

**KOCAELİ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ**

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

ARAŞTIRMA PROBLEMLERİ

**SORU DOKÜMANLARININ ANLAMSAL BENZERLİKLERİNE
DAYALI DERİN ÖĞRENME TABANLI KÜMELEME ANALİZİ**

ERAY YELMEN

KOCAELİ 2020

**KOCAELİ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
ARAŞTIRMA PROBLEMLERİ**

**SORU DOKÜMANLARININ ANLAMSAL BENZERLİKLERİNE
DAYALI DERİN ÖĞRENME TABANLI KÜMELEME ANALİZİ**

ERAY YELMEN

Dr. Öğr. Üyesi Alpaslan Burak İNNER
Danışman, Kocaeli Üniversitesi

.....

Prof. Dr. Nevcihan DURU
Jüri Üyesi, Kocaeli Üniversitesi

.....

Dr.Öğr. Üyesi Ersin KAYA
Jüri Üyesi, Konya Teknik Üniversitesi

.....

Tezin Savunulduğu Tarih: 17.08.2020

ÖNSÖZ VE TEŞEKKÜR

Bu tez çalışması kapsamında, x-band radarlar tarafından tespit edilen deniz hedeflerinin otomatik olarak sınıflandırılması için özgün bir füzyon yöntemi geliştirilmiştir. Önerilen yöntem sayesinde başarımların artışı ve otomatik sınıflandırma çalışması için SOA yaklaşımının kullanımı irdelenmiştir.

Tez çalışmamda desteğini esirgemeyen, çalışmalarına yön veren, bana güvenen ve yüreklendiren danışmanım Prof. Dr. Nevcihan DURU'ya sonsuz teşekkürlerimi sunarım.

Akademik çalışmalarım sırasında, birçok aşamada beni destekleyen Bilgisayar Mühendisliği Bölümü araştırma görevlilerine ve Fen Bilimleri Enstitüsü çalışanlarına teşekkür ediyorum.

Doktora öğrenimim boyunca desteklerini esirgemeyen başta Osman KARABAYIR ve Mehmet Zahid KARTAL olmak üzere çok sevgili TÜBİTAK BİLGEM RAPSİM çalışanlarına sonsuz teşekkürler sunarım.

Hayatım boyunca bana güç veren en büyük destekçilerim, her aşamada sıkıntılarımı ve mutluluklarımı paylaşan sevgili babam Fahrettin BATI, annem Hacer BATI, eşim Özge BATI ve kardeşlerim Mehmet Emre BATI ile Fatma BATI'ya sonsuz teşekkürlerimi sunarım.

Son olarak, gelecekte kendilerine ilham kaynağı olmasını ümit ederek bu tezi çocuklarım Elif Nil BATI ve Ege BATI'ya ithaf ediyorum.

Temmuz – 2020

Eray YELMEN

Bu dokümandaki tüm bilgiler, etik ve akademik kurallar çerçevesinde elde edilip sunulmuştur.

Ayrıca yine bu kurallar çerçevesinde kendime ait olmayan ve kendimin üretmediği ve başka

kaynaklardan elde edilen bilgiler ve materyaller (text, resim, şekil, tablo vb.) gerekli şekilde

referans edilmiş ve dokümanda belirtilmiştir.

Öğrenci No: 150201387

Adı Soyadı: Eray YELMEN

İmza:.....

İÇİNDEKİLER

ÖNSÖZ VE TEŞEKKÜR.....	i
İÇİNDEKİLER.....	ii
ŞEKİLLER DİZİNİ.....	iv
TABLolar DİZİNİ.....	v
SİMGELER VE KISALTMALAR DİZİNİ.....	vii
ÖZET.....	viii
ABSTRACT.....	ix
GİRİŞ.....	1
1. GENEL BİLGİLER.....	4
1.1. Endometrial Kanser.....	4
1.2. Radyolojik Tanısal Yöntemler.....	4
1.2.1. Manyetik rezonans görüntüleme.....	4
1.2.2. Bilgisayarlı tomografi.....	5
1.3. Radiomics.....	5
1.3.1. ROI ve segmentasyon.....	6
1.3.2. 3D Slicer.....	6
1.3.3. Doku(Texture) analizi.....	6
1.3.4. Pyradiomics.....	9
1.4. Makine Öğrenmesi.....	10
1.4.1. Destek vektör makineleri.....	10
1.4.2. K-En yakın komşu algoritması.....	12
1.4.3. Karar ağaçları (Decision Tree).....	13
1.4.4. Rasgele orman (Random forest).....	14
1.4.5. Çok katmanlı algılayıcı.....	14
1.4.6. Gradient boosting machines.....	15
1.4.7. XGBoost.....	16
1.4.8. LightGBM.....	16
1.4.9. CatBoost.....	16
1.5. Öznitelik Seçimi.....	17
1.5.1. Chi square test (Ki-kare yöntemi).....	17
1.5.2. Mutual information (Karşılıklı bilgi).....	18
1.5.3. MRMR.....	18
1.5.4. ReliefF.....	19
1.5.5. Step forward selection.....	19
1.5.6. Step backward selection.....	19
1.6. Sınıflandırma Performans Metrikleri.....	20
1.6.1. Karmaşıklık matrisi.....	20
1.6.2. Doğruluk.....	20
1.6.3. Recall (Sensitivity).....	21
1.6.4. Specificity (Özgüllük).....	21
1.6.5. Eğri altında kalan alan (AUC).....	21
1.6.6. Precision (Kesinlik).....	22
1.6.7. F-Score.....	22

1.6.8. Matthews correlation coefficient.....	22
2. MATERYAL VE YÖNTEM.....	24
2.1. Veri Seti.....	24
2.2. Hasta Seçimi.....	24
2.3. BT Parametreleri.....	25
2.4. Radiomics Verilerinin Çıkarılması.....	25
2.5. Veri Ön İşleme.....	26
2.6. Öznitelik Seçimi.....	27
2.7. Makine Öğrenmesinin Uygulanması.....	28
3. BULGULAR VE TARTIŞMA.....	30
3.1. Endometrioid- Seröz Alt-Tip İkili Sınıflandırma Sonuçları.....	30
3.2. Myom- NonMyom İkili Sınıflandırma Sonuçları.....	38
3.3. Myom-Endometrioid-Seröz Çok Sınıf Sınıflandırma Sonuçları.....	46
4. SONUÇLAR VE ÖNERİLER.....	57
KAYNAKLAR.....	58
KİŞİSEL YAYIN VE ESERLER.....	64
ÖZGEÇMİŞ.....	65

ŞEKİLLER DİZİNİ

Şekil 1.1. GLCM analizinin şematik çizimi, a) Gri Seviye Görüntü, b)Nümerik Gri Seviye Görüntü, c) Co-occurence Matrisi.....	8
Şekil 1.2. Destek vektör makineleri.....	11
Şekil 1.3. KNN algoritması için örnek veri dağılımı.....	13
Şekil 1.4. Karar Ağacı yapısı.....	14
Şekil 1.5. Çok Katmanlı Algılayıcı Modeli.....	15
Şekil 1.6. Karmaşıklık matrisi.....	20
Şekil 2.1. Üç kesitli BT görüntüsü üzerinde segmentasyon işlemi, a) Axial Plan, b) Sagittal Plan, c)Koronal Plan.....	26
Şekil 3.1. a) Decision Tree kullanarak yapılan sınıflandırma sonucu elde edilen karmaşıklık matrisi b) CatBoost karmaşıklık matrisi.....	47
Şekil 3.2. Karmaşıklık matrisleri a) GBM b) CatBoost.....	49
Şekil 3.3. Karmaşıklık matrisleri, a) SVM, b)GBM, c)LightGBM, d) CatBoost.....	50
Şekil 3.4. MLP'ye ait karmaşıklık matrisi.....	52
Şekil 3.5. Karmaşıklık matrisleri, a) KNN, b) Random Forest.....	53
Şekil 3.6. Karmaşıklık matrisleri, a) SVM, b)MLP.....	55
Şekil 3.7. Karmaşıklık matrisleri, a)SVM, b)XGBoost.....	56

TABLolar DİZİNİ

Tablo 3.1.	Tüm öznitelikler kullanılarak yapılan Endometrioid-Seröz sınıflandırma sonuçları.....	30
Tablo 3.2.	Endometrioid- Seröz sınıflandırması için kullanılan yöntemler ve elde edilen öznitelikler.....	31
Tablo 3.3.	Chi-Square Test ile öznitelik seçimi sonrası Endometrioid-Seröz sınıflandırma sonuçları.....	32
Tablo 3.4.	Mutual Information ile seçilen özniteliklerle elde edilen Endometrioid-Seröz sınıflandırma sonuçları.....	33
Tablo 3.6.	MRMR ile seçilen özniteliklerle elde edilen Endometrioid-Seröz sınıflandırma sonuçları.....	34
Tablo 3.7.	Endometrioid-Seröz sınıflandırması için SFS ve SBS algoritmalarıyla seçilen öznitelikler.....	36
Tablo 3.8.	SFS ile elde edilen Endometrioid-Seröz sınıflandırma sonuçları.....	37
Tablo 3.9.	SBS ile elde edilen Endometrioid-Seröz sınıflandırma sonuçları.....	37
Tablo 3.10.	Tüm Öznitelikler kullanılarak yapılan Myom-NonMyom sınıflandırma sonuçları.....	38
Tablo 3.11.	Myom-NonMyom Sınıflandırması için kullanılan öznitelik seçim yöntemleri ve elde edilen öznitelikler.....	39
Tablo 3.12.	Chi Square Test ile öznitelik seçimi sonrası Myom-NonMyom sınıflandırma sonuçları.....	40
Tablo 3.13.	Mutual Information ile seçilen özniteliklerle elde edilen Myom-Nonmyom sınıflandırma sonuçları.....	41
Tablo 3.14.	ReliefF ile seçilen özniteliklerle K=4 ve K=10 için yapılan Myom-NonMyom sınıflandırma sonuçları.....	42
Tablo 3.15.	MRMR ile seçilen özniteliklerle elde edilen Myom-NonMyom sınıflandırma sonuçları.....	43
Tablo 3.16.	Myom-NonMyom sınıflandırması için SFS ve SBS algoritmalarıyla seçilen öznitelikler.....	44
Tablo 3.17.	SFS kullanılarak elde edilen Myom-NonMyom sınıflandırma sonuçları.....	45
Tablo 3.18.	SBS ile elde edilen Myom-NonMyom sınıflandırma sonuçları.....	45
Tablo 3.19.	Tüm öznitelikler kullanılarak yapılan çoklu sınıflandırma sonuçları.....	46
Tablo 3.20.	Çok sınıflı sınıflandırma için seçilen öznitelikler.....	48
Tablo 3.21.	Chi Square Test ile öznitelik seçimi sonrası yapılan çoklu sınıflandırma sonuçları.....	48
Tablo 3.22.	Mutual Information ile seçilen özniteliklerle elde edilen çoklu sınıflandırma sonuçları.....	49
Tablo 3.23.	ReliefF ile seçilen özniteliklerle K=4 ve K=10 için elde edilen çoklu sınıflandırma sonuçları.....	51

Tablo 3.24.	MRMR yöntemiyle elde edilen çoklu sınıflandırma sonuçları.....	52
Tablo 3.25.	Çok sınıflı sınıflandırma için SFS ve SBS ile seçilen öznitelikler.....	53
Tablo 3.26.	SFS algoritması kullanılarak elde edilen çoklu sınıflandırma sonuçları.....	54
Tablo 3.27.	SBS ile elde edilen çoklu sınıflandırma sonuçları.....	55



SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

AUC	: Area Under the Curve (Eğri Altında Kalan Alan)
BT	: Bilgisayarlı Tomografi
CPTAC	: Clinical Proteomic Tumor Analysis Consortium (Klinik Proteomik Tümör Analiz Konsorsiyumu)
DICOM	: Digital Imaging and Communications in Medicine (Tıpta Dijital Görüntüleme ve İletişim)
DMI	: Depth of Myometrial Invasion (Miyometriyal invazyon derinliği)
EFB	: Exclusive Feature Bundling (Özel Değişken Paketi)
GBM	: Gradient Boosting Machines (Gradyan Artırma Makineleri)
GLCM	: Grey Level Co-occurrence Matrix (Gri Seviye Eş Oluşum Matrisi)
GLDM	: Grey Level Dependence Matrix (Gri Seviye Bağımlılık Matrisi)
GLRLM	: Grey Level Run Length Matrix (Gri Seviye Dizi Uzunluğu Matrisi)
GLSZM	: Grey Level Size Zone Matrix (Gri Seviye Boyutu Bölge Matrisi)
KNN	: K-Nearest Neighbours (K-En Yakın Komşu)
MCC	: Matthews Correlation Coefficient (Matthews Korelasyon Katsayısı)
MLP	: Multi Layer Perceptrons (Çok Katmanlı Algılayıcılar)
MRI	: Magnetic Resonance Imaging (Manyetik Rezonans Görüntüleme)
MRMR	: Minimum Redundancy Maximum Relevance (Minimum Fazlalık Maksimum Alaka)
NGTDM	: Neighborhood Grey Tone Difference Matrix (Komşuluk Gri Ton Fark Matrisi)
PET	: Positron Emission Tomography (Pozitron Emisyon Tomografi)
ROC	: Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)
ROI	: Region of Interest (İlgili Bölge)
SBS	: Step Backward Selection (Geri Yönlü Arama Seçimi)
SFS	: Step Forward Selection (İleri Yönlü Arama Seçimi)
SVM	: Support Vector Machines (Destek Vektör Makineleri)
TCGA	: The Cancer Genome Atlas (Kanser Genom Atlası)
TCIA	: The Cancer Imaging Archive (Kanser Görüntüleme Arşivi)
UCEC	: Uterine Corpus Endometrial Carcinoma (Rahim Yapısı Endometrial Karsinom)

SORU DOKÜMANLARININ ANLAMSAL BENZERLİKLERİNE DAYALI DERİN ÖĞRENME TABANLI KÜMELEME ANALİZİ

ÖZET

İnternet ortamında metinsel dokümanların miktarının büyük boyutlara ulaşması ile birlikte aranan doğru dokümana kolay ve hızlı bir şekilde ulaşmak zorlaşmıştır. Metin dokümanlarının benzerliklerine göre kümelenmesi manuel yöntemlerle oldukça zahmetlidir. Bu durumu otomatik hale getirerek kolaylaştırmak için gelişmiş yöntemlere ihtiyaç vardır. Belge kümelemede metin verileri yakınlık ve benzerlik ölçüsüne göre gruplandırılır. Kümelemede yüksek başarı elde etmek, belgelerin doğru bir şekilde keşfedilmesi için oldukça önemlidir.

Anahtar Kelimeler: Bulanık C-means, Derin Öğrenme, Doc2Vec, Doğal Dil İşleme, Doküman Kümeleme.