

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ TEKNİKLERİYLE MİKRODALGA MEME
KANSERİNİN TESPİTİ VE KONUMUNUN BELİRLENMESİ

Elif TÖLÜK

YÜKSEK LİSANS TEZİ

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Danışman

Doç. Dr. Birol ASLANYÜREK

Mart, 2022

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**MAKİNE ÖĞRENMESİ TEKNİKLERİYLE MİKRODALGA MEME
KANSERİNİN TESPİTİ VE KONUMUNUN BELİRLENMESİ**

Elif TÖLÜK tarafından hazırlanan tez çalışması 07.03.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Birol ASLANYÜREK
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Birol ASLANYÜREK, Danışman
Yıldız Teknik Üniversitesi

Prof. Dr. Hülya ŞAHİNTÜRK, Üye
Yıldız Teknik Üniversitesi

Dr. Öğr. Üye. Mustafa Zahid GÜRBÜZ, Üye
Doğuş Üniversitesi

Danışmanım Doç. Dr. Birol ASLANYÜREK sorumluluğunda tarafımca hazırlanan Makine Öğrenmesi Teknikleriyle Mikrodalga Meme Kanserinin Tespiti ve Konumunun Belirlenmesi 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.

Elif TÖLÜK

İmza



Aileme

TEŞEKKÜR

Yüksek Lisans tez çalışmalarım süresince bilgi ve birikimlerini benimle paylaşan ve bana yol gösteren değerli hocam ve tez danışmanım Doç.Dr. Birol ASLANYÜREK'e,

Ayrıca, bütün eğitim ve öğretim hayatım boyunca her koşulda yanımda olan ve desteklerini hiçbir zaman esirgemeyen çok sevdiğim aileme,

Sonsuz teşekkürler...

Elif TÖLÜK



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 Orijinal Katkı.....	4
2 VERİ MADENCİLİĞİ	5
2.1 Veri Madenciliğinin Tanımı	5
2.2 Veri Madenciliğinin Tarihçesi.....	5
2.3 Veri Madenciliğinin Faydaları	8
2.4 Veri Madenciliğinin Kullanıldığı Alanlar.....	8
2.5 Veri Madenciliği ile İlişkili Disiplinler.....	9
2.6 Veri Madenciliğinin Tıptaki Önemi ve Uygulamaları	10
2.7 Veri Madenciliğini Etkileyen Faktörler	10
2.8 Veri Madenciliğinde Karşılaşılan Sorunlar.....	11
2.9 Veri Madenciliği Süreci	11
2.9.1 Problemin Tanımlanması.....	12
2.9.2 Verilerin Hazırlanması.....	12
2.9.3 Modelin Kurulması ve Değerlendirilmesi	13
2.9.4 Modelin Kullanılması.....	14
2.9.5 Modelin İzlenmesi	14
2.10 Veri Madenciliği Modelleri	14
2.10.1 Tahmin Edici Modeller	15
2.10.2 Tanımlayıcı Modeller.....	21
3 YAPILAN ÇALIŞMA	22
3.1 Problemin Tanımlanması	22
3.2 Verilerin Hazırlanması	23

3.2.1 Elektromagnetik Alan Verilerinin Sentetik Olarak Üretilmesi	24
3.2.2 Verilerin Dönüştürülmesi.....	25
3.3 Modellerin Eğitilmesi ve Test Edilmesi	26
3.4 Modelin Oluşturulması.....	26
3.4.1 Karar Ağaçları.....	27
3.4.2 Diskriminant Analizi	28
3.4.3 Destek Vektör Makinesi	28
3.4.4 Naive Bayes	30
3.4.5 Topluluk Sınıflandırıcıları.....	30
3.4.6 Yapay Sinir Ağları.....	32
3.4.7 K-En Yakın Komşu Algoritması	33
3.5 Modellerin Değerlendirilmesi	34
4 SONUÇ VE ÖNERİLER	37
4.1 Sonuç ve Öneriler	37
KAYNAKÇA	39
TEZDEN ÜRETİLMİŞ YAYINLAR	42

SİMGE LİSTESİ

λ	Dalga boyu
m^2	Metrekare
cm	Santimetre
%	Yüzde



KISALTMA LİSTESİ

ABD	Amerika Birleşik Devletleri
DVM	Destek Vektör Makinesi
ENIAC	Electrical Numerical Integrator and Calculator
GHz	Gigahertz
KNN	K-Nearest Neighbors
LDA	Doğrusal Ayrımcılık Analizi
MR	Manyetik Rezonans
SVM	Support Vector Machine
TBA	Temel Bileşen Analizi
UWB	Ultra Wideband

ŞEKİL LİSTESİ

Şekil 2. 1	Veri Madenciliği tarihsel süreci [29]	7
Şekil 2. 2	Veri madenciliği ve disiplinler [29]	9
Şekil 2. 3	Bilgi keşfi sürecinde veri madenciliği [29]	12
Şekil 3. 1	İki boyutlu bir meme modelinin görelî dielektrik katsayısı dağılımı ...	23
Şekil 3. 2	Tümör içeren meme modelinin görelî dielektrik katsayısı dağılımı.....	24
Şekil 3. 3	Çizgisel kaynaklar ile aydınlatma ve ölçüm noktaları.....	25
Şekil 3. 4	Dönüştürülmüş veri seti	26



TABLO LİSTESİ

Tablo 2. 1	Veri Madenciliği gelişim süreci [30].....	7
Tablo 2. 2	Veri madenciliği sürecinde izlenen adımlar.....	11
Tablo 2. 3	Sınıflandırma ve regresyon kullanılan başlıca yöntemler	15
Tablo 3. 1	Karar ağaçları sınıflandırma sonuçları.....	27
Tablo 3. 2	Diskriminant analizi sınıflandırma sonuçları.....	28
Tablo 3. 3	Destek vektör makinesi sınıflandırma sonuçları	29
Tablo 3. 4	Naïve Bayes sınıflandırma sonuçları.....	30
Tablo 3. 5	Topluluk sınıflandırıcıları için sınıflandırma sonuçları	31
Tablo 3. 6	Yapay sinir ağları sınıflandırma sonuçları	32
Tablo 3. 7	K-En Yakın Komşu algoritması sonuçları.....	33
Tablo 3. 8	Medium Gaussian SVM karışıklık matrisi sonuçları.....	35
Tablo 3. 9	Coarse Gaussian SVM karışıklık matrisi sonuçları	35
Tablo 3. 10	Quadratic SVM karışıklık matrisi sonuçları	35
Tablo 3. 11	Quadratic Discriminant karışıklık matrisi sonuçları.....	36
Tablo 3. 12	Coarse KNN karışıklık matrisi sonuçları	36

Makine Öğrenmesi Teknikleriyle Mikrodalga Meme Kanserinin Tespiti ve Konumunun Belirlenmesi

Elif TÖLÜK

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Birol ASLANYÜREK

Meme kanseri diğer kanser türlerine göre kadınlarda en yaygın olarak görülen türdür. Erken teşhisi hastanın hayatının kurtulmasında önemli rol oynamaktadır. Meme kanserinin konumunun belirlenmesi ve erken teşhisi için geçmişte ve günümüzde çok sayıda çalışma yapılmaktadır. Elektromanyetik dalgaların kullanımı, uygulanma rahatlığı, tümörlere duyarlılık ve mikrodalga frekanslarında iyonlaştırıcı olmamaları nedeniyle, meme kanseri tespiti için umut verici bir tekniktir. Mikrodalga tomografisi en sık kullanılan mikrodalga meme görüntüleme yaklaşımlarından biridir ve bu yaklaşımda genellikle bütün meme görüntülenmeye çalışılır. Bu durumda, yüksek düzeyde matematiksel ve hesaplama karmaşıklığı olan bir ters saçılma problemi ile uğraşmak gerekir. Bu zorluklardan kaçınmak adına, bir tümörün varlığının tespiti veya konumunun belirlenmesi için sınıflandırma algoritmaları kullanılabilir.

Bu çalışmanın amacı, mikrodalgalar aracılığıyla meme kanserinin tespiti ve konumunun belirlenmesi için çeşitli makine öğrenmesi tekniklerini kullanarak inceleme yapmaktır. Bu yaklaşımda, meme modeli bir dizi verici anten aracılığı ile tek frekanslı zaman harmonik dalgalar ile ayrı ayrı aydınlatılır ve meme

evresinde bulunan alıcı antenler ile saılan alan llr. alıřmada, iki boyutlu gereki sayısal meme modelleri kullanılarak, moment yntemi aracılıėıyla dz saılma problemi zlmř ve saılan alanlar sayısal olarak hesaplanmıřtır. Farklı boyut ve konumlardaki tmrleri ieren veya tmr iermeyen yani saėlıklı meme modelleri kullanılarak sentetik olarak retilen saılan alanlar eėitim ve test veri setlerini oluřturmaktadır. alıřmada, veriler %70 oranında eėitim, %30 oranında test amalı kullanılmıřtır. Elde edilen sonular; destek vektr makinesi, naive bayes, karar aėaları, topluluk sınıflandırıcılar, k-en yakın komřu algoritması, diskriminant analizi ve yapay sinir aėları makine ėrenmesi algoritmalarının, tmrlerin hem varlıėının tespiti hem de konumlarının belirlenmesi aısından, farklı meme modelleri iin kullanılabileceėini gstermektedir.

Anahtar Kelimeler: Mikrodalga meme kanseri tespiti, meme tmrlerinin konum tespiti, makine ėrenmesi, sınıflandırma.

Detection and Localization of Microwave Breast Cancer by Machine Learning Techniques

Elif TÖLÜK

Department of Mathematics Engineering

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Birol ASLANYÜREK

Breast cancer is most common in women, rather than other types of cancer. Early diagnosis plays an important role in saving the patient's life. Today, many studies are carried out to determine the location and early diagnosis of breast cancer. The use of electromagnetic waves is a promising technique for breast cancer detection due to its ease of application, sensitivity to tumors, and non-ionizing at microwave frequencies. Microwave tomography is one of the most used microwave breast imaging approaches, and in this approach, the whole breast is usually tried to be imaged. In this case, it is necessary to deal with an inverse scattering problem with high mathematical and computational complexity. To avoid these difficulties, classification algorithms can be used to detect the presence or localization of a tumor.

The aim of this study is to examine using various machine-learning techniques to detect and locate breast cancer through microwaves. In this approach, the breast model is illuminated separately with single frequency time harmonic waves via a series of transmit antennas and the scattered field is measured with the receiving antennas around the breast. In the study, using two-dimensional realistic numerical breast models, the direct scattering problem was solved by the method

of moment and the scattering fields were calculated numerically. Scattered fields produced synthetically using breast models with or without tumors of different sizes and locations constitute the training and test datasets. In the study, 70% of the data were used for training and 30% for testing. Obtained results show that the machine learning algorithms such as support vector machine, naive bayes, decision trees, ensemble classifiers, k-nearest neighbor algorithm, discriminant analysis and artificial neural networks can be used for detection and localization of tumors in different breast models.

Keywords: Microwave breast cancer detection, localization of breast tumors, machine learning, classification.



1.1 Literatür Özeti

Meme kanseri günümüzde kadın sağlığı için en büyük tehditlerden biri olup, her yıl birçok kadın meme kanserinden dolayı hayatını kaybetmektedir. Kanserli dokunun erkenden teşhis edilmesi hasta kişinin hayatının kurtulmasında kritik rol oynamaktadır. Bu sebeple, meme kanserinin doğru olarak tespit edilmesine yönelik çeşitli alanlarda çok sayıda çalışma mevcuttur ve bu konuda çeşitli çalışmalar yapılmaktadır.

Mamografi, meme içerisinde yer alan tümörün tespitinde en sık kullanılan metottur. Fakat; mamografi yöntemi, tümörlü doku küçük ise veya glandular dokunun yoğun olduğu bölgelere yakın olduğu zaman başarılı olamamakla birlikte, bu yöntemde kullanılan X ışınları şimdilerde zararlılığı azaltılmasına rağmen insan sağlığı için tehlike arz etmektedir.

Meme kanserinin tespitinde en sık kullanılan tekniklerden biri de magnetik rezonans (MR) görüntülemedir. MR görüntüleme X ışınlarından daha az zararlı olmasına rağmen, bu tekniğin kullanımı kalp pili taşıyan veya erken hamilelik dönemindeki kişilerde risk teşkil etmektedir.

Yukarıda belirtilen dezavantajlar yanında, mamografi ve MR cihazlarının yüksek maliyeti düşünüldüğünde, son yıllarda meme kanserinin tespiti için çeşitli mikrodalga tekniklerinin kullanımı öne çıkmaktadır. Mikrodalga frekanslarında elektromagnetik malzeme özellikleri açısından sağlıklı doku ve kanserli doku arasında farklılıklar vardır [1-2]. Dokuların elektromagnetik malzeme özellikleri bakımından farklılık göstermesi mikrodalga tabanlı tekniklerin gelişiminde büyük rol oynamıştır. Ultra geniş bant radar tabanlı teknikler [3-6], hibrid mikrodalga akustik görüntüleme [7-8], mikrodalga tomografisi [9-10] meme içerisindeki kanserli dokuların tespitinde kullanılan temel tekniklerdendir. Hibrid mikrodalga – indüklenmiş akustik görüntüleme metodunda ilk olarak tümörlü doku ısıtılır,

sonrasında ısınan dokunun genişlemesi sonucunda oluşan basınç dalgalarının ultrason dönüştürücüleri ile belirlenmesine dayanmaktadır. Radar tabanlı tekniklerde, ultra geniş-bant mikrodalga sinyalleri ile meme aydınlatılır ve geriye saçılan enerji dağılımı ölçülür. Bu ölçüm kullanılarak tümörlerin varlığı ve konumu tespit edilmeye çalışılır. Mikrodalga tomografisinde ise, etrafına yerleştirilmiş verici mikrodalga antenleriyle meme aydınlatılır ve yine memenin etrafına yerleştirilmiş mikrodalga alıcı antenlerle memeden saçılan elektromagnetik alan ölçülür. Son olarak, çeşitli metotlar kullanılarak, bütün meme dokusunun ve meme içerisinde var ise tümörlü dokuların bilgisini taşıyan saçılan alandan memenin elektromagnetik profili görüntülenir. Mikrodalga görüntülenmesinde kullanılan geleneksel yöntemlerde yüksek düzeyde matematiksel ve hesaplama karmaşıklığı olan bir ters saçılma problemi ile uğraşmak gerekir. Bu zorluklardan kaçınmak adına, tüm profilin görüntülenmesi yerine tümör tipinin [11-16] ve tümör gelişiminin [17] belirlenmesi, tümörün varlığının tespiti [18-22] ve konumunun belirlenmesi [23] gibi amaçlar için sınıflandırma algoritmaları kullanılmıştır.

İyi huylu ve kötü huylu tümörler şekilsel farklılıklar içermektedir. Bu nedenle daha önceki dönemlerde yapılan çalışmalarda şekilsel özelliklerin farklılıkları dikkate alınarak, tümör tipinin belirlenmesi için radar tabanlı pek çok yöntem önerilmiştir [11-15]. Bu çalışmalarda, tümör tipinin belirlenmesinde destek vektör makineleri (DVM), lineer diskriminant analizi (LDA) ve kuadratik diskriminant analizi (QDA) gibi sınıflandırma yöntemleri kullanılmıştır. Tümör tipinin belirlenmesi amacıyla kullanılan yöntemlerden birisi de derin öğrenme algoritmasıdır [16]. Bu çalışmada sonlu farklar yöntemi kullanılarak, içine gömülmüş tümörlerle birlikte, yağ dokusundan oluşan homojen bir meme modelinden saçılan alan hesaplanmış, ardından elde edilen sonuçlar tümör tipinin sınıflandırılmasında kullanılmıştır.

[17]'de kanserli doku içeren meme modeli ultra geniş bant mikrodalga radarıyla ardışık olarak ölçülmüştür ve kanserli dokuda bir değişim olup olmadığını tespit eden bir yöntem önerilmiştir.

Tümör tipinin belirlenmesi dışında, literatürde tümörün varlığının tespit edilmesine dönük yaklaşımlar da ele alınmıştır. [18]'de tümörün tespiti için DVM

ve LDA yöntemlerine dayanan bir deneysel mikrodalga meme görüntüleme sistemi önerilmiştir. Bu çalışmada yarı küresel bir meme modeli kullanılmıştır ve tümör ile sağlıklı meme dokusu arasındaki elektromagnetik malzeme özellikleri arasındaki kontrast oldukça fazladır. [20]'de ise tümörün tespiti için LDA ve QDA yöntemlerini kullanan radar tabanlı bir yaklaşım önerilmiştir. Benzer şekilde, bu çalışmada da tümör ile sağlıklı meme dokusu arasındaki elektromagnetik malzeme özellikleri arasındaki karışıklık olması gerekenden [1-2] yüksek olarak belirlenmiştir. Radar teknolojisiyle tümör varlığını tespit etmek için sınıflandırma algoritmalarının uygulandığı daha gerçekçi meme modelleri de bazı çalışmalarda kullanılmıştır [19, 21]. Sınıflandırma işlemi için [19] DVM ve LDA kullanırken, [21] sadece DVM kullanmıştır. Ayrıca, gerçekçi meme modellerinde tümör varlığının sınıflandırma algoritmaları ile tespiti amacıyla mikrodalga tomografisini daha önce kullanan bir yaklaşım [22]'de sunulmuştur. Bu yaklaşımda, moment yöntemiyle düz saçılma problemi çözülmüş, ardından üretilen sentetik veri kullanılarak Naive Bayes, DVM ve C4.5 (karar ağacı) gibi çeşitli makine öğrenmesi yöntemleriyle tümörün varlığı tespit edilmeye çalışılmaktadır.

Tümör konumunu tespit etmek için sınıflandırma algoritmaları kullanılan çalışma sayısı ise son derece azdır. [23]'da mikrodalga meme kanseri tespiti ve konumunun belirlenmesi için SVM tabanlı bir yaklaşım önerilmiştir. Bu çalışmada meme modelleri homojen olarak seçilmiştir.

1.2 Tezin Amacı

Mikrodalga tomografisi en sık kullanılan mikrodalga meme görüntüleme yaklaşımlarından biridir ve bu yaklaşımda genellikle bütün meme görüntülenmeye çalışılır. Bununla birlikte, bu durum, yüksek düzeyde matematiksel ve hesaplama karmaşıklığı olan bir ters saçılma problemi ile çalışmayı gerektirir. Bu zorluklardan kaçınmak adına, bir tümörün varlığının tespiti veya konumunun belirlenmesi için sınıflandırma algoritmaları kullanılabilir. Bu tezin amacı, çeşitli makine öğrenmesi tekniklerini kullanarak mikrodalgalar aracılığıyla gerçekçi bir meme modelinde tümör varlığının tespiti ve tümör konumunun belirlenmesidir.

1.3 Orijinal Katkı

Literatürde meme kanserinin tespiti için mikrodalga yaklaşımına dayanan çok sayıda çalışma sunulmuştur. Bunların çok büyük kısmı karmaşık matematiksel modellerle uğraşmayı gerektiren ters saçılma problemlerinin çözümüne dönüktür. Son yıllarda ise bu problemin çözümünde makine öğrenmesi tekniklerinin kullanımı söz konusu olmaya başlamıştır.

Bu çalışmaların bir kısmı tümör tiplerinin (iyi huylu-kötü huylu) belirlenmesine yönelik iken, bir kısmı da tümör varlığını tespit etmeye çalışmaktadır. Sınıflandırma algoritmaları kullanarak meme tümörünün konumunu tespit etmeye çalışan tek yaklaşım ise [23]'de önerilmiştir. DVM yöntemini kullanan bu çalışmada gerçekçi olmayan homojen bir meme modeli kullanmıştır. Bu tezde ise, gerçekçi meme modelleri üzerinde tümör varlığı ve konumunu tespit eden bir mikrodalga yaklaşımı sunulmuştur.

Bu tez; [23]'de sunulan tek frekanslı zaman harmonik dalgalar ile gerçekçi meme modelleri içinde tümörün varlığının çok sayıda veri madenciliği algoritmasıyla incelendiği yaklaşımın geliştirilmiş bir halidir. [23]'de verilen çalışmada sadece tümörün varlığı incelenirken, bu çalışmada tümörün varlığı yanında tümörün konumunun da belirlenmesine dönük bir inceleme yapılmıştır. Elde edilen sonuçlar önerilen yaklaşımın tümör konumunun belirlenmesi amacıyla kullanılabileceğini göstermektedir.

2.1 Veri Madenciliğinin Tanımı

Teknolojinin hızlı gelişimi nedeniyle sürekli çoğalan ve işlenmediğinde mükerrer veriler oluşmaktadır. Bu veri yığınlarına ulaşılması için kullanılan temel yöntem veri madenciliği adı verilir.

Geçmişten bugüne kadar çeşitli kaynaklarda veri madenciliğinin birçok tanımıyla karşılaşmıştır. Bazı veri madenciliği tanımları aşağıdaki gibidir:

- Jacobs'a göre veri madenciliği, işlenmemiş bilgiyi meydana getiren veri analiz akışı olarak tanımlanmıştır [24].
- Veri madenciliği, devasa veri yığınları içerisinde tahminlerde bulunmayı sağlayacak ilişkilerin bilgisayar üzerinden araştırılmasıdır [25].
- Hand'a göre veri madenciliği, istatistik ve makine öğrenmesi gibi çeşitli disiplinler ile ilişkili olup, veri tabanlarında öngörülemeyen bağlantıların analizi olarak tanımlanmıştır [26].
- Kitle ve Wang'a göre veri madenciliği, fazlaca öngörülebilir anahtar değişkenlerin fazlaca potansiyel değişkenden ayrılmasının sağlanması olarak tanımlanmıştır [27].
- Shalvi ve DeClaris'e göre veri madenciliği; belirli alanda toplanan veriler arasındaki ilişkilerin bulunması olarak tanımlanır. [28]

Yukarıda belirtilen tanımlar baz alınarak veri madenciliği, devasa nicelikte bilginin depo edildiği veri tabanlarından, amaç üzerinden, gelecek ile ilgili öngörüler yapılmasını sağlayan, gerekli veriye erişme ve veriyi işleme olarak tanımlanabilir.

2.2 Veri Madenciliğinin Tarihçesi

Veri madenciliğinin kaynağı, ilk sayısal bilgisayar olarak bilinen ENIAC'a dayanmaktadır. ENIAC, 1946 yılında II. Dünya Savaşı'nda ABD ordusu için

geliştirilmiştir. İlk bilgisayar 30 ton ağırlığı ile 170 m²lik bir alanı kapsamakta olup, bilgisayarın 70 yıl içinde geçirmiş olduğu evrim günümüz masaüstü bilgisayarlarına bakarak anlaşılmaktadır. Bilgisayarın geçirdiği evrim, belirli proses ve koşullar altında gerçekleşmiştir. Donanımsal olarak hazırlanan bilgisayarlar, yazılımlar aracılığı ile var olup, bu bilgisayarların kullanıcılara ulaştırılması evrim dönüşümünün kilit noktasıdır.

İlgili kişilerin tasarlayıp, gelişimlerini ilerlettiği ürünler, kullanıcıların talepleri baz alınarak, süre içerisinde biçimlendirilerek şimdiki halini almıştır.

Döngü, donanımın geliştirilmesinin akabinde yazılımın ilgili donanıma entegre edilip, kullanıcıya ulaştırılması ile başlamaktadır. Kullanıcının gereksinimleri, talepleri doğrultusunda güncel yazılımlar geliştirilir. Geliştirilen bu güncel yazılımların doğru bir şekilde çalışabilmesi için donanım güncellemesi yapılması gerekir. Donanım güncellemesinin yapılabilmesi için de donanım uzmanlarına uyarıda bulunulur. Güncellenen bilgisayarlar tekrardan alıcılara iletilir.

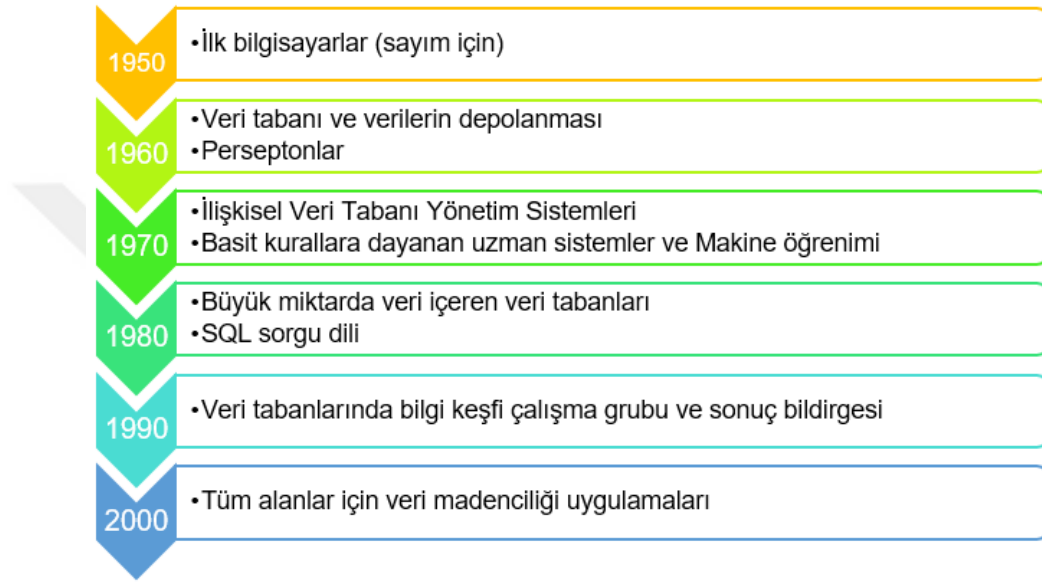
Verilerin depolanması ile bilgisayarın aktif kullanımına geçilmiştir. Bilgisayarlar, en başta kompleks hesaplamaların yapılması amacı ile geliştirilmiştir. Sonrasında, kullanıcı gereksinimleri aracılığıyla veri kayıtlarının tutulması için kullanılmıştır ve veri tabanları oluşmuştur.

Veri tabanları gelişmektedir. Dolayısıyla donanımsal olarak verilerin depo edileceği yerlerin de büyümesini gerektirmiştir. Böylece veri ambarı kavramı gün yüzüne çıkmıştır. Saklanması talep edilen veriler, tekrardan kullanılmak üzere saklanmaya başlamıştır. Gün geçtikçe büyüyen veri tabanlarının; yönetimi, planlaması ve düzenlemesi de zor olmaya başlamıştır. Bu durumda veri modellemesi tanımı gün yüzüne çıkmıştır.

Gereksinimler doğrultusunda biçimlenen veri tabanları ve veri modelleme türleri hızlı bir şekilde yaygınlaşırken, donanımlar da bu prosese uyum sağladılar. Şimdilerde milyarlarca bit veri küçük belleklerde tutulabilecek hale gelmiştir.

Verilerin gizlenmesi, biçimlendirilmesi, planlanması dışarıdan bir problem gibi görünmese de fazla veri ile talep edilen neticeye ulaşması bir problem olmuştur. Veri madenciliği, 1960'lı senelerde, bilgisayarların veri analiz sorunlarını çözümleyebilmek amacıyla kullanılmaya başlanması ile ortaya çıkmıştır. 1960'lı

yıllarda, bilgisayar kullanılarak talep edilen verilere erişilebileceği ortaya çıkmıştır. Bu araştırma veri madenciliği yerine geçmişte veri taraması, veri yakalanması olarak adlandırılmıştır. “Veri madenciliği” adı, ilk olarak 1990’lı yıllarda bilgisayar mühendisleri tarafından kullanılmıştır. Bilgisayar mühendislerinin hedefi, basmakalıp istatistiksel metotları kullanmadan veri analizinin algoritmik bilgisayar modülleri tarafından yapılmasına olanak sağlamaktır.



Şekil 2. 1 Veri Madenciliği tarihsel süreci [29]

Tablo 2. 1 Veri Madenciliği gelişim süreci [30]

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960’lar)	“Benim toplam karım geçen 5 yılda ne kadardı?”	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye dönük, statik veri dağılımı
Veri Erişimi (1980’ler)	“İngiltere’de geçen mart ayında birim satışları ne kadardı?”	İlişkisel Veri tabanları, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990’lar)	“İngiltere’de geçen mart ayında birim satışları ne kadardı?”	OLAP, Çok boyutlu veri tabanı sistemleri, Veri Ambarı	Pilot, Comshare, Arbor, Cognos, MicroStrategy	Çoklu düzeylerde geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	“Gelecek ay Boston’daki birim satışlar muhtemelen ne olabilir, neden?”	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veri tabanlar	Pilot, Lockheed, IBM, SGL, SPSS, SAS, Microsoft vs.	Geleceğe dönük, proaktif enformasyon dağıtımı

2.3 Veri Madenciliğinin Faydaları

Bu yöntem, çekişmenin oldukça kuvvetlendiği piyasalarda, şirketler için bazı değerlerin yönetilmesinde devasa rol oynamaktadır. Müşteri bilgileri bu değerlerin en başındadır.

Veri madenciliğinin karar verici için muhtemel faydaları aşağıda verildiği gibi sıralanabilir:

- Karar verici tarafından mevcut müşterilerin daha iyi bir şekilde tanınmasını sağlayabilir.
- Laboratuvar ya da bilgisayar ortamında sistemlerin simülasyonu ve incelenmesi aşamasında ortaya çıkan fazla nicelikte bilimsel veriler anlamlandırılabilir.
- Sağlık alanında tarama testlerinden ortaya çıkan verileri kullanarak türlü kanserlerin başlangıç teşhisi, kalp verilerini kullanarak kalp krizi riskinin belirlenmesi, acil servislerde hasta belirtileri baz alınarak risk ve öncelikler belirlenebilir.
- İnternet sunucularının ya da çevrimiçi servisten kullanıcı erişim desenlerinin incelenip bulunabilir. Belgeler arasında manuel bir işlem yapılmadan benzeşmeler belirlenebilir.

2.4 Veri Madenciliğinin Kullanıldığı Alanlar

Veri madenciliği, büyük ölçüde verinin bulunduğu yerlerde kullanılabilir. Günümüzde karar verme aşamasında gereksinim duyulan çeşitli alanlarda veri madenciliği uygulamaları çoğunlukla kullanılmaktadır. Örnek olarak, sağlık, bilim ve mühendislik, şifreleme, endüstri, istihbarat, sigortacılık, borsa, perakendecilik, telekomünikasyon, e-ticaret, pazarlama, biyoloji, bankacılık, genetik vb. birçok alanda başarılı uygulamaları görülmektedir. İncelemeler sonucunda veri madenciliğinin en fazla kullanıldığı alanlar tıp, genetik ve biyoloji olduğu bilinmektedir.

Veri madenciliğinin kullanıldığı alanların bazıları aşağıdaki gibi açıklanabilir:

- **Pazarlama:** Satış öngörülerinde ve mevcut müşterilerin elde tutulup, yeni müşterilerin kazanılmasında veri madenciliği kullanılabilir.

- **E-Ticaret:** Uygulamaların yönetimi, web sayfalarına yapılan saldırıların çözümlenmesinde, kullanıcıların taleplerine göre web sitelerinin yenilenmesinde veri madenciliği kullanılabilir.
- **Bankacılık:** Kredi isteklerinin değerlendirilmesi, kredi kartı kullanıcılarının harcamalarına bağlı olarak müşteri gruplandırılması ve kredi kartı dolandırıcılıklarının tespit edilmesinde veri madenciliği kullanılabilir.
- **Sigortacılık:** Risk bulunan müşteri gruplarının tespiti ve sigorta dolandırıcılarının teşhis edilmesinde veri madenciliği kullanılabilir.
- **Tıbbi Araştırmalar:** Hastalıkların teşhis edilmesi, protein analizi, hastalık şemalarının hazırlanmasında veri madenciliği kullanılabilir.

2.5 Veri Madenciliği ile İlişkili Disiplinler

Veri madenciliği, farklı bilim dallarından yararlanan bir yaklaşım sunar. Bu yaklaşım; Şekil 2.2’de verilen farklı disiplinlerden yararlanır. Veri madenciliğinde farklı disiplinlerde kullanılmakta olan tekniklerden faydalanılmaktadır. Veri madenciliği işlenen verinin yapısına göre, çeşitli algoritmalarından yararlanmaktadır.



Şekil 2. 2 Veri madenciliği ve disiplinler [29]

2.6 Veri Madenciliğinin Tıptaki Önemi ve Uygulamaları

Bu metodun bankacılık, sigortacılık ve endüstri alanlarında başarısı doğrulandığında, araştırmacılar bu metodu insanların sağlık problemlerini çözme amacı taşıyan tıp alanında da uygulamaya başlamıştır. Veri tabanı uygulamalarının hastanelerde ve laboratuvarlarda kullanılması ile tıpta veri madenciliği tekniklerinin uygulanması için gerekli zemin hazırlanmıştır. Bu durumda bilgisayarlar, hasta bakım hizmetlerinin desteklenmesi ve sağlık bakım hizmetlerinin kalitesinin değerlendirilmesi gibi direkt olarak sağlık bakım hizmetlerinin sunulmasında çokça tercih edilmeye başlamıştır. Ek olarak, bilimsel ve yönetsel fonksiyonlar için de kullanılmaktadır.

Temel veri tabanları; önceleri hastanın adı, soyadı, doğum tarihi ve yeri, adresi ve kan grubu gibi ana bilgilerini saklarken, şimdi ise hastaların hasta lokasyonundan finansal verilerine, ilaçlarından ziyaret planına kadar çeşitli detaylı verileri kayıt altında tutmaktadır.

2.7 Veri Madenciliğini Etkileyen Faktörler

Bu metod başlıca beş faktörden etkilenebilmektedir. Veri madenciliğini etkileyen etmenler ve açıklamaları aşağıda belirtilmektedir:

- **Veri:** Veri madenciliğinin gelişimindeki en mühim etmendir.
- **Donanım:** Büyüyen bellek ve işlem hızı kapasitesi dolayısıyla, geçmişte kullanılamayan veriler üzerinde çalışmayı olası kılmıştır.
- **Bilgisayar ağları:** Günümüzde internet, oldukça yüksek hızda kullanılabilir. Bu durumda, dağıtık verilerin analizi ve farklı algoritmaların kullanılması muhtemel olacaktır.
- **Bilimsel hesaplamalar:** Veri madenciliği ve bilgi keşfi, kuram, deney ve simülasyonu entegre etmede kritiktir.
- **Ticari eğilimler:** Kurumlar günümüzde çekişme şartlarında varlıklarını sürdürebilmek amacıyla çabuk aksiyon almalı, sunduğu hizmet kalitesi yüksek olmalı, tüm bunlar yapılırken de minimum maliyet ve minimum insan gücü dikkate alınmalıdır.

2.8 Veri Madenciliğinde Karşılaşılan Sorunlar

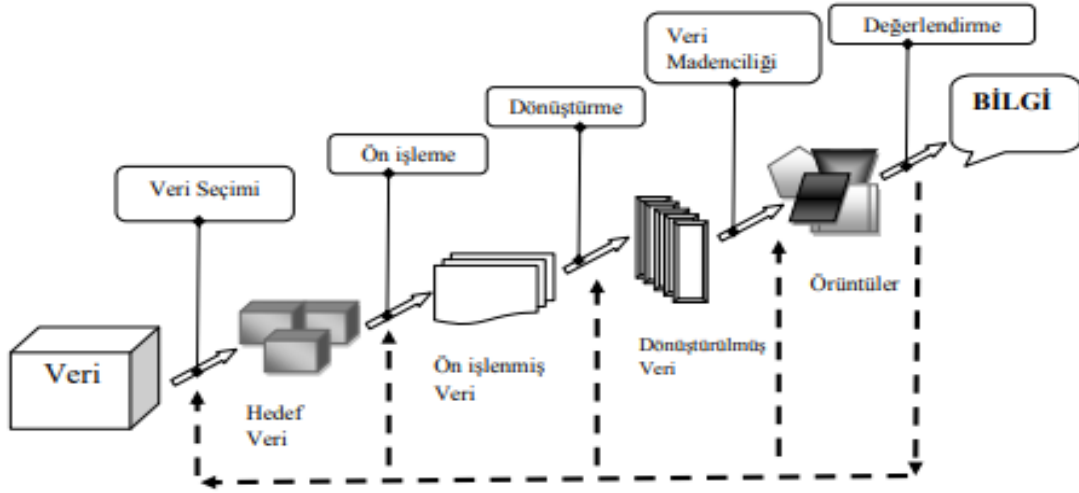
Büyük hacimli verilerin mevcut olduğu ortamlarda büyük problemler meydana çıkabilmektedir. Bu sebeple küçük veri setlerinde, simülasyon ortamlarında hazırlanmış veri madenciliği sistemleri, büyük hacimli, eksik, uğultulu, boş, mükerrer, aykırı ya da meçhul veri kümelerinin bulunduğu ortamlarda hatalı çalışabilmektedir. Bu yüzden, veri madenciliği sistemleri tasarlanırken bu problemlerin sonuca ulaştırılması gerekmektedir. Artık veri, belirsizlik, boş veri, dinamik veri, eksik veri, farklı tipteki verileri ele alma, gürültülü ve kayıp değerler, sınırlı bilgi ve veri tabanı boyutu veri madenciliği uygulamalarında karşılaşılabilecek problemlerdendir.

2.9 Veri Madenciliği Süreci

Bu metot, aynı zamanda süreç olarak da bilinmektedir. Veri madenciliği aşamasından önce, mutlaka bütün niteliklerin ayrıntılı olarak incelenmesi gerekmektedir. Bu süreçte gözetlenen adımlar Tablo 2.2’de verilmiştir [31].

Tablo 2. 2 Veri madenciliği sürecinde izlenen adımlar

Problemin tanımlanması
Verilerin hazırlanması
Modelin kurulması ve değerlendirilmesi
Modelin kullanılması
Modelin izlenmesi



Şekil 2. 3 Bilgi keşfi sürecinde veri madenciliği [29]

2.9.1 Problemin Tanımlanması

Veri madenciliği çalışmalarında başarılı olmak için birinci koşul, geliştirilen hangi örgüt amacı için yapılacağını ve ortaya çıkacak olan sonuçların başarının ölçümünde dikkat edilmesi gereken hususların tanımlanması gerekmektedir. Problemlerle tamamen örtüşmeyen veri madenciliği çalışması, problemin çözülmesinde yeterli olmayıp, bununla birlikte yeni sorunların çıkmasına sebep olabilir. Ek olarak, hatalı verilen karar veya kararlar maliyetin artması ve doğru karar veya kararlar ile oluşabilecek yararlar ile ilgili tahminler de problemin tanımlanması evresinde dikkate alınmalıdır.

2.9.2 Verilerin Hazırlanması

Modelin kurulması evresinde oluşabilecek problemler, sürekli olarak bu sürece geri gelinmesine ve bununla birlikte verilerin yeniden revize edilmesine sebep olabilecektir. Bu durum ise verilerin hazır edilmesi ve modelin kurulum süreçleri için analistin veri buluş evresinin tamamında zaman ve enerjisinin yaklaşık olarak %50 - %85 oranında harcamasına neden olmaktadır. Verilerin hazırlanması süreci; toplama, değer biçme, birleştirme ve temizleme, örneklem seçimi ve son olarak dönüştürme olarak tanımlanmaktadır.

2.9.2.1 Toplama ve Değer Biçme

Gerekli olan verilerle birlikte bu verilerin toplanacağı veri kaynaklarının belirlenmesini kapsayan bir evredir. Burada verilmesi gereken önemli karar, hangi veri kaynaklarından yararlanılacağıdır.

Veri uyumsuzluklarının sebebi, kullanılacak verilerin ayrı kaynaklardan derlenmesi olacaktır. Uyumsuzlukların başlıca faktörleri; zaman farkı, güncelleme hataları, veri formatlarının aynı olmaması, kodlama farklılıkları ve ayrı ölçü birimleri farklılıkları olarak açıklanır. İlaveten, verilerin hangi koşullarda, nerede ve nasıl derlendiği de oldukça ehemmiyetlidir. Güvenilir olmayan veri kaynaklarının kullanımı veri madenciliği sürecinin güvenilirliğini etkilemektedir. Bu durumda, verimli sonuçların alınabileceği veri madenciliği çalışmaları anlaşılabilir verilerin üzerine yapılıp, toplanan verilerin uyumluluk durumları da bu evrede analiz edilmelidir.

2.9.2.2 Birleştirme ve Temizleme

Bu adımda, ayrı kaynaklardan derlenen verilerde tespit edilen ve diğer süreçte belirlenen uyumsuzluklar ile problemler minimuma indirilerek, veriler sadece bir veri tabanında toplanır. Verilerin temizlenmesi süreci, uyumsuz verilerin düzeltilmesi, eksik değerlerin tamamlanması, çelişkili verilerin netleştirilmesi ya da veri setinden çıkartılması gibi işlemleri içermektedir. Bu süreçte eksik verilerin yerine yeni verilerin belirlenip eklenmesi için eksik değer içeren kayıtların veri setinden silinmesi, kayıtlı değerlerin yerine sabit değer kullanılması, değişkenin bütün verileri kullanılarak ortalama hesabının yapılıp, eksik değer yerine hesaplanan değer kullanılması ve değişkendeki bütün veriler yerine, yalnız bir sınıfa bağlı olan örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılması yöntemleri önerilebilir.

2.9.3 Modelin Kurulması ve Değerlendirilmesi

Bu evrede oluşturulacak modele göre seçim yapılır. Genel olarak yanlış veri girişi ya da bir defaya özgü olayın gerçekleşmesinden kaynaklanan veriler veri setinden kaldırılır. Kimlik numarası, sicil numarası, sıra numarası gibi anlam taşımayan

değişkenler modele girmemelidir. Zira bu türdeki değişkenler veriye ulaşma süresinin uzamasına neden olabilmektedir.

Tanımlanan problemde optimum modelin belirlenmesi için fazla miktarda modelin kurularak test edilmesi gerektiğinden, verinin hazırlanıp modelin kurulması süreci, en iyi sonucu veren model belirlenene kadar tekrarlanmaktadır.

Çapraz geçerlilik testi, sınırlı verinin mevcut olması durumunda kullanılabilecek bir yöntemdir. Bu yöntemde veri seti rassal olarak eşit olarak iki parçaya bölünür. Birinci aşamada bir parça üzerinde model eğitimi, diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi, birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.

Kurulan modelin doğruluk derecesi yüksek olmasına rağmen, reel dünyayı tam anlamıyla modellediği garanti edilememektedir. Testlerde geçerli modelin hatalı olmasının nedenleri, model kurulumunda kabul edilen varsayımlar ve kullanılan verilerin hatalı olmasıdır.

2.9.4 Modelin Kullanılması

Bazı modeller direkt olarak uygulama olabilmektedir ve diğer bir uygulamanın alt parçası olarak da kullanılabilmektedir.

2.9.5 Modelin İzlenmesi

İhtiyaca bağlı olarak sistemlerin özelliklerinde ve kayıtlarında değişiklikler olabilmektedir. Bu durumda, model sürekli olarak takip edilmeli ve yenilenmelidir.

2.10 Veri Madenciliği Modelleri

Veri madenciliği ile ilgili kullanılan modellere yeni metot ve algoritmalar eklenmektedir. Bu metot ve algoritmalarından bazıları uzun senelerdir kullanılan klasik istatistiksel yaklaşımlara dayanır. Diğer metotlar ise istatistiksel yaklaşımların yanında makine öğrenmesi ve yapay zekâ destekli güncel yaklaşımlara dayanır. Bu modeller, fonksiyonlarına bağlı olarak Sınıflama ve Regresyon, Kümeleme ve Birliktelik Kuralları olarak üç temel başlıkta toplanmaktadır.

2.10.1 Tahmin Edici Modeller

Bu modellerde sonuçlanmış olan veriler baz alınarak bir model geliştirilmesi ve geliştirilen bu modelden faydalanılarak neticeleri belli olmayan veri kümeleri için sonuçların öngörülebilmesi hedeflenmektedir. Sınıflama ve regresyon yöntemleri tahmin edici modeller içerisinde yer almaktadır.

Hayati veri sınıfları üretebilen ya da ilerideki eğilimi tahmin edebilecek modeller oluşturabilen veri analiz metotlarına sınıflandırma ve regresyon metotları denir. Sınıflandırma gruplu değişkenlerin tahmin edilmesinde kullanılırken; regresyon ise devamlı değişkenleri öngörmek için kullanılmaktadır. Sınıflandırma ve regresyonda kullanılan başlıca yöntemler aşağıdaki tabloda verilmiştir.

Tablo 2. 3 Sınıflandırma ve regresyon kullanılan başlıca yöntemler

Naive Bayes
Karar Ağaçları
K-En Yakın Komşu Algoritması
Yapay Sinir Ağları
Genetik Algoritmalar
Destek Vektör Makinesi
Topluluk Sınıflandırıcıları

2.10.1.1 Karar Ağaçları

Veri madenciliğinde sıkça kullanılan yöntemlerin başında karar ağaçları gelmekte olup, bu durumun sebepleri; ucuz olması, kolayca yorumlanabilir olması ve veri tabanı sistemleri ile entegre edilebilmesi olarak belirtilmektedir. Karar ağaçları, düğümler ve dallardan meydana gelen ve anlaşılabilir bir yöntemdir. Karar ağacında bulunan her bir dalın belli olasılığı mevcut olup, mevcut olan bu olasılık sayesinde son dallardan köke ya da ulaşılması istenen yere kadar olasılıkların hesaplanması mümkündür. Karar ağaçlarına dayanarak geliştirilen algoritmalar içinde en fazla kullanılan algoritmalar ID3, C4.5 ve C5 algoritmalarıdır [32].

En çok kullanılan yöntem olan karar ağaçları ismi kök olan ve girdi almayan bir düğümle başlar ve her bir düğüm birer girdiyi anlatan iç düğümlerden oluşur. Yaprak düğümlerinin çıktıları diğer düğüm için girdi olarak kullanılmaz; fakat test düğümlerinin çıktıları diğer bir düğüm için girdi olarak kullanılır. Her bir girdiyi anlatan bu iç karar düğümleri, suallerin sorulduğu, verilerin test edildiği ve dallanmanın nasıl şekilleneceğini gösteren düğümlerdir. Dallar, suallerin yanıtlarını verir. Uç yapraklarsa grubun olduğu sınıf etiketlerini kapsar. Özetle, ilk olarak, eğitilecek veri setinden ağaç yapısı oluşturulur ve oluşturulan bu ağaç yapısı test seti ile analiz edilir. Sonrasında, sınıfı belli olmayan verinin sınıfı diğerlerinden fazla başarı yüzdesiyle öngörülür.

2.10.1.2 Destek Vektör Makinesi

Bu yöntem, verileri optimum bir şekilde iki sınıfa bölen n boyutlu bir hiperdüzlem oluşturmaktadır. Bu model sigmoid bir kernel fonksiyonu kullanmakta olup, yapay sinir ağlarıyla yakından ilişkili olmakla birlikte; iki tabakalı verileri beslemeli bir yapay sinir ağına sahiptir. DVM'nin istatistiksel öğrenme kuramındaki yapısal risk minimizasyonu niteliğinde çalışması dikkat çeken bir özelliğidir. Eğitim kümesindeki örneklerin bağımsız ve homolog olarak dağılması DVM'nin esas varsayımlarındandır. DVM, sınıflandırma ve regresyon problemlerinde çalıştırılabilir. Regresyon yöntemindeki esas amacın elde bulunan eğitim verilerinin karakterini gerçeğe yakın bir şekilde yansıtan ve istatistiksel öğrenme teorisine uygun olan doğrusal ayırıcı fonksiyonun bulunması olduğu bilinmektedir. Regresyonda da sınıflandırmadaki gibi doğrusal olmayan durumların işlenebilmesi amacıyla çekirdek fonksiyonlarından yararlanılır.

Günümüzdeki en güçlü sınıflandırma algoritmalarından biri olan destek vektör makinesi, güçlü matematiksel tabanlara ve istatistiksel öğrenme teorisine dayanmaktadır. Ayrı sınıflara ait olan prototiplerini birbirlerinden ayıran fazlaca doğrusal düzleme erişmek mümkündür. Destek vektör makinesi, ayrı sınıflara ait destek vektörleri arasındaki mesafeyi maksimize edecek ayırma hiper düzleminin teşhisini amaçlar. Metodun tabanında ayrı neticelere sahip örnekleri birbirlerinden ayıran bir hiper düzlem vardır. Tercih edilen iki sınıflı problemler için planlanmış destek vektör makinesi, her iki sınıfın en yakın noktasına en büyük

uzaklıkta olan en uygun hiper düzlemi bulur. En uygun hiper düzlemi bulmak doğrusal sınıflandırmayı sağlar. Bulunan en uygun hiper düzleme en benzer olan örnek setine destek vektör denir. Genelde asıl veri setleri için ayrı sınıfların örneklerini alenen ayıran bir hiper düzlem yoktur. Bu problemi ortadan kaldırmak için veri kümesini bir sınıfın mümkün olan en yukarı seviyede geçerli olduğu iki kümeye ayıran gevşek sınır yöntemi tavsiye edilir.

2.10.1.3 Diskriminant Analizi

Bu yöntem, istatistiksel bir teknik olarak bilinmektedir. Ölçülebilir özelliklere sahip nesneleri gruplandırmak için kullanılmakta olan örüntü tanıma ve denetimli öğrenmedir. Sınıflandırma problemi bu nesnelerden ölçülerek elde edilen birçok özelliği kullanarak bu nesnelerin sınıfını tespit eder. Diskriminant analizinde, hangi özellik dizisi istenen nesnenin hangi gruba ait olacağının doğru olarak netleştirilmesine ve bu gruplara ayırmak için en doğru modelin veya sınıflandırma kuralının belirlenmesine dikkat edilmelidir. Bu bağlamda sınıflandırma için ilk olarak özellik seçimi yapıp, sonrasında da sınıflandırma ile devam edilir. Sınıflandırma için kullanılan yöntemler ilk olarak nitel değişkenin kategorilerinin olasılığını tahmin eder ve yeni bir gözlem değeri için kullanılabilir fonksiyonlar belirler.

2.10.1.4 Naive Bayes

Naive Bayes algoritması her kriterin sonuca olan etkilerinin olasılık olarak hesaplanmasına dayanmakta olup, genel olarak metin dokümanlarının sınıflandırılmasında yaygın olarak kullanılmaktadır. Performansı ve uygulanabilirliği ön plana çıkan bir algoritma olarak bilinmektedir. Naive Bayes algoritmasında istatistiksel yöntemler yardımıyla sınıflandırma yapılır.

Naive Bayes algoritması uygulanırken niteliklerin birbirinden bağımsız olduğu kabul edilmektedir. Burada eğer nitelikler birbirlerini etkilemekte ise olasılık hesaplamak zordur. Bu nedenle, bütün niteliklerin öneminin aynı derecede olduğu kabul edilmektedir.

Temelini Bayes savından alan Naive Bayes, bir sınıflandırma yöntemidir. Bu yöntem olasılık prensiplerine göre tanımlanmış hesaplamalar ile girilen verilerin sınıfını teşhis etmeye çalışır. Naive Bayes birden çok sınıftan oluşan veri setlerinde

kullanılır. Naive Bayes'in mantığı; her sınıfın alt örneğinden oluşan durumların gerçekleşme olasılıklarının çarpımlarının tüm sınıfın olasılığı ile çarpılıp normalleştirilmesidir. Yani standartlaştırmaya çalışır. Sisteme belli yüzdede öğretilmiş veri seti girilir ve bu verilerin sınıfları kesin seçilmiş olmalıdır. Öğretilmiş veri seti üzerinde yapılan işlemler neticesinde elde edilen olasılık değerleri ile güncel test verileri çalıştırılır. Çıkan test verilerinin sınıfı teşhis edilmeye çalışılır. Öğretilmiş veri sayısı ne kadar çoksa test verisinin sınıfını doğru öngörebilme yüzdesi bir o kadar çoktur. Algoritmanın uygulanmasında özelliklerin birbirinden bağımsız olması gibi birtakım kabuller yapılır. Zira eğer özellikler birbirini etkiliyorsa bu olasılıkları hesaplamak kolay değildir. Özelliklerin tümünün eş seviyede önemli olduğu varsayılır. Bu varsayım ise algoritmanın eksikliğidir. Mesela şeker hastalığını tespit etmede, hastanın tansiyon değeri ile kilosunun eş değerde birer nitelik olarak alınması mantık açısından doğru değildir. Bu yöntem makine öğreniminde öğreticili öğrenme sınıfındadır. Sınıflandırılması beklenen kümelerin ve örneklerin bağlı oldukları sınıflar vardır. Geniş, bitmemiş ve kompleks veri setlerinin sınıflandırılması için uygundur. Metin belgelerinin sınıflandırılmasında sıklıkla kullanılır. Yapılabilirlik ve performansı nedeniyle öncelik verilen algoritmalarından biridir. Performansı kategorik verilerde nicel verilere oranla optimumdur. Daha iyi sonuç verir. Bu algoritmanın uygulanmasındaki başka bir güçlük ise olasılıkların giriş değerlerine gereksinim duyulmasıdır. Bu olasılıkların giriş değerlerinin bilinmemesi hâlindeyse genellikle mevcut verilerle alakalı belli başlı bilgiler ve veri dağılımları baz alınarak tahmin edilebilir.

2.10.1.5 K-En Yakın Komşu Algoritması

KNN, bağımsız değişkenlerin meydana getirdiği vektörün en yakın komşularının hangi sınıfta ağırlıklı olduğu bilgisinden yararlanılarak sınıfı öngörmeye dayanan bir yöntemdir. Bu algoritmada, Distance (Uzaklık) ve K (komuşuluk sayısı) olmak üzere iki temel değer üzerinden tahmin yapılır. KNN algoritması ile üretilmiş olan bir modelin başarısının ölçülmesi amacıyla kullanılan Jaccard Index, F1-Score ve LogLoss olmak üzere 3 indikatör vardır.

Makine öğrenimi algoritmalarının diğerlerinden anlaşılır ve kolay olanı “En Yakın Komşu Algoritmasıdır”. Sınıfı belli olan eğitim setindeki gözlem verilerinin her birinin, sınıfı teşhis edilmesi istenen test verilerine olan uzaklıkları hesaplanır ve hesaplanan bu mesafelerden diğerlerine göre düşük olan k adeti seçilir. Seçilen mesafesi düşük olan K adet komşu niteliklerin sınıfları tespit edilir ve test verisinin de bu sınıfa bağlı olması istenir. $K = 1$ belirlendiğinde belirlenen komşu gürültülü veri ise model çok etkilenir. K çok büyük seçilirse algoritma tek sınıfmış gibi algılar ve sınıftaki verilerin mod değeri belirlenmiş olur. Mesafelerin hesaplanmasında, Öklid mesafe test sınıfı öğelerinin konumunu belirlemek üzere kullanılan Öklid mesafe formülü aşağıdaki gibidir:

$$d(i, j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (2.1)$$

KNN, diğerlerinden kolay sınıflandırma yöntemlerinden biri olduğu için fazlaca tercih edilmektedir. KNN'nin nicel verilerde uygulanması kategorik verilere oranla daha kolaydır. İlaveten gürültülü verilere karşı sağlam olması da sınıflandırma sorunlarında özel olarak tercih edilmesinin nedenlerinden birdir. Fakat milyonlarca verinin bulunduğu veri setlerinde hesaplanacak mesafe sayısı da aşırı fazla olacağından işlem süresinin çok olması hesaplama maliyetini olumsuz etkiler. Ancak kimi indeksleme metotlarıyla fazla olacak maliyet de düşürülebilir.

2.10.1.6 Topluluk Sınıflandırıcıları

Bu yöntemde, tek bir modele kıyasla daha güçlü ve genellenebilir sonuçlar elde etmek amacıyla birden fazla baz modelin tahmin sonuçları birleştirilip, bu yöntemlerin başarısı temel olarak öğrenme başarısı ile ve birbirinden farklılıklarıyla ölçülür.

2.10.1.7 Yapay Sinir Ağları

Bu sınıflandırma, ağırlıkların hesaplanmasına dayanmaktadır. Eğitim veri seti üzerinde hesaplanan ağırlıklar, test veri seti üzerinde kullanılarak öğrenmenin gerçekleşme oranını belirlemektedir. Elde edilen ağırlıkların etkinliği doğrulanamaz ise ağırlıklar üzerinde düzenleme ve yeniden hesaplama yapılır.

Öğrenme süreci tamamlandığında ağırlıklar yardımı ile yeni bir verinin hangi sınıfa ait olduğu belirlenebilir.

İnsan beyninin başlıca fonksiyonlarından biri olan öğrenme niteliğini gerçekleştirebilen bilgisayar sistemleri yapay sinir ağlarıdır. Öğrenmeyi deneyimler yani girdi olarak verilen örnekler vasıtasıyla gerçekleştirirler. Yapay sinir ağları, öğrenme yoluyla bulma, güncel bilgiler oluşturabilme gibi etkinlikleri otomatik olarak gerçekleştirebilmek hedefiyle geliştirilmişlerdir. Öğrenmeye ilave olarak bilgiler arasında irtibat kurma niteliğine de sahiptir. Yapay sinir ağlarının temel fonksiyonları: Sınıflandırma, kümeleme, tahminleme, kontrol, veri birleştirme ve kavramsallaştırmadır. Mali ve endüstriyel çalışmalar, askeri ve defans uygulamaları, görüntü işleme ve örüntü tanıma, robotbilim, tıp ve sağlık sektörü vs. alanlarda kullanılmaktadır. Yapay sinir ağları birbirine ilişik yapay sinir dokularından ve katmanlardan oluşur. O aradaki bütün irtibatlar bir ağırlık değerine sahiptir. Bilgiler bu ağırlık değerlerinde korunup bütün ağa yayılır. Yapay sinir ağları aşağıda açıklandığı gibi 3 katmandan oluşmaktadır.

- **Girdi katmanı:** Bilgiler kullanıcıdan alınır. Girdi katmanında alınan bilgiler işleme girmez.
- **Ara katmanlar:** Ara katmanda kullanıcıdan gelen bilgiler işlenir. Bir tane ara katman birden fazla soruna çare olabilir. Şayet ağa öğretilmek istenen sorunun karmaşıklığı çoğalırsa veya girdi ile çıktı arasındaki bağ doğrusal değilse birçok sayıda ara katman kullanılabilir.
- **Çıktı katmanı:** Girdi katmanından gelen bilgiler ara katmanda işlenir. Çıktı katmanı ise ara katmandan halihazırda gelen bilgileri işler. Ağın çoğalttığı çıktıyı bulur ve bu bulduğu çıktıyı kullanan kişiye gönderir.

Yapay Sinir Ağlarının bir problemde kullanılabilmesi için ilk önce girdi ve çıktı değişkenleri tespit edilir. Sonrasında ise veriler toplanır. Veri setindeki bütün durumlar girdi ve çıktı değişkenlerinin değerlerinden meydana gelir. Eğitim ve test evrelerinde kullanılır. Veri düzenlenip değişkenler tespit edildikten sonra kullanılacak ağ modeli ve nasıl yapılandırılacağı belirlenir. Her katmanda belli sayıda nöronun sonraki katmanlara bağlanmasıyla ağ topolojisi oluşturulur. Problemlle ilişkili bir sinir ağı tespit etmek performans açısından önemlidir. Ağın

yapısını oluşturmak zor bir işlemdir. Ağın yapısını oluştururken deneme-yanılma yapılır. Ağ eskiye ait çeşitli girdi ve çıktı değerlerine sahip veri setini kullanıp eğitilerek girdi ve çıktı arasındaki ilişki öğretilir. Bu ilişkinin öğretilmesi sırasında bir eğitim algoritması kullanılarak yanlışı minimuma düşürmek hedeflenir.

2.10.2 Tanımlayıcı Modeller

Bu modellerde; karar vermeye yönelik işlenebilecek verilerin örüntülerinin belirlenmesi sağlanmaktadır. Kümeleme ve birliktelik kuralları tanımlayıcı modellerdir.

2.10.2.1 Kümeleme Yöntemi

Kümeleme, nesneleri nesnelerin benzerlikleri bakımından gruplanması süreci olarak tanımlanır. Kümeleme analizinin esas amacı, nesneleri sahip oldukları karakteristik özelliklerine göre gruplamaktır. Mevcut veriler ile güvenilir ve verimli analiz yapılabilmesi için kümeleme analizi büyük bir öneme sahiptir.

Kümelerin oluşturulması aşamasında, aynı kümenin içerisindeki nesneler arasındaki benzerliğin en fazla, farklı kümeler arasındaki nesnelerin benzerliğinin ise en az olması gerekmektedir.

2.10.2.2 Birliktelik Kuralı

Bu kuralda büyük veri setleri arasındaki birliktelik bağıntıları bulunur. Saklanan ve toplanan veri gün geçtikçe büyümektedir. Bu yüzden, veri tabanlarındaki birliktelik kurallarını ortaya çıkarma bir amaç haline gelmiştir. Örneğin, büyük miktardaki işlem kayıtlarından farklı birliktelik ilişkilerinin bulunması firmaların karar alma işlemlerini daha verimli hale getirmektedir.

Bu bölümde, öncelikle, iki boyutlu sağlıklı ve kanserli meme modelleri için elektromagnetik düz saçılma problemi çözülmüştür. Düz saçılma probleminden elde edilen sayısal veriler kullanılarak çeşitli makine öğrenmesi teknikleriyle meme modeli içinde kanserli dokunun varlığı ve (varsa) kabaca konumu tespit edilmeye çalışılmıştır. Bu çalışmada verilerin üretilmesi, düzenlenmesi ve makine öğrenmesi teknikleriyle sınıflandırma aşamalarının tümünde Matlab programlama dili kullanılmıştır.

Aşağıda sırasıyla problemin tanımı, verilerin sentetik olarak üretilmesi ve düzenlenmesi, çeşitli makine öğrenmesi yöntemlerinin uygulanması ve elde edilen sonuçların sunumu verilmektedir.

3.1 Problemin Tanımlanması

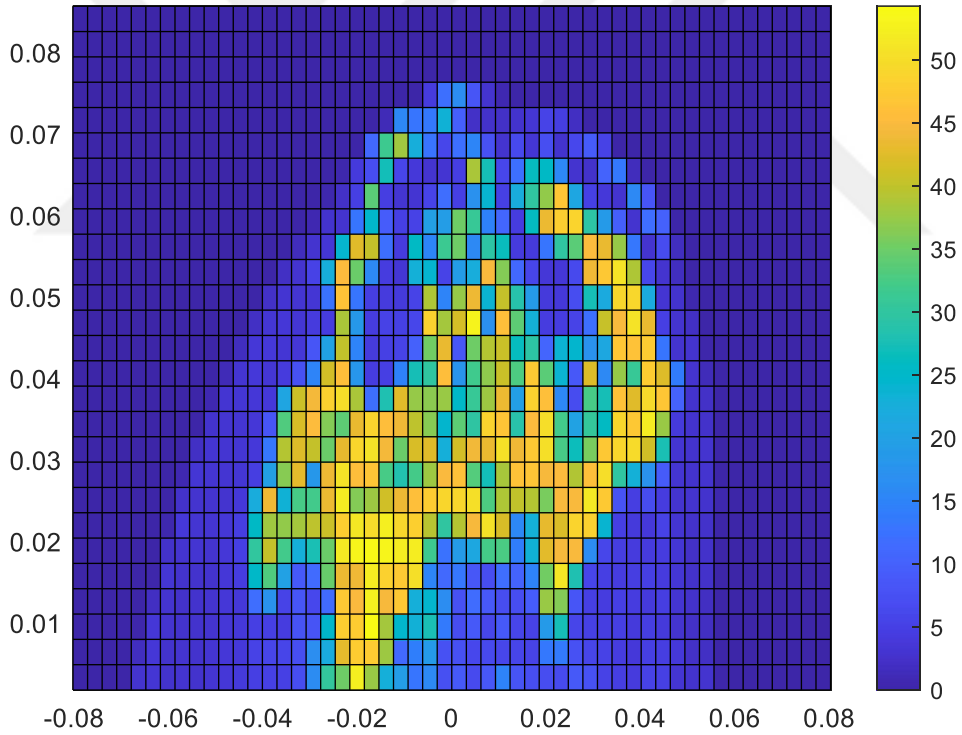
Mikrodalga tomografisi yaklaşımında; meme, etrafına yerleştirilmiş verici antenlerle aydınlatılır ve bunun sonucunda ortamda bir elektromagnetik alan dağılımı oluşur. Yine memenin etrafına yerleştirilmiş alıcı antenlerle oluşan elektromagnetik alanlar ölçülür. Ölçülen bu veri, memenin içindeki sağlıklı dokuların ve varsa kanserli dokunun geometrik (şekil ve konum) ve malzeme özelliklerinin (dielektrik katsayısı, iletkenlik katsayısı ve magnetik geçirgenlik katsayısı) bilgisini içermektedir. Bu olgu, ölçülen verilerden sınıflandırma algoritmaları kullanarak tümörün varlığını ve konumunu tespit edebilme olanağını sağlamaktadır.

Bu çalışmada, ilk aşamada elektromagnetik alan saçılımını elde etmek için söz konusu problem bilgisayar ortamında simüle edilmiştir. Böylece sayısal olarak üretilen elektromagnetik alan verileri sınıflandırma algoritmalarında kullanılmış ve meme içinde tümörün varlığı ve konumu incelenmiştir.

3.2 Verilerin Hazırlanması

İlk olarak, erişime açık olan [33] veri tabanındaki magnetik rezonans (MR) ile üretilmiş üç boyutlu meme görüntüleri, [34]'de verilen teknik aracılığıyla sayısal meme modellerine dönüştürülmüştür. Toplamda farklı boyut ve şekillere sahip, içindeki yağ ve kas doku dağılımlarının oldukça farklı olduğu 9 adet üç boyutlu meme modeli üretilmiştir. Üretilen bu sayısal meme modelleri, meme içindeki tüm yapının dielektrik katsayısı ve iletkenlik katsayısı dağılımını içermektedir.

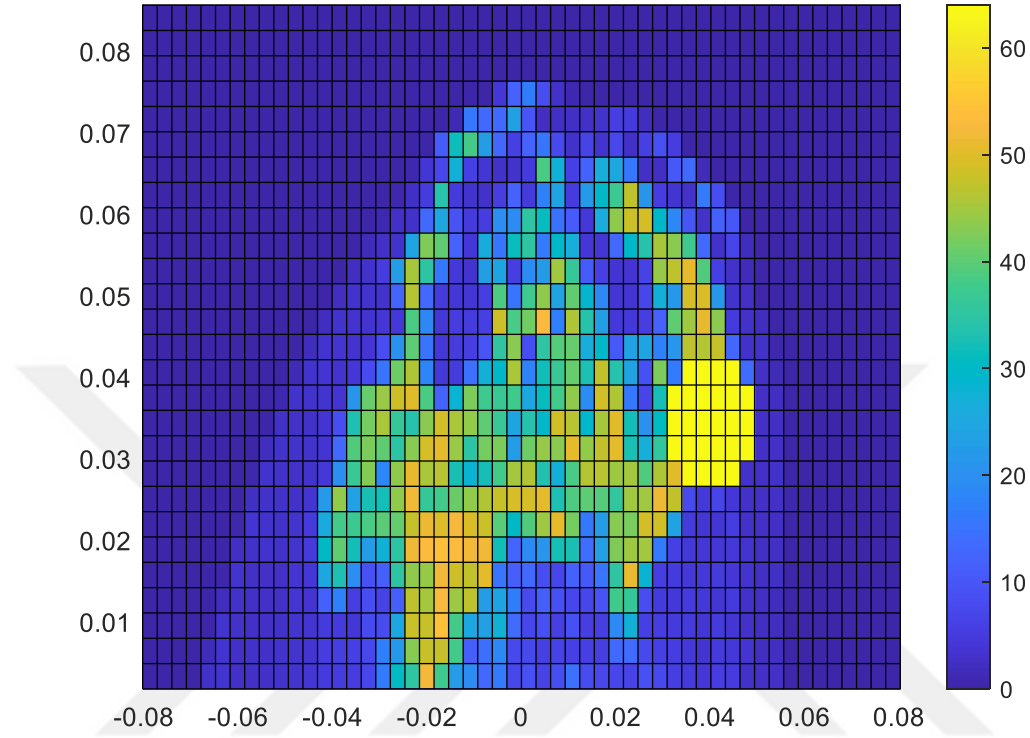
Üç boyutlu meme modellerinin her birinden, sagittal düzlemde iki boyutlu meme kesitleri alınmıştır. Örnek bir meme kesitinin göreceli dielektrik katsayısı dağılımı Şekil 3.1'de verilmiştir. Ayrıca iki boyutlu meme modellerinin düşey eksen etrafında 180 derece döndürülmesiyle meme modeli sayısı iki katına çıkarılmıştır.



Şekil 3. 1 İki boyutlu bir meme modelinin göreceli dielektrik katsayısı dağılımı

Tümör içeren meme modelleri üretmek için, çapı 1 *cm* ve 2 *cm* arasında rastgele değişen dairesel tümör örnekleri sağlıklı meme modelleri içine rastgele dağıtılmıştır. Tümörlerin dielektrik katsayısı $65\epsilon_0$, iletkenlik katsayısı ise 1.3 (S/m) olarak alınmıştır [1]. Her bir model için 50 adet tümör meme içinde

rastgele dağıtılmış olup, bunlardan meme sınırları dışına taşmış olanlar çıkarılmıştır. Kanserli doku içeren bir meme modelinin görelî dielektrik katsayısı dağılımı Şekil 3.2’de gösterilmektedir.



Şekil 3. 2 Tümör içeren meme modelinin görelî dielektrik katsayısı dağılımı

Sonuç olarak 2759 adet sağlıklı, 123392 adet kanserli dokuya sahip meme modeli üretilmiştir.

3.2.1 Elektromagnetik Alan Verilerinin Sentetik Olarak Üretilmesi

Homojen olmayan yapılardan saçılan alanların hesaplanması için Moment yöntemi, sonlu elemanlar yöntemi ve sonlu farklar yöntemi gibi bir çok sayısal yöntem geliştirilmiştir. Bu çalışmada sentetik veriler Moment yöntemi [34] kullanılarak üretilmiştir. Moment yöntemi saçıcının küçük hücrelere ayrıklaştırmasına dayanan bir yöntemdir. λ minimum dalga boyu olmak üzere, her bir hücrenin boyutları $\lambda/14 \times \lambda/14$ olarak seçilmiştir.

Her bir meme modeli yarıçapı 18 cm olan yarım çember üzerinde 10 farklı konumda bulunan çizgisel kaynak ile tek tek aydınlatılmış olup, yarıçapı 16 cm olan yarım çember üzerindeki 40 noktada bulunan alıcı antenler ile saçılan alan

Tümörün varlığı ve konumunu temsil eden sınıf bilgisi ve saçılan alan değerlerini içeren nihai veri setinin bir kısmı Şekil 3.4’de verilmiştir. Tümörün varlığı ve konumunu temsil eden sınıf bilgileri ‘Class No’ adıyla tanımlanmıştır.

CLASS_NO	T1M1_Re	T1M2_Re	T1M3_Re	T1M4_Re	T1M5_Re	T1M6_Re	T1M7_Re
2	-0,02582	-0,0405	-0,05729	-0,07521	-0,09752	-0,12148	-0,15107
3	-0,03116	-0,0337	-0,04077	-0,04097	-0,04776	-0,04891	-0,05218
2	-0,10357	-0,12556	-0,15636	-0,17964	-0,21412	-0,23936	-0,26642
4	-0,18436	-0,19695	-0,20087	-0,21109	-0,21424	-0,22072	-0,22498
2	-0,09594	-0,11672	-0,13249	-0,15184	-0,17472	-0,19333	-0,21423
4	-0,03116	-0,03968	-0,04515	-0,0543	-0,05662	-0,0645	-0,06735
4	-0,16839	-0,18195	-0,19206	-0,1982	-0,20325	-0,21096	-0,21404
4	-0,19359	-0,20655	-0,21247	-0,22313	-0,23758	-0,24507	-0,24748
2	0,117666	0,117638	0,117011	0,111877	0,105174	0,097145	0,089022
4	-0,09424	-0,1129	-0,12792	-0,14398	-0,15609	-0,17695	-0,19168

Şekil 3. 4 Dönüştürülmüş veri seti

3.3 Modellerin Eğitilmesi ve Test Edilmesi

Veri seti 27581 sağlıklı, 123392 kanserli olmak üzere olmak üzere toplam 150973 kayıttan oluşmaktadır. Dokuz üç boyutlu meme modelinden altı tanesi makine öğrenmesinde kullanılan modelin eğitimi, üç tanesi ise modelin testi için ayrılmıştır. Eğitim aşaması için kullanılan altı adet üç boyutlu meme modelinden tüm verilerin %70’i oluşturulurken, test aşaması için kullanılan üç adet üç boyutlu meme modelinden tüm verilerin %30’u oluşturulmuştur.

Modellerin eğitimi aşamasında 10 katlamalı çaprazlama tekniği kullanılmıştır. Karar ağaçları, diskriminant analizi, destek vektör makinesi, Naive Bayes, topluluk sınıflandırıcıları (ensemble classifiers), yapay sinir ağları ve KNN gibi makine öğrenmesi algoritmaları kullanılarak modeller geliştirilmiş ve ardından bu modeller test verileri ile sınanmıştır. Ayrıca temel bileşen analizi (TBA) kullanılmadan ve özelliklerin %95’i için TBA kullanılarak sonuçlar karşılaştırılmıştır. Elde edilen sonuçlar aşağıda sunulmuştur.

3.4 Modelin Oluşturulması

Çeşitli makine öğrenmesi teknikleri kullanılarak oluşturulan modeller için eğitim ve test verisine ait elde edilen sonuçları aşağıda verilmiştir.

3.4.1 Karar Ağaçları

Karar ağacı sınıflandırmasında, Matlab programı ağaçların derinliğini ve dal sayısını dikkate alarak çeşitli ağaç modelleri kullanmaktadır. Aşağıdaki tabloda bu ağaç modelleri ve optimize edilebilir karar ağacı için test ve eğitim başarı oranları verilmiştir.

Tablo 3. 1 Karar ağaçları sınıflandırma sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Fine Tree	Eğitim	53.7	50.0
	Test	50.1	50.4
Medium Tree	Eğitim	47.3	44.2
	Test	45.3	45.0
Coarse Tree	Eğitim	34.4	33.3
	Test	33.6	35.2
Optimizable Tree	Eğitim	76.4	76.3
	Test	52.5	54.0

Yukarıdaki tabloda verilmiş olan karar ağaçları için TBA kullanılmadan yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %76.4 oranı ile '*Optimizable Tree*'; test için en başarılı sonucu ise yine %52.5 oranı ile '*Optimizable Tree*' vermiştir. Bununla birlikte, karar ağaçları için TBA oranı %95 seçilerek yapılan sınıflandırmalarda, eğitim ve test için en başarılı sonuçları yine '*Optimizable Tree*' yöntemi kullanılarak elde edilmiştir. Başarı oranları eğitim ve test verisi için sırasıyla %76.3 ve da %54.0 olarak bulunmuştur.

3.4.2 Diskriminant Analizi

Doğrusal (linear), quadratik (ikinci mertebeden) ayırıcılar için ve optimize edilebilir diskriminat ayırıcılar için modeller geliştirilip, test edilmiştir. Tablo 3.2’de elde edilen test ve eğitim başarı oranları verilmiştir.

Tablo 3. 2 Diskriminant analizi sınıflandırma sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Linear Discriminant	Eğitim	62.4	48.9
	Test	58.0	49.3
Quadratic Discriminant	Eğitim	73.2	57.5
	Test	67.0	63.1
Optimizable Discriminant	Eğitim	73.2	57.5
	Test	67.0	63.1

Yukarıdaki tabloda verilmiş olan discriminant analizi için TBA kullanılmadan yapılan sınıflandırmalarda, eğitim için en başarılı sonuçları %73.2 oranı ile ‘*Quadratic Discriminant*’ ve ‘*Optimizable Discriminant*’, test için en başarılı sonuçları da %67.0 oranı ile yine ‘*Quadratic Discriminant*’ ve ‘*Optimizable Discriminant*’ vermiştir. Bununla birlikte, bu algoritma için TBA oranı %95 seçilerek yapılan sınıflandırmalarda, eğitim için en başarılı sonuçları %57.5 oranı ile ‘*Quadratic Discriminant*’ ve ‘*Optimizable Discriminant*’, test için de en başarılı sonuçları %63.1 oranı ile yine ‘*Quadratic Discriminant*’ ve ‘*Optimizable Discriminant*’ vermiştir.

3.4.3 Destek Vektör Makinesi

Destek vektör makinesinin çeşitli çekirdekleri için elde edilen test ve eğitim başarı oranları aşağıdaki tabloda belirtilmektedir.

Tablo 3. 3 Destek vektör makinesi sınıflandırma sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Linear SVM	Validation	61.0	49.9
	Test	56.6	48.4
Quadratic SVM	Validation	79.7	66.9
	Test	71.0	63.9
Fine Gaussian SVM	Validation	89.5	67.0
	Test	46.5	67.3
Medium Gaussian SVM	Validation	79.3	51.9
	Test	72.1	51.4
Coarse Gaussian SVM	Validation	64.4	47.3
	Test	67.0	49.3

Yukarıdaki tabloda verilmiş olan destek vektör makinesi algoritması için TBA kullanılmadan yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %89.5 oranı ile *Fine Gaussian SVM*; test için en başarılı sonucu ise %72.1 oranı ile *Medium Gaussian SVM* vermiştir. Bununla birlikte, bu algoritma için TBA oranı %95 seçilerek yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %67.0 oranı ile *Fine Gaussian SVM*; test için en başarılı sonucu ise %67.3 oranı ile *Fine Gaussian SVM* vermiştir.

3.4.4 Naive Bayes

Naive Bayes sınıflandırıcısının çeşitli çekirdekleri için elde edilen test ve eğitim başarı oranları aşağıdaki tabloda belirtilmektedir. Büyük miktarda verinin modeli eğitmek için aşırı zaman aldığı ya da sıradan ev bilgisayarlarının donanım nedeniyle yetersiz kaldığı durumlar için sonuçlar verilmemiştir.

Tablo 3. 4 Naive Bayes sınıflandırma sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Gaussian Naive Bayes	Eğitim	46.6	48.6
	Test	48.8	54.0
Kernel Naive Bayes	Eğitim	-	51.8
	Test	-	54.7
Optimizable Naive Bayes	Eğitim	-	51.8
	Test	-	54.7

Yukarıdaki tabloda verilmiş olan Naive Bayes algoritması için TBA kullanılmadan yapılan sınıflandırmalarda, “Gaussian Naive Bayes” algoritmasıyla elde edilen model, eğitim verisi için %46.6, test verisi için %48.8 başarı oranıyla sınıflandırmaları doğru tahmin etmiştir. Bununla birlikte, karar ağaçları için TBA oranı %95 seçilerek yapılan sınıflandırmalarda, eğitim için en başarılı sonuçları %51.8 oranı ile “Kernel Naive Bayes” ve “Optimizable Naive Bayes”; test için en başarılı sonuçları ise %54.7 oranı ile yine “Kernel Naive Bayes” ve “Optimizable Naive Bayes” vermiştir.

3.4.5 Topluluk Sınıflandırıcıları

Çeşitli topluluk sınıflandırıcıları için elde edilen test ve eğitim başarı oranları aşağıdaki tabloda belirtilmektedir. Büyük miktarda verinin modeli eğitmek için

aşırı zaman aldığı ya da sıradan ev bilgisayarlarının donanım nedeniyle yetersiz kaldığı durumlar için sonuçlar verilmemiştir.

Tablo 3. 5 Topluluk sınıflandırıcıları için sınıflandırma sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Boosted Tree	Eğitim	53.6	45.5
	Test	49.5	46.5
Bagged Tree	Eğitim	90.0	88.0
	Test	65.0	62.2
Subspace Discriminant	Eğitim	62.6	48.9
	Test	58.7	49.0
Subspace KNN	Eğitim	-	88.8
	Test	-	59.4
RUSBoosted Tree	Eğitim	48.8	45.3
	Test	48.3	46.8

Yukarıdaki tabloda verilmiş olan topluluk sınıflandırıcıları algoritması için TBA kullanılmadan yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %90.0 oranı ile *'Bagged Tree'*; test için en başarılı sonucu ise %65.0 oranı ile yine *'Bagged Tree'* vermiştir. Bununla birlikte, bu algoritma için TBA oranı %95 seçilerek yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %88.8 oranı ile *'Subspace KNN'*; test için en başarılı sonucu ise %62.2 oranı ile *'Bagged Tree'* vermiştir.

3.4.6 Yapay Sinir Ağları

Matlab tarafından büyüklük ve mimarisine göre tanımlanan çeşitli yapay sinir ağları için elde edilen test ve eğitim başarı oranları aşağıdaki tabloda verilmiştir. Büyük miktarda verinin modeli eğitmek için aşırı zaman aldığı ya da sıradan ev bilgisayarlarının donanım nedeniyle yetersiz kaldığı durumlar için sonuçlar tabloda gösterilmemiştir.

Tablo 3. 6 Yapay sinir ağları sınıflandırma sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Narrow Neural Network	Validation	72.4	59.8
	Test	65.5	58.3
Medium Neural Network	Validation	82.4	67.1
	Test	64.0	62.4
Wide Neural Network	Validation	90.0	78.7
	Test	66.0	61.2
Bilayered Neural Network	Validation	73.7	-
	Test	63.9	-
Trilayered Neural Network	Validation	74.1	-
	Test	63.8	-

Yukarıdaki tabloda verilmiş olan yapay sinir ağları algoritması için TBA kullanılmadan yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %90.0

oranı ile “Wide Neural Network”; test için en başarılı sonucu ise %66.0 oranı ile yine “Wide Neural Network” vermiştir. Bununla birlikte, bu algoritma için TBA oranı %95 seçilerek yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %78.7 oranı ile “Wide Neural Network”; test için en başarılı sonucu ise %62.4 oranı ile “Medium Neural Network” vermiştir.

3.4.7 K-En Yakın Komşu Algoritması

Matlab tarafından tanımlanan KNN tipleri için elde edilen test ve eğitim başarı oranları aşağıdaki tabloda verilmiştir. Veri sayısının büyüklüğü nedeniyle modelin oluşturulmadığı durumlara ait sonuçlar verilmemiştir.

Tablo 3. 7 K-En Yakın Komşu algoritması sonuçları

Model Tipi	Veri Seti	TBA Kullanmadan (%)	TBA %95 (%)
Fine KNN	Validation	89.8	89.2
	Test	63.9	60.0
Medium KNN	Validation	82.5	81.3
	Test	66.8	63.2
Coarse KNN	Validation	70.0	68.5
	Test	67.8	64.8
Cosine KNN	Validation	82.7	81.6
	Test	67.0	63.6
Cubic KNN	Validation	-	81.1

	Test	-	63.2
Weighted KNN	Validation	86.0	85.7
	Test	66.8	63.0
Optimizable KNN	Validation	-	88.4
	Test	-	60.1

Yukarıdaki tabloda verilmiş olan K-en yakın komşu algoritması için TBA kullanılmadan yapılan sınıflandırmalarda, eğitim için en başarılı sonucu %89.8 oranı ile *Fine KNN*; test için en başarılı sonucu ise %67.8 oranı ile *Coarse KNN* vermiştir. Bu algoritma için TBA oranı %95 seçilerek yapılan sınıflandırmalarda ise, eğitim için en başarılı sonucu %89.2 oranı ile *Fine KNN*; test için en başarılı sonucu ise %64.8 oranı ile *Coarse KNN* vermiştir.

3.5 Modellerin Değerlendirilmesi

Kullanılan sınıflandırma algoritmaları başarı oranlarına göre sıralandığında, veri setleri ve modeller için en uygun sınıflandırma algoritmasının destek vektör makinesi algoritması olduğu görülmektedir. Medium Gaussian çekirdeği kullanan destek vektör makinesi 10 katlamalı çaprazlama tekniğinde %79.3 başarı oranına çıkmıştır. Eğitilen model test verisi için kullanıldığında %72.3 başarı oranı elde edilmiştir. Eğitilen modeller test verisi için kullanıldığında en başarılı beş sonucu 'Medium Gaussian SVM', 'Coarse Gaussian SVM', 'Quadratic SVM', 'Quadratic Discriminant' ve 'Coarse KNN' modelleri vermiştir. Bu beş modelin karışıklık matrisi (confusion matrix) sayı ve yüzde bazında hesaplanmıştır ve hesaplanan sonuçlar aşağıdaki tablolarda verilmiştir.

Tablo 3. 8 Medium Gaussian SVM karışıklık matrisi sonuçları

Doğru Sınıf	0	3908 %45.12	499 %5.76	648 %7.48	2134 %24.64	1471 %16.98
	1	150 %1.70	6310 %71.55	308 %3.49	834 %9.45	1217 %13.79
	2	125 %1.44	197 %2.27	6275 %72.60	1513 %17.50	533 %6.16
	3	61 %0.52	131 %1.11	492 %4.20	9710 %82.89	1319 %11.26
	4	65 %0.56	458 %3.97	217 %1.88	1404 %12.19	9367 %81.37
		0	1	2	3	4
		Tahmin Edilen Sınıf				

Tablo 3. 9 Coarse Gaussian SVM karışıklık matrisi sonuçları

Doğru Sınıf	0	3990 %46.07	338 %3.90	312 %3.60	2234 %25.79	1786 %20.62
	1	439 %4.97	5438 %61.66	269 %3.05	1180 %13.38	1493 %16.92
	2	407 %4.70	248 %2.86	5362 %62.03	1953 %22.59	673 %7.78
	3	168 %1.43	137 %1.16	302 %2.57	9516 %81.24	1590 %13.57
	4	169 %1.46	318 %2.76	135 %1.17	2146 %18.64	8743 %75.95
		0	1	2	3	4
		Tahmin Edilen Sınıf				

Tablo 3. 10 Quadratic SVM karışıklık matrisi sonuçları

Doğru Sınıf	0	3388 %39.12	713 %8.23	758 %8.75	1809 %20.88	1992 %23.00
	1	87 %0.98	6812 %77.24	437 %4.95	517 %5.86	966 %10.95
	2	60 %0.69	373 %4.31	6698 %77.49	953 %11.02	559 %6.46
	3	77 %0.65	399 %3.40	795 %6.78	8976 %76.63	1466 %12.51
	4	54 %0.46	659 %5.72	491 %4.26	1148 %9.97	9159 %79.56
		0	1	2	3	4
		Tahmin Edilen Sınıf				

Tablo 3. 11 Quadratic Discriminant karışıklık matrisi sonuçları

Doğru Sınıf	0	2031 %23.45	2181 %25.18	1856 %21.43	1483 %17.12	1109 %12.80
	1	3 %0.03	7355 %83.39	886 %10.04	113 %1.28	462 %5.23
	2	8 %0.09	692 %8.00	7258 %83.97	482 %5.57	203 %2.34
	3	12 %0.10	1397 %11.92	1306 %11.15	8302 %70.87	696 %5.94
	4	10 %0.08	1430 %12.42	1283 %11.14	679 %5.89	8109 %70.44
		0	1	2	3	4
		Tahmin Edilen Sınıf				

Tablo 3. 12 Coarse KNN karışıklık matrisi sonuçları

Doğru Sınıf	0	5903 %68.16	530 %6.12	463 %5.34	954 %11.01	810 %9.35
	1	948 %10.74	5561 %63.05	443 %5.02	696 %7.89	1171 %13.27
	2	915 %10.58	513 %5.93	5297 %61.28	1357 %15.70	561 %6.49
	3	757 %6.46	512 %4.37	570 %4.86	8489 %72.47	1385 %11.82
	4	667 %5.79	587 %5.09	448 %3.89	1613 %14.01	8196 %71.20
		0	1	2	3	4
		Tahmin Edilen Sınıf				

Tablo 3.8’de verilen Medium Gaussian SVM karışıklık matrisi sonuçlarında en başarılı sınıf 3’tür. Sınıfı 0 olan verilerin başarı oranı görece daha düşük çıkmıştır ve bu veriler daha çok 3.sınıfa ait olarak tahmin edilmiştir. Tablo 3.9’da verilen Coarse Gaussian SVM’ye ait karışıklık matrisi için de benzer sonuçlar elde edilmiştir. Tablo 3.10’da verilen Quadratic SVM karışıklık matrisine göre en başarılı sınıf 4 iken, en başarısız sınıf yine 0 olmuştur. Tablo 3.11’de verilen Quadratic Discriminant karışıklık matrisi sonuçlarında en başarılı sınıf 2 olup, bu sefer 0.sınıfa ait veriler en çok 2.sınıfla karışmaktadır. Tablo 3.12’de verilen sonuçlar, Coarse KNN yönteminin tüm sınıflar için daha dengeli doğruluk oranları verdiğini göstermektedir.

4.1 Sonuç ve Öneriler

Bu çalışmada, iki boyutlu gerçekçi sayısal meme modelleri kullanılarak, moment yöntemi aracılığıyla düz saçılma problem çözülmüş ve saçılan alanlar sayısal olarak hesaplanmıştır. Farklı boyut ve konumlardaki tümörleri içeren veya tümör içermeyen meme modelleri kullanılarak sentetik olarak üretilen saçılan alanlar eğitim ve test veri setlerini oluşturmaktadır. Elde edilen sonuçlar; karar ağaçları, diskriminant analizi, destek vektör makinesi, Naive Bayes, topluluk sınıflandırıcıları, yapay sinir ağları ve KNN gibi çeşitli makine öğrenmesi algoritmalarının, tümörlerin hem varlığının tespiti hem de konumlarının belirlenmesi açısından, farklı meme modelleri için anlamlı sonuçlar verdiğini göstermektedir. Kullanılan algoritmalar arasında en başarılı sonuç veren destek vektör makinesi olmuştur. Kullanılan algoritmaların eğitim aşamasında başarı oranları yüksek olup, test aşamasında başarı oranlarının eğitim aşamasına göre azaldığı durumlar da olmuştur. Ayrıca, karar ağaçlarında TBA kullanılabileceği görülmüştür. Diğer makine öğrenmesi algoritmaları kullanıldığında ise başarı oranları düşmüştür.

Gelecekteki çalışmalarda meme modellerinin şekil bilgileri ve çeşitli meme tipleri öznitelik olarak veri setine eklendiğinde başarı oranlarının ciddi şekilde artacağı düşünülebilir. Diğer yandan, tümörlerin konumu meme modelinde sınıfları ayıran sınıra yakın olduğunda sınıflandırma hataları artacaktır. Bu durumda tümörlerin konumunu gösteren sınıf sayısı artırılabilir. Böylece tümör konumunu tespit etmede yöntemler daha hassas bir öğrenme gerçekleştirebilir.

Eğitim için ayrılan veri setinde sınıfları 3 ve 4 olan veri sayısı görece fazla olduğu için, yöntemlerin bu sınıfları tahmin etme yönünde bir eğilimi olmuştur. Veri setindeki dengesizlikten kaynaklanan bu sorunu çözmek için diğer sınıfların yanlış bulma maliyeti artırılarak modeller eğitilebilir. Ayrıca sınıfı 0 olan veriler, yani tümörün bulunmadığı durumlara ait veriler için sınıflandırma başarısı Coarse KNN yöntemi dışında düşük çıkmıştır. Sınıfı 0 olan verilere ait başarıyı artırmak için,

nce tmrn varlıęı-yokluęu řeklinde ikili bir sınıflandırma yapılabilir. Bylece tmrn var olarak sınıflandırıldıęı durumlar iin, ikinci ařamada, sadece konuma ait bir sınıflandırma yapmak mmkndr.



- [1] Lazebnik, M., Popovic, D., McCartney, L. ve Watkins, B., (2007). "A large-scale study of the ultrawideband microwave dielectric properties of normal, benign, and malignant breast tissues obtained from cancer surgeries", *Physics in Medicine and Biology*, 20:6093-6115.
- [2] Lazebnik, M., Popovic, D., Lindstrom, J., McCartney, L. ve Watkins, B., (2007). "A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries", *Physics in Medicine and Biology*, 10:2637-2656.
- [3] Fear, E., Li, X., Hagness, S. ve Stuchly, M., (2002). "Confocal microwave imaging for breast cancer detection: Localization of tumors in three dimensions", *IEEE Trans. Biomed.*, 8:812-822.
- [4] Kurrant, D.J., Fear, C. ve Westwick, T., (2008). "Tumor response estimation in radar based microwave breast cancer detection", *IEEE Trans. Biomed.*, 12: 2801-2822.
- [5] Klemm, M., Craddock, I., Leendertz, J. ve Benjamin, R., (2008). "Experimental and clinical results of breast cancer detection using UWB microwave radar", *IEEE Antennas Propag Soc. Int. Symp.*, 4:916-925.
- [6] Meaney, P.M., Fanning, M.W., Poplack, S.P. ve Paulsen, K.D., (2000). "A clinical prototype for activemicrowave imaging of the breast", *IEEE Trans.Microw. Theory Tech.*, 11:1841-1853.
- [7] Kruger, A., Kopecky, K., Reinecke, D.R., Aisen, A.M. ve Kiser, W.L., (1999). "A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries", *Physics in Medicine and Biology*, 1:275-278.
- [8] Wang, L.W., Zho, X., Sun, H. ve Ku, G., (1999). "Microwave-induced acoustic imaging of biological tissues", *Review of Scientific Instruments*, 9:3744-3748.
- [9] Gürbüz, T.U., Aslanyürek, B., Yapar, A., Şahintürk, H. ve Akduman, I., (2014). "A Nonlinear Microwave Breast Cancer Imaging Approach Through Realistic Body–Breast Modeling," in *IEEE Transactions on Antennas and Propagation*, sayı. 62, no. 5, ss. 2596-2605
- [10] Li, D., Meaney, P. M. ve Paulsen, K. D., (2003). "Conformal microwave imaging for breast cancer detection," *IEEE Trans. Microw. Theory Tech.*, sayı. 51, no. 4, ss. 1179–1186.

- [11] Conceição, R. C., Medeiros, H., O'Halloran, M., Rodriguez-Herrera, D., Flores-Tapia, D., Pistorius, S., (2013). "Initial classification of breast tumour phantoms using a UWB radar prototype", International Conference on Electromagnetics in Advanced Applications (ICEAA), 55(8):2011-21.
- [12] Conceição, R., O'Halloran, M., ve Jones, E., (2011). "EVALUATION OF FEATURES AND CLASSIFIERS FOR CLASSIFICATION OF EARLY-STAGE BREAST CANCER", Journal of Electromagnetic Waves and Applications, ss. 1-14.
- [13] Conceição, R., O'Halloran, M., Byrne, D., Jones, E. ve Glavin, M., (2010). "Tumor Classification Using Radar Target Signatures", PIERS Proceedings, ss. 1-14.
- [14] Conceição, R., O'Halloran, E. Jones ve Glavin, M., (2010). "Support vector machines for the classification of early-stage breast cancer based on radar target signatures", Progress In Electromagnetics Research B, Sayı. 23, 311-327.
- [15] Conceição, R., Medeiros, H., O'Halloran, M., Rodriguez-Herrera, D., Flores-Tapia, D. ve Pistorius, S., (2014). "SVM-Based Classification of Breast Tumour Phantoms Using a UWB Radar Prototype System", IEEE.
- [16] Conceição, R ve Gerazov, B., (2017). "Deep learning for tumour classification in homogeneous breast tissue in medical microwave imaging", IEEE.
- [17] Byrne, D., O'Halloran, M., Jones, E. ve Glavin, M., (2011). "BREAST CANCER DETECTION BASED ON DIFFERENTIAL ULTRAWIDEBAND MICROWAVE RADAR", Progress in Electromagnetics Research M, 231-242.
- [18] Santorelli, A., Porter, E., Kirshin, E., Liu, Y.J. ve Popovic, M., (2014). "Investigation of Classifiers for Tumor Detection with an Experimental Time-Domain Breast Screening System", Progress In Electromagnetics Research, Sayı. 144, 45-57.
- [19] Byrne, D., O'Halloran, M., Jones, E. ve Glavin, M., (2011). "SUPPORT VECTOR MACHINE-BASED ULTRAWIDEBAND BREAST CANCER DETECTION SYSTEM", Journal of Electromagnetic Waves and Applications, ss. 1807-1816.
- [20] Conceição, R., Byrne, D., Noble, J.A., Craddock, J., (2016). "Initial Study for the Investigation of Breast Tumour Response with Classification Algorithms Using a Microwave Radar Prototype", IEEE.
- [21] Li, Y., Santorelli, A., Laforest, O. ve Coates, M. (2015). "COST-SENSITIVE ENSEMBLE CLASSIFIERS FOR MICROWAVE BREAST CANCER DETECTION", IEEE.

- [22] Çınar, Ç. (2015). "MİKRODALGA DOKU ANALİZİ İLE MEME KANSERİ TEŞHİSİ".
- [23] Kerhet, A., Raffetto, M., Boni, A. ve Massa, A., (2006). "A SVM-based approach to microwave breast cancer detection", Engineering Applications of Artificial Intelligence, sayfa 807–818.
- [24] Jacobs, P., (1999). "Data Mining: What General Managers Need to Know", Harvard Management Update, Cilt 4, No 10, 8.
- [25] Doğan, Ş., ve Türkoğlu, İ., (2007). "Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees", Doğu Anadolu Bölgesi Araştırmaları Dergisi, Cilt 5, No 2, 163-169.
- [26] Hand, D.J., (1998). "Data Mining: Statistics and More?", The American Statistician, Cilt 52, 112-118.
- [27] Kitler R. ve Wang W., (1998). "The Emerging Role of Data Mining", Solid State Technology, Cilt 42, No 11, 45.
- [28] Kaya, E., Bulun, M. ve Arslan, A., (2009). "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları", Akademik Bilişim 2005, Adana.
- [29] Serkan Savaş, N. T. ve M. Y. (2012). "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 1-23.
- [30] Baykasoğlu, A., (2003). "Veri Madenciliği ve Çimento Sektöründe Bir Uygulama", Akademik Bilişim 2005, Gaziantep.
- [31] Shearer, C., (2000). "The Crisp-DM Model: The New Blueprint for Data Mining" Journal of Data Warehousing, Cilt 5 No 4, 13-23.
- [32] Cios, K.J ve Moore, G.W., (2002). "Uniqueness of Medical Data Mining" Intelligence in Medicine journal, 26: 1-24.
- [33] "UWCEM numerical breast phantom repository," Comput. Electromagn. Lab., Univ. Wisconsin, (accessed: 20 January 2022). [Online]. Available: <http://uwcem.ece.wisc.edu>
- [34] E. Zastrow, S. K. Davis, M. Lazebnik, F. Kelcz, B. D. Van Veen, ve S. C. Hagness, "Development of anatomically realistic numerical breast phantoms with accurate dielectric properties for modeling microwave interactions with the human breast," IEEE Trans. Biomed. Eng., sayı. 55, no. 12, ss. 2792–2800, Dec. 2008.

TEZDEN ÜRETİLMİŞ YAYINLAR

- [1] Tölük E., Aslanyürek B., (2021). “Mikrodalga Meme Kanserinin Tespiti ve Konumunun Belirlenmesi için Makine Öğrenmesi Algoritmalarının İncelenmesi”, VI. International Scientific and Vocational Studies Congress – Engineering, ss.9-10.

