T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

PAP-SMEAR GÖRÜNTÜLERİ ÜZERİNDE RAHİM AĞZI KANSERİ TESPİTİ

Fatma Betül AKYOL

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Danışman

Dr. Öğr. Üyesi Oğuz ALTUN

Ağustos, 2020

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

PAP-SMEAR GÖRÜNTÜLERİ ÜZERİNDE RAHİM AĞZI KANSERİ TESPİTİ

Fatma Betül AKYOL tarafından hazırlanan tez çalışması 04.08.2020 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Dr. Öğr. Üyesi Oğuz ALTUN
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Dr. Öğr. Üyesi Oğuz ALTUN, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Songül VARLI, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Mustafa DAĞTEKİN, Üye
İstanbul Üniversitesi- Cerrahpaşa _____

Danışmanım Dr. Öğr. Üyesi Oğuz ALTUN sorumluluğunda tarafımca hazırlanan Pap-Smear Görüntüleri Üzerinde Rahim Ağzı Kanseri Tespiti başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Fatma Betül AKYOL

İmza

Aileme

TEŞEKKÜR

Tez çalışmam boyunca desteğini esirgemeyen danışman hocam Dr. Öğr. Üyesi Oğuz Altun'a teşekkürlerimi sunarım.

Rahim ağzı kanseri ve hastalığın teşhisi hakkında bilgi veren ve çalışmam üzerinde fikirlerimi paylaştığım İstanbul Zeynep Kâmil Kadın ve Çocuk Hastalıkları Eğitim ve Araştırma Hastanesi Patoloji Bölümü'ne teşekkür ederim.

Yüksek lisans eğitimim sürecinde beni hep destekleyen sevgili aileme sonsuz teşekkürlerimi sunarım.

Fatma Betül AKYOL

İÇİNDEKİLER

SİMGE LİSTESİ vii

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

TABLO LİSTESİ x

ÖZET xi

ABSTRACT xiii

1 GİRİŞ 1

1.1 Literatür Özeti 1
1.2 Tezin Amacı 3
1.3 Hipotez 3

2 MATERYAL VE METOTLAR 4

2.1 Rahim Ağzı Kanseri 4
2.1.1 Pap-Smear Testi 4
2.2 Sayısal Görüntü İşleme 5
2.2.1 Görüntü İyileştirme 5
2.2.2 Görüntü Segmentasyonu 7
2.3 Sınıflandırma 8
2.3.1 Derin Yapay Sinir Ağları 9
2.3.2 Evrişimli Sinir Ağları 10
2.3.3 Transfer Öğrenme 12
2.3.4 Çok Katmanlı Algılayıcılar 14
2.3.5 Destek Vektör Makineleri 15
2.3.6 K- En Yakın Komşu 16
2.3.7 Naive Bayes Yöntemi 16
2.3.8 Rastgele Ormanlar Yöntemi 17
2.3.9 Lojistik Regresyon Yöntemi 17

3 VERİ SETİ VE ÖN İŞLEMLER 19

3.1 Herlev Veri Seti 19
3.2 Veri Önişleme 19
3.3 Çekirdek Segmentasyonu 22
3.4 Sitoplazma Segmentasyonu 25
3.5 Özellik Çıkarımı 26

4 DENEYSEL SONUÇLAR 28

4.1 Değerlendirme Ölçütleri 28
4.2 Deneysel Sonuçlar 29
4.2.1 İki-Sınıflı Sınıflandırma 33
4.2.2 Üç-Sınıflı Sınıflandırma 36

5 SONUÇ VE ÖNERİLER 39

KAYNAKÇA 41

TEZDEN ÜRETİLMİŞ YAYINLAR 45

SİMGE LİSTESİ

d	Çekirdek-sitoplazma arası uzaklık
a	Mavi renk alt sınırı
t	Mavi renk üst sınırı
σ	Sigmoid fonksiyonu
Rnd	Yuvarlaklık oranı

KISALTMA LİSTESİ

ADD	Ayrık Dalgacık Dönüşümü
AKD	Ayrık Kosinüs Dönüşümü
CIN	Cervical Intraepithelial Neoplasia
ÇKA	Çok Katmanlı Algılayıcı
DD	Düşük Düşük
DN	Doğru Negatif
DP	Doğru Pozitif
DSIL	Düşük Dereceli Skuamöz İntraepitelyal Lezyon
DVM	Destek Vektör Makineleri
DY	Düşük Yüksek
ESA	Evrişimli Sinir Ağları
IPV	İnsan Papilloma Virüsü
KNN	K-En Yakın Komşu
LD	Doğrusal Ayırımcı
LR	Lojistik Regresyon
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
RF	Random Forest
RO	Rastgele Orman
SGD	Stochastic Gradient Descent
SIN	Servikal Intraepitelyal Neoplazi
SVM	Support Vector Machine
YD	Yüksek Düşük
YN	Yanlış Negatif
YP	Yanlış Pozitif
YSA	Yapay Sinir Ağları
YSIL	Yüksek Dereceli Skuamöz İntraepitelyal Lezyon
YY	Yüksek Yüksek

ŞEKİL LİSTESİ

Şekil 2.1 ANN nöron modeli..... 9
Şekil 2.2 3x3 Konvolüsyon işlemi.....11
Şekil 2.3 2x2 maksimum havuzlama işlemi.....11
Şekil 2.4 Transfer öğrenme mimarisi 12
Şekil 2.5 Inception mimarisi.....13
Şekil 2.6 InceptionV2 mimarisi.....13
Şekil 2.7 ÇKA yapısı.....14
Şekil 2.8 Veriyi doğrusal olarak ayıran düzlem.....15
Şekil 2.9 Sigmoid fonksiyonu 18
Şekil 3.1 Çalışmada izlenen adımlara ait akış diyagramı 20
Şekil 3.2 Serviks hücresine ait giriş görüntüsü.....21
Şekil 3.3 AKD sonucu elde edilen görüntü 22
Şekil 3.4 ADD sonucu elde edilen görüntü 22
Şekil 3.5 Hücre çekirdeklerinin bulunması 23
Şekil 3.6 Hücre çekirdeklerinin giriş görüntüsü üzerinde işaretlenmesi.....24
Şekil 3.7 Farklı K değerleri için (a) Giriş görüntüsü (b)K=3, (c)K=5 ve (d)K=7 için K-ortalama yönteminin sonuçları.....25
Şekil 3.8 Segmentasyon işlemi sonucu sitoplazma.....26
Şekil 4.1 Çekirdek-Sitoplazma eşleştirme.....30
Şekil 4.2 Yoğunluk değeri büyük olan bölgeler.....31
Şekil 4.3 İki çekirdeğe sahip hücrenin görüntü üzerinde görünümü.....32
Şekil 4.4 (a) Giriş görüntüsü (b) Çekirdek segmentasyonu (c) Sitoplazma segmentasyon.....32

TABLO LİSTESİ

Tablo 4.1 Karışıklık matrisi....28
Tablo 4.2 Normal ve anormal hücrelere ait bazı veriler....33
Tablo 4.3 DVM 2-Sınıf için karışıklık matrisi 34
Tablo 4.4 RF 2-Sınıf için karışıklık matrisi 34
Tablo 4.5 KNN yöntemi K sayısına göre sınıflandırma sonuçları 34
Tablo 4.6 2-Sınıflı sınıflandırma işlemi sonuçları 35
Tablo 4.7 Derin öğrenme 2-Sınıflı sınıflandırma işlemi sonuçları 36
Tablo 4.8 3-Sınıflı veri seti örneği 37
Tablo 4.9 DVM 3-Sınıf için karışıklık matrisi 37
Tablo 4.10 2-Sınıflı sınıflandırma işlemi sonuçları 38

ÖZET

Pap-Smear Görüntüleri Üzerinde Rahim Ağzı Kanseri Tespiti

Fatma Betül AKYOL

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Dr. Öğr. Üyesi Oğuz ALTUN

Dünya çapında kadınlarda görülen kanser hastalıkları sıralamasında dördüncü sırada yer alan rahim ağzı kanserini erken aşamada tespit edebilmek hastalığın tedavisinde önemli bir yapı taşıdır. Hastalığın teşhisi için pap-smear testi kullanılmaktadır. Bu çalışma kapsamında rahim ağzı kanserini erken aşamada tespit edebilecek bilgisayar temelli bir karar sisteminin tasarlanması amaçlanmıştır. Pap-smear testi sonucunda elde edilmiş olan serviks görüntülerinde normal ve anormal özellikli hücreler bulunarak anormal olan hücreler görüntü üzerinde işaretlenmiştir. Çalışma genel olarak segmentasyon ve sınıflandırma olmak üzere iki bölümden oluşmaktadır. Segmentasyon işleminde çekirdek ve sitoplazma ayrı olarak ele alınmış ve birbirinden bağımsız olarak bölütlenmiştir. Çekirdek segmentasyonu aşamasında eşikleme ve K-means olmak üzere iki farklı yöntem kullanılmıştır. Sitoplazma segmentasyonu işleminde eşikleme ve iki görüntünün farkı yöntemleri kullanılarak sitoplazma bulunmuştur. Sınıflandırma aşamasında ise görüntülerden çıkarılmış olan bazı özellikler sınıflandırma kriteri olarak belirlenmiştir. Görüntülerden çıkarılan özellikler patoloji uzmanları ile incelenmiş

ve bir veri seti oluşturulmuştur. Herlev veri setinde bulunan 917 görüntünün her biri için bu özellikler çıkarılmış ve bir veri setine kaydedilmiştir. Çalışmanın sınıflandırma aşamasında ise oluşturulan veri setine makine öğrenmesi yöntemlerinden derin öğrenme, Destek Vektör Makinesi (DVM), Naive Bayes (NB), Rastgele Orman (RO), Çok-Katmanlı Algılayıcı (ÇKA), Lojistik Regresyon (LR), K-En Yakın Komşu (KNN) ve Derin öğrenme yöntemleri uygulanarak sınıflandırma işlemi gerçekleştirilmiştir ve %81 ve %93 arasında doğruluk değerleri elde edilmiştir. Gelecek çalışmalarda bilgisayar temelli karar sistemini tamamlamak ve yapay zekâ modelini eğitebilmek için bu veri seti kullanılacaktır.

Anahtar Kelimeler: Rahim ağzı kanseri, pap-smear, segmentasyon, sınıflandırma, derin öğrenme.

ABSTRACT

Cervical Cancer Detection over Pap-Smear Images

Fatma Betül AKYOL

Department of Computer Engineering

Master of ScienceThesis

Advisor: Dr. Öğr. Üyesi Oğuz ALTUN

Early detection of cervical cancer is an important building block in the treatment of the disease. Cervical cancer which ranks fourth in the ranking of cancer diseases seen in women worldwide. Pap-smear tests are used to diagnose the disease. In this study, it is aimed to design a computer-based decision system that can detect cervical cancer at an early stage. Normal and abnormal cells are found in the cervix images obtained as a result of the pap-smear test and the abnormal cells are marked on the image. The study generally consists of two parts: segmentation and classification. In the segmentation process, the nucleus and cytoplasm are handled separately and partitioned independently. In the nucleus segmentation phase, two different methods, thresholding and K-means were used. In the cytoplasm segmentation process, cytoplasm was found by using the threshold and the difference of the two images. In the classification phase, some features extracted from the images were determined as the classification criteria. The features extracted from the images were examined with pathologists and a dataset was created. For each of the 917 images in the Herlev dataset, these features were

extracted and stored in a dataset. In the classification stage of the study, the classification process was performed by applying deep learning, Support Vector Machines, Naive Bayes, Random Forest, Multi-Layer Perceptron, Logistic Regression, K-Nearest Neighbor, and Deep Learning methods to the dataset created and accuracy values between 81% and 93% were obtained. In future studies, this dataset will be used to complete the computer-based decision system and to train the artificial intelligence model.

Keywords: Cervical cancer, pap-smear, segmentation, classification, deep learning.

1
GİRİŞ

Kanser, günümüzde yaygın olarak görülmektedir. Dünya genelinde yaşayan birçok kadın kanser türlerinden biri olan rahim ağzı kanseri riski taşımaktadır. Kanser hastalığı için henüz etkin bir tedavinin olmaması bilim insanlarının kanser üzerinde yoğun çalışmalar yapılmasına yöneltmiştir. Hastalığı erken teşhis etmek için bilgisayar temelli sistemler de geliştirilmektedir. Bu bölümde literatürde hastalık teşhisi için tasarlanan bilgisayar temelli çalışmalar belirtilmiştir.

1.1 Literatür Özeti

Kanser hastalığında ve kanser kaynaklı ölümlerde meydana gelen artış bu hastalık üzerinde yoğun çalışmalar yapılmasını beraberinde getirmiştir. Yapay zekâ ile birlikte bilgisayar temelli karar sistemleri tasarlanmaya başlanmıştır ve bu karar sistemleri günümüzde tıp dâhil birçok alanda kullanılmaktadır. Makine öğrenmesi yöntemlerini kullanarak tıbbi görüntü analizi alanında yapılan çalışmalar yaygın şekilde devam etmektedir. Literatürde bu alanda yapılan birçok çalışma bulunmaktadır.

Rahim ağzı kanseri teşhisinde serviks hücrelerinin çekirdek ve sitoplazmaları ayrı olarak ele alınmıştır. Bu nedenle hücreye çekirdek ve sitoplazma segmentasyonu işlemleri uygulanmıştır. Segmantasyon işlemi için literatürde farklı yöntemler kullanılarak yapılmış olan çalışmalar mevcuttur.

Plissiti ve diğerleri [1], pap-smear görüntülerinde hücre çekirdeği ve sitoplazmasını otomatik olarak sınıflandırmak için bulanık c-ortalama (fuzzy c-means) ve DVM tekniklerini kullandılar. Çalışmalarında hücre çekirdeklerini %99,25 doğruluk oranı ile sınıflandırdılar.

Wang ve diğerleri [2], hücre çekirdeğini bulma ve segmantasyonu için matematiksel morfolojiye sahip ortalama kaydırma kümeleme algoritmasını ve sınıflandırıcı olarak DVM yöntemini kullanmışlardır. Çalışmaları kapsamında serviks hücre çekirdeğini %94,25 duyarlılık oranı ile bölütlemişlerdir.

Diğer bir çalışmada, Marinakis ve diğerleri [3], çalışmalarında özellik seçimi için genetik algoritmaları, sınıflandırma için KNN yöntemini kullanmışlardır. Çalışmalarının sonucunda iki sınıflı sınıflandırma işleminde %100 başarı oranı elde etmişlerdir.

Genctav ve diğerleri [4], çalışmalarında Herlev [5] ve Hacettepe veri setlerini kullanarak pap-smear görüntüleri üzerinde serviks hücresi segmentasyonu ve sınıflandırması üzerinde çalışmışlardır. Segmentasyon için watershed algoritmasını ve hücreleri sınıflandırmada ise denetimsiz öğrenme algoritmalarını kullanmışlardır. Phoulady vd. [6], genişletilmiş alan derinliği algoritmasını kullanarak serviks hücrelerinden çekirdek segmentasyonu işlemi yapmışlardır. %96 hassasiyet ile diğer metotlardan daha iyi bir sonuç elde etmişlerdir.

Kale ve Aksoy [7], iki aşamadan oluşan bir rahim hücresi segmentasyonu üzerinde çalışmışlardır. İlk aşamada hiyerarşik segmentasyon algoritması kullanarak görüntüyü bölümlendirmişlerdir. Çalışmalarının ikinci aşamasında ise şekil ve spektral özellikleri kullanarak çekirdek ve sitoplazma segmentasyonunu gerçekleştirmişlerdir. Çalışmalarında sınıflandırıcı olarak DVM kullanmışlardır ve %96 doğru sınıflandırma oranı elde etmişlerdir.

Win ve diğerleri [8] pap-smear görüntüleri üzerinde rahim ağzı kanseri tespiti için sınıflandırıcı olarak LD, DVM, KNN, yükseltilmiş ağaçlar (boosted trees) ve torbalı ağaçlar (bagged trees) yöntemlerinin sonuçlarının kombinasyonu olan bir yöntem kullandılar ve bu yöntem ile iki sınıflı sınıflandırma işleminde %98,27 doğruluk değeri elde ettiler.

William ve diğerleri [9] çalışmalarında pap-smear görüntülerinden serviks kanseri tespiti için bir araç geliştirme üzerine çalışma yapmışlardır. Çalışmalarında sınıflandırıcı olarak bulanık c-ortalama (fuzzy c-means) yöntemini kullanmışlardır ve %98,8 doğruluk değeri elde etmişlerdir.

Taha ve diğerleri [10], pap-smear görüntülerinin analizinde özellik çıkarımı için önceden eğitilmiş AlexNet mimarisini ESA modellerine uygulamışlardır. Görüntüleri sınıflandırmak için ise DVM yöntemini kullandılar. Çalışmalarında, Herlev veri seti üzerinde %99,19 oranında doğruluk değeri elde etmişlerdir.

1.2 Tezin Amacı

Kanser, hücrelerde meydana gelen anormal değişimler sonucu ortaya çıkan bir hastalık türüdür. Birden çok çeşidi olan kanser etkilediği hücre çeşidine göre farklı isimler almaktadır. Rahim ağzı kanseri, serviks hücrelerinde meydana gelen değişimler sonucunda ortaya çıkmaktadır.

Bu tez çalışması kapsamında pap-smear testi sonucunda elde edilmiş serviks görüntülerinde normal ve anormal görünümde olan hücrelerin bulunması ve sınıflandırılması amaçlanmıştır. Anormal olarak bulunan hücrelerin Sınıflandırma aşamasından önce segmentasyon işlemi yapılarak çekirdek ve sitoplazma ayrı olarak ele alınmıştır. Sınıflandırma işleminde DVM [11], RO [12], ÇKA [13], KNN [14], NB [15], LR [16] ve Derin öğrenme [17] yöntemleri kullanılmıştır.

1.3 Hipotez

Dünya genelinde binlerce kadın rahim ağzı kanserine yakalanmaktadır. Artan hastalık ve ölüm sayıları ve kanser hastalığı için etkin ilacın olmaması erken teşhis ve tedavinin hastalığın seyrindeki önemini artırmıştır. Hastalığın tespiti sırasında rahimden alınan sürüntünün patoloji uzmanları tarafından incelenmesi ve uzman doktorun raporu sonucunda teşhis konulmaktadır. Sürüntü alınarak incelenme işlemi pap-smear testi olarak adlandırılmaktadır.

Kanserin dünya genelinde hızlı yayılımı ve etkin tedavisinin olmaması hastalık üzerine yapılan bilimsel çalışmaları artırmıştır. Son yıllarda gelişen teknoloji ile tıp ve teknoloji birleştirilerek bilgisayar temelli teşhis ve tedavi sistemleri geliştirilmektedir. Bilgisayar destekli karar sistemleri ile geliştirilen teşhis işlemi kısa süre içerisinde birçok hastanın pap-smear görüntülerini inceleyerek teşhis koyarak doktorlara yardımcı olabilir. Hastalık, pap-smear görüntüleri üzerinde makine öğrenmesi yöntemleri kullanılarak tespit edilebilir ve erken teşhis için önemli olan zamanı kısaltabilir.

2
MATERYAL VE METOTLAR

Bu çalışmada rahim ağzı kanseri hastalığı incelenmiştir. Hastalığın teşhisi için kullanılan pap-smear testi sonucu elde edilen görüntüler kullanılarak sınıflandırma işlemi yapılmıştır. Sınıflandırma işleminde kullanılan makine öğrenmesi yöntemleri bu başlık altında incelenmiştir.

2.1 Rahim Ağzı Kanseri

Rahim ağzı kanseri kadınlarda en sık görülen dördüncü kanser türüdür. 2018 yılında dünya çapında yaklaşık olarak 570.000 kadında rahim ağzı kanseri görülmüş ve 311.365 kadın bu hastalık nedeniyle hayatını kaybetmiştir [18]. Servikal intraepitelyal neoplazi (SIN) anormallikleri ile ilişkilendirilen İnsan Papilloma Virüsü (IPV) rahim ağzı kanseri vakalarının %50'sinde hastalık kaynağı olarak gözlenmiştir [19]. Bethesda Sistemi [20], SIN'i, düşük dereceli skuamöz intraepitelyal lezyon (DSIL), yüksek dereceli skuamöz intraepitelyal lezyon (YSIL) ve skuamöz hücre karsinoması olmak üzere üç gruba ayırır. Anormal görünümde olan fakat kanser hücresi olmayan kanser öncüsü olarak görülen hücrelere displazi denir. Bethesda Sistemi'ne göre yukarıda belirtilen hücre anormalliklerinden, DSIL hafif displazi, YSIL ise orta displazi olarak sınıflandırılmaktadır.

Hastalığın teşhisinde farklı yöntemler mevcuttur. Mevcut yöntemler arasında en yaygın olarak kullanılan tekniklerden biri görüntülemedir. Günümüzde rahim ağzı kanseri görüntüleme yöntemi olarak "Pap-Smear Testi" yaygın olarak kullanılmaktadır [21].

2.1.1 Pap-Smear Testi

Rahim ağzı kanseri teşhisinde hastalıklı hücrelerin tespit edilebilmesi önemli bir yer tutar. Hastalıklı serviks hücresini tespit edebilmek için serviks görüntülerine ihtiyaç duyulmaktadır ve serviks hücrelerini görüntüleyebilmek için Pap-smear testi yaygın olarak yapılmaktadır. Pap-smear testi ilk olarak Georgios Papanikolaou tarafından 1940'lı yılların başında öne sürülmüştür [22]. Test bir çubuk tarafından serviksten sürüntü alınarak yapılmaktadır. Alınan sürüntü örneği patoloji

uzmanları tarafından mikroskop ile incelenmektedir ve hastanın sağlıklı veya kanser hücresine sahip olduğu raporlanmaktadır.

2.2 Sayısal Görüntü İşleme

Görüntü, bir ana kaynak tarafından aydınlatılan bir nesnenin insan görme sistemi tarafından algılanan temsilidir. İnsan görme sistemi tarafından bir bütün olarak algılanan bir görüntü aslında her biri koyuluk değerlerini ifade eden piksellerden oluşmaktadır. Algılanan bir görüntü tek renkli (monochromatic) veya renkli olabilir. Tek renkli bir görüntü tek renk bandına sahip iken renkli bir görüntüde 3 renk kanalı vardır. Bir görüntünün bilgisayarlar ve dijital makineler tarafından algılanabilmesi için görüntü sayısal formata dönüştürülmektedir. Bu dönüşüm iki yolla yapılmaktadır:

1. Örnekleme (Sampling)
2. Niceleme (Quantization)

2.2.1 Görüntü İyileştirme

Görüntü kaynakları yardımı ile elde edilen görüntüler üzerinde gürültü denilen parazitler içerebilir. Bu görüntüler üzerinde görüntü işleme yöntemlerini uygulayabilmek için öncelikle bu gürültünün temizlenmesi gerekmektedir. Görüntü iyileştirme görüntü elde etme adımından sonra uygulanan ilk işlemdir ve görüntü işlemede ön işlem olarak uygulanmaktadır [23]. Gaussian, tuz ve karabiber gürültüsü olmak üzere gürültü çeşitleri mevcuttur. Görüntü üzerinde yapılan çalışmaların daha başarılı sonuçlar vermesi için gürültü temizleme (denoising) işlemi yapılarak bu siyah ve beyaz parazitlerin temizlenmesi gerekmektedir. Gürültü temizleme işleminde kullanılan en yaygın yöntemlerden biri görüntüye filtre uygulama yöntemidir. Filtreleme işleminde görüntü üzerinde boyutu belirlenmiş olan bir filtre gezdirilir. 3x3'lük bir ortanca (median) filtreleme işleminde gürültü içeren görüntü üzerinde 3x3 boyutunda bir filtre gezdirilir ve görüntüde bu alanda bulunan piksel değerlerinin ortancası bulunur ve yeni değer olarak bu değer yazılır.

Görüntü iyileştirme işleminin diğer bir kullanım amacı ise görüntü üzerinde zıtlığın artırılmasıdır. Görüntü üzerinde bulunan piksellerin koyuluk değerlerinin dağılımını gösterildiği histogramlarda meydana gelen dengesiz dağılımlar zıtlık seviyesi bakımından net olmayan görüntüler oluşturur. Dengeli bir histogram dağılımı elde etmek için histogram eşitleme yöntemi kullanılmaktadır [24]. Bu yöntemde amaç belirli bir alanda yoğunlaşmış olan piksel değerlerini histogram geneline yayarak daha iyi bir görüntü elde etmektir.

2.2.1.1 Ayrık Kosinüs Dönüşümü

Ayrık kosinüs dönüşümü (AKD) sinyal işleme ve görüntü sıkıştırma alanında kullanılan ve bu alanda yüksek performans gösteren bir yöntemdir [25]. AKD yönteminde amaç görüntüyü uzaysal alandan frekans alanına çevirerek iyileştirmektir. Bu yöntemde sinyal kosinüs fonksiyonları şeklinde gösterilir [26].

$$X(m) = \sum_{n=0}^{N} \cos(\frac{\pi \mathrm{mn}}{N}) x(n) \quad m = 0,1, \ldots\ldots, N \tag{2.1}$$

AKD katsayılarını hesaplayabilmek için kullanılan denklem (2.1)'de belirtilmiştir. x sütun vektörüne $\cos(\frac{\pi \mathrm{mn}}{N})$ ayrık fourier kosinüs çekirdeği uygulandığında X vektörü elde edilmektedir.

2.2.1.2 Ayrık Dalgacık Dönüşümü

Fourier dönüşümü sonucunda ortaya çıkan sinyalin zaman bilgisi elde edilememektedir. Ayrık Dalgacık Dönüşümü (ADD) sonucunda elde edilen sinyalin hem zaman hem de frekans bilgisi elde edilebilir. Ayrık bir işarete yüksek ve alçak geçiren filtrelerin uygulanması ile ADD elde edilmektedir [27].

$$y_{yüksek}(k) = \sum_{n} x(n).g(2k - n) \tag{2.2}$$

$$y_{alçak}(k) = \sum_{n} x(n).h(2k - n)$$

$X(n)$ giriş sinyaline ilk olarak yüksek $g(n)$ yüksek geçiren filtre ve ardından $h(n)$ alçak geçiren filtre uygulanmaktadır [27]. Uygulanan yüksek alçak filtre işlemleri (2.2)' de belirtilmiştir.

ADD, giriş olarak verilen görüntüyü ilk olarak D ve Y olmak üzere 2 banda böler. 2 boyutlu ADD ise görüntüyü DD (düşük- düşük), DY (düşük -yüksek), YD (yüksek-düşük), YY (yüksek-yüksek) olmak üzere dört banda böler [28]. DY, YD ve YY bantları görüntünün yüksek frekansa sahip alanlarını içerirken DD ise düşük frekanslı alanını içerir [29].

2.2.2 Görüntü Segmentasyonu

Örüntü tanıma ve görüntü işlemede uygulanan ilk ve önemli adımlardan birisi olan segmentasyon, görüntüyü homojen bölümlere ayırma işlemidir [30]. Segmentasyon işlemi görüntülerin kenarları gibi yoğunluk değerlerinde ani değişimlerin olduğu bölgelerin bulunması veya görüntü üzerindeki objelerin benzerliklerine göre gruplandırılması ile gerçekleştirilebilmektedir [23]. Görüntü segmentasyonu üzerinde yapılmış çalışmalardan Fu ve Mui [31], segmentasyon işlemini eşikleme veya kümeleme, kenar tespiti ve bölge çıkarımı olmak üzere üç kategoriye ayırmışlardır.

2.2.2.1 Eşikleme (Thresholding)

Eşikleme yöntemi en basit ve yaygın olarak kullanılan segmentasyon yöntemlerinden biridir [32].

Global eşikleme yönteminde sabit bir eşik değeri seçilir ve giriş görüntüsüne bu değere bağlı olarak eşikleme uygulanır [33].

$$f(x,y) = \begin{cases} 0, & f(x,y) < t \\ 1, & f(x,y) \geq t \end{cases} \tag{2.3}$$

(2.3)'te gösterildiği gibi f giriş görüntüsünün her piksel değeri sabit olarak belirlenen t eşik değerinden büyük ise 1, küçükse 0 olarak güncellenir. Böylece giriş görüntüsünde eşik değerinden daha küçük değere sahip olan bölgeler siyah, daha büyük veya eşit değere sahip olan yerler beyaz olarak gösterilir.

2.2.2.2 K-Ortalama (K-Means) Yöntemi

K-ortalama kümeleme, bir veri setini k kümeye bölmek için kullanılan bir kümeleme yöntemidir [34]. Çekirdek segmentasyonu işleminde kullanılan diğer bir yöntem ise K- ortalama (K-means) olmuştur. Algoritma temel olarak aşağıda verilen adımları takip eder:

1-) Veri setinin kaç kümeye ayrılacağını belirleyen K değeri seçilir,

2-) Veri setinde bulunan her örnek ile belirlenen K küme merkezi arasındaki mesafe hesaplanır ve her örnek kendisine en yakın olan kümeye ait olduğu varsayılır.

3-) Her küme için, tüm örneklerin ortalaması ile yeni kümeleme merkezleri hesaplanır [35].

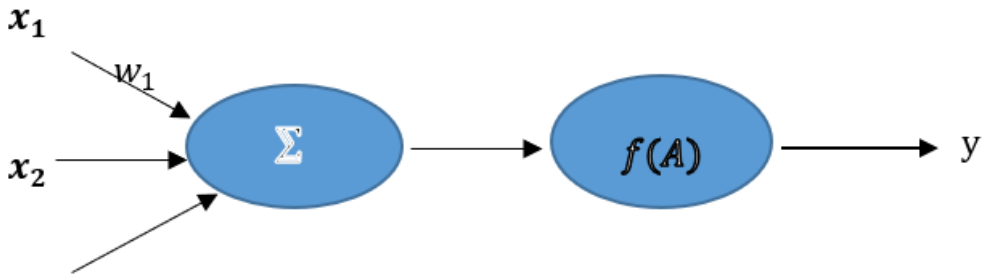
2.3 Sınıflandırma

Sınıflandırma, bir nesneyi ait olduğu sınıf ile etiketleme işlemidir. Veri madenciliği başlığı altında kullanımı oldukça yaygındır. Sınıflandırma işleminde amaç bir veri setinde bulunan her bir örneği bir sınıf ile etiketlemektir. Makine öğrenmesinin gelişmesi ile sınıflandırma işlemi birçok alanda yaygın şekilde kullanılmaktadır.

2.3.1 Derin Yapay Sinir Ağları

Yapay sinir ağları (YSA), nöron, akson ve dendritlerden oluşan insan sinir sistemini yapay olarak modelleyen bir modeldir. İlk YSA mimarisi McCulloch ve Pitts tarafından 1943 yılında geliştirilmiştir. Nöronlar hücreleri, sinapslar ağırlık değerlerini, aksonlar çıktıyı belirtmektedir. YSA modelini kullanarak yapılan eğitimlerde amaç sinapsları değiştirerek hatayı en aza düşürmektir. Bir hücreden diğer bir hücreye sinyaller aksonlar yardımı ile iletilmektedir.

Öğrenme, denetimli (supervised) ve denetimsiz (unsupervised)olmak üzere iki gruba ayrılmaktadır. Denetimli öğrenmede, veri setinde bulunan her örnek bir sınıf ile etiketlenmiştir. Denetimsiz öğrenmede ise veriler bir sınıf ile etiketlenmemektedir. Denetimsiz öğrenmede bir öğretmen yoktur ve öğrenme işlemi deneyimler yardımı ile gerçekleştirilmektedir. Kümeleme işlemi denetimsiz öğrenmeye örnek olarak gösterilebilir [36].



Şekil 2.1 ANN nöron modeli [37]

Şekil 2.5'te gösterilen bir ANN modelinde her bir $(x_1, x_2, .., x_n)$ girdisi için $(w_1, w_2, .., w_n)$ o girdiye ait ağırlık değerlerini belirtmektedir. Nörona gelen girdilerin ve ağırlıkların çarpımı bir aktivasyon fonksiyonundan geçirilerek bir çıktı üretilmektedir. Üretilen çıktı değerini hesaplamak için kullanılan eşitlik (2.4)'te verilmiştir.

$$y = A\left(\sum x_1.w_1 + x_2.w_2 + .. + x_n.w_n\right) \tag{2.4}$$

Her bir katman için hesaplanan ağırlık çarpımlarının toplamına doğrusal olmayan bir aktivasyon fonksiyonu uygulanır. Aktivasyon fonksiyonu olarak, ReLu, sigmoid veya hiperbolik tanjant (tanh) kullanılmaktadır. ReLu aktivasyon fonksiyonu (2.5) ile belirtilmiştir [17]. ReLu aktivasyon birimine gelen giriş değeri 0'dan küçük ise $f(x)$ fonksiyonu çıktı olarak 0, pozitif giriş değeri için ise çıktı olarak bu pozitif x değerini üretmektedir.

$$f(x) = \max(x, 0) \tag{2.5}$$

Sigmoid aktivasyon fonksiyonu (2.6) ile belirtilmiştir [17].

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.6}$$

Hiperbolik tanjant fonksiyonu (2.7)' de gösterildiği üzere sigmoid fonksiyonu yardımı ile hesaplanabilmektedir. (2.6)' da belirtilen sigmoid aktivasyon fonksiyonunda x yerine $2x$ yazıldığında (2.8) ile verilen tanjant aktivasyon fonksiyonu denklemi bulunmaktadır.

$$\tanh(x) = 2\,\sigma(2x) - 1 \tag{2.7}$$

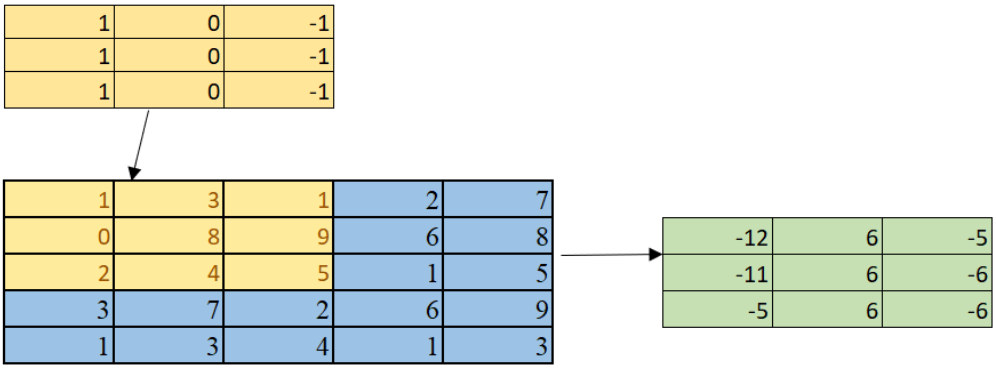
$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1 \tag{2.8}$$

2.3.2 Evrişimli Sinir Ağları

Evrişimli sinir ağları (ESA) [38], örüntü tanıma, segmentasyon ve sınıflandırma işlemlerinde kullanılan çok katmanlı bir yapay sinir ağı modelidir. Çoklu katman yapısına sahip olan ESA Evrişim, havuzlama ve tam bağlı katmanlarını içermektedir.

2.3.2.1 Evrişimli Katman (Convolution Layer)

Bu katmanda filtreleme yöntemi kullanılarak bir özellik haritası (feature map) çıkartılır. Görüntü üzerinde belirlenen boyutta bir filtre gezdirilir ve görüntüye konvolüsyon işlemi uygulanır.

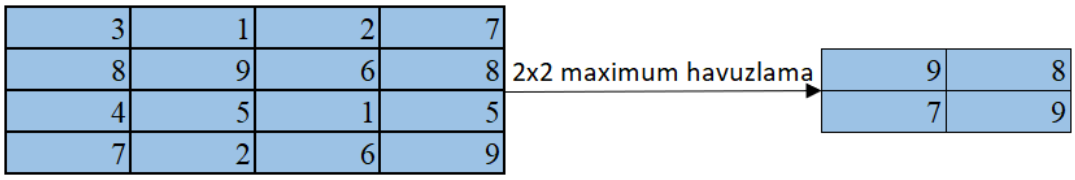


Şekil 2.2 3x3 Konvolüsyon işlemi

Görüntü üzerinde gezdirilen filtre sonucu bir özellik matrisi çıkartılır. Bu matrisin her bir elemanının değeri hesaplanırken görüntünün piksel değeri ile görüntü üzerinde o piksel değeri üzerine denk gelen filtrelerin çarpımlarının toplamına eşittir.

2.3.2.2 Havuzlama Katman (Pooling Layer)

Havuz katmanında boyut azaltma işlemi yapılmaktadır [36]. Maksimum havuzlama (max_pooling) yaygın olarak kullanılan havuzlama yöntemlerinden biridir. Maksimum havuzlama yönteminde görüntü üzerinde bir filtre gezdirilir ve gezilen bölgede bulunan maksimum değerler alınarak görüntünün boyutu azaltılır.



Şekil 2.3 2x2 maksimum havuzlama işlemi

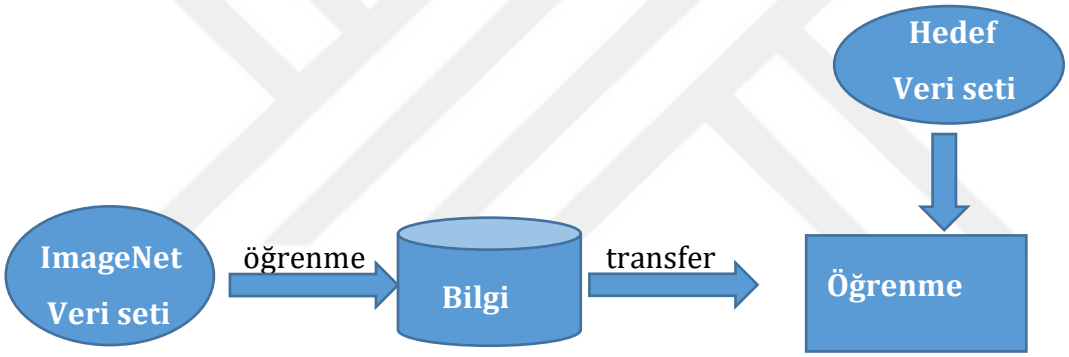
Şekil 2.3 'de giriş görüntüsüne uygulanmış olan maksimum havuzlama yöntemi gösterilmiştir. 4x4 boyutunda olan görüntü üzerinde 2x2 boyutunda bir filtre gezdirilmiştir.

2.3.2.2 Tamamen Bağlı Katman (Fully-Connected Layer)

Tamamen bağlı katman, ESA mimarisinde bulunan son katmandır. Bu katmanda, sınıflandırma ve tanıma işleminde kullanılmaktadır. Kendisinden önce bulunan konvolüsyonel katmanının çıktılarını 1 boyutlu özellik vektörüne dönüştürür. Her sınıf için bir çıkış birimi bulunur ve aktivasyon fonksiyonu kullanılarak her bir nöronun çıkış değeri bir sınıf ile ilişkilendirilmektedir [39].

2.3.3 Transfer Öğrenme

Makine öğrenmesi öğrenme işlemini başlangıçtan başlayarak gerçekleştirmektedir. Transfer öğrenme kullanılan makine öğrenmesi yöntemleri ise daha önceden öğrenilmiş olan bilgileri kullanmaktadır. Öğrenme işlemi sıfırdan başlayarak yapılmamaktadır [40]. 14 milyondan fazla görüntü içeren ImageNet [41] veri seti ile eğitilmiş olan modelden çıkarılan bilgiler Şekil 2.4 ile gösterilen transfer öğrenme mimarisine göre hedef veri setinde öğrenme aşamasında kullanılmak üzere başka modellere transfer edilmektedir.

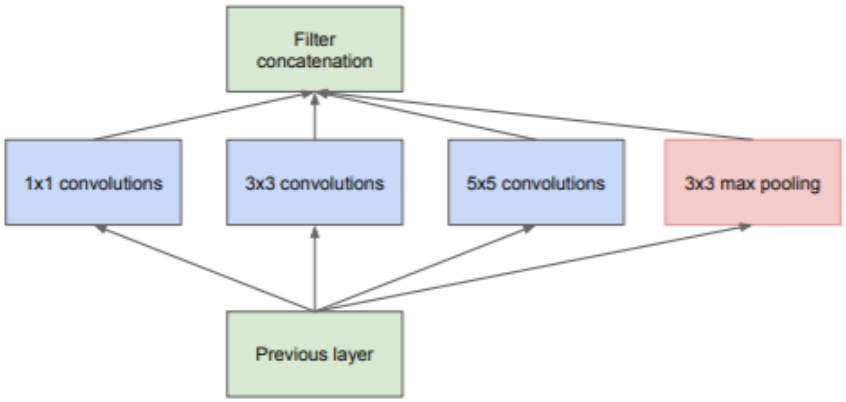


Şekil 2.4 Transfer öğrenme mimarisi [42]

2.3.3.1 InceptionV3 Mimarisi

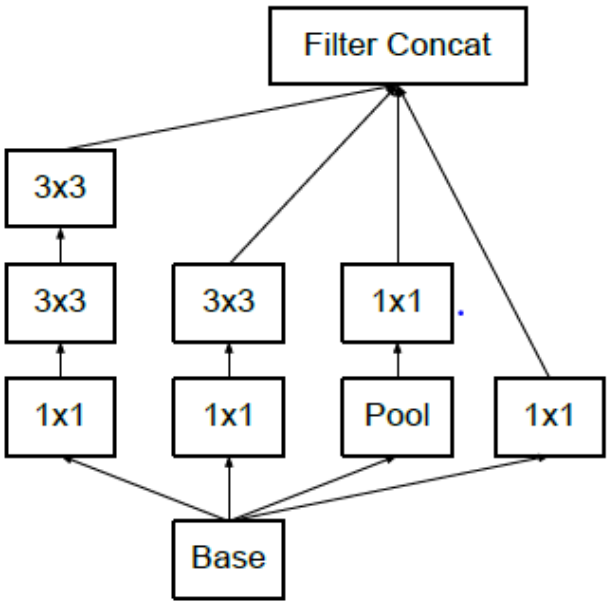
Inception, ILSVRC14 (ImageNet Büyük Ölçekli Görsel Tanıma Zorluğu 2014) yarışmasında sınıflandırma alanında iyi bir başarıya ulaşan bir derin sinir ağı mimarisidir [43]. Hesaplama maliyeti olarak VGG'den [44] daha düşük olması bu mimarinin büyük veriler ile kullanımını kolaylaştırmıştır [45].

Inception mimarisi Şekil 2.6'da gösterilmiştir.



Şekil 2.5 Inception mimarisi [43]

Şekil 2.5 ile verilen Inception mimarisi 1x1, 3x3 ve 5x5 olmak üzere 3 farklı boyutta konvolüsyona sahiptir. Bir önceki konvolüsyon katmanlarının çıktıları bir sonraki katmanın girdileri ile birleştirilerek yeni bir çıktı üretilmektedir. Ayrıca her katman konvolüsyona ek olarak havuzlama katmanına sahiptir.



Şekil 2.6 InceptionV2 mimarisi [45]

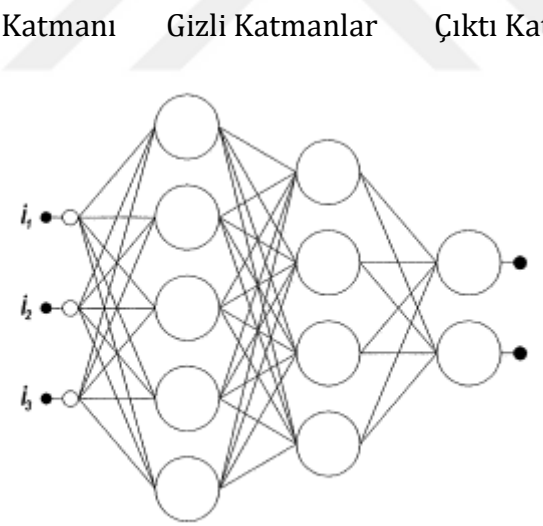
5x5 gibi büyük filtreli konvolüsyon işlemleri hesaplama maliyeti olarak daha pahalıdır. Bu nedenle InceptionV2 mimarisinde, Inception mimarisinde 5x5 boyutunda olan konvolüsyon işlemleri 3x3 boyutuna düşürülmüştür.

InceptionV3 mimarisinde ise InceptionV2 mimarisine ek olarak etiket yumuşatma (label smoothing) ve toplu normalleştirme (batch- normalization) gibi ek özellikler eklenmiştir. InceptionV3 mimarisinde girdi boyutu 299x299x3 iken çıktı boyutu 8x8x2048' dir.

2.3.4 Çok Katmanlı Algılayıcılar

Tek katmanlı yapay sinir ağları ile XOR probleminin çözülememesi nedeniyle yapısında birden fazla katman bulunduran bu problemi çözebilen çok katmanlı algılayıcılar (ÇKA) ortaya çıkmıştır.

ÇKA yapısında, giriş, gizli ve çıkış katmanları olmak üzere üç katman yapısı bulunmaktadır.



Şekil 2.7 ÇKA yapısı [13]

Şekil 2.7' de bir ÇKA modeli gösterilmiştir. Bu modelde ağırlıklar ile birbirine bağlı her bir düğüm, ağırlıklar ile giriş değerlerinin çarpımlarının toplamının bir aktivasyon fonksiyonundan geçirilmesi ile bir çıktı değeri üretilir. ÇKA modelinde verinin akış yönü giriş katmanından çıkış katmanına doğru yapılır. Bu duruma İleri beslemeli ağ (Feed-forward network) adı verilmektedir. Eğitim aşamasından sonra modelin ürettiği hata değeri ağırlıkların değiştirilmesi ile yeniden hesaplanmaktadır. Eğitim aşamasında Geri yayılım algoritması (Backpropagation algorithm) yaygın olarak kullanılmaktır. Algoritmada üretilen hata çıkış katmanından giriş katmanına doğru gidilerek ağırlıklar değiştirilerek azaltılmaktadır.

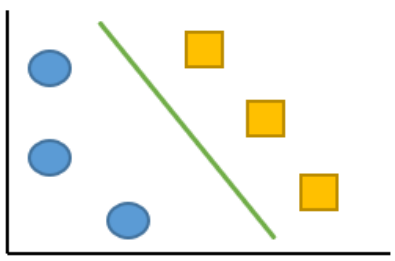
2.3.5 Destek Vektör Makineleri

DVM, vektör tabanlı sınıflandırma ve regresyon uygulamalarında kullanılan bir algoritmadır. DVM algoritmasında amaç iki sınıfı en az hata ile ayıran bir hiper-düzlem bulabilmektir. İdeal bir hiper-düzlemin iki sınıf vektörü arasında bulunan mesafe fazla olmalıdır [11]. Eğitim veri setinde bulunan her bir x için y değerleri x örneklerinin sınıf etiketidir. $(x_1, y_1), (x_2, y_2), .., (x_n, y_n)$ şeklinde verilen etiketlenmiş veri kümleri için optimum hiper - düzlemi bulmak için kullanılan denklem (2.9)' da verilmiştir.

$$w^T . x + b = 0 \tag{2.9}$$

$$\begin{cases} w.x + b \geq 1 \\ w.x + b \leq -1 \end{cases} \tag{2.10}$$

Bir (w, b) ikilisi (2.10) ile verilen eşitlikleri sağlıyorsa, hiper düzlem, iki sınıfa ait verileri Şekil 2.8'te gösterildiği gibi doğrusal olarak ayırabilmektedir [46].



Şekil 2.8 Veriyi doğrusal olarak ayıran düzlem

2.3.6 K-En Yakın Komşu Yöntemi (KNN)

KNN yöntemi, benzerlik tabanlı bir sınıflandırma algoritmasıdır. KNN yöntemi eğitim aşamasını içermediği için tembel bir denetimli öğrenme yöntemidir. Bu yöntemde amaç sınıf etiketi verilmeyen bir örneği K en yakın komşusu arasında çoğunluğu sağlayan sınıfa atamaktır [47]. Sınıflandırma işleminde K en yakın komşuyu bulmak uzaklık hesaplaması yapılmaktadır. Öklid, Minkowski ve Manhattan yaygın olarak kullanılan uzaklık metrikleridir. (x_1, x_2..,x_n) ve (y_1, y_2..,y_n) noktaları arasında bulunan Öklid, Minkowski ve Manhattan uzaklığına ait eşitlikler, (2.11), (2.12) ve (2.13) ile verilmiştir[48].

$$d_{\text{öklid}} = \sqrt{(x_i - y_i)^2} \tag{2.11}$$

$$d_{minkowski} = (\sum_{i=1}^{n} |x_i - y_i|^{p})^{1/p} \tag{2.12}$$

$$d_{manhattan} = \sum_{i=1}^{n} |x_i - y_i| \tag{2.13}$$

2.3.7 Naive Bayes (NB)

Bayes yöntemi istatistik tabanlı bir sınıflandırma yöntemidir ve bilinen bir ilk olasılık değerinden son olasılık değerini hesaplama işleminde kullanılmaktadır [49]. Bayes teoremi adı verilen bu yöntemde koşullara bağlı olarak bir durumun gerçekleşme olasılığı hesaplanmaktadır.

$$p(x|y) = \frac{p(y|x).p(x)}{p(y)} \tag{2.14}$$

$$sonraki\ olasılık = \frac{olasılık * önceki\ olasılık}{kanıt} \tag{2.15}$$

(2.14)' te verilen Bayes formülü, gözlemlenen x durumu yardımı ile önceki olasılık değeri olan $p(x)$ değerini sonraki olasılık değeri olan $p(x|y)$ değerine dönüştürmek için kullanılmaktadır [49]. Bayes teoremi (2.15) ile verilen eşitlik gibi genellenmektedir.

2.3.8 Rastgele Ormanlar

Rastgele orman (Random Forest), çok sayıda ağaçtan oluşan ve en yüksek oy puanına göre veri sınıflandırması yapan, ağaç yapısına dayalı bir sınıflandırma algoritmasıdır [50]. RO, aykırı değerler ve gürültüye karşı dirençlidir ve büyük veri setlerinde iyi sonuçlar üretmektedir. Hesaplama açısından diğer ağaç tabanlı yöntemlere göre kolaydır [51].

RO algoritması, özellik seçiminde Gini indeksini kullanmaktadır. Gini indeksi T eğitim setinde rastgele seçilen bir pikselin C_i sınıfına ait olma olasılığını söyler [12].

$$g = \sum \sum_{j \neq i} (f(C_i, T)/|T|)(f(C_j, T)/|T|) \tag{2.16}$$

2.3.9 Lojistik Regresyon

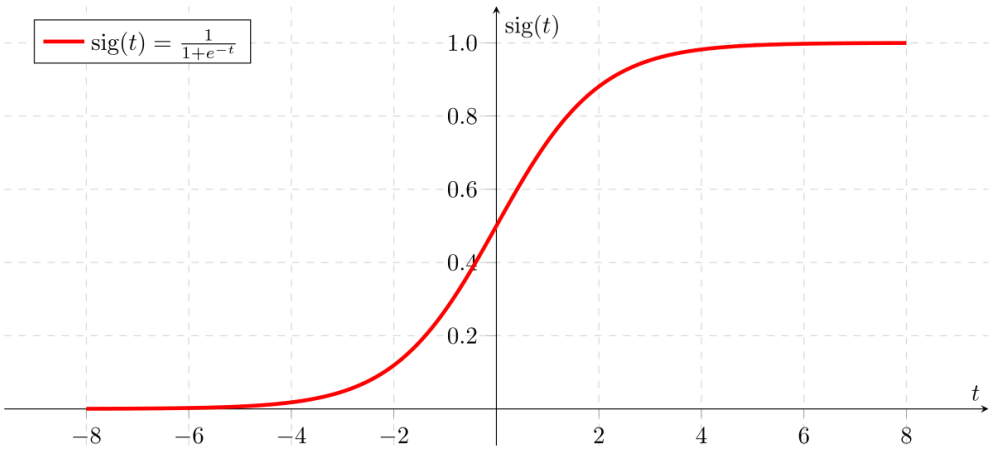
Lojistik regresyon modeli veri analizinde yaygın olarak kullanılan bir modellemedir. Lojistik regresyon sonucunda elde edilen çıktı ikili veya dikotom olabilir [52]. Lojistik regresyon bağımsız değişkenler ile bağımlı değişken arasında bulunan ilişkiyi inceleyen matematiksel bir modeldir. Bağımlı değişkenin kategorik olma durumunda lojistik regresyon modeli tercih edilmektedir.

Lojistik regresyon sonucu elde edilen olasılık değeri [0,1] aralığında olmalıdır. N tane bağımsız değişken için (2.17) ile verilen u doğrusal regresyon katsayılarını [0,1] arasına getirebilmek için Şekil 2.9 ile gösterilen lojistik bir fonksiyon olan sigmoid kullanılmaktadır.

$$u = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_n x_n \tag{2.17}$$

$$p(x) = \frac{e^u}{1+e^u} \tag{2.18}$$

(2.18) ile verilen eşitlikte $p(x)$ regresyon sonucu elde edilen olasılık değerini belirtmektedir ve bu eşitlik regresyon çıktısını [0,1] arasında bir değere getirebilmek için kullanılmaktadır.



Şekil 2.9 Sigmoid fonksiyonu [53]

3
VERİ SETİ VE ÖNİŞLEMLER

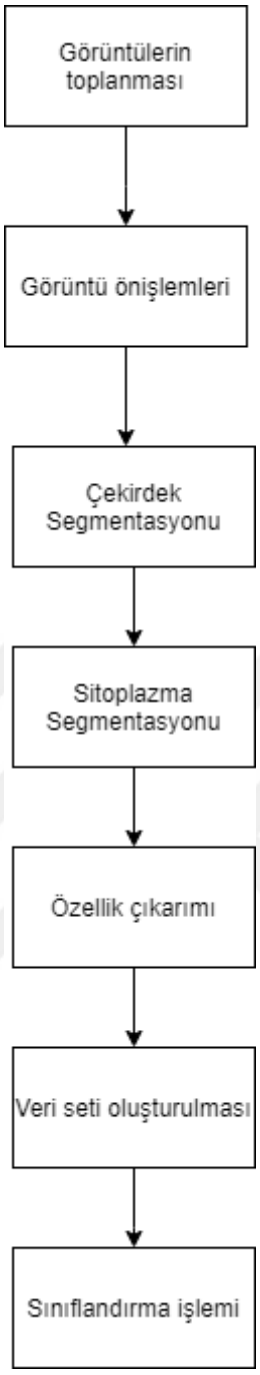
Bu çalışmada pap-smear testi sonucu elde edilen görüntüler kullanılmıştır. Kullanılan her görüntü üzerinde bazı önişlemler uygulanarak görüntü iyileştirme işlemi yapılmıştır.

3.1 Herlev Veri Seti

Çalışmada ele alınan yöntemler Herlev veri seti [5] üzerinde uygulanmıştır. Herlev Üniversitesi Hastanesi tarafından hazırlanmış olan Herlev veri seti 4 tanesi anormal ve 3 tanesi normal olarak belirlenmiş 7 sınıfa ait 917 pap-smear görüntüsünden oluşmaktadır. Anormal olarak belirlenen kategorilerden carcinoma_in_situ'ya ait 150, light_dysplastic'te 182, moderate_dysplastic'te 146 ve severe_dysplastic' te 192 pap-smear görüntüsü bulunmaktadır [54]. Normal olarak sınıflandırılan kategorilerden normal_superficial'da 74, normal_intermadiate'ta 70 ve normal_columnar'da 98 adet görüntü bulunmaktadır.

3.2 Veri Önişleme

Çalışma segmentasyon, özellik çıkarımı ve sınıflandırma olmak üzere üç temel adımdan oluşmaktadır. Görüntüleri segmentasyon işlemine hazır hale getirmek için görüntü üzerinde bazı önişlemler uygulanmıştır. Çalışma kapsamında takip edilen adımlar Şekil 3.1'de gösterilmiştir. Çalışma, Python programlama dili ve Keras kütüphanesi kullanılarak geliştirilmiştir.



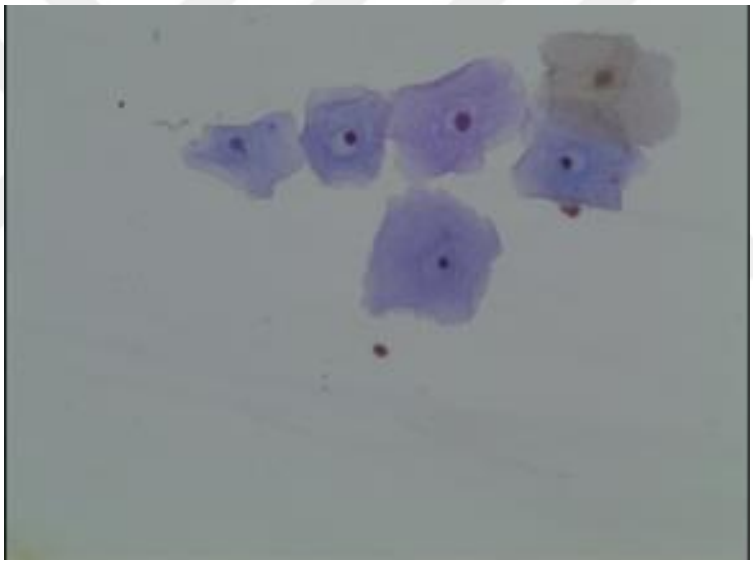
Şekil 3.1 Çalışmada izlenen adımlara ait akış diyagramı

Rahim ağzı kanseri tespitinde hücre çekirdeğinin önemli bir yeri vardır. Bu nedenle ilk olarak görüntü üzerinde hücre çekirdeği kısmı incelenmiştir. Çekirdeği tek olarak ele alabilmek için sitoplazma kısmı görüntüden çıkarılmıştır. Hücre çekirdeğini sitoplazmadan ayırabilmek için hücreye segmentasyon işlemi uygulanmıştır.

Bu çalışmada veri olarak pap-smear testi sonucu elde edilen serviks görüntüleri kullanılmıştır. Segmentasyon işlemine hazırlık için görüntüler bazı ön işlemlerden

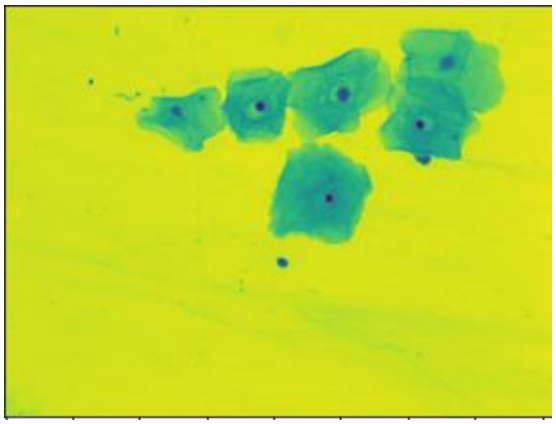
geçirilmiştir. İyileştirme işleminde amaç görüntü üzerinde bulunan gürültünün temizlenmesi ve kontrastlığın artırılmasıdır. Kontrastlığın artırılma amacı ise çekirdek ve sitoplazma arasındaki koyuluğun artırılarak çekirdeğin daha koyu sitoplazmanın daha açık hale getirilmesidir. Gürültüyü gidermek için boyutu 5 olan bir medyan filtre uygulanmıştır. Görüntüyü iyileştirmek için ise AKD ve ADD yöntemleri uygulanmıştır.

AKD yönteminde her pikselin grilik seviyelerinin tutulduğu 8×8' lik bir piksel matrisi oluşturulmuştur ve bu matrise AKD yöntemi uygulanmıştır. Girdi olarak verilen görüntüyü yeniden elde etmek için ters ayrık kosinüs dönüşümü uygulanmıştır. SIPAKMED [55] veri setinden alınan Şekil 3.2 'de gösterilen giriş görüntüsü için AKD işlemi sonucu elde edilen çıkış görüntüsü Şekil 3.3' de gösterilmiştir.



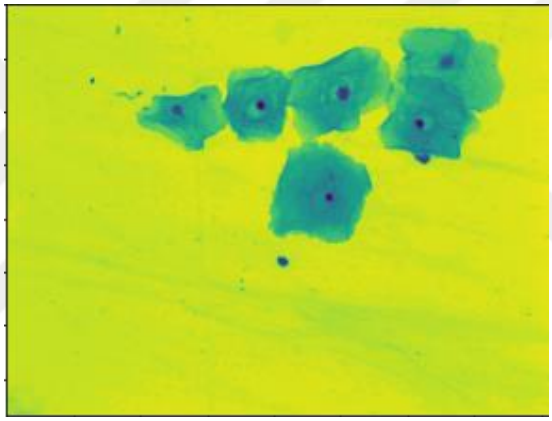
Şekil 3.2 Serviks hücresine ait giriş görüntüsü

Çekirdek ve sitoplazmayı birbirinden daha iyi ayırabilmek için bu iki görüntü arasındaki zıtlık oranı, çekirdek daha koyu, sitoplazma daha açık hale getirilerek artırılmıştır.



Şekil 3.3 AKD sonucu elde edilen görüntü

AKD işlemi sonucu elde edilen görüntü üzerinde ADD yöntemi uygulanmıştır ve ADD işlemi sonucu elde edilen görüntü Şekil 3.4' de gösterilmiştir.



Şekil 3.4 ADD sonucu elde edilen görüntü

3.3 Çekirdek Segmentasyonu

Bir serviks hücresini normal veya anormal olarak sınıflandırabilmek için hücrenin çekirdeği ve sitoplazması ayrı olarak ele alınmıştır. Bu iki yapıyı birbirinden ayırabilmek için segmentasyon işlemi uygulanmıştır.

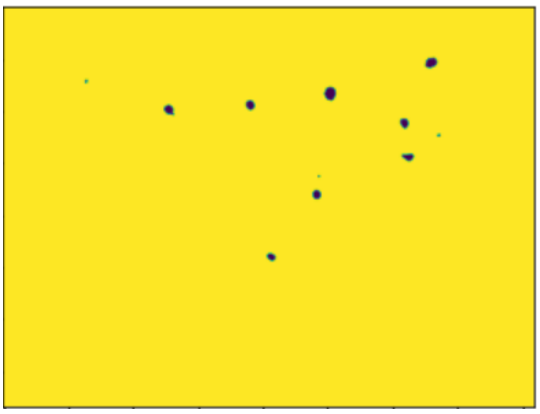
Şekil 3.4' de gösterilen ayrık dalgacık dönüşümü ile iyileştirilen görüntüde hücre çekirdeklerinin yaygın olarak mavi renk spektrumunda olduğu görünmektedir. Bu nedenle hücre çekirdeklerini tespit edebilmek için giriş görüntüsünde mavi renkli alanların bulunması gerekmiştir. Bu alanları bulabilmek için ilk olarak RGB renk uzayında bulunan giriş görüntüsü HSV (Hue, Saturation, Value) renk uzayına dönüştürülmüştür. Daha sonra mavi renkli alanları filtrelemek için (3.1)' de verilen

denklem ile belirtilen mavi renk alt sınırı (3.2)' de t ile gösterilen mavi renk üst sınırı belirlenmiştir ve görüntüde bu aralıklarda bir maske oluşturulmuştur. Alt ve üst sınır üç değerden oluşmaktadır. Her bir değer HSV renk uzayının üç kanalına (h, s ve v) denk gelmektedir. H değeri rengi, s değeri rengin doygunluğunu ve v değeri rengin parlaklığını ifade eder. Renk uzayında bulunan tüm renkler belirli bir alt ve üst sınır renk değerleri arasında bulunmaktadır. a ve t dizilerinin ilk elemanları (0-120) renk uzayında mavi rengin alt (0) ve üst (120) renk değerini, ikinci elemanları rengin canlılık değerinin alt ve üst sınırını ve üçüncü elemanlar ise mavi renginin belirlenen parlaklık alt ve üst sınırlarıdır. Bu çalışmada mavi renkli bölgeler filtreleneceği için renk uzayında mavi renk alt ve üst sınırı [0,120] arası olarak belirlenmiştir. RGB renk uzayında mavi renk şeklinde ifade edilmektedir. [0,0,255] değeri HSV renk uzayına çevrilmiştir ve h değeri olarak 0 elde edilmiştir ve bu değer alt sınır olarak kullanılmıştır. Oluşturulan maske ile görüntüye (3.3)'te belirtilen bit bazında mantıksal işlemler uygulanarak elde edilen maske görüntüsü Şekil 3.5' de gösterildiği gibi maskelenmiştir.

$$a = [0, 0{,}55] \tag{3.1}$$

$$t = [120, 255, 255] \tag{3.2}$$

$$g = f \,\&\, m \tag{3.3}$$



Şekil 3.5 Hücre çekirdeklerinin bulunması

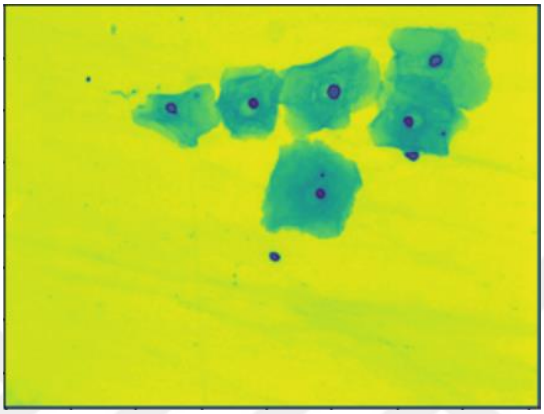
Maskeleme işlemi ile bulunan Şekil 3.4 ile gösterilen f görüntüsü için sırasıyla morfolojik açma ve yayma işlemleri uygulanmıştır.

$$m = f \circ s \tag{3.4}$$

$$g = f \ominus s \tag{3.5}$$

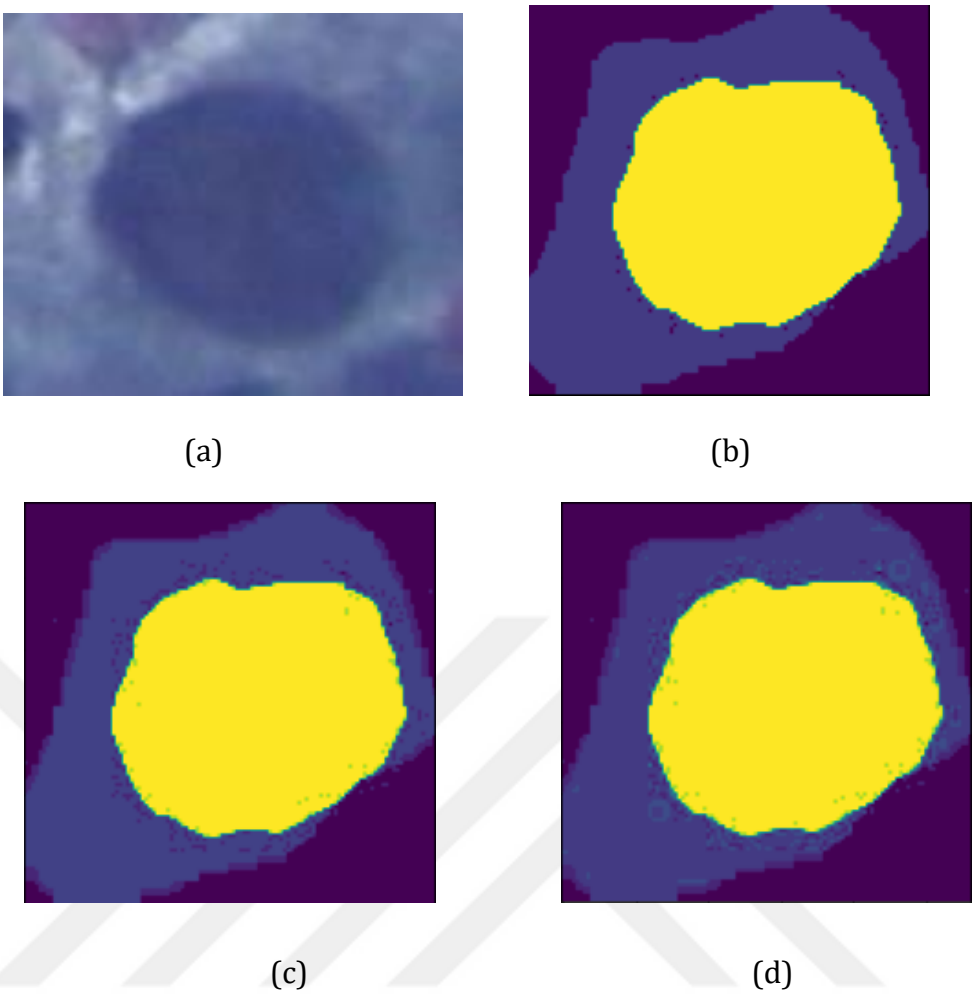
(3.4)' te HSV renk uzayında maske uygulanan f imgesi için s yapı elemanı ile uygulanan morfolojik açma işlemi belirtilmiştir. (3.5)' te ise açma işlemi ve s yapı elemanı ile uygulanan morfolojik yayma işlemi verilmiştir.

Yayma işlemi sonucu elde edilen g görüntüsü Şekil 3.5' de gösterilmiştir. Elde edilen çekirdek maskeleri f giriş görüntüsü üzerinde Şekil 3.6' da gösterildiği gibi işaretlenmiştir.



Şekil 3.6 Hücre çekirdeklerinin giriş görüntüsü üzerinde işaretlenmesi

Serviks hücresinde çekirdek segmentasyonu için eşikleme ve morfolojik işlemlerin dışında K-ortalama yöntemi kullanılmıştır. Bu yöntemde $k = 3$ alınmıştır ve Şekil 3.7 (a)'da verilen giriş görüntüsü için k-ortalama işlemi sonucu elde edilen görüntü Şekil 3.7 (b)'de gösterilmiştir. Farklı k değerleri uygulandığında ise görüntü üzerinde yapılan kümeleme işleminde bazı noktaların yanlış sınıflara kümelendiği görülmüştür. Şekil 3.7(c)'de görüldüğü gibi K-ortalama yönteminde k=5 değeri kullanıldığında çekirdek içerisinde bulunan bazı noktaların yanlış kümelere atandığı gözlemlenmiştir. K değeri 7 olarak alındığında Şekil 3.7 (d)'de görüldüğü gibi kümeleme işlemi sonucu elde edilen görüntünün bozulduğu ve yanlış kümelere atanan noktaların sayısının arttığı gözlemlenmiştir.



Şekil 3.7 Farklı K değerleri için K-ortalama yöntemi sonuçları (a) Giriş görüntüsü (b) K=3 (c) K=5 (d) K=7 için K-ortalama yönteminin sonuçları

3.4 Sitoplazma Segmentasyonu

AKD ve ADD yöntemleri ile iyileştirilen pap-smear görüntülerinde hücre ve sitoplazma arasındaki kontrast artırılmıştır. Çekirdek daha koyu, sitoplazma daha açık hale getirilerek segmentasyon aşamasında bu iki yapının birbirinden daha kolay ayrılabilmesi sağlanmıştır.

Hücrenin sitoplazma bölümlerini bulmak için uygulanan adımlar (3.6)' da gösterilmiştir.

$$S_{stp} = G - f \tag{3.6}$$

Kontrastlığı artırılmış G giriş görüntüsünden eşikleme yöntemi kullanılarak bulunan çekirdek kısımları çıkarılmıştır. Kalan bölümlere OTSU eşikleme yöntemi

ve (3.4) ve (3.5)' te belirtilen morfolojik işlemler uygulanmıştır. Yapılan işlemler sonucu elde edilen görüntü Şekil 3.8' de gösterilmiştir. Görüntüde sitoplazma alanları beyaz olarak görünmektedir ve her biri kontur olarak kaydedilip özellik çıkarımı alanında belirtilen sınıflandırma kriterleri hesaplanmıştır.



Şekil 3.8 Segmentasyon işlemi sonucu sitoplazma

3.5 Özellik Çıkarımı

Bir serviks hücresini anormal veya normal olarak sınıflandırabilmek için görüntüden çıkarılan bazı özellikler sınıflandırma kriteri olarak kullanılmıştır. Anormal hücrelerde genellikle; konturda düzensizlik ve çekirdekte genişleme gözlemlenir. Koyu renk çekirdeğe sahip olup yüksek çekirdek-sitoplazma oranına sahiptirler [56].

Yukarıda belirtilen bilgilere dayanarak görüntü üzerinde aşağıdaki özellikler hesaplanmış ve sınıflandırma kriteri olarak belirlenmiştir:

Çekirdek alanı: Çekirdek olarak belirlenen her bir konturun alanı

Sitoplazma alanı: Sitoplazma olarak bulunan her konturun alanı

Çekirdek koyuluk değeri: Çekirdek konturlarının grilik seviyeleri

Çekirdek çevresi: Çekirdek konturlarının çevre hesabı

Sitoplazma çevresi: Sitoplazmaların hesaplanan çevre değeri

Çekirdeğin merkez koordinatları: Her bir çekirdeğin merkezinin bulunduğu (x, y) koordinatları

Sitoplazmanın merkez koordinatları: Her bir sitoplazmanın merkezinin bulunduğu (x, y) koordinatları

Çekirdek/ Sitoplazma (Ç/S) oranı: Hücrede bulunan çekirdek alanının sitoplazma alanına oranı

Çekirdek sitoplazma arası mesafe: Her bir hücre çekirdeğinin o hücrenin sitoplazması ile arasındaki uzaklık

Çekirdek yuvarlaklığı: Çekirdek konturunun yuvarlaklığının hesaplanması (3.7) ile gösterilmiştir.

$$NRnd = \frac{4*\pi*cekirdekalan}{cekirdekcevre^2} \tag{3.7}$$

Sitoplazma yuvarlaklığı: Sitoplazmanın yuvarlaklık değerinin (3.8)' de göre hesaplanan değeri

$$SRnd = \frac{4*\pi*stpalan}{stpcevre^2} \tag{3.8}$$

Her hücrede bulunan çekirdek sayısı: Bazı anormal hücreler birden fazla çekirdekten oluşabildiği için sınıflandırma işleminde her bir hücrede kaç tane çekirdek bulunduğu hesaplanmıştır.

4
DENEYSEL SONUÇLAR

Bu tez çalışması kapsamında Herlev veri setinde bulunan pap-smear görüntüleri için sınıflandırma kriterleri belirlenmiştir ve bu kriterler bir veri setine kaydedilmiştir. Oluşturulan veri seti sınıflandırma işleminde kullanılmıştır. Sınıflandırma işleminde kullanılan her bir yöntem için başarı değerleri elde edilmiştir.

4.1 Değerlendirme Ölçütleri

Bir sınıflandırıcının, sınıflandırma işlemi sonucunda veri setinde bulunan örnekleri doğru sınıflar ile eşleştirme oranı sınıflandırma işleminin başarısı olarak kabul edilmektedir. Bu başarı oranını ölçmek için bazı değerlendirme ölçütlerine ihtiyaç duyulmaktadır. Değerlendirme ölçütlerini hesaplama işleminde Tablo 4.1'de gösterilen karışıklık matrisi (confusion matrix) kullanılmaktadır. Karışıklık matrisi sınıflandırma sonucunda gerçek sınıfların ve tahmin edilen sınıfların gösterildiği bir matristir. Bu çalışma kapsamında 2 sınıflı bir sınıflandırma işlemi yapıldığı için DP (Doğru-Pozitif), DN (Doğru- Negatif), YP (Yanlış- Pozitif) ve YN (Yanlış- Negatif) olmak üzere 2x2 boyutunda bir karışıklık matrisi elde edilmiştir.

DP: veri setinde kanser hücresine ait bir görüntünün sınıflandırma sonucu kanser olarak sınıflandırılması

DN: kanser olmayan bir hücrenin kanser değil olarak sınıflandırılması

YN: kanser hücresine ait bir görüntünün kanser değil olarak sınıflandırılması ve

YP: kanser olmayan bir hücreyi kanser olarak sınıflandırmaktır.

Tablo 4.1'de verilen karışıklık matrisinden sınıflandırıcının sınıflandırma başarısının ölçümünde kullanılan bazı ölçütler hesaplanmaktadır. Bu ölçütler:

Doğruluk (Accuracy): Gerçek pozitif değerlerin tüm değerlere oranıdır.

$$doğruluk = \frac{DP}{DP+YP+DN+YN} \tag{4.1}$$

Tablo 4.1 Karışıklık matrisi

	Gerçek Değerler		
Tahmin edilen değerler		Pozitif	Negatif
	Pozitif	DP	YP
	Negatif	YN	DN

Hassasiyet (Precision): Gerçek pozitif olan değerlerin tüm pozitif değerlendirmelere oranıdır.

$$hassasiyet = \frac{DP}{DP+YP} \tag{4.2}$$

Duyarlılık (Sensitivity): Duyarlılık, gerçekte anormal olan hücrelerin anormal olarak sınıflandırılma olasılığıdır.

$$duyarlılık = \frac{DP}{DP+YN} \tag{4.3}$$

Özgünlük (Specifity): Bir sınıflandırma sonucunda elde edilen Gerçek-Negatif (TN) oranını belirtir.

$$özgünlük = \frac{DN}{DN+YP} \tag{4.4}$$

4.2 Deneysel Sonuçlar

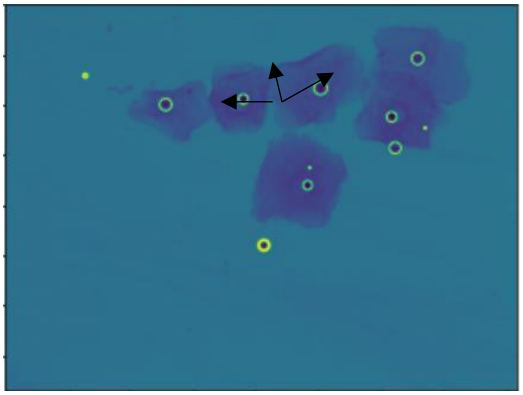
Çekirdek ve sitoplazma alanının hesaplanması çekirdek genişlemesi durumunun ölçülebilmesine olanak sağladığı için hücreyi anormal olarak sınıflandırmada büyük önem taşır. Bu nedenle her bir sitoplazma ve çekirdek konturu için görüntüden alan özelliği çıkarılmıştır. Elde edilen alan bilgilerinden önemli bir teşhis kriteri olan Ç/S oranı hesaplanmıştır. Segmentasyon işlemi sonucunda görüntüde birden fazla çekirdek ve sitoplazma konturları bulunduğu için her çekirdeğin içerisinde bulunduğu sitoplazması ile eşleştirilmesi gerekmiştir. Bu eşleştirme işleminin doğru

yapılmaması sonucunda yanlış Ç/S oranı hesaplanacağı için eşleştirme işlemi doğru oran bulmada önemli yer tutmaktadır. Eşleştirme işlemi aşamasında her bir çekirdek ve sitoplazma için çıkarılan merkez koordinatları kullanılmıştır.

Her bir çekirdek için $(x_n,\ y_n)$ ikilisi ve her sitoplazma için $(x_s,\ y_s)$ ikilisi konturun merkezinin bulunduğu x ve y koordinatlarını belirtmektedir. Her sitoplazma içinde çekirdek bulundurduğu için sitoplazmanın merkez koordinatları ile içinde bulundurduğu çekirdeğin merkez koordinatları arasında bulunan mesafenin az olması gerektiği düşünülmüştür. Bu iki nokta arasındaki mesafeyi ölçmek için (4.5)' te gösterilen Euclidean mesafesi yöntemi kullanılmıştır.

$$d = \sqrt{(x_n - x_s)^2 + (y_n - y_s)^2} \tag{4.5}$$

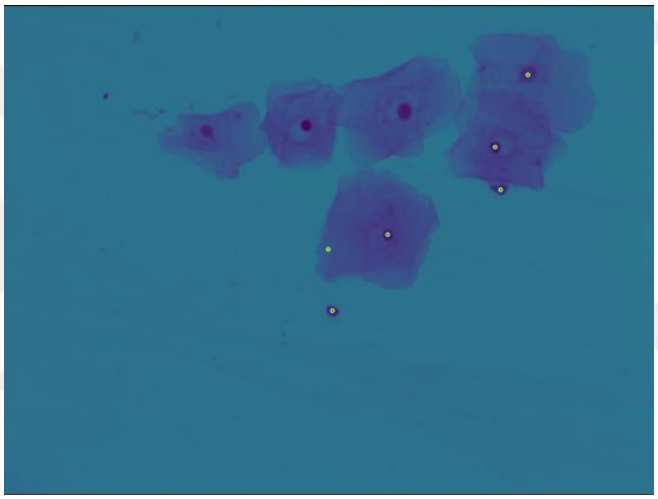
Şekil 4.1' de verilen referans görüntüsü için (4.5)' te belirtilen d uzaklık değerinin 1.00 değerine eşit veya daha küçük olması doğru çekirdek-sitoplazma eşleştirme sonucu verdiği gözlemlenmiştir. d uzaklık değeri olarak hesaplama işleminde doğru eşleştirme sonucunu veren minimum değer kabul edilmiştir. Minimum değer alınmasının nedeni ise d mesafesinin yüksek bir değer alınması durumunda Şekil 4.1' de üzerinde örnek olarak alınan bir çekirdek oklar ile gösterilen üç sitoplazma ile de eşleştirilebilir ve bu durum yanlış çekirdek-sitoplazma eşleştirme sonucunu ortaya çıkarmaktadır. Sitoplazması ile eşleştirilen her bir çekirdek yuvarlak içine alınmıştır.



Şekil 4.1 Çekirdek-Sitoplazma eşleştirme

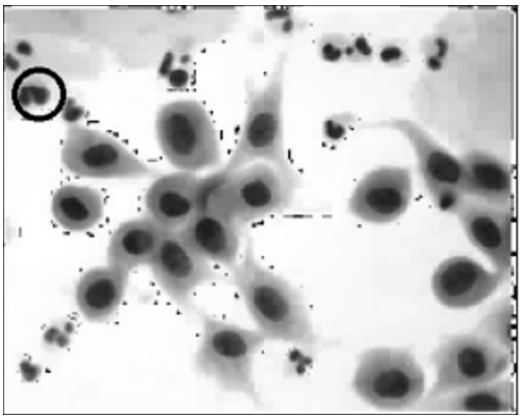
Eşleştirilen her bir çekirdek ve sitoplazma için Ç/S oranı hesaplanmıştır ve veri setine kaydedilmiştir. Veri setinde mesafesi 1.2 ve altında olan hücrelerde Ç/S oranı olarak hesaplanan değer, diğer durumlarda 0 olarak alınmıştır.

Koyu renk çekirdeğe sahip olan hücreleri bulmak için her bir çekirdek konturunun ortalama koyuluk değeri hesaplanmış ve ilgili çekirdek ile ilgili verilere kaydedilmiştir. Daha sonra veri setinde yoğunluk özelliğinin altında bulunan tüm değerlerin ortalaması bulunmuştur. Bir çekirdeğin konturu hesaplanan ortalama değerinden büyükse hücre koyu renk çekirdeğe sahip olduğu düşünülmüştür. Ortalama yoğunluk değerinden büyük yoğunluğa sahip olan bölgeler Şekil 4.2'de sarı işaretler ile gösterilmiştir.



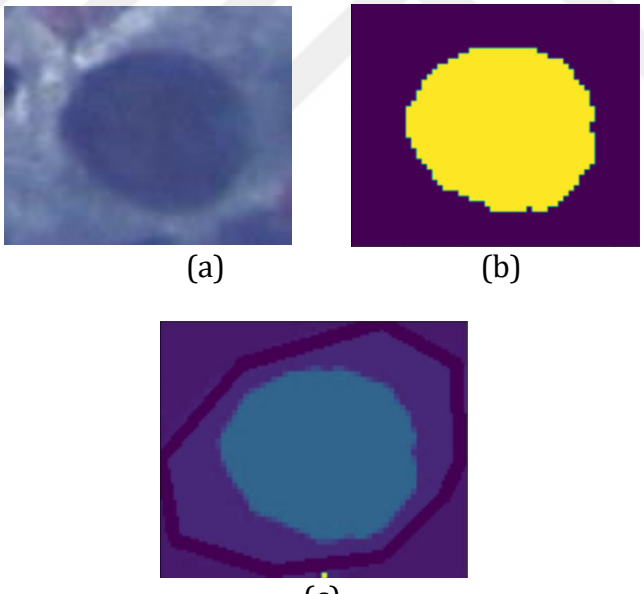
Şekil 4.2 Yoğunluk değeri büyük olan bölgeler

Anormal hücre sınıflandırma kriterlerinden biri olan sitoplazmada bulunan hücre sayısı hesaplanmıştır. Eğer bir sitoplazmada birden fazla çekirdek bulunuyorsa bu hücreler anormal olarak değerlendirilmiştir. Bu hesaplama işleminde çekirdek konturlarının merkezlerinin arasındaki mesafe bulunmuştur. Mesafe hesaplama işleminde her $(x_n,\ y_n)$ çekirdek merkez koordinatları için (4.5) ile verilen mesafe bulunmuştur ve bu d mesafesinin 5.00'ın altında olması durumunda sitoplazma içerisinde bulunan çekirdek sayısı artırılmıştır. Şekil 4.3' de verilen görüntü [27] için birden fazla çekirdeğe sahip olan hücre bulunmuştur ve görüntü üzerinde yuvarlak içerisine alınmıştır.



Şekil 4.3 İki çekirdeğe sahip hücrenin görüntü üzerinde görünümü

Makine öğrenmesi yöntemleri ile başarı oranı yüksek bir sınıflandırma yapabilmek için veri sayısının fazla olması gerekmektedir. Bu nedenle Herlev veri seti kullanılarak bu veri setinde bulunan 917 görüntü için ilk olarak çekirdek ve sitoplazma segmentasyon işlemleri yapılmıştır. Şekil 4.4 (a)'da belirtilen görüntü için çekirdek ve sitoplazma segmentasyonu işlemleri sonucunda elde edilen görüntüler Şekil 4.4 (b) ve (c)'de gösterilmiştir.



(a) (b)

(c)

Şekil 4.4 (a) Giriş görüntüsü (b) Çekirdek segmentasyonu (c) Sitoplazma segmentasyonu

4.2.1 İki- Sınıflı Sınıflandırma İşlemi

Çekirdek ve sitoplazma segmentasyonu işlemlerinden sonra sınıflandırma kriteri olarak belirlenmiş özellikler veri setinde bulunan 917 görüntü için bulunmuştur.

Görüntülerden çıkarılan özellikler ile oluşturulan veri setinden normal ve anormal hücrelere ait bazı veriler Tablo 4.2'de gösterilmiştir. Herlev veri setinde bulunan 7 kategori bu çalışma kapsamında normal ve anormal olmak üzere iki kategoriye indirgenmiştir. Tablo 4.2' de görüldüğü gibi normal hücrelerin normal kategorisi 1 olarak, anormal hücrelerin ise 0 olarak etiketlenmiştir. Anormal hücrelerin ise anormal kategorileri 1, normal kategorileri ise 0 olarak belirlenmiştir. Herlev veri setinde, Şekil 4.4 (a)' da görüldüğü gibi her sitoplazma bir çekirdek içerdiği için sitoplazmada bulunan çekirdek sayısı ve çekirdek sitoplazma arası mesafe sınıflandırma kriterleri içinde kullanılmamıştır.

Tablo 4.2 Normal ve anormal hücrelere ait bazı veriler

NC_Oran	CAlan	StpAlan	CÇevre	StCevre	NRnd	SRnd	Normal	Anormal
1.18	4825.5	4057	267.86	393.88	0.84	0.32	0	1
0.80	3415.5	4243	228.55	341.06	0.82	0.45	0	1
0.01	569	36487	92.76	752.05	0.89	0.81	1	0
0.15	3639.5	23356	599.16	691.71	0.69	0.61	1	0

Tablo 4.2'de gösterilen CÇevre ve StCevre özellikleri çekirdek ve sitoplazma konturlarının çevre hesabını ifade etmektedir. Nrnd ve SRnd özellikleri ise (3.7) ve (3.8)' de verilen eşitlikler ile hesaplanmış olan çekirdek ve sitoplazmaların yuvarlaklık oranını belirtmektedir. Anormal hücrelerin çekirdeklerinde genişleme görüldüğü için CAlan özelliği ile belirtilen çekirdek alanları ile StpAlan özelliği belirtilen sitoplazma alanının birbirine yakın olduğu bu nedenle NC_Oranı olarak belirtilen ÇAlan/StpAlan oranlarının yüksek çıktığı görülmüştür.

Oluşturulan 2-Sınıflı veri setinde sınıflandırma işlemi yapmak için makine öğrenmesi yöntemlerinden DVM, KNN, LR, NB, RO, ÇKA ve derin öğrenme uygulanmıştır.

DVM yönteminde çekirdek fonksiyonu olarak linear kullanılmıştır ve düzenleme parametresi (C) 1.00 olarak alınmıştır.

Tablo 4.3 DVM 2-Sınıf için karışıklık matrisi

		Tahmin edilen	
		Anormal	**Normal**
Gerçek	**Anormal**	42	9
	Normal	11	114

RO yönteminde ormanda bulunan ağaç sayısını belirten n_estimators değişkeni 100 olarak alınmıştır.

Tablo 4.4 RO 2-Sınıf için karışıklık matrisi

		Tahmin edilen	
		Anormal	**Normal**
Gerçek	**Anormal**	129	5
	Normal	9	33

KNN yönteminde, farklı K değerleri için elde edilen doğruluk değerleri Tablo 4.5 'te gösterilmiştir. K=7 ve K=9 değerleri için doğruluk değerleri aynı gözükmektedir fakat K=7 değeri için eğitim başarısı %89 iken k=9 değeri için eğitim başarısı %90 olarak sonuçlanmıştır.

Tablo 4.5 KNN yöntemi K sayısına göre sınıflandırma sonuçları

K	**Doğruluk (%)**
K=3	86
K=5	88

Tablo 4.5 KNN yöntemi K sayısına göre sınıflandırma sonuçları (devamı)

K=7	90
K=9	90

ÇKA yönteminde aktivasyon fonksiyonu olarak ReLU (Rectified Linear Unit) kullanılmıştır ve gizli katman sayısı 3 olarak belirlenmiştir. Her bir gizli katmanda bulunan düğüm sayısı 10 ve 100 değerleri arasında değiştirilerek başarı değerleri ölçülmüştür. Ölçümler sonucunda düğüm sayısının 40 ve üstüne çıktığı sınıflandırma işlemlerinde eğitim setinde başarı artarken test setinde başarı azalmıştır. Bu nedenle her katmanda bulunan düğüm sayısı 30 olarak alınmıştır. 2-Sınıflı sınıflandırma işlemi sonucu elde edilen doğruluk, hassasiyet, duyarlılık ve özgünlük değerleri Tablo 4.6' da belirtilmiştir.

Tablo 4.6 2-Sınıflı sınıflandırma işlemi sonuçları

	DVM	**RO**	**ÇKA**	**LR**	**KNN (k=9)**	**NB**
Doğruluk (%)	90.3	92.04	90	91.5	90	82.9
Hassasiyet (%)	87.8	80	93.1	93.5	85.2	87.0
Duyarlılık (%)	90.9	95.5	85.7	91	90.8	67.5
Özgünlük (%)	87.8	80	93.1	93.5	85.2	87.0

Veri setinde bulunan görüntü sayısının az olması nedeniyle öğrenme başarısını artırmak için derin öğrenme yönteminde transfer öğrenme kullanılmıştır. Transfer öğrenme yönteminde InceptionV3 modeli kullanılarak ImageNet veri seti ile eğitilen bir model çalışmada oluşturulan veri setine uygulanmıştır. ImageNet'te bulunan

görüntülerin kategorileri ile serviks görüntüleri kategori olarak farklı olduğu için transfer öğrenme modelinde sınıflandırma işleminin yapıldığı tam bağlı katman modele eklenmemiştir.

Derin yapay sinir ağı modelinde InceptionV3 mimarisi ile ImageNet üzerinde eğitilmiş olan modele ek olarak 512 nörona sahip gizli katman kullanılmıştır. Giriş katmanında alınan her girdi için boyut 299x299x3 olarak belirlenmiştir. Gizli katmanda aktivasyon fonksiyonu olarak ReLu kullanılırken çıkış katmanında sigmoid kullanılmıştır.

Derin öğrenme yönteminde aktivasyon fonksiyonu olarak sigmoid, kayıp fonksiyonu olarak binary_crossentropy kullanılmıştır. İyileştirici (optimizer) olarak ise SGD kullanılmıştır. Meydana gelebilecek aşırı öğrenme (overfitting) durumunu önlemek için katmanlar arasına 0.7 olasılık değeri ile Bırakma (Dropout) katmanı eklenmiştir. Devir (Epoch) sayısı 100 ve yığın boyutu (batch-size) 32 olarak alınmıştır. Transfer öğrenme ve derin öğrenme tekniklerinin birlikte kullanımı sonucu elde edilen doğruluk değeri ve sadece derin öğrenme kullanılarak elde edilen doğruluk değeri Tablo 4.7'de belirtilmiştir.

Sadece derin yapay sinir ağları kullanılan teknikte sırasıyla 512 ve 128 nörona sahip iki gizli katman kullanılmıştır ve bu katmanlarda aktivasyon fonksiyonu olarak ReLu kullanılmıştır. Çıkış katmanında ise sigmoid aktivasyon fonksiyonu kullanılmıştır. Devir sayısı ve yığın boyutu transfer öğrenmede kullanılan devir sayısı ve yığın boyutu ile aynı kullanılmıştır. İyileştirici fonksiyonu olarak Adam ve kayıp fonksiyonu olarak binary_crossentropy kullanılmıştır.

Tablo 4.7 Derin öğrenme 2-Sınıflı sınıflandırma işlemi sonuçları

Metot	**Doğruluk (%)**
Transfer öğrenme+ derin öğrenme	92.61
Derin öğrenme	93.1

4.2.2 Üç- Sınıflı Sınıflandırma İşlemi

Herlev veri setinde bulunan yedi kategori CIN, Normal ve Cancer olmak üzere üç kategoriye indirgenmiştir. Bu kategoriler ve içerdikleri sınıflar:

1. CIN: Serviks görüntüsünde bulunan displazi durumunu içeren kategoridir. Herlev veri setinde bulunan light_dysplastic, moderate_dysplastic ve severe_ dysplastic kategorileri veri setinde CIN sınıfı altına toplanmıştır.
2. Normal: Normal hücreleri ifade etmektedir. Herlev veri setinde bulunan normal_superficial, normal_intermadiate ve normal_columnar kategorileri veri setinde Normal sınıfı altına toplanmıştır.
3. Cancer: Rahim ağzı kanseri bulunan hücreleri ifade etmektedir. Herlev veri setinde bulunan carcinoma_in_situ veri setinde Cancer sınıfı altına toplanmıştır.

3-Sınıflı veri setine ait bazı veriler Tablo 4.8'de belirtilmiştir. Tabloya göre CAlan çekirdek alanını, StAlan sitoplazma alanını belirtirken Cevre özelliği ile çekirdeğin çevresi belirtilmiştir. Veri setinde belirlenen üç kategori için N/C oranlarının ve NRoundness özelliklerinin ortalama değeri hesaplanmıştır. Hesaplanan ortalama N/C değerleri CIN kategorisi için 0.74, Normal kategorisi için 0.28 ve Cancer kategorisi için 4.38 olarak bulunmuştur. NRoundness özelliği CIN kategorisi için 0.77, Normal kategorisinde 0.59 ve Cancer kategorisinde 0.72 olarak ortalama değer hesaplanmıştır.

Tablo 4.8 3-Sınıflı veri seti örneği

N/C	CAlan	StAlan	Solidity	Cevre	StCevre	NRnd	SRnd	Class
1.18	4825.5	4057	1	267.86	393.88	0.84	0.32	CIN
1.64	2407.5	1462.5	0.95	220.5	264.79	0.62	0.26	CIN
0.01	569	36487	0.98	92.76	752.05	0.89	0.81	Normal
1.80	3570	1980.5	0.95	276.93	293.37	0.80	0.28	Cancer

Tablo 4.9 DVM 3-Sınıf için karışıklık matrisi

		Tahmin edilen		
Gerçek		**CIN**	**Cancer**	**Normal**
	CIN	86	9	5
	Cancer	12	11	2
	Normal	9	2	40

Tablo 4.9' da belirtilen 3-Sınıflı karışıklık matrisinde ise gerçekte normal kategorisinde bulunan bazı verilerin CIN olarak sınıflandırıldığı görülmektedir. Cancer kategorisinde yanlış sınıflandırılma oranının arttığı gözlenmiştir. Yanlış sınıflandırılmanın nedeni olarak ise CIN sınıfı içinde bulunan üç alt kategoriden light_dysplastic kategorisinde bulunan bazı görüntülerin normal olarak moderate_dysplastic ve severe_dysplastic kategorisinde bulunan bazı görüntülerin Cancer olarak sınıflandırılması düşünülmektedir.

Oluşturulan veri seti üzerinde DVM, LR, ÇKA, RO NB, KNN ve derin öğrenme yöntemleri uygulanmıştır ve elde edilen sonuçlar Tablo 4.10'da gösterilmiştir. Değerlendirme işleminde ölçüt olarak doğruluk, hassasiyet, duyarlılık ve özgünlük değerleri kullanılmıştır.

Tablo 4.10 3-Sınıflı sınıflandırma işlemi sonuçları

	DVM	**RO**	**ÇKA**	**LR**	**KNN (k=9)**	**NB**
Doğruluk (%)	80.1	75	78	74.0	74	66.9
Hassasiyet (%)	80.1	77.2	77.8	74.4	74.4	66.9
Duyarlılık (%)	91.9	88.0	92.5	91.2	87.2	91
Özgünlük (%)	58.6	46.	54.8	55.1	46.1	50

3-Sınıf için derin öğrenme sonucu %56,82 oranında doğruluk elde edilmiştir.

3-Sınıflı sınıflandırma sonuçları incelendiğinde DVM yönteminin en iyi sonucu verdiği ve 2-Sınıflı sınıflandırma işleminde olduğu gibi NB' nin diğer yöntemlerin gerisinde kalmış olduğu gözlemlenmiştir. NB yönteminin başarı oranının düşük olması bu yöntemin istatistiksel veriler ile daha iyi sonuç vermesi olarak düşünülmüştür.

5
SONUÇ VE ÖNERİLER

Bu çalışma kapsamında pap-smear görüntüleri kullanılarak serviks hücrelerinin normal veya anormal olma durumu bulunmuştur. Normal ve anormal hücre sınıflandırmasında pap-smear görüntülerinin keskinliği artırılarak çekirdek ve sitoplazmanın ayrımı belirgin hale getirilmiştir. Keskinleştirilen görüntüde çekirdek ve sitoplazma kısımları ayrı olarak bölütlenmiştir. Çekirdek segmentasyonu için eşikleme ve K-means olmak üzere iki farklı yöntem kullanılmıştır. Sitoplazma segmentasyonu için eşikleme ve iki görüntünün farkını alma işlemleri uygulanarak gerçekleştirilmiştir. Bir hücrede rahim ağzı kanserini tespit edebilmek için görüntüde bulunan bazı özellikler belirleyici rol oynamaktadır. Bu özellikler ayrı ayrı bulunan çekirdek ve sitoplazma için çıkarılmıştır ve sınıflandırma kriteri olarak belirlenmiştir. Herlev veri setinde bulunan her görüntü için çekirdek ve sitoplazma segmentasyonu işlemlerinin ardından bu sınıflandırma kriterleri bulunmuş ve Tablo 4.2'de gösterilen veri setine kaydedilmiştir. Herlev veri seti iki sınıfın ardından CIN, Normal ve Cancer olmak üzere üç sınıfa ayrılmıştır ve Tablo 4.8'de gösterilen veri setine kaydedilmiştir. Oluşturulan veri setleri makine öğrenmesi yöntemlerinden DVM, LR, ÇKA, RO, NB, KNN ve Derin Yapay Sinir Ağları kullanılarak sınıflandırılmıştır. Sınıflandırma işlemi sonucu her bir yöntem ile elde edilen doğruluk değerleri karşılaştırılmıştır. 2-Sınıflı sınıflandırma başarısı olarak derin öğrenme yöntemi en iyi sonucu verirken 3-Sınıflı sınıflandırmada DVM yöntemi en iyi sonucu vermiştir. ÇKA yönteminde başarı oranının %90 olarak kalması veri sayısının yeterli miktarda olmaması olarak düşünülmüştür. Üç-sınıflı sınıflandırma işleminde daha düşük bir doğruluk değeri elde edilmesinin nedeni olarak CIN kategorisine ait verilerin Normal ve Cancer kategorilerine benzerlik göstermesi bu kategoride bulunan bazı verilerin Normal ve Cancer olarak yanlış sınıflandırılması olarak düşünülmüştür.

Önerilen renk-tabanlı segmentasyon ve matematiksel işlemler ile yapılan özellik çıkarımı yöntemleri doğru çekirdek- sitoplazma eşleşmesi yaptığı için ve hücrelerde bulunan fazla çekirdek sayılarını tespit edip anormal hücreleri görüntü üzerinde

bulduğu için birden fazla hücre içeren pap-smear test sonuçlarında patoloji uzmanlarına hastalığın teşhisinde katkı sağlayacaktır.

Oluşturulan veri seti genişletilerek DSIL, YSIL için daha belirleyici ve sayıca fazla kriterler ve daha fazla pap-smear görüntüsü eklenerek rahim ağzı kanserini daha erken ve daha başarılı sonuçlar ile tespit eden bilgisayar-temelli bir arayüz geliştirilerek çalışma tekrar yapılabilir.

KAYNAKÇA

[1] M.E. Plissiti, N. Christophoros, and A. Charchanti, "Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering", *IEEE Transactions on information technology in biomedicine*, vol. 15, no.2, pp. 233-241, 2010.

[2] P. Wang, L. Wang, Y. Li, Q. Song, S. Lv, X. Hu, "Automatic cell nuclei segmentation and classification of cervical Pap smear images", *Biomedical Signal Processing and Control*, vol. 48, pp. 93-103, 2019.

[3] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification", *Computers in Biology and Medicine*, vol. 39, no.1, pp. 69-78, 2009.

[4] A. GençTav, S. Aksoy, and S. Önder, "Unsupervised segmentation and classification of cervical cell images", Pattern recognition, vol. 45, no.12, pp. 4151-4168, 2012.

[5] MDE-lab, "MDE-lab downloads page,"2011. Accessed on: 30-Sep-2019. [Online]. Available: http://mde-lab.aegean.gr/downloads.

[6] H.A. Phoulady, "A framework for nucleus and overlapping cytoplasm segmentation in cervical cytology extended depth of field and volume images", *Computerized Medical Imaging and Graphics*, vol. 59, pp. 38-49, 2017.

[7] A.Kale, & S.Aksoy, "Segmentation of cervical cell images", In *2010 20th International Conference on Pattern Recognition*, pp. 2399-2402, IEEE, 2010.

[8] K.P. Win, Y. Kitjaidure, K. Hamamoto, T. Myo Aung, "Computer-Assisted Screening for Cervical Cancer Using Digital Image Processing of Pap Smear Images", *Appl. Sci.*, vol. 10, no.5, pp.1800, 2020.

[9] W. William, A. Ware, A.H. Basaza-Ejiri, & J. Obungoloch, "A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images", *Biomedical engineering online*, vol.18, no.1, pp.16, 2019.

[10] B. Taha, J. Dias, & N. Werghi, "Classification of cervical-cancer using pap-smear images: a convolutional neural network approach", In *Annual Conference on Medical Image Understanding and Analysis, Springer, Cham, Jul. 2017*, pp. 261-272.

[11] V. Vapnik, *The nature of statistical learning theory*. New York: Springer-Verlag, 1995.

[12] M. Pal, "Random forest classifier for remote sensing classification", *International Journal of Remote Sensing*, vol.26, no. 1, pp.217–222, 2005.

[13] M.W. Gardner, & S.R. Dorling, "Artificial neural networks (the multilayer perceptron) —a review of applications in the atmospheric sciences", *Atmospheric environment*, vol. 32, no.14-15, pp. 2627-2636, 1998.

[14] Y. Liao, V. R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection", *Computers & security*, vol. 21, no.5, pp. 439-448, 2002.

[15] M. M. Saritas, A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification", *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88-91, 2019.

[16] S.K. Shevade, S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression", *Bioinformatics*, vol. 19, no. 17, pp. 2246-2253, 2003.

[17] Y. LeCun, Y. Bengio, & G. Hinton, "Deep learning", *Nature*, vol. 521no.7553, pp. 436-444, 2015.

[18] Cancer Today, "Global Cancer Observatory homepage", 2018. Accessed on: Feb.2, 2020. [Online]. Available: http://gco.iarc.fr/today/online-analysis-table.

[19] DB. Cooper, C.E. McCathran, "Cervical Dysplasia". StatPearls: StatPearls Publishing, 2019. Accessed on: Oct. 20, 2019. [Online].

[20] RJ. Kurman, D. Solomon, *The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnoses*. New York: Springer- Verlag, 1994.

[21] S.E. Waggoner, "Cervical cancer", *The Lancet*, vol. 361, no. 9376, pp. 2217-2225, 2003.

[22] Y. Marinakis, M. Magdalene, and D. Georgios, "Particle swarm optimization for pap-smear diagnosis.", *Expert Systems with Applications*, vol. 35, no.4 pp.1645-1656, 2008.

[23] R.C. Gonzalez & R.E. Woods, *Digital image processing (preview)*. New Jersey: Prentice Hall, 2002.

[24] M. Sonka, V. Hlavac, & R. Boyle, *Image processing, analysis, and machine vision*. Stanford: Cengage Learning, 2014.

[25] N. Nill, "A visual model weighted cosine transform for image compression and quality assessment", *IEEE Transactions on communications*, vol.33, no. 6, pp. 551-557, 1985.

[26] Z. Karhan and B. Ergen, "Classification of face images using discrete cosine transform," *2013 21st Signal Processing and Communications Applications Conference (SIU)*, Haspolat, 2013, pp. 1-4, doi: 10.1109/SIU.2013.6531364.

[27] M. Uyar, S. Yıldırım, & M.T. Gençoğlu, "Güç kalitesi bozulmalarının sınıflandırılmasında dalgacık dönüşümüyle enerji dağılımına dayalı özelliklerin incelenmesi", *Elektrik-Elektronik Bilgisayar Biyomedikal*, 2007.

[28] H. Demirel, G. Anbarjafari, "Discrete Wavelet Transform-Based Satellite Image Resolution Enhancement", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 1997-2004, 2011.

[29] H. Demirel, G. Anbarjafari, "Image resolution enhancement by using discrete and stationary wavelet decomposition*", IEEE transactions on image processing*, vol. 20, no. 5, pp. 1458-1460, 2010.

[30] N.R. Pal, & S.K. Pal, "A review on image segmentation techniques", *Pattern recognition*, vol. 26, no. 9, pp.1277-1294,1993.

[31] K.S. Fu, and J. K. Mui. "A survey on image segmentation", *Pattern recognition*, vol. 13, no.1, pp.3-16,1981.

[32] R. M. Haralick, L. G. Shapiro, "Image segmentation techniques", *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 1, pp. 100-132, 1985.

[33] E. Anjna, E. Rajandeep, Kaur, "Review of image segmentation technique", *International Journal of Advanced Research in Computer Science*, vol. 8, no. 4, 2017.

[34] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, "Constrained k-means clustering with background knowledge", *Icml*, vol.1, pp. 577-584, 2001.

[35] Yuheng, Song, and Yan Hao. "Image segmentation algorithms overview.", arXiv preprint arXiv:1707.02051 (2017).]

[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[37] R.M. Hristev, *The ANN book*.1998.

[38] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, & L.D. Jackel," Backpropagation applied to handwritten zip code recognition", *Neural computation*, vol.1, no.4, pp. 541-551,1989.

[39] D. CireşAn, U. Meier, J. Masci, & J. Schmidhuber," Multi-column deep neural network for traffic sign classification", *Neural networks*, vol.32, pp.333-338, 2012.

[40] S.J. Pan, & Q. Yang," A survey on transfer learning", *IEEE Transactions on knowledge and data engineering*, vol. 22, no.10, pp.1345-1359,2009.

[41] "ImageNet Web page", 2016. Accessed on: Feb. 3, 2020. [Online]. Available: http://www.image-net.org/

[42] D. Sarkar, R. Bali, & T. Ghosh, *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Birmingham: Packt Publishing Ltd., 2018.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov,... & A. Rabinovich, "Going deeper with convolutions", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[44] K. Simonyan, & A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, & Z. Wojna, "Rethinking the inception architecture for computer vision", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[46] C.F. Lin, & S.D. Wang, "Fuzzy support vector machines*", IEEE transactions on neural networks*, vol. 13, no.2, pp.464-471, 2002.

[47] I. Saini, D. Singh, & A.Khosla, "QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases*", Journal of advanced research*, vol.4,no.4,pp. 331-344, 2013.

[48] E. Taşcı, & A. Onan, "K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi", *Akademik Bilişim*, 2016.

[49] R. O. Duda, P. E. Hart, & D. G. Stork, *Pattern classification*. New York: John Wiley & Sons, 2012.

[50] L. Breiman, "Random Forests", *Machine Learning*, vol.45, no. 1, pp.5–32, 2001.

[51] V.F.Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, & J.P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93-104, 2012.

[52] D.W. Hosmer Jr, S. Lemeshow, & R.X. Sturdivant, *Applied logistic regression* (Vol. 398). New York: John Wiley & Sons, 2013.

[53] S. Swaminathan, (2018). Logistic Regression—Detailed Overview. *Towards Data Science*.

[54] J. Jantzen, J. Norup, Jonas, G. Dounias, B. Bjerregaard, "Pap-smear Benchmark Data For Pattern Classification", *Nature Inspired Smart Information Systems (NiSIS)*, pp. 1-9, 2005.

[55] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A. Charchanti, SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images, *IEEE International Conference on Image Processing (ICIP) 2018*, Athens, Greece, 7-10 October 2018.

[56] E.S. Cibas, B.S. Ducatman, *Cytology E-Book: Diagnostic principles and clinical correlates*, Elsevier Health Sciences, 2013.

TEZDEN ÜRETİLMİŞ YAYINLAR

İletişim Bilgisi: akyolfbetul@gmail.com

Makaleler

1. F.B. Akyol,O.Altun, "Detection of Cervix Cancer from Pap-smear Images", *Sakarya University Journal of Computer and Information Sciences,* vol.3, no.2, pp.99-111,2020.