye

Highlights:

- Ottoman OCR of printed naksh font with deep neural network
- Letter, ligature and word frequency of Ottoman
- Comparison of Ottoman OCR tools: Tesseract, Google Docs, Abby Finereader, Miletos

Keywords:

- Ottoman
- OCR
- Optical Character Recognition
- Printed naksh font
- Deep neural networks

Article Info:

Research Article
Received: 24.01.2022
Accepted: 24.12.2022

DOI:

10.17341/gazimmfd.1062596

Correspondence:

Author: Atakan Kurt
e-mail: atakan.kurt@iuc.edu.tr
phone: +90 534 483 4533

Graphical/Tabular Abstract

Purpose: Ottoman was a written language used in the Ottoman Empire between 13th and 20th centuries. There are literally millions of Ottoman documents in various archives and libraries all over the world. Only a tiny portion of these pages have been converted to text. The purpose of this study is to develop an Ottoman OCR model with deep neural networks that can recognize characters in printed Ottoman documents in the naksh font. The number of studies on Ottoman OCR is limited and the results aren't satisfactory in general.

Theory and Methods:
We used a deep neural network architecture combining CNN and LSTM. We prepared 3 data sets - original, synthetic and hybrid - to train 3 different deep learning models. Original data set roughly consists of 1000 page images. Synthetic data set consists of 23.000 pages. Hybrid data set contains both. A set of 21 pages was used as a test set. We conducted experiments to compare our model with those of Google Docs, Abby FineReader, Miletos, and Tesseract's Arabic and Persian models (Figure A). We used a normalization algorithm to clean and correct the user errors and the software errors in the ground truth and OCR output texts. We computed the character, connected-component (joined letter sequence), and word recognition accuracies on raw, normalized and joined texts. We also conducted a frequency analysis of Ottoman on the 1000-page original data set. We produced character level, component level and word level language models that can be utilized in a number of NLP tasks including OCR. We grouped letters in the Ottoman alphabet based on several features including connectedness, shape of letter body or stroke, the position and the number of dots, the type and origin of letter etc. and calculated frequencies for each group. We presented a detailed error analysis on character recognition errors in terms of insertion, deletion, substitution operations at character level providing valuable insights into the kinds of errors occurring during recognition for all models used in the comparisons.

Model	Raw	Normalized	Joined
Hybrid	88.86	96.12	97.37
Original	87.73	94.87	96.16
Synthetic	73.16	77.64	78.10
Google Docs	83.86	92.02	91.43
Abby FineReader	71.98	80.19	81.05
Tesseract Arabic	76.92	82.37	81.27
Tesseract Persian	75.30	83.85	83.48
Miletos	75.76	86.46	86.88

Figure A. Character recognition accuracy of OCR models

Results:
The hybrid model produced 88.86% raw, 96.12% normalized and 97.37% joined text character recognition rate, 80.48% raw, 91.60% normalized, and 97.37% joined text connected-component recognition rate, and 44.08% raw, and 66.45% normalized text word recognition rate clearly outperforming the other models.

Conclusion:
The experimental study on Ottoman OCR conducted with a deep neural network architecture combining the object recognition strength of CNNs and the sequence recognition strength of bidirectional LSTMs reveals that this model outperforms the well known tools/models on character recognition, connected-component recognition, and word recognition metrics with a clear margin. The model was converted into an OCR tool and made available at osmanlica.com.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:4 (2023) 2579-2593

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Derin sinir ağlarıyla Osmanlıca optik karakter tanıma

İshak Dölek[1], Atakan Kurt[2]*
[1]Mina Ar-Ge Bilişim, Entertech ofis:322, İstanbul-Ün Cerrahpaşa, 34320, İstanbul, Türkiye
[2]İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 34320, İstanbul, Türkiye

ÖNEÇIKANLAR

- Derin sinir ağlarıyla Osmanlıca matbu nesih hattının OCR'ı
- Osmanlıca harf, bağlı karakter katarı ve kelime sıklıkları
- Tesseract, Google docs, Abby FineReader, Miletos OCR araçlarının karşılaştırılması

Makale Bilgileri
Araştırma Makalesi
Geliş: 24.01.2022
Kabul: 24.12.2022

DOI:
10.17341/gazimmfd.1062596

Anahtar Kelimeler:
Osmanlıca,
optik karakter tanıma,
OCR,
matbu nesih hattı,
derin öğrenme,
CNN, RNN, LSTM

ÖZ

Bu makalede nesih hattıyla basılmış Osmanlıca doküman görüntülerini CNN+RNN tabanlı derin sinir ağı modelleriyle metne dönüştüren web tabanlı bir optik karakter tanıma (OCR) sistemi sunulmuştur. Eğitim için *orijinal*, *sentetik* ve *hibrit* olmak üzere 3 veri kümesi hazırlanmış ve 3 farklı OCR modeli oluşturulmuştur. Orijinal veri seti yaklaşık 1.000 sayfadan, sentetik veri seti ise yaklaşık 23.000 sayfadan oluşmaktadır. Eğitilen modeller Tesseract'ın Arapça ve Farsça, Google Docs'ın Arapça, Abby FineReader'ın Arapça ve Miletos'un OCR model/araçlarıyla 21 sayfalık bir test setiyle karşılaştırılmıştır. Karşılaştırma *ham*, *normalize* ve *bitişik* olmak üzere 3 farklı metin ve *karakter*, *katar* ve *kelime* tanıma olmak üzere 3 farklı ölçüt ile yapılmıştır. Osmanlica.com Hibrit modeli karakter tanımada %88,86 *ham*, %96,12 *normalize* ve %97,37 *bitişik* doğruluk oranlarıyla; bağlı karakter katarı tanımada %80,48 *ham*, %91,60 *normalize* ve %97,37 *bitişik* doğruluk oranlarıyla; kelime tanımada %44.08 *ham* ve %66.45 *normalize* doğruluk oranlarıyla diğerlerinden belirgin şekilde daha iyi sonuçlar üretmiştir. Alfabenin kendine özgü karakteristiklerinin OCR'a etkilerini araştırmak için Osmanlıcanın karakter, katar ve kelime sıklık analizleri yapılmıştır. Bu analizde alfabedeki karakterler bitişebilme, harf gövdesi, noktaların konumu ve sayıları, karakterin türü, kaynak dil vb. ayırt edici özelliklere göre gruplandırılmış grup bazında sıklıklar ve tanıma doğruluk incelenmiştir. OCR sonuçları ayrıca harf bazında ortaya konulmuştur.

Ottoman Optical Character Recognition with deep neural networks

HIGHLIGHTS

- Ottoman OCR of printed naksh font with deep neural networks
- Letter, ligature and word frequency of Ottoman
- Comparison of Ottoman OCR tools: Tesseract, Google Docs, Abby FineReader, Miletos

Article Info
Research Article
Received: 24.01.2022
Accepted: 24.12.2022

DOI:
10.17341/gazimmfd.1062596

Keywords:
Ottoman,
OCR,
optical character Recognition,
printed naksh font,
deep neural networks,
CNN, RNN, LSTM

ABSTRACT

In this paper, we present a web-based optical character recognition (OCR) system that converts images of Ottoman documents printed with naskh font into text using CNN+RNN-based deep neural network models. For training, three datasets - original, synthetic, and hybrid - were prepared and three different OCR models were created. The original data set consists of 1,000 pages and the synthetic data set consists of 23,000 pages. Hybrid data set contains both. The trained models were compared with Tesseract's Arabic and Persian, Google Docs' Arabic, Abby FineReader's Arabic, and Miletos OCR model/tools with a 21-page test set. The comparison was made with 3 different texts (raw, normalized, and joined) and using 3 different criteria (character, ligature, and word recognition). The Osmanlica.com Hybrid model produced significantly better results than the others with 88.86% raw, 96.12% normalized, and 97.37% joined accuracy in character recognition; 80.48% raw, 91.60% normalized, and 97.37% joined accuracy in ligature recognition; and 44.08% raw and 66.45% normalized accuracy in word recognition. To investigate the effects of the characteristics of the alphabet on OCR, character, ligature, and word frequency analyses of Ottoman was performed. In this analysis, the characters in the alphabet were grouped according to distinctive features such as connectedness, letter body, position and number of dots, type of character, and source language; and frequencies and recognition accuracies were examined for each group. OCR results are also reported for each character.

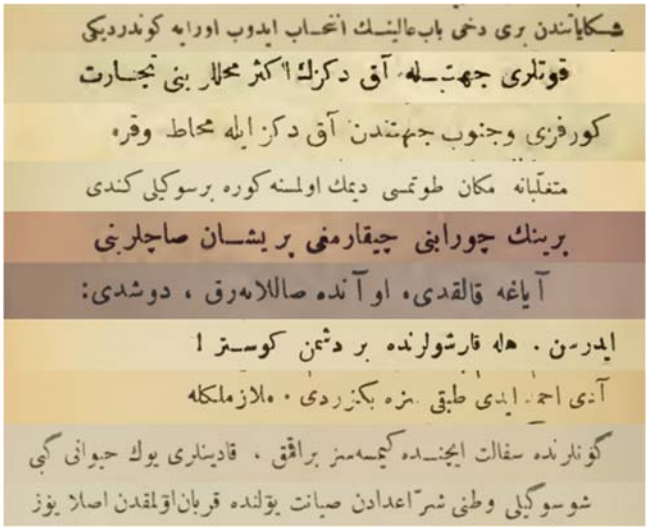
*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : ishakdolek54@gmail.com, *atakan.kurt@iuc.edu.tr / Tel: +90 534 483 4533

1. Giriş (Introduction)

Osmanlıca 13. yüzyıldan 20. Yüzyıllar arasında Osmanlı İmparatorluğunda kullanılan ve Arap alfabesiyle yazılan bir yazı dilidir [1]. Osmanlıca Arapça ve Farsça ifadelerin yoğun olarak kullanıldığı bir dildir. Günümüzde Latin (Roman) alfabesine geçiş yapıldığı ve kelimelerin çoğu kullanımdan kalktığı için Osmanlıcayı hem okumak hem de anlamak güçtür. Millî kültürümüzün temelini oluşturan eserlerin büyük bir kısmı Osmanlıcada yazılmıştır. Osmanlı arşiv ve kütüphanelerindeki kitap, dergi, gazete, defter, kayıt ve belgeler yüzlerce yıllık kültür, sanat ve tarih mirası içinde önemli bir yer tutar. Bu kaynaklarda saklı bilgiye hızlı ve etkin bir şekilde erişilmesi için başta OCR olmak üzere teknolojinin yardımına ihtiyaç vardır. TÜBİTAK destekli *Osmanlıcadan Günümüz Türkçesine Yapay Zeka Destekli Uçtan Uca Aktarım Projesi* bu amaçla geliştirilen bir projedir. Projenin çıktıları Osmanlica.com sitesinden kullanıcılara sunulmaktadır. Projede Tablo 1'de görüldüğü gibi bütüncül bir yaklaşımla Osmanlıca dokümanların bir uçtan diğer uca yani *Osmanlıca dokümandan Türkçe metne* dört adımda bilgisayarlı aktarımı amaçlanmaktadır: (i) Doküman-görüntü dönüşümü, (ii) Görüntü-metin dönüşümü (OCR) (iii) Osmanlıca-Türkçe alfabe çevirisi (harfçevrim), (iv) Osmanlıca-Türkçe dil çevirisi. Şekil 1'de örnekleri verilen nesih hattı özellikle Osmanlının son dönemlerinde kitap, gazete, dergi vb. basılı eserlerde yaygın olarak kullanılan yazı hattıdır. Bu makalede aktarımın 2. adımında kullanılacak OCR modeliyle bu modelin benzer araçlarla deneysel karşılaştırılması sunulmaktadır.

Tablo 1. Osmanlıca-Türkçe aktarım örneği
(Ottoman-Turkish conversion example)

#	Çıktı
1	ادراك معالى بو كوچك عقله كركز زيرا بو ترازو او قدر ثقلى چكمز
2	ادراك معالى بو كوچك عقله كركمز زيرا بو ترازو او قدر ثقلتى چكمز
3	İdrâk-i me'âlî bu küçük akla gerekmez Zîrâ bu terâzû o kadar sıkleti çekmez
4	İdrâk-i me'âlî bu küçük akla gerekmez Zîrâ bu terâzû o kadar sıkleti çekmez



Şekil 1. Osmanlıca matbu nesih hattı örnekleri
(Ottoman documents in printed naksh font)

Makale şu şekilde organize edilmiştir: Bölüm 2'de Osmanlıca bağlamında OCR problemi, Bölüm 3'te benzer çalışmalar, Bölüm 4'te derin öğrenme modeli mimarisi, Bölüm 5'te OCR modelinin eğitilmesi ve testinde kullanılan veri kümeleri, Bölüm 6'te deneyler ve karşılaştırma sonuçları anlatılmaktadır. Özet ve kısa bir değerlendirme Bölüm 7'de verilmiştir.

2. Osmanli Alfabesi ve Harf, Katar, Kelime Sikliklari (Ottoman Alphabet and Letter, Ligature, Word Frequencies)

Başarılı bir OCR çalışması için öncelikle Osmanlı alfabesindeki harf, rakam ve noktalama işaretlerini yakından tanımak ve Osmanlıca imla (yazım, ortography) kurallarının bilmek gereklidir [1]. Osmanlıcada harfler birbirine çok benzediği ve imla sıkı kurallara bağlandığından bu konu önem arz etmektedir. Bu bölümde Osmanlı alfabesi ve imlası üzerine kısa bir değerlendirme yapılacaktır.

2.1. Özelliklerine göre Osmanlıca harfler
(Ottoman Letters by Distinctive Features)

Osmanlı alfabesi Arapçadan 28 (ابتثجحخدذرزسشصضطظعغفقكلمنهوى), Farsçadan 4 (پچژگ), Türkçeden 3 (ڭۇۆ) ses alınarak oluşturulmuş Arap alfabesi tabanlı bir alfabedir. Rakamlar hariç olmak üzere harfler sağdan sola (RTL), bitişik (joined) ve harekesiz (diacritcs) yazılır. Sadece 8 harf (ادذرزژوه) kendinden sonra gelen yani solundaki harfle bitişmez. Bitişmeyen (ayrık, non-joiner) harflerin münferit ve sonda (önceki harfle bitişen), bitişen (joiner) harflerin ise münferit, başta (sonraki harfle bitişen), ortada (önceki ve sonraki harfle bitişen) ve sonda farklı yazım şekilleri vardır. Tablo 2'de 2., 3., 4. ve 5. sütunlarda harflerin farklı yazımları verilmiştir. Tablodaki harfler baskı nesih fontunun yazım kurallarına göre verilmiştir.

Tablo 4'te görüldüğü gibi OCR çalışması için alfabedeki harfler şu temel ayırt edici özelliklere göre gruplanabilir: Bitişken/ayrık, noktalı/noktasız, bir/iki/üç noktalı, noktası altta/üstte, gövdesi satır çizgisinde/üstünde/altında harfler. Harflerin özelliklerine göre gruplandırılması, gruptaki harf sayısı ve 1000 sayfalık orijinal eğitim veri kümesindeki sıklıkları Tablo 3'de verilmiştir. En önemli ayırt edici özellik harf gövdesinin (stroke) şeklidir. Çünkü alfabedeki çoğu harfin gövdesi ya birebir aynı ya da benzerdir. Harf gövdesi nokta, med, hemze gibi bazı özel diakritikler çıkınca geriye kalan eğri veya doğru parçalarıdır. Gövdelerin kelimedeki ya da katardaki konumlarına göre farklı yazımları Tablo 2'de 6.-9. sütunlarda verilmiştir. Gövdelerine göre harfler genellikle 15 grupta toplanır. Tablo 2'de bu gruplar 7. ve 8. sütunlarda listelenmiştir. Bu 15 grup ta benzerliklerine göre 4 ana grupta toplanabilir: Halkalı (looped), dişli (toothed), düz (straight), ve diğer. Tablo 2'de ana gruplar 1. sütunda verilmiştir. Harf gövdelerinin en belirgin ve yaygın ayırt edici özellikleri halkalı veya dişli olmalarıdır. Daha ileri çalışmalarda harfler gövdelerindeki *alt parçalar dizisi*, *kesişim noktaları dizisi veya uç noktaları dizisi* gibi ayırt edici geometrik özelliklere göre gruplanabilmektedir. Bitişik veya ayrık yazım şekillerine göre harflerin farklı biçimde gruplanması söz konusu olabilir. Bitişik yazıma göre bir gruplama örneği aşağıda verilmiştir:

بتثپنيند ,سشـ ,جححچ ,رزژ ,صضطظ ,عغفقمو ,ككگ ,لا١١)(,٥٥ ,ھ ,ء

Osmanlıcadaki harflerin 16 adedi (احدرسصطعكلمهوگ٥ه) noktasızdır. Tek noktalı 10 (بجخذضظغفن), iki noktalı 3 (تقي), ve üç noktalı 6 (ثشپچژڭ) olmak üzere 19 adet noktalı harf vardır. Noktalı harflerin 14 tanesinin noktası üstte (تثخذزشضظغفقژڭ) beşinin noktası alttadır (بجپيچ). Harflerin bazıları (ورزژ vb.) satır çizgisinin (baseline) altına taşmakta (descenders) diğerleri (ككگا١ل vb.) çizginin üstünde (ascender) yazılmaktadır.

Tablo 2. Harf gövdeleri ve sıklıkları (Letter stems and frequencies)

Tür	Harf				Sıklık	%	Gövde				Kod
	ب ت	بـ تـ	ـب ـت	ب ت	65031	6,91					
							ـں			ں	066E
	ٹ پ	ٹـ پـ	ـٹ ـپ	ٹ پ	6417	0,68		ـہ	ہ		
	ن	نـ	ـن	ن	61415	6,53	ں			ں	06BA
	ی	یـ	ـی	ی	108850	11,57	ی			ی	0649
dişli	س ش	سـ شـ	ـس ـش	س ش	41715	4,43	س	سـ	ـس	س	0633
	ڈ ذ			ڈ ذ	58188	6,19	ـڈ			ڈ	062F
	ر ز ژ			ر ز ژ	87188	9,27	ـر			ر	0631
diğer	ج ح خ چ	جـ حـ خـ چـ	ـج ـح ـخ ـچ	ج ح خ چ	35506	3,78	ح	حـ	ـح	ح	062D
	ص ض	صـ ضـ	ـص ـض	ص ض	10442	1,11	ص	صـ	ـص	ص	0635
	ط ظ	طـ ظـ	ـط ـظ	ط ظ	8888	0,94	ط	طـ	ـط	ط	0637
	ع غ	عـ غـ	ـع ـغ	ع غ	20473	2,18	ع	عـ	ـع	ع	0639
	ق ف	قـ فـ	ـق ـف	ق ف	35576	3,78	ق	قـ	ـق	ق	066F
	و			و	70114	7,45	و			و	0648
halkalı	م	مـ	ـم	م	46842	4,98	م	مـ	ـم	م	0645
	ه ہ	هـ ہـ	ـه ـہ	ه ہ	53535	5,69	ه ہ	هـ ہـ	ـه ـہ	ه ہ	0647
	ك ڭ گ	كـ ڭـ گـ	ـك ـڭ ـگ	ك ڭ گ	42570	5,02	ك ک	كـ کـ	ـك ـک	ك ک	0643
	ل	لـ	ـل	ل	63288	6,73	ل	لـ	ـل	ل	0644
düz	ا			ا	117719	12,52	ا			ا	0627

Yukarıdaki ayırt edici özellik ve gruplamalar OCR'da yönlendirici ya da yardımcı rol oynamaktadır. Örneğin bazı çalışmalarda harfler yukarıdaki gibi gruplandırılarak OCR üç aşamada gerçekleştirilmektedir [2]. İlk aşamada metindeki harfler harf gövdelerine dönüştürülür ve OCR adımında harfler yerine gövdeler tanınmaya çalışılır. Üçüncü adımda gövdelerden harflere geri dönüşüm yapılır. Bu yaklaşımın avantajı gövde sayısının harflere göre daha az ve dolayısıyla tanımanın daha kolay olması, dezavantajı ise harflerin ön işlemeyle gövdelenmesi ve OCR sonrası harfe geri dönüşümünde yaşanan güçlüktür.

2.2. Osmanlıca harf sıklık dağılımları
(Ottoman Letter Frequency Distribution)

Alfabedeki harflerin metin içinde kullanım şekilleri, geçiş yerleri, sıklıkları ve diğer harflerle komşulukları ve birlikte geçişleri OCR öncesi, sırası ve sonrası temizleme, normalizasyon, düzeltme ve hata ayıklama algoritmaları geliştirilmesi vb. birçok uygulamada yol gösterici ve faydalıdır. Osmanlıca orijinal eğitim verisindeki metinlerin normalizasyonu sonrası hesaplanan harf sıklıkları Tablo 3'te verilmiştir. Tablodaki en dikkat çekici nokta metindeki her dört harften birinin ی ya da ا seslilerinde biri olmasıdır. Sonrasında harfler ر 'dan başlayarak düzgün olarak azalarak sıralanmaktadır. Hem sesli hem sessiz olabilen bir diğer okutucu و harfi %7,5'luk yüksek bir sıklığa sahiptir. En sık 20 harfin toplam sıklığı metindeki tüm harflerin %94'e karşılık gelmektedir. Sıklıkları toplamı %6 olan son 15 harf içinde Arapça kelimelerde kullanılan غ ط ص ذ ظ ث ض خ gibi harfler ve Farsça/Türkçe seslere özel ء ك گ ژ چ پ gibi harflerin bulunması dikkat çekicidir.

Tablo 3. Osmanlıca harf sıklıkları (Ottoman letter frequencies)

Harf	Sıklık	%
ا	117719	12,52
ی	108850	11,57
ر	74131	7,88
و	70114	7,45
ل	63288	6,73
ن	61415	6,53
د	56699	6,03
ه	53535	5,69
م	46842	4,98
ك ک	42282	4,50
ب	35743	3,80
ت	29288	3,11
ق	25336	2,69
ش	24968	2,65
س	16747	1,78
ز ع	13446	1,43
ج	12964	1,38
ح	11625	1,24
ف	10277	1,09
ص	10240	1,09
ط	7726	0,82
خ	7264	0,77
ث	7027	0,75
ہ	6844	0,73
ض	6760	0,72
پ	6394	0,68
ظ	4643	0,49
ڤ ڵ	2716	0,29
ظ	1774	0,19
ذ	1624	0,17
غ	1489	0,16
ء	362	0,04
ڭ	297	0,03
ژ	93	0,01
گ	91	0,01

Harflerin bitişik harf katarları içindeki konumuna göre sıklık dağılımları Tablo 5'de verilmiştir. Tablonun en alt satırında katarların içerdiği harf sayısına göre dağılımları bulunmaktadır. Tablo 7'daki verilere göre katarların %46'sı tek harf, %29'u 2 harf ve %16'sı 3 harften oluşmaktadır. Tablo 5'teki 1. 2. ve 3. harf sıklık toplamı yaklaşık olarak tüm katarların %91'ine dağılmaktadır. Hemen hemen tüm harflerde konum ilerledikçe harf sıklığının hızla düştüğü gözlenmektedir. ژ, ء, ہ, ہ, gibi bazı harflerde olağan dışı sıralamalar gözlenmiştir. ہ, ہ harfleri sesli harf olarak kullanıldıkları için 3. 4. ve 5. harf olarak yüksek sıklıkta görülmektedir.

Kelime içindeki konumlarına göre harf sıklıkları Tablo 6'te verilmiştir. Tablonun son satırında Tablo 11'ten alınan içerdikleri harf sayısına göre kelimelerin dağılımları bulunmaktadır. Kelimelerin büyük çoğunluğu 4 veya 5 harften oluşmaktadır. Harfler arasında kelime içi konuma göre net bir ayrışma olmasa da benzer gruplar oluştuğu gözlenmektedir. Bazı harfler kelime içinde en sık 1. konumda, bazı harfler en sık 2. konumda, bazıları ise en sık 3. konumda geçmektedir. ہ ژ ء ve ڭ gibi bazı harflerde ise daha homojen bir dağılım gözlenmektedir.

Tablo 4. Osmanlıca harf grupları ve sıklıkları (Ottoman letter groups and frequencies)

Grup	Harf	Adet	%	Frekans
Halkalı	ص ض ط ظ ع غ ف ق م ه ھ و	12	28	4570
Dişli	پ ب ت ث س ش ن ي	8	30	3745
Düz	ا ك ل ڭ گ	5	24	3897
Diğer	ج چ ح خ د ذ ر ز ژ	9	19	3102
Bitişen	ب ت ث ج ح خ س ش ص ض ط ظ ع غ ف ق ك ل م ن ھ ي پ چ گ ڭ	26	61	9289
Bitişmeyen	ا د ذ ر ز و ۋ ه	8	39	6025
Noktalı	ب ت ث ج خ ذ ز ش ض ظ غ ف ق ن ي پ چ ژ ڭ	19	32	4907
Noktasız	ا ح د ر س ص ط ع ك ل م ھ ه و گ	15	68	10407
1 noktalı	ب ج خ ذ ز ض ظ غ ف ن	10	56	2768
2 noktalı	ت ق ي	3	34	1662
3 noktalı	ث ش پ چ ژ ڭ	6	10	477
Nokta üstte	ت ث خ ذ ز ش ض ظ غ ف ق ن ژ ڭ	14	67	3286
Nokta altta	پ ب ج چ ي	5	33	1621
Ortada	د ذ ب ت ث پ ف ھ ه	9	16	2505
Alçalan	و ر ز ژ ج ح خ چ ع غ ق م ن س ش ص ض ى	18	51	7815
Yükselen	ا ك ل گ ڭ ط ظ	7	33	4998
Arapça	ى و ھ ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ت ب ا	28	78	15115
Osmanlıca	ه ڭ گ ژ چ پ	6	1	203
Noktalama	ـ ، ؛ : ؟ ( ) –	8	21	4002
Rakam	۰ ۱ ۲ ۳ ٤ ٥ ٦ ۷ ۸ ۹	10	0	54
Harf	ه ڭ گ ژ چ پ ى و ھ ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ت ب ا	34	98	15314
Diğer	ـ ، ؛ ؟ ( ) – ۰ ۱ ۲ ۳ ٤ ٥ ٦ ۷ ۸ ۹	16	2	296

Tablo 5. Katardaki konuma göre harf sıklığı (Letter frequency by position in ligature) (%)

Harf	1	2	3	4	5
ا	51,85	29,71	13,50	3,94	0,81
ب	78,76	14,29	5,53	1,09	0,31
ت	45,58	36,91	11,99	3,21	2,17
ث	44,08	39,97	13,36	2,48	0,11
ج	63,79	19,97	11,40	4,07	0,62
ح	69,21	24,42	4,51	1,62	0,24
خ	76,97	16,55	5,61	0,80	0,06
د	42,29	28,86	17,94	7,70	2,36
ذ	54,26	39,22	4,63	1,41	0,20
ر	41,63	34,55	16,13	4,73	2,05
ز	54,29	22,55	14,09	5,85	2,30
س	61,82	25,70	7,72	3,50	0,89
ش	53,46	26,81	11,58	6,50	1,36
ص	72,18	19,03	7,25	1,07	0,45
ض	50,88	31,15	15,65	2,21	0,11
ط	66,71	16,55	15,47	1,12	0,15
ظ	33,87	45,69	16,44	3,39	0,18
ع	57,99	30,23	9,08	2,31	0,39
غ	55,50	33,60	8,94	1,30	0,58
ف	64,96	23,16	7,43	3,26	1,15
ق	64,83	19,32	11,67	3,13	0,97
ك	54,55	28,56	10,41	4,58	1,31
ل	47,56	33,47	12,88	3,87	1,75
م	57,53	22,85	14,89	3,54	1,00
ن	55,27	28,87	10,74	3,45	1,21
ه	28,35	27,06	22,71	16,80	4,35
ھ	38,88	22,53	22,68	11,02	3,43
و	45,80	39,72	9,48	3,75	0,69
ى	50,96	26,02	14,35	6,11	1,81
ء	74,86	5,80	8,29	5,52	5,25
ڭ	42,09	33,33	16,16	6,06	1,68
گ	78,02	16,48	5,49	0,00	0,00
چ	60,17	30,57	7,90	1,23	0,12
ژ	67,74	25,81	0,00	6,45	0,00
پ	82,73	13,70	3,10	0,47	0,00
Katar	45,62	29,40	16,06	6,19	1,98

Tablo 6. Kelimedeki konuma göre harf sıklığı (Letter frequency by position in word) (%)

Harf	1	2	3	4	5	6	7	8
ا	38,18	20,88	15,38	14,75	6,35	2,58	1,18	0,41
ب	57,61	10,67	11,31	8,69	7,37	2,29	1,24	0,49
ت	18,44	10,19	*31,55*	19,08	13,16	4,18	2,61	0,47
ث	15,33	*42,45*	26,21	11,89	2,93	0,68	0,06	0,11
ج	30,96	16,89	20,19	10,83	11,57	4,69	2,55	1,45
ح	46,86	29,79	11,82	7,62	2,45	0,90	0,25	0,18
خ	47,84	26,97	12,22	5,19	5,96	1,12	0,44	0,12
د	18,48	7,18	*20,24*	19,77	13,44	10,02	4,81	3,16
ذ	38,75	35,12	9,94	9,74	4,50	0,94	0,74	0,20
ر	3,31	25,39	22,39	14,68	13,68	7,70	7,02	2,74
ز	11,56	18,78	*34,56*	8,77	9,71	6,89	4,17	2,92
س	31,86	20,05	16,32	11,47	10,17	5,68	2,18	1,31
ش	19,90	16,00	26,49	13,23	14,03	4,51	3,42	1,31
ص	48,59	23,27	17,20	6,52	2,49	1,11	0,38	0,14
ض	13,95	*34,17*	31,48	14,03	4,12	1,88	0,26	0,07
ط	47,82	14,85	24,11	8,51	2,20	0,95	1,31	0,17
ظ	21,55	*35,22*	23,03	7,51	8,99	3,02	0,43	0,06
ع	39,60	26,28	16,86	10,58	4,72	1,29	0,34	0,13
غ	15,57	16,08	24,31	8,50	8,54	16,42	5,12	3,33
ف	27,41	26,32	21,36	13,28	8,39	2,18	0,59	0,25
ق	34,16	15,16	19,17	11,51	10,29	4,85	2,36	1,38
ك	29,38	16,37	17,22	10,56	9,59	7,75	3,95	2,48
ل	4,71	16,45	*35,53*	17,24	11,30	7,97	3,51	2,06
م	*28,30*	11,73	18,18	20,25	9,88	5,45	3,13	1,63
ن	8,60	14,66	15,76	16,96	*18,99*	10,65	6,56	3,92
ه	0,13	3,03	8,09	*37,38*	24,60	14,34	7,30	2,35
ه	5,33	12,34	6,98	*20,95*	19,19	14,83	10,13	5,28
و	16,33	49,69	9,14	14,37	4,38	3,23	1,42	0,90
ى	7,70	*26,25*	13,81	17,58	13,47	8,79	6,20	3,22
ء	0,55	1,10	24,03	*29,83*	24,03	15,19	3,04	1,10
ڭ	3,70	*18,86*	17,85	11,11	14,81	13,47	7,74	6,40
گ	*68,13*	10,99	9,89	2,20	3,30	3,30	2,20	0,00
چ	*40,56*	11,72	39,00	5,29	2,00	0,82	0,25	0,19
ڑ	*35,48*	21,51	18,28	13,98	6,45	3,23	0,00	1,08
پ	*53,80*	11,93	25,07	5,36	2,80	0,52	0,26	0,04
Kelime	2,76	10,73	13,48	20,87	20,91	12,56	9,02	4,67

2.3. Osmanlıca katar sıklık dağılımları
(Ottoman Ligature Frequency Distribution)

Osmanlıcada harfler bitiştirilerek yazılır. Bunun istisnası yukarıda verilen bitişmeyen 8 harftir. Kelime içinde bu harflerden biri geldiğinde yazıda bir kopma meydana gelir. Bu noktada bir bitişik harf katarı biter ve yeni bir katar başlar. Bu iki katar arasındaki boşluğa kelime içi boşluk (zero width non space) denir. OCR'da kelime için boşluk kelimeler arasındaki boşlukla karıştırıldığı zaman kelime bölümleme ya da kelime bitişme problemleri ortaya çıkar. Katarların yapısını, çeşitlerini ve özelliklerini anlamak için yaptığımız katar sıklık dağılımı analizinin bazı sonuçları Tablo 7-Tablo 10'da paylaşılmıştır. Uzunluklarına yani içerdikleri harf sayısına göre katarlar incelendiğinde (Tablo 7) bunların yaklaşık %95'inin 1, 2, 3 veya 4 harften oluştuğu, bunun da 1/2'sinin 1 harf, 1/3'ünün 2 harf, 1/6'sının 3 harften oluştuğu görülmektedir.

Katarlar son harfine göre incelendiğinde (Tablo 8) genellikle bitişmeyen 4 harften birisiyle sonlandıkları görülmektedir: ا katarların 1/3'ünü, ر 1/5'ini, و 1/5'ini, د 1/6'sını sonlandırmaktadır. İlk harfine göre tetkik edildiğinde (Tablo 9) katarların çoğunlukla bitişmeyen ve sesli harflerle başlıyor olması dikkat çekicidir: Katarların 1/3'ü ا, 1/5'i ر, 1/5'i و, 1/5'i ى, 1/6'sı د, 1/6'sı ه / ہ ile başlamaktadır.

Metinde geçen katarların büyük çoğunluğunun bir, iki ve üç harften oluştuğu, bir harfli katarların da ekseriyetle ا و ر د ه ن ى harflerinden birisi olduğu anlaşılmaktadır. Burada da sesli harflerin yoğun olarak bir katar halinde yazıldığı anlaşılmaktadır. İki ve daha fazla harften oluşan bazı katarların tek başına bir kelime olduğu bunların çoğunluğunun da Türkçede sık kullanılan kelimeler olduğu görülmektedir.

Tablo 7. Harf sayılarına göre katarlar (Ligatures by letter count)

Harf #	Sıklık	%
1	223558	45,62
2	144076	29,40
3	78683	16,06
4	30308	6,19
5	9699	1,98
6	2828	0,58
7	734	0,15
8	89	0,02
9	31	0,01
10	3	0,00

2.4. Osmanlıca kelime sıklık dağılımları
(Ottoman Word Frequency Distribution)

Osmanlıca kelimelerin harf sayısına göre sıklıkları Tablo 11'de verilmiştir. Kelimelerin %42'si dört veya beş harften, %92'si 2,3,4,5,6,7 veya 8 harften oluşmaktadır. Kelimelerin katar sayısına göre sıklıkları ise Tablo 12'te paylaşılmıştır. Kelimelerin 1/4'ü 1 katar,

1/3'ü 2 katar, 1/4'ü 3 katar, 1/8'i 4 katar ve 1/10'u 5 katardan oluşmaktadır. Kelimelerin hemen hemen tamamı (%98) 1-5 katardan oluştuğu için daha uzun kelimelerle çok nadir karşılaşılmaktadır. Osmanlıca kelime sıklıklarının Tablo 13'de listelenmiştir. Sık geçen kelimelerin genellikle *ve, bir, bu, de, ile, ye, ne, gibi, olan, o, için, kadar*, vb. bağlaç, edat, zamir sıfat ve fiillerden oluşmaktadır. İlk beş kelime tüm kelimelerin yaklaşık %5'ini oluşturmaktadır.

Tablo 8. Son harfe göre katarlar (Ligatures by last letter)

Son harf	Sıklık	%
ا	117719	34,62
د	56699	16,68
ذ	1489	0,44
ر	74131	21,80
ژ	93	0,03
ز	12964	3,81
و	70114	20,62
ه	6394	1,88
ء	362	0,11

Tablo 9. İlk harfe göre katarlar (Ligatures by first letter)

İlk harf	Sıklık	%
ا	178752	36,48
ب	32917	6,72
ت	19215	3,92
ث	1036	0,21
ج	7983	1,63
ح	7603	1,55
خ	5480	1,12
د	80677	16,46
ذ	2297	0,47
ر	104990	21,43
ز	20002	4,08
س	16470	3,36
ش	12031	2,46
ص	5842	1,19
ض	1930	0,39
ط	5624	1,15
ظ	651	0,13
ع	8838	1,80
غ	4129	0,84
ف	7988	1,63
ق	21610	4,41
ك	35122	7,17
ل	35223	7,19
م	34607	7,06
ن	52104	10,63
ه	8207	1,67
ه	66501	13,57
و	102228	20,86
ى	89935	18,35
ء	633	0,13
ڭ	335	0,07
گ	71	0,01
ۋ	4768	0,97
ژ	156	0,03
ې	4063	0,83

2.5. Osmanlıca OCR'da Problemler (Problems in Ottoman OCR)

Temelde bir karakter sınıflandırma problemi olan OCR bir görüntüdeki karakterlerinin görüntü işleme ve makina öğrenmesi teknikleriyle tespiti ve tanınması işlemidir. OCR süreci (i) görüntü ön işlemleri, (ii) görüntü bölümleme, (iii) özellik çıkarma, (iv) karakter tanıma, (v) son işlemler ve hata düzeltme olmak üzere genellikle 5 adımda gerçekleşir. Yukarıda sunulan Osmanlıca sıklık analizleriyle elde edilen Osmanlıca harf, katar ve kelime modelleri OCR sürecinin tüm adımlarında iyileştirme için kullanılabilir. Osmanlı alfabesinin OCR yönünden öne çıkan önemli bazı karakteristiklerine aşağıda değinilmiştir. Bu karakteristiklerin çoğu Arap, Fars, Urdu vb. alfabeleri için de söz konusudur. Osmanlıca OCR'da dikkat edilmesi gereken konuların başında kavislilik, el yazısı, bitişme, harflerin birden çok yazımı olması, harfler arası benzerlik, noktalar, hemze ve med gibi özel harekeler, yığılarak yazılan harf katarları, kelime bölümleme ve bitişme hataları, referans ve çıktı metinlerdeki kullanıcı ve yazılım kaynaklı hatalar sayılabilir.

Tablo 10. Katar sıklıkları (Ligature frequencies)

Katar	%	Katar	%	Katar	%
ا	12,46	ه	0,37	يې	0,19
و	6,55	شا	0,36	كې	0,19
ر	6,30	نك	0,36	قر	0,19
د	4,89	نا	0,32	كند	0,18
ه	3,28	پا	0,32	سند	0,18
ن	2,80	قد	0,31	نو	0,17
ى	2,45	طو	0,31	تو	0,17
ز	1,44	سى	0,30	يلر	0,17
بر	1,37	كر	0,29	ها	0,17
بو	1,04	لى	0,28	فر	0,17
م	0,89	ينه	0,28	چا	0,17
لر	0,87	صو	0,28	يتد	0,17
با	0,85	عا	0,27	د	0,16
كو	0,84	تا	0,26	كه	0,16
يد	0,83	مو	0,26	ۋ	0,16
ك	0,76	سو	0,25	غا	0,16
ب	0,72	خا	0,25	حر	0,15
لا	0,68	حا	0,25	ف	0,15
ند	0,66	له	0,25	مى	0,15
با	0,65	كا	0,24	سر	0,14
قا	0,60	نلر	0,24	سنه	0,14
ير	0,59	نى	0,22	ين	0,14
يو	0,59	طر	0,22	ش	0,14
ما	0,58	مر	0,22	جو	0,13
ل	0,56	يكى	0,21	تر	0,13
تا	0,54	جا	0,20	يو	0,13
ق	0,51	سا	0,20	ينك	0,13
لد	0,51	طا	0,20	يچو	0,13
يله	0,51	صا	0,19	تد	0,13
قو	0,48	هر	0,19	كى	0,13
يه	0,46	ينى	0,19	مه	0,13
لو	0,42	چو	0,19	خى	0,12
نه	0,42	مد	0,19	يغى	0,12
				مز	0,12

OCR'da yukarıdaki karakteristiklerden bağımsız olarak doğruluk oranına etki eden görüntü çözünürlüğü, görüntü kalitesi, tanınacak birim (gövde, harf, katar, kelime vb.), yazı hattı, eğitim veri kümesi, sözlüğün büyüklüğü ve kalitesi gibi başka önemli faktörler de söz konusudur.

3. Benzer Çalışmalar (Related Work)

Benzer çalışmalar Osmanlıca ve diğer Arap-tabanlı alfabelerde yapılan çalışmalar olarak iki bölümde incelenebilir. Osmanlıca OCR çalışmaları 2000'li yılların başına doğru başlamıştır [3-5]. OCR'da başarı elde etmek güç olduğundan OCR'a bir alternatif olarak görüntü içinde görüntü arama ve eşleştirme yöntemi ile arşivdeki dokümanlara harf, harf gövdesi veya harf katarı ile içerik tabanlı erişim çalışmaları da yapılmaya başlanmıştır [6-10]. 2000'li yıllarda klasik makine

öğrenmesi yöntemlerinden Support Vector Machines [11], Hidden Markov Models [3, 12] Linear Discrimant Analysis [13], Yapay Sinir Ağları [14] kullanarak yapılan çalışmalara ek olarak çizge-tabanlı algoritmalar [15] kullanarak ta OCR çalışmaları yapılmıştır. Çalışmalardan çoğunluğu karakterlerin tanınmasına yoğunlaşırken [3, 4, 5, 11, 12], içeriğe erişimi hedefleyen çalışmalar karakter gövdesi veya parçası denilen alt birimlere ya da birden fazla bağlı karakterden oluşan üst birimlerin tanınmasına yoğunlaşabilmektedir [6, 7, 15]. Bunların yanı sıra sadece satır veya karakter bölümleme gibi belli bir alt probleme odaklanan çalışmalar da yapılmıştır [4, 16].

Tablo 11. Harf sayısına göre kelime sıklığı
(Word frequency by letter count)

Harf #	Sıklık	%
1	5436	2,76
2	21130	10,73
3	26560	13,48
4	41110	20,87
5	41189	20,91
6	24754	12,56
7	17775	9,02
8	9210	4,67
9	5428	2,76
10	2524	1,28
11	1143	0,58
12	462	0,23
13	179	0,09
14	72	0,04
15	24	0,01
16	3	0,00
17	7	0,00
18	3	0,00
19	3	0,00
20	1	0,00
21	1	0,00

Tablo 12. Katar sayısına göre kelime sıklığı
(Word frequency by ligature count)

Katar #	Sıklık	%
1	46829	23,77
2	63077	32,02
3	49028	24,89
4	25177	12,78
5	9297	4,72
6	2747	1,39
7	667	0,34
8	145	0,07
9	35	0,02
10	10	0,01
11	1	0,00
12	1	0,00

Son zamanlarda derin öğrenme modelleri kullanılarak ta OCR çalışmaları yapılmıştır [17] [18] [19]. Bununla beraber son yıllarda endüstri [20], medikal [21] [22] vb. birçok alanda çeşitli problemlere çok başarılı bir şekilde uygulanan derin öğrenme OCR'a da başarıyla uygulanmıştır ve Arapça için oldukça iyi sonuçlar alınmıştır [23]. Yakın geçmişte derin sinir ağlarıyla yapılan Arap-tabanlı OCR çalışmalarına [24] [25] [26] [27] örnek verilebilir. [28]'de Kuran harflerini tanımak için CNN ve RNN mimarilerini birleştiren bir derin sinir ağı modeli kullanılmıştır. [25]'de 7-12 yaş aralığındaki çocukların el yazısını içeren yeni bir veri seti kullanılarak Arapça el yazısı tanımaya yönelik bir CNN mimarisi geliştirilmiş ve testlerde %88'lik bir karakter tanıma oranı elde edilmiştir. Arapça el yazısı tanıma üzerine son zamanlarda derin sinir ağlarıyla yapılan çalışmaların özeti ve kısa bir karşılaştırması [26]'te verilmiştir. [27] çalışmasında Arapça harflerdeki dikey ve yatay doğru parçalarını tanımlayan özellikler kullanmak yerine, parça enterpolasyonu (segment interpolation) yöntemiyle belirli pencere aralıklarında en iyi çizgi parçasını bulmayı sağlayan bir model geliştirilmiştir. Osmanlıca üzerinde yapılan diğer iki çalışmada [29, 24] ise OCR yerine Osmanlıca-Türkçe alfabe çevirisine odaklanılmıştır.

Tablo 13. Osmanlıca kelime sıklıkları (Ottoman word frequencies)

Kelime	%	Kelime	%	Kelime	%
و	2,34	صوكره	0,15	لى	0,10
بر	2,15	اولور	0,15	بتون	0,10
بو	0,95	اك	0,15	بكا	0,10
ده	0,80	ايچنده	0,15	سلطان	0,10
ايله	0,74	اولديغى	0,14	خان	0,10
يه	0,54	ايدن	0,14	برابر	0,10
نه	0,47	سنى	0,13	بويله	0,10
كبى	0,45	كون	0,13	اوچ	0,09
اولان	0,45	جمله	0,13	على	0,09
او	0,41	كه	0,13	بيله	0,09
ايچون	0,31	اى	0,13	الله	0,09
قدر	0,30	ايده	0,13	اون	0,09
نك	0,29	سنده	0,13	حالده	0,09
دخى	0,28	بيك	0,12	برى	0,09
ايدى	0,26	اولدى	0,12	ك	0,09
سى	0,26	چوق	0,12	ايتمش	0,09
هر	0,25	لر	0,12	دولت	0,09
ايكى	0,24	ايتمك	0,12	پك	0,09
ايسه	0,23	زمان	0,12	كندى	0,09
بى	0,23	ايدر	0,11	احسان	0,09
اولوب	0,23	اما	0,11	بيوك	0,09
فقط	0,22	محمد	0,11	افندى	0,08
سنه	0,21	داها	0,11	عليه	0,08
عايشه	0,20	بك	0,11	جك	0,08
اول	0,20	لرى	0,11	سن	0,08
بن	0,20	اوزرينه	0,11	اوله	0,08
پاشا	0,19	سنك	0,11	صكره	0,08
كى	0,19	اويله	0,10	بى	0,08
وار	0,16	هيچ	0,10	يعنى	0,08
ايدوب	0,16	ايتدى	0,10	شو	0,08
دن	0,16	ينه	0,10	ايلدى	0,08
اوزره	0,16	بنى	0,10	كيجه	0,08
در	0,15	رك	0,10	يالكز	0,08
				عدد	0,08

4. Derin Öğrenme Mimarisi (Deep Learning Architecture)

Bu bölümde OCR için kullanılan derin sinir ağları tabanlı derin öğrenme modeli tanıtılacaktır. Bu problem için kullanılan derin sinir ağı CNN ve RNN mimarilerini birleştiren bir CRNN mimarisidir.

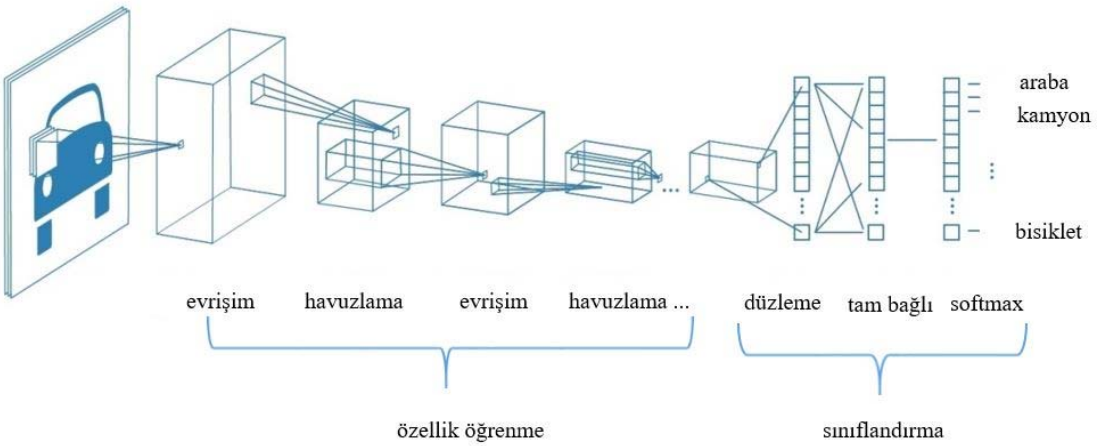
CNN'in ana amacı görüntüdeki görsel örüntüleri tanımaktır. Şekil 2'de görüldüğü gibi girdi verileri CNN'de katman denilen adımlardan geçerek işlenir. İlk katman özelliklerin belirlendiği evrişim (convolutional) katmanıdır. Bu katman çıkışında sisteme doğrusal olmayanlığın (non-linearity) tanımlandığı ReLU gibi fonksiyonlar kullanılmaktadır. Sonra boyut ve özellik sayısının azaltıldığı (down sampling) havuzlama (pooling) katmanı gelir. Girdi görüntülerinin büyüklüğü veya karmaşıklığına bağlı olarak CNN mimarisinde birden çok evrişim+havuzlama katmanı olabilir. Daha sonra iki boyutlu verinin bir sonraki katmanda kullanılabilmesi için tek boyuta dönüştürüldüğü normalizasyon ya da düzleme (flattening) katmanı gelir. Son adımda sınıflandırmanın yapılacağı tam bağlı (fully-connected) katmanlar gelir. Tam bağlı katmanların sonuncusunda genellikle softmax fonksiyonu ile sınıflandırma tamamlanır.

CNN mimarisinde sinir ağının her katmanında girdi verisini biraz daha soyut ve biraz daha üst düzeyli olarak temsil edilir. Örneğin ham veri bir dokümanın bir satırının 2 boyutlu gösterildiği bir piksel matrisi olsun. Bu durumda ilk katmanda soyut piksel ve kenarlar temsil edilebilir, ikinci katmanda kenar grupları ve komposizyonları temsil edilebilir, üçüncü katmanda harf gövdeleri ve dördüncü katmanda harflerin kendisi temsil edilebilir. Hangi özelliklerin seçileceği ve hangi özelliklerin hangi katmanda temsil edileceği de eğitim sırasında belirlenir. Bu sayede özellik belirleme ve özellik seçimi adımları derin öğrenme içinde çözümlenir. Verinin katmanlarda küçük ara yapılara dönüştürülmesi ve çok katmanlı mimaride tekrarlardan izole edilip sıkıştırılarak saklanması derin sinir ağlarının önemli özelliklerinden biridir. Derin öğrenmedeki "derinlik" terimi verinin dönüşüme uğradığı ara katmanları ifade etmektedir.

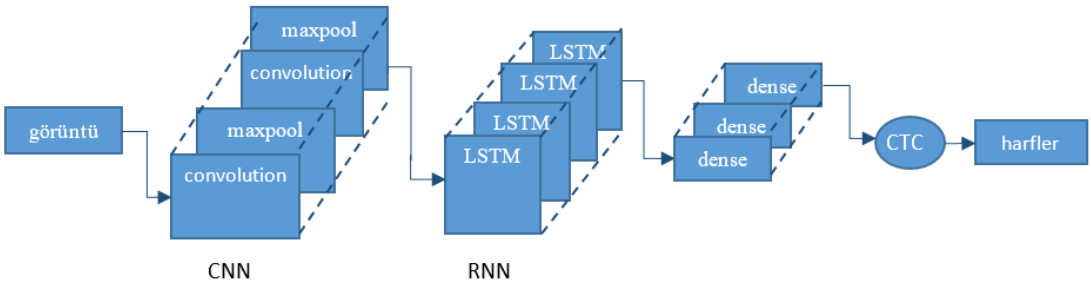
Derin sinir ağları kabaca CNN ve RNN olmak üzere iki çeşit mimariye sahiptir. CNN mimarisinde veri bir katmandan bir defa geçerken, RNN mimarisinde veri bir katmandan birden fazla geçebilmektedir. CNN mimarisi daha ziyade görüntü sınıflandırmada kullanılırken, RNN mimarisi otomatik konuşma sınıflandırması, dil tanıma, özetleme, kelime tahmini gibi zaman boyutu olan NLP problemlerine uygulanmaktadır. Metin bir karakter, katar ya da kelime dizisi (sequence) olarak modellendiğinde, yaygın bir RNN türü olan iki yönlü LSTM (Bidirectional Long Short-Term Memory) mimarilerinin bağlam bilgisini yani dizideki önceki ve sonraki öğeleri başarılı olarak hatırlayabildiği yani model belleğinde saklayabildiği, başka bir deyişle (zaman) dizilerini değişkenlikleriyle birlikte başarıyla öğrenebildikleri görülmektedir. Bu yüzden mimarimizde CNN'in peşi sıra iki yönlü LSTM kullanılmıştır. LSTM'in CTC (Connectionist Temporal Classification) ile birlikte kullanımı birçok NLP probleminde başarılı sonuçlar vermektedir. Bu yüzden -bir satırdaki harfleri dizi olarak düşünerek- LSTM'in peşi sıra dizideki harfleri etiketlemek için bir CTC katmanı kullanılmıştır. CTC katmanı girdilere bakarak bir çıktı dizisinin olasılığını hesaplamaktadır. Böylece dizideki her bir harf tahmin edilmektedir.

Öncelikle sayfa görüntüleri satır, kelime ve karakter bölütleme aşamalarından geçirilerek işaretlenir. Bölütleme için imagemagick ve openCV kütüphanelerinden yararlanılmıştır. Orijinal veride yoğun olarak elle işaretleme ve/veya düzeltme yapılması gerekmektedir. Sentetik veri hazır araçlar kullanılarak algoritmik olarak üretildiğinden ve görüntüler kaliteli olduğundan bölütleme problemsiz olarak ilerlemektedir. Mimariye girdi olarak satır görüntüsü verilir ve satırdaki karakterler dizi olarak teker teker tanınır. CNN katmanlarında satır görüntüsü soldan sağa yukarıdan taranarak özellikler çıkarılır.

Deneylerde kullanılan CRNN mimarisi bir CNN, bir RNN, bir yoğun katman ve bir CTC katmanından veya fonksiyonundan oluşmaktadır. Mimarinin genel yapısı Şekil 3'te verilmiştir. Mimarinin CNN bölümü 16 filtreli 3x3 convolution içeren bir katman ve 3x3'lük bir maxpool katmanından oluşmaktadır. Mimarinin RNN bölümü 4 katmanlıdır ve her katman bir adet LSTM içermektedir. Birinci katman y ekseninde 64 çıktılı bir LSTM, ikinci ve üçüncü katman x ekseninde 128 çıktı ile özetleme yapan biri ileri yönlü diğeri geri yönlü olmak üzere iki yönlü (bidirectional) bir LSTM ve dördüncü katman x ekseninde 256 çıktı üreten ileri yönlü bir LSTM'den oluşmaktadır. RNN bölümünden sonra sınıflandırmanın yapıldığı bir yoğun (dense) katman ve bu katman çıktısını yorumlayan bir CTC işlevi kullanılmaktadır. Eğitim sırasında doküman satırlara bölümlendikten sonra satır görüntüsü CRNN'e girdi olarak, satır görüntüsündeki metin ise çıktı olarak kullanılır. Eğitim sırasında; learning rate = 0.002, momentum = 0.5, epochs = 3.000.000 olarak alınmıştır. Eğitim sırasında kullanılan CRNN mimarisinin hiper parametre özeti Tablo 14'te verilmiştir.



Şekil 2. Görüntü tanımada kullanılan standart CNN mimarisi [30] (Conventional CNN architecture used in image recognition)



Şekil 3. Osmanlıca OCR için CRNN mimarisi (CRNN architecture for Ottoman OCR)

5. Veri Kümesi (Data SET)

Bu bölümde Osmanlica.com derin öğrenme modelini eğitmek için kullanılan eğitim veri kümesi ve karşılaştırma için kullanılan test veri kümesi tanıtılacaktır. Veri setlerinin sayfa, satır, kelime ve harf sıklıkları Tablo 15'te verilmiştir.

5.1. Eğitim Veri Kümesi (Training Data Set)

Eğitim verisi orijinal, sentetik ve hibrit olmak üzere üç farklı kümeden oluşmaktadır:

- *Orijinal veri*: Değişik Osmanlıca eserlerden yaklaşık 1000 sayfa görüntü toplanmış ve bunlar yarı otomatik yöntemlerle metin dosyasına dönüştürülmüştür. Bu veri kümesi yaklaşık 18 bin satır, 35 bin kelime, 252 bin karakterden oluşmaktadır. Bu veri kümesindeki sayfalar yaklaşık olarak 1400 x 2000 pixel boyutlarında ve 300 dpi çözünürlüktedir. Bir sayfada ortalama 20 satır bulunur. Font boyu yaklaşık 12 nokta ve satır yüksekliği 48 noktadır.
- *Sentetik veri*: Orijinal veri hazırlamak uzun ve zahmetli olduğundan metin-görüntü dönüşüm araçları kullanarak sentetik bir veri kümesi hazırlanmıştır. Önce değişik eserlerden metinler toplanmış ve peşi sıra bu metinler 70 farklı Arapça fontuyla görüntü dosyalarına dönüştürülmüştür. Bu veri kümesi yaklaşık 26 bin sayfa, 1.3 milyon satır, 263 bin kelime, 78 milyon karakterden oluşmaktadır. Bu veri kümesindeki sayfalar 2500 x 4800 nokta boyutlarında ve 300 dpi çözünürlükte olup font boyu 12 noktadır. Her sayfada 42 satır olup satır yüksekliği 48 noktadır.
- *Hibrit veri*: Orijinal ve sentetik veri kümelerinin birleşimidir.

5.2. Test Veri Kümesi (Test Data Set)

Test için 8 farklı eserden 21 orijinal sayfa görüntüsü kullanılarak bir veri kümesi hazırlanmıştır. Test kümesi eğitim kümesinde kullanılmamış orijinal sayfaları içermektedir. Test kümesi hazırlanırken eserlerden kalitesi düşük, harfleri silik ve farklı kâğıt rengine sahip örnekler seçilmeye çalışılmıştır. Her sayfa ortalama 20 satırdan, her satır 7 kelime ve 55 karakterden oluşmaktadır. Test kümesi osmanlica.com/test adresinde paylaşılmıştır. Paylaşım görüntü dosyaları, altışar adet OCR test çıktısı, doğru metni dosyaları ve difflib ile doğruluk oranı hesaplayan bir adet Python dosyası içermektedir.

Tablo 14. CRNN hiper parametreleri (CRNN hyperparameters)

Katman	Çıktı	Parametre
Conv2D	3x3 16 filtre + tanh	160
MaxPool2D	3x3	0
LSTM1	64 y ekseninde	20736
LSTM2	128 x ekseninde ileri	98816
LSTM3	128 x ekseninde geri	131584
LSTM4	256 x ekseninde	394240
Yoğun katman	71 düğüm	18247
Toplam		663703

Tablo 15. Veri kümesi sıklıkları (B: Bin, M: Milyon) (Dataset frequencies) (B: Thousand, M: Million)

Küme	Sayfa	Satır	Kelime	Karakter
Sentetik	26B	1.3M	263B	78M
Orijinal	1B	18B	35B	252B
Eğitim	27B	1.3M	298B	78M
Test	21	420	3B	23B

6. Deney, Karşılaştırma ve Sonuçlar (Experiment, Comparison and Results)

Bu bölümde önce karşılaştırma deneylerinde kullanılan OCR araçları tanıtılacak, sonrasında karşılaştırmada kullanılan metin türleri tanımlanacaktır. Sonrasında karşılaştırma deneyleri sonuçları paylaşılacak ve sonuçlar kısaca yorumlanacaktır.

6.1. Karşılaştırmada kullanılan OCR Araçları (OCR Tools Used in Comparison)

Bu çalışmada Osmanlica.com projesinde geliştirilen OCR modeli bazıları ticari, bazıları ücretsiz, bazıları açık kodlu çok bilinen OCR araçları ya da modelleriyle karşılaştırılmıştır. Karşılaştırmada kullanılan test kümesinin OCR araçlarının model eğitiminde kullanılmadığı var sayılmıştır. Test sonuçları bu varsayımı destekler mahiyettedir:

- *Tesseract Arapca ve Farsça*: Tesseract 4.0 ile Arapça ve Farsça için eğitilmiş hazır modellerdir. Bu modeller yaklaşık 400000 satır ve 4500 farklı fontla eğitilmiştir. Tesseract açık kodlu ve ücretsiz bir yazılımdır. Tesseract birçok farklı dil için yüksek başarımlı OCR modelleri olan ve yaygın olarak kullanılan bir sistemdir.
- *Abby FineReader*: Finereader piyasada kullanılan en yaygın ticari OCR araçlarından bir tanesi olup Arapça ve Farsça dâhil çok sayıda dil için OCR hizmeti sağlamaktadır. Bu karşılaştırmada versiyon 15 kullanılmıştır.
- *Google Docs*: Google'ın Google Documents kapsamında kullanıcılarına ücretsiz olarak sunduğu kodu kapalı bir OCR hizmetidir. Online olarak kullanıma açıktır. Yaygın olarak kullanılan bir OCR aracıdır.
- *Miletos*: Miletos adlı firma tarafından Osmanlıca için geliştirilen ve online OCR hizmeti sağlayan ticari bir OCR aracıdır. Miletos Osmanlıca için özel olarak geliştirildiğinden karşılaştırmaya dâhil edilmiştir.

6.2. Doğruluk oranı için kullanılan metin çeşitleri (Text types used for recognition accuracy)

Tanıma işleminde doğruluk oranları OCR çıktısı metin ile doğru metindeki birimleri yani karakter, katar ya da kelimelerin karşılaştırılmasıyla hesaplanır. Doğruluk oranı Python difflib kütüphanesinin SequenceMatcher.ratio fonksiyonu kullanılarak hesaplanmıştır: $doğruluk = 2.0*M/T$ formülünde T referans ve hesaplanan (OCR çıktısı) metinlerindeki toplam birim sayısını, M eşleşen yani doğru tanınan birimlerin sayısını ifade eder. Eğer metinler birebir aynıysa doğruluk 1.0, eğer eşleşen birim yoksa doğruluk 0.0 olarak hesaplanır. İki metin arasındaki farklılıklar ekleme (insertion), silme (deletion) ve yer değiştirme (substitution) işlemleri cinsinden ifade edilir. OCR çıktısındaki *eklemeler* doğru metinde yer almayan birimlere, *silinenler* atlanılan ya da tanınmayan birimlere ve *yer değiştirenler* de farklı bir birim olarak tanınan birimlere karşılık gelmektedir.

Bu çalışmada doğruluk oranı 3 farklı metin türü için hesaplanmıştır: Ham metin (raw text), normalize metin (normalized text) ve bitişik metin (joined text).

- *Ham metin:* Metnin herhangi bir işleme tabi tutulmamış yani olduğu gibi tabir edilen haline ham metin denir. Bu metin ile hesaplanan doğruluk oranları metin çok sayıda hata içerdiğinden olması gerekenden daha yüksek ve yanıltıcıdır.
- *Normalize metin:* Osmanlı alfabesinin yukarıda bahsedilen kendine özgü karakteristikleri, program ve kullanıcı hataları OCR çıktılarında metin düzenleme ve hata düzeltme işlemlerini zorunlu

kılmaktadır. Bu yüzden OCR çıktısındaki hataları gidermek için metinlerin normalize edilmesi zorunludur. Metin normalizasyonu 3 adımda yerine getirilir:

- *Boşluk normalizasyonu:* Standart dışı boşluk karakterlerinin normal boşluk karakterine dönüştürülmesi ve çoklu boşluk karakterlerinin tekil boşluk karakterine dönüştürülmesi bu aşamada gerçekleşir.
- *BIDI normalizasyonu:* BIDI algoritması bir metin içerisinde kullanılan alfabe ve karakterlerin diziliş ya da yazılış yönüne göre metni biçimlendiren ve bir standarda bağlanmış bir algoritmadır. Diziliş yönü Osmanlıcadaki gibi sağdan sola (RTL), rakamlardaki gibi soldan sağa (LTR) ve çiftler halinde kullanılan (), [], {}, … vb. karakter çiftleri için tarafsız (neutral) olabilir. Görüldüğü gibi Osmanlıcada metin içinde her üç diziliş yönü de kullanılmaktadır. Bu durumlarda üretilen ya da oluşturulan metin ile gösterilen metin arasında farklılıklar söz konusu olabilmektedir. Özellikle metinlerin üretildiği ve tüketildiği yazılımlar farklı olduğunda bu farklar görünür hale gelebilmektedir. Metin üretilirken (yazılırken) ve tüketilirken (okunurken) kullanılan BIDI algoritmalarının versiyonları ya da implementasyonları farklı ise elde edilen metinler farklı olabilmektedir. BIDI normalizasyonda metinde yanlış dizilmiş veya yanlış gösterilen karakterlerin düzeltilmesi yapılır. Özellikle yönü yanlış olan ve eşleşmeyen parantez ve tırnak çiftlerinin düzeltilmesi bu aşamada yapılır.
- *Karakter normalizasyonu:* Osmanlıca için henüz standart bir alfabe tanımlanmamış olduğu için OCR, editör gibi farklı yazılımlarda Arap veya Fars alfabesi kullanılmaktadır. Bu alfabelerde bir harf ya da noktalama işareti için birbirinden küçük farklılıklarda ayrışan birden fazla karakter bulunmaktadır. Bu yüzden veri setlerindeki doğru metinlerde ya da OCR çıktılarındaki hesaplanan metinlerde Osmanlıcadaki karakterler için farklı birçok alternatif karakterler ortaya çıkmaktadır. Bu karakterlerin normalize edilerek Tablo 19'deki karakterlere dönüştürülmesi bu aşamada yapılır.
- *Bitişik metin:* Metindeki tüm boşlukların silinerek kelimelerin bitiştirilmesiyle oluşan boşluksuz uzun karakter dizisine bitişik metin denir. Bitişik metin ile doğruluk oranı hesaplanmasının amacı Osmanlıcadaki kelime bölümleme ve bitişme probleminden bağımsız olarak doğruluk oranını hesaplayabilmektir. Yukarıda bahsedildiği gibi kelime içi ve kelimeler arası boşluklar birbiriyle sıkça karışmaktadır.

6.3. Karakter tanıma oranları (Character Recognition Rates)

Karakter tanıma doğruluk oranları ve hata dağılımları Tablo 16'te ham ve normalize OCR çıktıları için verilmiştir. Ham metinle hesaplanan doğruluk oranları %73 ile %89 arasında değişiklik göstermektedir. En düşük tanıma %73 ile Osmanlica Sentetik modelinde, en yüksek tanıma %89 ile Osmanlıca Hibrit modelindedir. Hatalardan arındırılmış normalize metinde doğruluk oranları ise %78 ile %96 arasında değişmektedir. En düşük tanıma %78 ile Osmanlica Sentetik modelinde, en yüksek tanıma %96 ile Osmanlıca Hibrit modelindedir. Bitişik doğruluk oranı yönünden araçlar arasındaki sıralama değişiklik göstermemektedir: Tesseract Farsça (%80), Finereader (%81), Tesseract Arapça (%81), Miletos (%87), Google Docs (%93), Osmanlica.com (%97). Bitişik doğruluk oranıyla %97'nin üzerinde doğruluk oranı hesaplanmaktadır. Bitişik doğruluk oranında normalize oranlara göre en yüksek iyileşme yaklaşık %1 ile Google Docs ve Osmanlica.com'da gözlenmiştir. Özet olarak Osmanlica.com'un %88,64 (ham), %95,92 (normalize) ve %97,18 (bitişik) doğruluk oranlarıyla en yakın alternatif modelden yaklaşık %4 daha iyi sonuç verdiği görülmüştür. Karakter tanıma hatalarının Tablo 17'deki kelime içindeki konumlarına göre dağılımları incelendiğinde hatalarının büyük çoğunluğunun 2. ve 8. konumlar arasında ortaya çıktığı görülmektedir. Karakter tanıma hatalarının bitişik karakter katarı içinde 2., 3., 4. ve 5. konumlarda oluştuğu görülmektedir (Tablo 18).

Osmanlica Hibrit modeli için normalize metinler kullanılarak hesaplanan hata oranları karakter bazında Tablo 19'de verilmiştir. Tabloda karakterler Arapça harfler, Farsça ve Türkçe harfler, noktalama işaretleri ve sayılar olarak gruplanmıştır. Her satırda bir karakter, Unicode değeri ve model bazında hata oranı verilmiştir. Karakter hatalarının Osmanlıca karakterlerde, noktalama işaretlerinde ve sayılarda yüksek olduğu görülmektedir.

Tablo 17. Kelimedeki konuma göre karakter tanıma hatası
(Character recognition errors by position in word)

Konum	Sıklık	%
1 harf	30	1,12
2 harf	189	7,08
3 harf	470	17,60
4 harf	407	15,24
5 harf	459	17,19
6 harf	377	14,12
7 harf	288	10,79
8 harf	205	7,68
9 harf	120	4,49
10 harf	78	2,92
11 harf	26	0,97
12 harf	10	0,37
13 harf	6	0,22
14 harf	2	0,07
15 harf	1	0,03
16 harf	1	0,03

Bu karakter kümelerinin sıklığı diğerlerine göre çok daha düşük olması hata oranını yükselten sebeplerden birisidir. Karakter hata oranları tüm modellerde benzer bir örüntü göstermektedir. Karakter hatalarını ayrıntılı incelemek için burada yeterli alan olmamakla birlikte Miletos'un rakamları tanımada nerdeyse sıfır hata yaptığı, Tesseract Arapça Osmanlıca harfler ve rakamlarda %100 hata yaptığı hemen göze çarpan gözlemlerdendir.

Tablo 16. Karakter tanıma doğruluk oranı ve normalize metin hata dağılımları (%)
(Character recognition accuracy rate and normalized text error distribution)

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 18. Kelimedeki konuma göre karakter tanıma hatası
(Character recognition errors by position in word)

Konum	Sıklık	%
1 katar	29	1,43
2 katar	789	39,50
3 katar	558	27,60
4 katar	358	17,70
5 katar	174	8,61
6 katar	66	3,26
7 katar	29	1,43
8 katar	5	0,24
9 katar	2	0,09
10 katar	1	0,04

6.4. Katar tanıma oranları (Ligature Recognition Accuracy)

Osmanlıca OCR'da bağlı karakter katarları kelimeyi oluşturan alt birimlerdir. Bu birimlerin doğru tanınması kelimelerin doğru tanınabilmesi için önemlidir. Osmanlıcada geçen en sık katarlar yukarıda verilmişti. Tablo 20'de katar tanıma doğruluk oranları ve katar tanıma hataları paylaşılmıştır. Bir katarın doğru tanınması için içindeki tüm karakterlerin doğru tanınması gereklidir. Bir katarda en az bir karakterin hatalı tanınması o katarın hatalı tanınması için yeterlidir. Katar tanıma oranı ham metinde %51 (Fine Reader) ile %80 (Osmanlıca Hibrit) arasında, normalize metinde %73 (Miletos) ile %91 (Osmanlıca Hibrit) arasında değişmektedir. Osmanlıca Sentetik modelin katar tanımada çok zayıf kaldığı hemen görülmektedir.

Tablo 19. Normalize metinle modellerin karakter tanıma hatası, %) (Character recognition errors by model using normalized text %)

Harf	Unicode	Hibrit	Orijinal	Sentetik	G.Docs	TessAr.	TessPe.	Miletos
ا ااا	0627	1,5	2,7	11,0	3,4	6,5	4,5	4,8
ب بببب	0628	6,2	7,1	52,1	13,7	40,4	43,9	21,3
ت تتتت	062a	5,8	7,5	35,4	12,6	25,7	34,8	21,9
ث ثثث	062b	20,0	35,1	80,0	34,3	38,9	62,5	83,3
ج ججج	062c	2,9	9,9	33,5	4,7	33,3	32,8	65,7
ح حححح	062d	2,1	3,6	22,8	4,3	13,7	14,9	25,0
خ خخخ	062e	2,0	2,0	35,3	7,7	26,9	18,4	16,0
د ددد	062f	1,1	2,1	10,7	2,9	6,5	3,8	9,3
ذ ذذذ	0630	2,7	10,8	45,1	0,0	24,4	18,8	75,0
ر ررر	0631	0,7	2,2	10,3	3,4	8,8	7,0	4,7
ز زززز	0632	1,8	7,1	17,5	16,0	28,3	30,6	14,7
س سسس	0633	3,6	4,8	32,8	8,0	19,1	13,4	25,8
ش ششش	0634	1,9	1,9	34,7	3,8	20,6	16,5	20,0
ص صصص	0635	0,8	3,0	48,1	3,8	9,8	7,3	39,1
ض ضضض	0636	1,7	3,4	39,7	6,8	8,5	13,0	33,3
ط طططط	0637	0,0	2,9	28,0	5,8	5,8	7,7	13,0
ظ ظظظ	0638	2,0	0,0	30,0	4,1	10,2	13,7	35,7
ع ععع	0639	0,7	5,4	45,4	3,7	15,6	19,1	22,4
غ غغغ	063a	2,7	6,4	40,4	23,9	43,9	48,4	51,9
ف ففف	0641	2,6	5,5	43,1	3,4	19,1	17,5	27,9
ق قققق	0642	3,1	2,0	32,7	8,5	18,1	19,5	28,7
ك ككك	0643	0,8	1,1	24,3	5,6	12,9	14,7	9,2
ل لل	0644	1,7	3,5	31,9	10,0	30,3	24,8	14,9
م ممم	0645	2,3	5,3	47,9	9,0	22,9	24,9	25,1
ن نnn	0646	4,7	5,5	31,1	10,0	34,4	22,0	14,8
ه/ ه	06d5-0647	4,8	8,2	25,4	5,2	21,0	9,2	9,5
و ووو	0648	2,6	3,8	9,9	3,2	5,1	3,9	5,4
ى ییی	06cc	4,6	5,8	29,4	15,9	34,6	39,8	19,3
ء ءءء	0621	50,0	66,7	96,9	25,0	57,1	33,3	0,0
ڭ ڭڭڭ	06ad	3,9	36,7	22,2	100,0	100,0	100,0	100,0
گ گگگ	06af	25,0	54,2	40,7	29,2	100,0	39,3	0,0
چ چچچ	0686	5,8	7,2	45,1	50,6	100,0	71,6	20,0
ژ ژژژ	0698	0,0	100,0	100,0	0,0	100,0	50,0	0,0
پ پپپ	067e	9,8	19,5	63,2	54,8	100,0	89,2	84,6
/.-	002e-06d4	17,6	18,8	39,5	18,3	31,4	41,4	18,9
،	060c	29,0	29,4	28,2	23,5	98,9	100,0	12,1
؛	061b	53,8	50,0	38,9	28,6	37,5	35,7	0,0
؟	061f	0,0	29,4	51,7	5,9	14,3	0,0	100,0
:	003a	0,0	21,4	52,4	0,0	29,4	18,8	0,0
-	002d	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	0020	16,1	16,8	23,6	18,1	16,4	16,8	22,9
()	0028-0029	5,9	4,0	32,7	6,9	9,1	11,7	15,2
٠	0660	40,0	60,0	89,7	66,7	70,0	75,0	0,0
١	0661	11,1	0,0	58,8	33,3	83,3	50,0	100,0
٢	0662	25,0	22,2	50,0	16,7	100,0	0,0	100,0
٣	0663	0,0	40,0	28,6	40,0	100,0	0,0	0,0
٤	0664	0,0	25,0	83,3	100,0	100,0	100,0	0,0
٥	0665	28,6	37,5	80,8	100,0	100,0	75,0	0,0
٦	0666	40,0	20,0	42,9	100,0	100,0	100,0	0,0
٧	0667	0,0	0,0	20,0	25,0	100,0	20,0	0,0
٨	0668	16,7	28,6	44,4	33,3	100,0	20,0	0,0
٩	0669	0,0	0,0	77,8	0,0	0,0	50,0	100,0

Orijinal veri setiyle eğitilen modelin tanımada gayet başarılı olduğu, hibrit model sentetik modelin yaklaşık %2'lik bir katkı sağladığı görülmektedir.

6.5. Kelime tanıma oranları (Word Recognition Accuracy)

Veri kümesindeki kelimelerin karakter sayısına göre normalizasyon sonrası dağılımları Tablo 13'da verilmişti. Tablo 21'da ise ham ve normalize metinlerde kelime tanıma oranları verilmiştir. Bir kelimenin hatalı tanınması için kelimedeki en az bir karakterin hatalı tanınması, doğru tanınması içinse kelimedeki tüm karakterlerin doğru tanınması gerekmektedir. Ham metinde kelime tanıma oranları %15 (Miletos) ile %44 (Osmanlıca Hibrit) arasında, normalize metinde ise %24 (Fine Reader) ile %66 (Osmanlıca Hibrit) arasında değişmektedir. Kelime tanımada Osmanlıca Hibrit, Osmanlıca Orijinal modelin diğer araçlardan belirgin şekilde daha iyi oranlar ürettiği göze çarpmaktadır. Osmanlıca modeller haricinde sadece Google Docs aracı %50 kelime tanıma oranını geçebilmiştir.

Kelime tanıma hata oranları Tablo 21'de 2., 3. ve 4. sütunlarda değişen (yer değişimi/substitution) kelime sayısı, tanınmayan (silme/deletion) kelime sayısı ve sonradan ortaya çıkan (ekleme/insertion) kelime sayıları olmak üzere yüzde olarak verilmiştir. Ham metindeki tanıma oranları alan kısıtından dolayı verilmemiştir. Kelime hata oranları genelde karakter hata oranlarına göre çok daha yüksek çıkmaktadır. Bir kelimenin doğru tanınması için kelimedeki harflerin hepsinin doğru tanınması gerektiğinden kelime tanıma oranları karakter tanımaya göre çok daha düşüktür.

Hataların çok büyük çoğunluğunun karakter tanıma kaynaklı olarak ortaya çıkan kelime değişim hatası olduğu görülmektedir. Hataların çok küçük bir çoğunluğu ise kelime bölümleme ve bitişme hatalarına sebep olan karakter tanıma hatalarından kaynaklanan kelime ekleme ve silinme hatalarıdır. Çeşit yönünden incelendiğinde en çok kelime değişim hatası Hatalar kendi içinde incelendiğinde en çok değişim (%60-75) hatası, sonra ekleme (~%1-15) ve silme (~%1-15) hataları oluşmaktadır. Kelime değişim hataları karakterlerin yanlış tanınmasından, ekleme ve silme hataları ise kelime bölümleme ve bitişme hatalarından kaynaklanmaktadır.

6.6. Harf türüne göre karakter tanıma oranları (Character Recognition Accuracy by Letter Type)

Osmanlıca harflerin ayırt edici bazı temel özelliklere göre gruplanması yukarıda Tablo 3'te verilmişti. Bu harf gruplarına göre karakter tanıma hataları Tablo 22'te verilmiştir. Alan kısıtından dolayı tabloda sadece normalize metin kullanılarak hesaplanan hata oranları verilmiştir. Tabloda her harf türü için 2 farklı hata oranı paylaşılmıştır. Grup sütununda metinde o gruptaki karakterlerin ne kadarının hatalı tanındığı, Genel sütununda ise metindeki tüm karakter içinde o

Tablo 20. Katar tanıma doğruluk oranı ve hata dağılımları (Ligatura recognition accuracy and error distributions) (%)

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 21. Kelime tanıma doğruluk oranı ve hata dağılımları (Word recognition accuracy and error distributions) (%)

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 22. Normalize metinle harf türüne göre grup içi ve genel karakter hataları (%)
(In-group and global character errors by letter type using normalize text (%)

	Orijinal		Sentetik		Hibrit		GoogleD.		FineRead.		T. Arapça		T. Farsça		Miletos	
Türü	Grup	Genel	Grup	Genel	Grup	Genel	Grup	Genel	Grup	Genel	Grup	Genel	Grup	Genel	Grup	Genel
Arapça	6,0	60	27,6	78	4,7	59	9,8	67	25,2	76	28,3	78	21,7	79	18,5	77
Osmanlıca	22,8	3	44,4	2	8,4	1	57,8	5	100	3	100	3	74,9	3	29,0	1
Noktalı	6,7	22	38,4	35	5,0	20	14,3	34	40,4	38	40,7	36	34,5	41	29,6	36
Noktasız	6,0	42	22,7	44	4,6	39	8,2	37	19,5	41	23,8	45	16,6	41	14,0	41
Nokta üstte	5,7	12	34,1	21	4,0	11	11,0	16	28,4	18	34,9	20	26,7	21	21,5	19
Nokta altta	8,6	9	46,9	14	7,1	10	19,2	18	66,4	20	52	15	49,8	20	50,7	17
1 nokta	6,0	11	37,4	19	4,3	10	9,7	12	28,8	16	38,1	19	30,5	20	20,4	15
2 nokta	6,7	7	39,3	12	6,1	8	17,3	17	55,6	17	41,2	13	38,4	16	48,5	18
3 nokta	10,2	3	41,2	3	5,0	2	27,2	6	58,8	5	54,2	4	43,5	5	25,7	3
Harf	6,2	63	27,7	79	4,7	60	10,3	71	26,0	79	29,1	81	22,3	83	18,6	78
Diğer	19,3	4	44,4	3	16,2	4	22,4	3	46,6	3	59,9	3	33,1	2	25,6	3

gruptaki karakterlerin meydana getirdiği hata oranı verilmiştir. Örneğin Arapça satırında Orijinal/Grup sütunundaki 6.0 Arapça türündeki karakterlerin %6.0'ının hatalı tanındığını, Orijinal/Genel sütunundaki 60 tüm hataların içinde bu hataların payının %60 olduğunu ifade etmektedir.

Arapça/Osmanlıca grubu incelendiğinde Osmanlıca harflerinin frekanslarının Arapçalara göre çok daha az olmasına rağmen, Osmanlıca karakterlerde hataların göreceli olarak çok yüksek olduğu görülmektedir. Buna rağmen Osmanlıca Hibrit modeldeki %59'luk oranla Arapça harflerdeki tanıma hatalarının OCR toplam hatasının çok büyük bir kısmını oluşturduğu görülmektedir. FineReader, Tesseract Arapça ve Farsça modellerde Osmanlıca kelimelerin hemen hemen tamamının tanınamadığı görülmektedir. Noktalı/noktasız grubuna bakıldığından noktalı harflerdeki hata oranının noktasızlardakinden 2 kat daha yüksek olduğu, noktasızların tüm hataların yaklaşık %50-45'ini oluşturduğu, noktalılardaki hataların ise tüm hataların %35-40 oluşturduğu göze çarpmaktadır. Noktası altta ve üstteki harflerin tüm hatalardaki payı yaklaşık olarak %10-20 arasında değişmekte olup genel hataya benzer oranda katkı yaptıkları görülmektedir. Noktası altta olan harflerin kendi içindeki tanıma hatası noktası üstte olan harflere nazaran ortalama 2 kat daha yüksektir. Google Docs, FineReader, Tesseract Arapça modelleri göz önüne alındığında harflerde nokta sayısı arttıkça tanıma hatasının arttığı görülmektedir. Toplam hataya katkı düşünüldüğünde ise genel anlamda bu durumun zıddı bir durum ortaya çıkmaktadır. Yani 3 noktalı harflerdeki hata toplam hataya en düşük katkıyı, 1 noktalı harfler ise en yüksek katkıyı sağlamaktadır. Harf/diğer (noktalama+rakam) gruplamasında harf tanıma hatalarının diğer karakter hatalarına nazaran toplam hataya kat kat daha fazla katkı yaptığı hatta toplam hatanın %60-80'ini teşkil ettiği görülmektedir. Harflerin kendi için %5-30'u hatalı tanınırken bu oran diğer karakterlerde yaklaşık iki üç kat daha fazla gözlenmektedir.

6.7. Hiper parametre kestirimi (Hyper-Parameter Tuning)

Derin sinir ağlarında uygun hiper parametre kestirimi önem arz eden bir konudur. Derin sinir ağlarından önce bir problemi çözmek için önce öznitelik kümesinin belirlenmesi, sonrasında bunların içinden modelin performansını optimize edecek en iyi özniteliklerin bulunması (feature engineering) üzerinde durulurdu. Derin sinir ağlarıyla beraber yapay sinir ağının mimarisi, ağın kaç katmandan oluşacağı, her katmandaki düğüm sayısı, başlangıç ağırlık değerleri, seçilen optimizasyon algoritması (stochastic gradient descent, adagrad, adadelta, adam, adamax), seçilen aktivasyon fonksiyonu (sigmoid, tanh, ReLu, PReLu), seyreltme (dropout) değeri, CNN'de filtre (kernel) boyutu ve pooling fonksiyonu, öğrenme hızı (learning rate), momentum katsayısı, eğitim turu (epoch) sayısı, vb. hiper parametrelerin kestirimi veya seçimi daha önemli hale geldi. Derin sinir ağı oluşturulurken hiper parametre seçimi sezgisel bir yaklaşımla genel geçer kurallara ve probleme hakkındaki ön bilgiye bağlı olarak yapılır. En iyi parametre değerlerini baştan seçmek çoğu zaman mümkün değildir. En iyi değerleri hesaplayan belirli bir algoritma da söz konusu değildir.

Bununla birlikte hiper parametre kestirimi için aralık-tabanlı arama (GridSearch), rastgele arama (RandomSearch) ve manuel arama (ManuelSearch) gibi yaklaşımlar kullanılmaktadır. Bu çalışmada zaman kısıtından dolayı manuel arama yaklaşımı kullanılmış ve sadece 4 hiper parametre için deneyler *hibrit veri seti* için tekrar edilmiştir. Deneyler yukarıdakiler gibi detaylı olarak yapılmış olmasına rağmen, sayfa kısıtından dolayı detaylı sonuçlar yerine yalnızca genel doğruluk oranları özet olarak Tablo 23-Tablo 25'te verilmiştir. Tabloların birinci satırındaki *orijinal deney* ifadesi Tablo 14'deki hiper parametre değerleriyle yapılan Osmanlıca.com modeline işaret etmektedir. Hiper parametre kestirim çalışmasındaki birinci deneyde (*filtre/kernel boyu*) CNN Conv2D katmanında filtre boyutu 3x3 ten 5x5'e değiştirilmiş, ikinci deneyde *tanh* fonksiyonu yerine *ReLU* aktivasyon fonksiyonu kullanılmış, üçüncü deneyde LSTM katmanlarında 64, 128, 128, 256 değerleri yerine 64, 64, 64, 128 değerleri kullanılmış ve dördüncü deneyde 0.0002 olan öğrenme katsayısı 0.0005 olarak değiştirilmiştir. Bu dört deneyin karakter, katar ve kelime tanıma doğruluk oranları tabloların sonraki 4 satırında paylaşılmıştır.

Deney sonuçları incelendiğinde *Öğrenme hızı* parametresi ham metin ile katar tanıma deneyi hariç tüm deneylerde doğruluk oranlarında herhangi bir artış elde edilememiştir. Doğruluk oranlarında sağlanan artışa göre deney sonuçları sıralandığı zaman en kötü değer seçiminin LSTM parametresinde, sonra aktivasyon fonksiyonu seçiminde, sonra da filtre boyutu seçiminde olduğu gözlenmiştir. Her ne kadar yukarıdaki deneylerde belirgin bir başarı elde edilmemiş olsa da, farklı hiper parametre değerleri ve bu değerlerin farklı kombinasyonlarıyla yapılacak çok fazla sayıda deney vardır. İleride hiper parametre kestirim yöntemleri kullanılarak yapılacak yeni deneylerle daha yüksek doğruluk oranı veren derin sinir ağı modellerinin eğitilmesi mümkün olacaktır.

Tablo 23. Hiper parametre kestirim deneyi I: Karakter tanıma
(Hyperparameter estimation experiment I: Character recognition)

Deney	Parametre	Ham	Normalize	Bitişik
Orijinal deney	663703	88,86	96,12	97,37
Öğrenme hızı	663783	88,63	95,63	96,80
LSTM boyutu	194919	88,01	95,26	96,43
Aktivasyon fonk.	663783	88,26	95,33	96,51
Filtre boyutu	664039	88,44	95,64	96,89

Tablo 24. Hiper parametre kestirim deneyi II: Katar tanıma
(Hyperparameter estimation experiment II: Ligature recognition)

Deney	Parametre	Ham	Normalize	Bitişik
Orijinal deney	663703	80,48	91,60	92,14
Öğrenme hızı	663783	80,56	91,53	91,39
LSTM boyutu	194919	78,73	89,86	89,62
Aktivasyon fonk.	663783	79,08	90,03	89,56
Filtre boyutu	664039	79,77	90,77	91,15

Tablo 25. Hiper parametre kestirim deneyi III: Kelime tanıma
(Hyperparameter estimation experiment III: Word recognition)

Deney	Parametre	Ham	Normalize
Orijinal deney	663703	44,08	66,45
Filtre boyu	664039	43,01	65,53
Aktivasyon fonk.	663783	42,04	63,40
LSTM boyutu	194919	40,64	63,05
Öğrenme hızı	663783	42,33	64,47

7. Sonuçlar (Conclusions)

Osmanlıca OCR konusunda birçok çalışma yapılmış olsa da bir karşılaştırma yapabilmek için an itibarıyla Osmanlıcaya özel sadece Miletos'un OCR aracı erişilebilirdir. Karşılaştırmada kullandığımız diğer araçlar ise piyasada bu konuda bulunabilecek en gelişmiş ve en güçlü araçlardır. Şimdiye kadar hem paylaşıma açık hem de orijinal bir veri setiyle bu kadar OCR araç/modelini, özellikle de Google Docs ve Tesseract araçlarını, Osmanlıca OCR için karşılaştıran bir çalışma söz konusu olmamıştır. Benzer çalışmalarda yayınlanan test sonuçlarına metindeki karakterler standart bir formata normalize edilemediği için dikkatli yaklaşılması gerektiği kanaatindeyiz. Osmanlıca metinlerin normalize edilmesi zorunluluğu ilk defa bu çalışmada dile getirilmiş ve metinleri normalize eden bir Python programı paylaşıma açılmıştır. Çalışmamızda benzerlerinden farklı

olarak sadece karakter tanıma oranı değil, katar ve kelime tanıma oranları da ham, normalize ve bitişik metin üzerinde hesaplanmıştır. İlk defa olmak üzere Osmanlıcanın karakter, katar ve kelime sıklıkları yani kısıtlı da olsa istatistik dil modelleri ortaya konulmuştur. Yine ilk defa olmak üzere Osmanlıca harfler ayırt edici bazı özelliklerine göre gruplanmış, bu harf gruplarının metin içindeki dağılımları ortaya konmuş ve bu harf grupları üstünden OCR oranları hesaplanmıştır. Makalemizde Osmanlıca matbu nesih doküman resimlerini derin öğrenme modelleriyle metne dönüştüren web tabanlı bir optik karakter tanıma sistemi sunulmuştur. Geliştirilen OCR aracı online kullanıma açılmıştır. Sistem sentetik ve orijinal verilerden oluşan bir veri kümesiyle eğitilip 21 sayfalık bir veri kümesiyle test edilmiştir. Test veri kümesi ve test sonuçları osmanlica.com/test adresinde paylaşılmıştır. Geliştirilen OCR aracı Tesseract'ın Arapça ve Farsça, Google Docs'ın Arapça, Finereader'ın Arapça ve Miletos'un Osmanlıca OCR araç/modeliyle deneysel olarak karşılaştırılmıştır. Karakter tanımada Osmanlica.com Hibrit OCR modeli %88,86 ham, %96,12 normalize ve %97,37 bitişik doğruluk oranıyla diğerlerinden belirgin şekilde daha yüksek bir performans sağlamıştır. Osmanlica.com Hibrit modeli %80.48 ham, %91.60 normalize ve %97.37 bitişik katar tanıma oranıyla diğer araçlardan daha başarılı sonuçlar üretmiştir. Kelime tanımada ise Osmanlica.com Hibrit OCR modeli %44.08 ham ve %66.45 normalize doğruluk değerleriyle diğer araçlardan çok daha yüksek bir doğruluk oranı vermektedir. Hiper parametre kestirim çalışmasında *filtre boyutu, öğrenme hızı, LSTM boyutu,* ve *aktivasyon fonksiyonu* olmak üzere 4 adet parametre değiştirilerek yeni deneyler yapılmıştır. Bu deneylerden sadece öğrenme hızının değiştirildiği deneyde doğruluk oranında artış gözlenmiştir. Bu çalışma, sadece matbu nesih hattı için tasarlanmış olması, hemze ve med işareti taşıyan harflerin tanınmaması, OCR sonrası karakter düzeltme adımına sahip olmaması gibi bazı kısıtlara sahiptir. İleride bu kısıtların giderilmesine yönelik çalışmalar yapılması planlanmaktadır. Osmanlıca-Türkçe uçtan-uca aktarım süreci (i) sayısallaştırma, (ii) Osmanlıca OCR, (iii) Osmanlıca-Türkçe alfabe çevirisi ve (iv) Osmanlıca-Türkçe dil çevirisi adımlarından oluşmaktadır. Bu sürecte Osmanlıca OCR'da elde edilen metin sonraki adımda girdi olarak kullanılacağından, OCR'daki hatalar sonraki adımlardaki hataların artmasına sebep olmaktadır. Bu yönüyle Osmanlıca OCR adımının başarı seviyesi diğer adımlara göre daha kritik bir rol oynamaktadır.

Kaynakça (References)

1. Ergin M, Osmanlıca Dersleri, Boğaziçi yayınları, İstanbul, 2020.
2. Akram Q. U. A., Hussain S., Niazi A., Anjum U., Irfan F., Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique, 11th IAPR International Workshop on Document Analysis Systems (DAS), Tours-France, 191-195, 2014.
3. Atici A. A., Yarman-Vural F. T., A heuristic algorithm for optical character recognition of Arabic Script, Signal Processing, 62 (1), 87-99, 1997.
4. Öztop E., Mülayim A. Y., Atalay V., Yarman-Vural F., Repulsive Attractive Network for Baseline Extraction on Document Images, Signal Processing, 75 (1), 1-10, 1999.
5. Ozturk A., Güneş S., Özbay Y., Multifont Ottoman Character Recognition, 7th IEEE Int. Conf. on Electronics Circuits and System (ICECS), Jounieh-Lebanon, 945-949,2000.
6. Şeykol E., Sinop A. K., Güdükbay U., Ulusoy Ö., Content Based Retrieval of Historical Ottoman Documents Stored as Textual Images, IEEE Transactions on Image Processing, 13 (3), 314-325, 2004.
7. Ataer E., Duygulu P., Matching ottoman words: an image retrieval approach to historical document indexing, Proceedings of the 6th ACM International conference on Image and Video Retrieval, Amsterdam-Netherlands, 341-347, 2007.
8. Yalniz I. Z., Sengor Altingovde I., Güdükbay U., Ulusoy Ö., Ottoman Archives Explorer: A Retrieval System for Digital Ottoman Archives, Journal on Computing and Cultural Heritage (JOCCH), 2 (3), 1-20, 2010.
9. Can E. F., Duygulu P., A line based representation for matching words in historical manuscripts, Pattern Recognition Letters, 32 (8), 1126-1138, 2011.
10. Duygulu P., Arifoglu D., Kalpaklı M., Cross-document word matching for segmentation and retrieval of Ottoman divans, Pattern Analysis and Applications, 19 (3), 647-663, 2016.
11. Kilic N., Gorgel P., Ucan O. N., Kala A., Multifont Ottoman character recognition using support vector machine, 3rd Int. Sym. on Communications, Control and Signal Processing, Saint Julian's-Malta, 328-333, 2008.
12. Onat A., Yildiz F., Gündüz M., Ottoman Script Recognition Using Hidden Markov Model, World Academy of Science, Engineering and Technology, 2, 630-632, 2008.
13. Kurt Z., Turkmen H. I., Karsligil E., Linear Discriminant Analysis in Ottoman Alphabet Character Recognition, Proceedings of the European Computing Conference, 2, 601-607, 2009.
14. Gorgel P., Kilic N., Ucan B., Kala A., Ucan O. N., A Backpropagation Neural Network Approach for Ottoman Character Recognition, Intelligent Automation & Soft Computing, 15 (3), 451-462,2009.
15. Yalniz I. Z., Altingovde I. S., Güdükbay U., Ulusoy Ö., Integrated segmentation and recognition of connected Ottoman script, Optical Engineering, 48 (11), 117205-117205, 2009.
16. Adigüzel H., Şahin P. D., Kalpaklı M., Line Segmentation of Ottoman Documents, Signal Processing and Communications Applications Conference, s Conference (SIU), Muğla-Turkey, 1-4, 2012.
17. Küçükşahin N., Design of an offline ottoman character recognition system of translating printed documents to modern turkish, Doktora Tezi, İzmir Institute of Technology, İzmir, 2019.
18. Kirmizialtin S., Wrisley D., Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive, arXiv preprint arXiv:2011.01139, 2020.
19. Doğru M., Ottoman-Turkish Optical Character Recognition and Latin Transcription, Yüksek Lisans Tezi, Yıldırım Beyazıt University, Fen Bilimleri Enstitüsi, Ankara, 2016.
20. Elmas B., Identifying species of trees through bark images by convolutional neural networks with transfer learning method, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (3), 1253-1270, 2021.
21. Yıldız O., Melanoma detection from dermoscopy images with deep learning methods: A comprehensive study Journal of the Faculty of Engineering and Architecture of Gazi University, 34 (4), 2241-2260, 2019.
22. Gurkahraman K., Karakiş R., Brain tumors classification with deep learning using data augmentation, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (2), 997-1012, 2021.
23. Al-Khatib W. G., Shahab S. A., Mahmoud S. A., Digital Library Framework for Arabic Manuscripts, IEEE/ACS International Conference on Computer Systems and Applications, Amman-Jordan, 458-465, 2007.
24. Jaf A. A., Koç Kayhan S., Machine-Based Transliterate of Ottoman to Latin-Based Script, Scientific Programming, 2021, 1-8, 2021.
25. Altwaijry N., Al-Turaiki I., Arabic handwriting recognition system using convolutional neural network, Neural Computing & Applications, 33 (7), 2249–2261, 2021.
26. Lamtougui H., Moubtahij H. E., Fouadi H., Yahyaouy A., Satori K., Offline Arabic Handwriting Recognition Using Deep Learning: Comparative Study, 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez-Morocco, 1-8, 2020.
27. Ghadhban H. Q., Othman M., Samsudin N., Kasim S., Mohamed A., Aljeroudi Y., Segments Interpolation Extractor for Finding the Best Fit Line in Arabic Offline Handwriting Recognition Words, IEEE Access, 9, 73482-73494, 2021.
28. Mohd M., Qamar F., Al-Sheikh I., Salah R., Quranic Optical Text Recognition Using Deep Learning Models, IEEE Access, 9, 38318-38330, 2021.
29. Bilgin E. F., Machine transliteration of Ottoman Turkish texts to modern Turkish, Yüksek Lisans Tezi, İstanbul Fatih Üniversitesi, Fen Bilimleri Enstitüsü, Istanbul, 2012.
30. Albelwi S., Mahmood A., A Framework for Designing the Architectures of Deep Convolutional Neural Networks, Entropy,19 (6), 242, 2017.