



**KÜTLE SPEKTROMETRESİ TEMELLİ  
KLİNİKSEL GLİKOMİK ÇALIŞMALARINA  
YÖNELİK BİR MAKİNE ÖĞRENMESİ  
UYGULAMASI**

**Deniz Baran DEMİRHAN**

**2022  
YÜKSEK LİSANS TEZİ  
BİYOMEDİKAL MÜHENDİSLİĞİ**

**Tez Danışmanı  
Dr. Öğr. Üyesi Hacı Mehmet KAYILI**

**KÜTLE SPEKTROMETRESİ TEMELLİ KLİNİKSEL GLİKOMİK  
ÇALIŞMALARINA YÖNELİK BİR MAKİNE ÖĞRENMESİ UYGULAMASI**

**Deniz Baran DEMİRHAN**

**T.C.  
Karabük Üniversitesi  
Lisansüstü Eğitim Enstitüsü  
Biyomedikal Mühendisliği Anabilim Dalında  
Yüksek Lisans Tezi  
Olarak Hazırlanmıştır**

**Tez Danışmanı  
Dr. Öğr. Üyesi Hacı Mehmet KAYILI**

**KARABÜK  
Haziran 2022**

Deniz Baran DEMİRHAN tarafından hazırlanan “KÜTLE SPEKTROMETRESİ TEMELLİ KLİNİKSEL GLİKOMİK ÇALIŞMALARINA YÖNELİK BİR MAKİNE ÖĞRENMESİ UYGULAMASI” başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Dr. Öğr. Üyesi Hacı Mehmet KAYILI .....  
Tez Danışmanı, Biyomedikal Mühendisliği Anabilim Dalı

Dr. Öğr. Üyesi Hakan YILMAZ .....  
İkinci Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından Oy birliği ile Biyomedikal Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 17/06/2022

Ünvanı, Adı SOYADI (Kurumu) İmzası

Başkan : Prof. Dr. Bekir SALİH (HÜ) Çevrimiçi

Üye : Dr. Öğr. Üyesi Hacı Mehmet KAYILI (KBÜ) .....

Üye : Dr. Öğr. Üyesi Hakan YILMAZ (KBÜ) .....

Üye : Dr. Öğr. Üyesi Mehmet ATAKAY (HÜ) Çevrimiçi

Üye : Dr. Öğr. Üyesi Daver ALİ (KBÜ) .....

KBÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Prof. Dr. Hasan SOLMAZ .....  
Lisansüstü Eğitim Enstitüsü Müdürü

*“Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim.”*

Deniz Baran DEMİRHAN

## ÖZET

**Yüksek Lisans Tezi**

### **KÜTLE SPEKTROMETRESİ TEMELLİ KLİNİKSEL GLİKOMİK ÇALIŞMALARINA YÖNELİK BİR MAKİNE ÖĞRENMESİ UYGULAMASI**

**Deniz Baran DEMİRHAN**

**Karabük Üniversitesi**

**Lisansüstü Eğitim Enstitüsü**

**Biyomedikal Mühendisliği Anabilim Dalı**

**Tez Danışmanı:**

**Dr. Öğr. Üyesi Hacı Mehmet KAYILI**

**Dr. Öğr. Üyesi Hakan YILMAZ**

**Haziran 2022, 55 Sayfa**

Mide kanseri dünyada en çok görülen kanser türlerinden biridir. Mide kanseri dokuları üzerinde *N*-glikozilasyon yapılarının incelenip makine öğrenmesi algoritmaları yöntemleriyle değerlendirilmesine yönelik literatürdeki çalışma sayısı çok azdır. Bu tez kapsamında, mide kanseri glikan profilleri MALDI-MS ile incelenip oluşturulan veri setinin makine öğrenmesi algoritmalarıyla değerlendirilmesi amaçlanmıştır. Bu kapsamda Karabük Eğitim ve Araştırma Hastanesi Patoloji bölümünden parafilm ile kaplanmış mide kanseri ve kontrol dokuları alınıp bu çalışmada kullanılmıştır.

İlk başta dokulardan parafilm uzaklaştırılmıştır. Dokulardan proteinlerin ekstraksiyonu kloform/methanol yöntemi ile gerçekleştirilmiştir. Ekstrakte edilen glikoproteinlerden *N*-glikanların salınımı sonrası glikanlar 2-AA ile etiketlenmiştir. Sırasıyla selüloz ve gözenekli grafitize karbon katı faz kartuşları ile 2-AA etiketli glikanların saflaştırılmaları yapılmıştır ve 2-AA ile etiketlenmiş *N*-glikanların MALDI-MS analizleri gerçekleştirilmiştir. Analizler sonucu 59 adet *N*-glikan yapısı tespit edilmiştir. Daha sonra *N*-glikanların yapısal doğrulamaları tandem MS/MS analizleri ile gerçekleştirilmiştir ve yapısal özelliklerine göre glikanlar gruplandırılmıştır. Bu analizler sonucunda tespit edilen glikanların göreceli alanları ve analit alanları hesaplanmıştır. Elde edilen veriler makine öğrenmesi algoritmaları için düzenlenerek veri seti haline getirilmiştir. Nötr *N*-glikanlar için göreceli ve analit alanı olarak 2 ayrı veri seti ve tüm *N*-glikanlar için göreceli ve analit alanı olarak 2 ayrı veri seti ve sonuç olarak toplamda 4 farklı veri seti oluşturulmuştur. Her bir kanser ve kontrol örneğine ait maksimum 4 teknik tekrardan elde edilen veriler analizlere dahil edilmiştir. KNN, RF, DT, SVM, LR ve MLPC algoritmaları ile model değerlendirme parametreleri kullanılarak verilerden anlamlı sonuçlar elde edilmiştir. Bu bağlamda uygulanan algoritmalar ve model değerlendirmeleri ile *N*-glikanlarda kanser ve kontrol dokuları arasında anlamlı farklılıklar olduğu görülmüştür. MLPC algoritması her veri setinde en yüksek doğruluk skoruna sahip olan algoritma olmuştur. Nötr *N*-glikanlar göreceli alan veri setinde MLPC algoritmasının doğruluk skoru %93.6+-0.8 ve ROC analizi sonrası AUC skoru 0.99 olarak bulunmuştur.

**Anahtar Sözcükler :** Mide kanseri, *N*-Glikan, glikozilasyon, kütle spektrometresi, MALDI, makine öğrenmesi.

**Bilim Kodu :** 92504

## **ABSTRACT**

**M. Sc. Thesis**

### **A MACHINE LEARNING APPLICATION FOR MASS SPECTROMETRY BASED CLINICAL GLYCOMIC STUDIES**

**Deniz Baran DEMİRHAN**

**Karabük University**

**Institute of Graduate Programs**

**Department of Biomedical Engineering**

**Thesis Advisor:**

**Asist. Prof. Dr. Hacı Mehmet KAYILI**

**Asist. Prof. Dr. Hakan YILMAZ**

**June 2022, 55 Pages**

Stomach cancer is one of the most common types of cancer in the world. The number of studies in the literature on the examination of *N*-glycosylation structures on gastric cancer tissues and their evaluation with machine learning algorithms is very few. Within the scope of this thesis, it is aimed to evaluate the data set created by examining the gastric cancer glycan profiles with MALDI-MS with machine learning algorithms. In this context, parafilm-embedded gastric cancer and control tissues were taken from the Pathology Department of Karabuk Training and Research Hospital and used in this study.

Initially, parafilm was removed from the tissues. Extraction of proteins from tissues was carried out with the chloroform/methanol extraction method. After the release of *N*-glycans from the extracted glycoproteins, the glycans were labeled with 2-AA. Purification was achieved with cellulose and porous graphitized carbon-containing solid-phase cartridges, respectively. MALDI-MS analyzes of 2-AA-labeled *N*-glycans were performed. As a result of the analysis, 59 *N*-glycans were detected. Then, structural confirmations of *N*-glycans were performed by tandem MS/MS analyzes and grouped according to their structural properties. The relative areas and analyte areas of the glycans detected as a result of these analyzes were calculated. The obtained data were arranged for machine learning algorithms and turned into a suitable data set for analysis. For neutral *N*-glycans, two separate data sets as relative and analyte domains, and two separate data sets as relative and analyte domains for all *N*-glycans, and as a result, four different data sets in total were created. Data obtained from a maximum of 4 technical replicates of each cancer and control sample were included in the analyses. Significant results were obtained from the data by using KNN, RF, DT, SVM, LR, and MLPC algorithms and model evaluation parameters. In this context, it has been observed that there are significant differences between cancer and control tissues in *N*-glycans with applied algorithms and model evaluations. The MLPC algorithm was the algorithm with the highest accuracy score in each data set. In the neutral *N*-glycans relative area dataset, the accuracy score of the MLPC algorithm was 93.6%±0.8%, and after the ROC analysis, the AUC score was found to be 0.99.

**Key Word** : Stomach cancer, *N*-glycan, glycosylation, mass spectrometry, MALDI, machine learning.

**Science Code** : 92504



## TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın danışman hocalarım Dr. Öğr. Üyesi Hacı Mehmet KAYILI ve Dr. Öğr. Üyesi Hakan YILMAZ'a sonsuz teşekkürlerimi sunarım.

Kanserli dokuların temin edilmesinde yardımcı olan Karabük Üniversitesi Eğitim ve Araştırma Hastanesi Patoloji birimi çalışanlarına ve Dr. Harun Erol hocama teşekkür ederim. Çalışmada standart MALDI analizlerinin elde edilmesinde Hacettepe Üniversitesi, Kimya Bölümü, Kütle spektrometresi Araştırma Laboratuvar imkanlarını kullanmamızı sağlayan değerli hocam Prof. Dr. Bekir Salih'e, aynı laboratuvarda analizlerde yardımcı olan başta Serhat Öztürk olmak üzere tüm araştırmacılara teşekkür ederim.

Bu tez çalışması, Karabük Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü tarafından FYL-2020-2243 kodlu tez projesi kapsamında desteklenmiştir. Bundan dolayı Karabük Üniversitesi BAP Koordinatörlüğü'ne ve Karabük Üniversitesi'ne teşekkür ederim.

Sevgili aileme manevi hiçbir yardımı esirgmeden her daim yanımda oldukları için tüm kalbimle teşekkür ederim.

## İÇİNDEKİLER

|   | <u>Sayfa</u> |
|---|--------------|
| KABUL.....  | ii           |
| ÖZET.....   | iv           |
| ABSTRACT.....   | vi           |
| TEŞEKKÜR.....   | viii         |
| İÇİNDEKİLER .....   | ix           |
| ŞEKİLLER DİZİNİ.....  | xii          |
| ÇİZELGELER DİZİNİ .....   | xiv          |
| SİMGELER VE KISALTMALAR DİZİNİ .....                                  | xv           |
| <br>  |              |
| BÖLÜM 1 .....   | 1            |
| GİRİŞ .....   | 1            |
| <br>  |              |
| BÖLÜM 2 .....   | 4            |
| 2.1. POST-TRANSLASYONEL MODİFİKASYONLAR.....                          | 4            |
| 2.2. PROTEİN GLİKOZİLASYONU VE TIPLERİ.....                           | 5            |
| 2.2.1. N-Bağlı Glikozilasyon .....                                    | 6            |
| 2.2.2. O-Bağlı Glikozilasyon.....                                     | 7            |
| 2.3. MİDE KANSERİ VE GÖRÜLME SIKLIĞI.....                             | 8            |
| 2.3.1. Kanser ve Glikozilasyon İlişkisi.....                          | 9            |
| 2.4. KÜTLE SPEKTROMETRESİ.....  | 12           |
| 2.4.1. İyonlaştırma Teknikleri .....                                  | 13           |
| 2.4.1.1. Elektrosprey İyonlaştırma.....                               | 14           |
| 2.4.1.2. Matris Destekli Lazer Desorpsiyonu / İyonizasyonu (MALDI)... | 15           |
| 2.4.2. Kütle Analizörleri .....                                       | 16           |
| 2.4.2.1 Uçuş Zamanlı Kütle Analizörü (TOF–Time-of-Flight).....        | 17           |
| 2.5. MAKİNE ÖĞRENMESİ.....  | 18           |
| 2.5.1. Gözetimli Öğrenme .....  | 19           |
| 2.5.2. Gözetimsiz Öğrenme .....                                       | 19           |

|  | <u>Sayfa</u> |
|--|--------------|
| 2.5.3. Takviyeli Öğrenme .....   | 19           |
| 2.6. K-EN YAKIN KOMŞU ALGORİTMASI (KNN, K-NEAREST<br>NEIGHBORS ALGORITHM).....                   | 20           |
| 2.7. DECISION TREE (KARAR AĞAÇLARI).....   | 21           |
| 2.8. RANDOM FORESTS (RASSAL - RASTGELE ORMANLAR) .....   | 23           |
| 2.9. SUPPORT VECTOR MACHINES (SVM) .....   | 24           |
| .....  | 25           |
| 2.10. LOJİSTİK REGRESYON .....   | 25           |
| 2.11. MLP.....   | 26           |
| 2.12. MODELİN DEĞERLENDİRİLMESİ (SINIFLANDIRILMANIN<br>DEĞERLENDİRİLMESİ).....                   | 27           |
| 2.12.1 Karışıklık Matrisi (Confusion Matrix) .....   | 27           |
| 2.12.2. ROC ve AUC .....   | 28           |
| <br>BÖLÜM 3 .....  | <br>30       |
| 3.1. KULLANILAN MALZEMELER.....  | 30           |
| 3.2. DOKULARDAN PARAFİNİN UZAKLAŞTIRILMASI.....  | 30           |
| 3.3. PROTEİNLERİN EKSTRAKSİYONU (KLOROFORM/METHANOL<br>EKSTRAKSİYON YÖNTEMİ) .....               | 31           |
| 3.4. N-GLİKANLARIN SALINIMI .....  | 31           |
| 3.5. 2-AA İLE N-GLİKANLARIN ETİKETLENMESİ.....   | 31           |
| 3.6. SELÜLOZ İÇEREN KATI FAZ KARTUŞLAR İLE 2-AA ETİKETLİ N-<br>GLİKANLARIN SAFLAŞTIRILMASI ..... | 32           |
| 3.7. GÖZENEKLİ GRAFİTİZE KARBON KATI FAZ EKSTRAKSİYONU İLE<br>SAFLAŞTIRMA (PGC- SPE) .....       | 32           |
| 3.8. MALDI-MS ANALİZLERİ .....   | 32           |
| 3.9. MALDI-MS VERİ ANALİZLERİ.....   | 33           |
| 3.10. MAKİNE ÖĞRENMESİ.....  | 33           |
| <br>BÖLÜM 4 .....  | <br>34       |
| 4.1. DOKU PROTEİNLERİNE AİT N-GLİKANLARIN MALDI-MS İLE<br>ANALİZLERİ .....                       | 34           |

|  |    |
|--|----|
| 4.2. <i>N</i> -GLİKAN VERİLERİ ÜZERİNE MAKİNE ÖĞRENMESİ                              |    |
| ALGORİTMALARININ TEST EDİLMESİ .....   | 39 |
| 4.2.1. Tüm <i>N</i> -Glikanlar Analit Alan Veri Setinin Makine Öğrenmesi Analizleri  |    |
| .....  | 41 |
| 4.2.2. Tüm <i>N</i> -Glikanlar Göreceli Alan Veri Setinin Makine Öğrenmesi           |    |
| Analizleri .....   | 43 |
| 4.2.3. Nötr <i>N</i> -Glikanlar Analit Alan Veri Setinin Makine Öğrenmesi Analizleri |    |
| .....  | 44 |
| 4.2.4. Nötr <i>N</i> -Glikanlar Göreceli Alan Veri Setinin Makine Öğrenmesi          |    |
| Analizleri .....   | 46 |
| <br>   |    |
| BÖLÜM 5 .....  | 48 |
| SONUÇLAR VE ÖNERİLER .....   | 48 |
| KAYNAKLAR .....  | 50 |
| <br>   |    |
| EK AÇIKLAMALAR A. ETİK KURUL İZİNİ .....   | 53 |
| <br>   |    |
| ÖZGEÇMİŞ .....   | 55 |

## ŞEKİLLER DİZİNİ

### Sayfa

|  |    |
|--|----|
| Şekil 2.1. Bazı post-translasyonel modifikasyonların gösterimi .....   | 4  |
| Şekil 2.2. Proteinlerde <i>N</i> -bağlı glikolizasyon tipi .....   | 6  |
| Şekil 2.3. Proteinlerde <i>O</i> -bağlı glikozilasyon tipi .....   | 6  |
| Şekil 2.4. <i>N</i> -glikan tipleri .....  | 7  |
| Şekil 2.5. <i>O</i> -glikan tipleri .....  | 8  |
| Şekil 2.6. Normal hücrelerde ve kanserli hücrelerdeki glikan değişimleri .....                                 | 10 |
| Şekil 2.7. <i>N</i> - ve <i>O</i> - glikanların ve Lewis antijenlerinin özel yapıları.....                     | 11 |
| Şekil 2.8. Kütle spektrometresi şeması.....  | 13 |
| Şekil 2.9. Elektrosprey İyonizasyon işleminin şeması .....   | 14 |
| Şekil 2.10. Elektrosprey İyonlaştırma mekanizması .....  | 15 |
| Şekil 2.11. MALDI kaynağının şematik gösterimi .....   | 16 |
| Şekil 2.12. Kütle spektrometresinde kullanılan analizör çeşitleri .....  | 17 |
| Şekil 2.13. MALDI uçuş zamanlı kütle analizörü çalışma prensibi.....   | 18 |
| Şekil 2.14. Gözetimli Öğrenme Algoritmaları .....  | 20 |
| Şekil 2.15. K-En Yakın Komşu Algoritması.....  | 21 |
| Şekil 2.16. Karar ağaçları şeması .....  | 22 |
| Şekil 2.17. RF'nin çalışma prensibi .....  | 23 |
| Şekil 2.18. SVM .....  | 24 |
| Şekil 2.19. Çok katmanlı ağlar .....   | 26 |
| Şekil 2.20. Aktivasyon fonksiyonu .....  | 27 |
| Şekil 2.21. Confusion Matrix .....   | 28 |
| Şekil 2.22. ROC eğrisi.....  | 29 |
| Şekil 4.1. Kanser (A) ve Kontrol (B) dokularına ait MALDI-MS Spektrumu.....                                    | 34 |
| Şekil 4.2. MALDI MS/MS analizleri sonucu yapıları doğrulanmış glikanlar (A) H5N4S1, (B) H8N2, (C) H7N6F1. .... | 35 |
| Şekil 4.3. Karışıklık Matrisi Oranları (TPR, TNR, FNR ve FPR). ....  | 42 |
| Şekil 4.4. ROC Analizi.....  | 42 |
| Şekil 4.5. Karışıklık Matrisi Oranları.....  | 43 |
| Şekil 4.6. ROC Analizi.....  | 44 |
| Şekil 4.7. Karışıklık Matrisi Oranları.....  | 45 |

|   | <b><u>Sayfa</u></b> |
|---|---------------------|
| Şekil 4.8. ROC Analizi.....                 | 46                  |
| Şekil 4.9. Karışıklık Matrisi Oranları..... | 47                  |
| Şekil 4.10. ROC Analizi.....                | 47                  |
| Şekil Ek A.1. Etik Kurul İzni.....          | 54                  |

## ÇİZELGELER DİZİNİ

### Sayfa

|  |    |
|--|----|
| Çizelge 4.1. Mide kanseri dokularına ait tespit edilen <i>N</i> -glikanlar.....              | 36 |
| Çizelge 4.2. Tüm <i>N</i> -glikanlar analit alanı, “1” olarak işaretlenmiş kanserli dokular. | 40 |
| Çizelge 4.3. Tüm <i>N</i> -glikanlar analit alanı, “0” olarak işaretlenmiş kontrol dokular.  | 40 |
| Çizelge 4.4. Karışıklık Matrisi (TP, FN, FP, TN).....  | 41 |
| Çizelge 4.5. Karışıklık Matrisi (TP, FN, FP, TN).....  | 43 |
| Çizelge 4.6. Karışıklık Matrisi (TP, FN, FP, TN).....  | 44 |
| Çizelge 4.7. Karışıklık Matrisi (TP, FN, FP, TN).....  | 46 |

## SİMGELER VE KISALTMALAR DİZİNİ

### SİMGELER

mL : Mililitre

mg : Miligram

m/z : Kütle – Yük oranı

Da : Dalton

H : Hidrojen

dH<sub>2</sub>O: Deiyonize su



## KISALTMALAR

|          |  |
|----------|--|
| PTM      | : Post-Translasyonel Modifikasyon  |
| MS       | : Mass Spectrometry (Kütle Spektrometresi)   |
| HPLC     | : High Performance Liquid Chromatography (Yüksek Performanslı Sıvı Kromatografisi) |
| LC-MS/MS | : Sıvı Kromatografisi-Tandem Kütle Spektrometrisi                                  |
| QTOF-MS  | : Dört Kutuplu Uçuş Zamanlı Kütle Spektrometresi                                   |
| MALDI    | : Matriks Yardımlı Lazer Desorpsiyon/İyonlaştırma                                  |
| GC       | : Gas Chromatography (Gaz Kromatografisi)  |
| ESI      | : Elektrospray Ionization (Elektrosprey İyonizasyon)                               |
| SNR      | : Signal to Noise Ratio (Sinyal Gürültü Oranı)                                     |
| TOF      | : Time-of-flight (Uçuş-zamanlı)  |
| Asn      | : Asparajin  |
| DHB      | : Dihidroksibenzoik asit   |
| Dol-P    | : Dolikol-Fosfat   |
| DMSO     | : Dimetil Sulfoksit  |
| Ser      | : Serin  |
| Thr      | : Treonin  |
| 2-AA     | : 2-Aminobenzoik Asit  |
| 2-AB     | : 2-Aminobenzamid  |
| KNN      | : K-En Yakın Komşu Algoritması   |
| RF       | : Random Forest (Rassal Ormanlar)  |
| SVM      | : Support Vector Machines (Destek Vektör Makineleri)                               |
| LR       | : Lojistik Regresyon   |
| MLP      | : Multi Layer Perceptron (Çok Katmanlı Algılayıcı)                                 |
| MLPC     | : Çok Katmanlı Algılayıcı Sınıflandırıcı   |
| ROC      | : Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)                 |
| AUC      | : Area Under Curve (Eğri Altındaki Alan)   |
| TP       | : True Positive (Doğru Pozitif)  |
| FP       | : False Positive (Yanlış Pozitif)  |
| TN       | : True Negative (Doğru Negatif)  |
| FN       | : False Negative (Yanlış Negatif)  |

TPR : True Positive Rate (Dođru Pozitif Oranı)  
TNR : True Negative Rate (Dođru Negatif Oranı)  
FPR : False Positive Rate (Yanlıř Pozitif Oranı)  
FNR : False Negative Rate (Yanlıř Negatif Oranı)  
YSA : Yapay Sinir Ađı

## BÖLÜM 1

### GİRİŞ

Kanser, bakıldığı zaman temelde hem genetik problemlerden hem de çevresel etkilerden dolayı hücrelerin anormal olarak çoğalmasıdır. Metastaz dolayısıyla da diğer doku ve organlara yayılıp çoğalmaktadır. Mide dokularında da kanser belirtileri görülebilmektedir. Mide kanseri dünya genelinde ölüm nedenleri arasında ilk sıralarda yer almaktadır. Bilim insanları mide kanseri de dahil olmak üzere tüm kanser türlerinin ölümcül etkilerini azaltmak için birçok farklı tanı ve tedavi geliştirmiştir ve geliştirmeye de devam etmektedirler. Bu çözüm yollarından biri de biyobelirteç olarak değerlendirilen moleküllerden olan glikanlardır.

Şeker parçalarının proteinlere bağlanması olan glikozilasyon, diğer post-translasyonel modifikasyonlardan daha fazla proteomik çeşitlilik sağlayan bir post-translasyonel modifikasyondur. Glikozilasyon, hücre dışı matrise hücre bağlanması ve hücre içindeki protein-ligand etkileşimleri dahil olmak üzere çok çeşitli biyolojik süreçler için kritiktir. Glikozilasyon sırasında meydana gelen değişimlerin tespiti edilmesi önem arz etmektedir. Protein yapılarında anormal şekilde meydana gelen glikozilasyonun, kanser, genetik hastalıklar ve bağışıklık sistemi hastalıklarıyla ilişkili olduğu bildirilmiştir. Glikanların analizlerinde en çok kullanılan tekniklerden biri kütle spektrometrisidir.

Kütle spektrometrisi atomlardan, moleküllerden veya bunların karışımından iyonlar meydana getiren ve sonra da bu iyonları kütle/yük ( $m/z$ ) oranına göre ayırıp ayrı ayrı saptayan bir cihazdır. Kütle spektrometrisi cihazları kütle ayırıcılığına, doğruluğa ve hassasiyete sahiptir. Bu özelliklere sahip olmaları onları glikoproteomik ve glikomik çalışmalarda vazgeçilmez kılmaktadır. Kütle spektrometrisi kullanılarak yapılan çalışmalar her glikozilasyon bölgesindeki glikan mikro çeşitliğini ortaya çıkarmaktadır.

MALDI (Matrix-Assisted Laser Desorption Ionization-Matriks-yardımlı Lazer Desorpsiyon/iyonlaştırma), MS bazlı proteomikte anahtar bir tekniktir. MALDI-MS glikanların analizlerinde sıklıkla kullanılmaktadır. Minimum parçalanma ile büyük moleküllerden iyonlar oluşturmak için bir lazer enerjisi kullanan iyonlaştırma tekniğidir.

Yapay Zekâ (Artificial Intelligence), insanlar gibi çalışan ve tepki veren akıllı makinelerin oluşturulmasını vurgulayan bir bilgisayar bilimi alanıdır. Makine öğrenimi (Machine Learning) ise, bilgisayarların geçmiş örneklerden öğrenmesini ve büyük, karmaşık veri setlerinden ayırt edilmesi zor kalıpları tespit etmesini sağlayan çeşitli istatistiksel ve optimizasyon tekniklerini kullanan yapay zekanın bir dalıdır. Derin öğrenmeye kıyasla veri sayısı daha az olursa makine öğrenimi kullanılabilir. Derin Öğrenme (Deep Learning) tekniği, bir makine öğrenme yöntemidir ve verilen bir veri kümesi ile çıktıları tahmin edecek yapay zekayı eğitmeye olanak sağlar. Derin öğrenmenin diğer bir adı ise yapay sinir ağlarıdır. Derin öğrenmedeki veri sayısı sistemin çalışmasını etkileyen en önemli faktördür. Tıp ve biyomedikal alanında makine öğrenmesi algoritmaları sıklıkla kullanılmaktadır. Özellikle son yıllarda görüntü işleme teknikleri ile birlikte kanser ve kontrol teşhisleri için makine öğrenmesi kullanılmaktadır.

KNN, SVM, DT, RF, LR ve MLPC gibi çeşitli sınıflandırma algoritmaları bulunmaktadır. Makine öğrenmesi algoritmaları uygulandıktan sonra elde edilen sonuçları değerlendirmek için çeşitli sınıflandırma modellerini değerlendirme parametreleri bulunmaktadır. Bunlar karışıklık matrisi, (TN, FN, FP, TP), doğruluk, kesinlik, f1 skor ve ROC-AUC değerleridir.

Tez içeriğinde aşağıdaki bilgiler sırasıyla verilmiştir:

Giriş bölümünde çalışma hakkında kısa bilgilendirme yapılmıştır.

İkinci bölümde *N*- ve *O*-glikozilasyon, glikozilasyon ve kanser ilişkisi, kütle spektrometrisi ve yöntemleri. *N*-glikanlar için kütle spektrometrik analizler ve mide kanseri hakkında literatür çalışmaları verilmiştir.

Üçüncü bölümde kullanılan malzemelerden, mide kanser kontrol dokularına uygulanan deneysel işlemlerden, MALDI-MS ve makine öğrenmesi analizlerinden bahsedilmiştir.

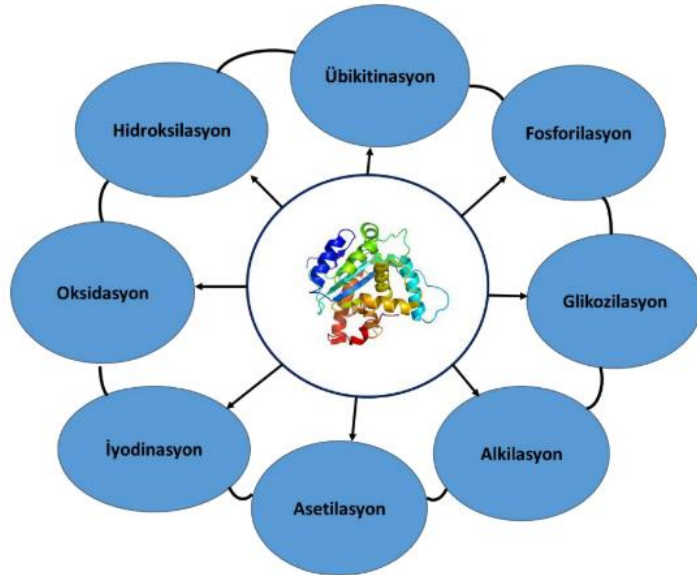
Dördüncü bölümde *N*-glikanların MALDI-MS analizlerden, MALDI-MS analizleri sonucu elde edilen veri setlerini analiz etmek için uygulanan makine öğrenmesi algoritmalarından ve model değerlendirmelerinden bahsedilmiştir.

Beşinci ve son bölümde sonuçlar tartışılmış ve öneriler sunulmuştur.

## BÖLÜM 2

### 2.1. POST-TRANSLASYONEL MODİFİKASYONLAR

Bazı proteinler moleküler görevlerini modifiye olmamış polipeptit formunda gösteremez. Post-translasyonel modifikasyonlar (PTM) translasyon sonrası proteinlerin kimyasal düzenlemeleridir ve bu proteinlerin doğru biyolojik fonksiyonlarını doğru bir şekilde tespit edebilmek için çok önemlidir. Post-translasyonel modifikasyonlar bugüne kadar analiz edilen neredeyse tüm proteinlerde gerçekleşmektedir. Proteinler kendi görevlerini yerine getirebilmeleri amacıyla kalıcı ya da geçici moleküler özellikler edinmesi gerekir. Proteinlerin post-translasyonel modifikasyonları genel olarak proteolitik bölünme olayları veya belirli amino asit kalıntılarındaki kovalent modifikasyonlar şeklinde gözlemlenir (Şekil 2.1.). Hemen hemen 20 doğal amino asitten herhangi biri, bir tür post-translasyonel modifikasyon ile değiştirilebilir [1].



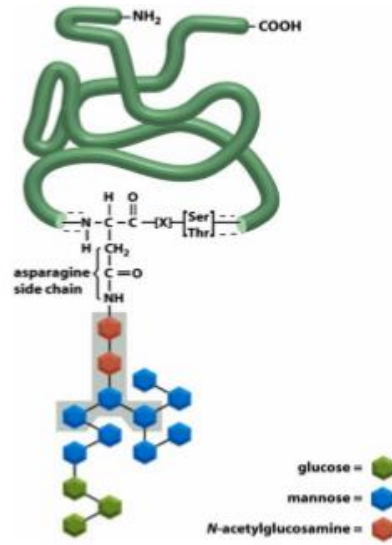
Şekil 2.1. Bazı post-translasyonel modifikasyonların gösterimi [2].

Modifikasyona uğramış bu proteinlerin işlevleri tamamen bu modifikasyonlardan etkilenmektedir. Proteinler, post-translasyonel modifikasyona uğrayınca birçok hücrel faaliyetleri yönetir ve bu nedenle bu modifikasyonların hücre düzenlemesi işleyişlerindeki rollerinin incelenmesi çok önemlidir. Bu gibi değişiklikler sinirsel iletimde rol alan peptitleri aktive etmek için gereken yapısal parçalanmalarla beraber, lizine metal grubu bağlanması gibi aminoasitlere fonksiyonel grupların kovalent olarak bağlanması biçiminde de görülebilmektedir [2].

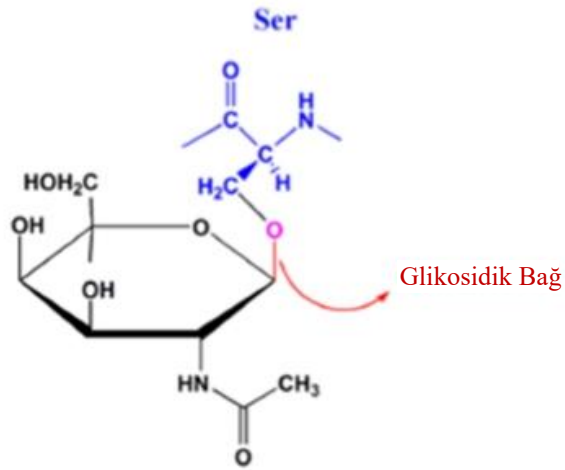
## 2.2. PROTEİN GLİKOZİLASYONU VE TİPLERİ

Şeker parçalarının birbirine bağlanması şekli olan glikozilasyon, diğer post-translasyonel modifikasyonlardan çok daha fazla proteomik çeşitlilik sağlayan bir post-translasyonel modifikasyondur. Proteine özgü glikozilasyon, son zamanlarda hastalık gelişimini gösterme ve hücrelerin fizyolojik davranışını yansıtmada konusunda ciddi endişeler uyandırmıştır. Hücre zarı ve ekstrasellüler matriks ile temas eden hücre yüzeyi proteinleri üzerinde glikozilasyon, hücrel iletişim, göç ve moleküler alışveriş ve açıklık gibi bir dizi hücrel davranış içerir. Buna ek olarak, anormal glikozilasyon çoğunlukla kardiyovasküler hastalık, bakteriyel enfeksiyon ve özellikle kanser de dahil olmak üzere bazı hastalık işlemlerine aracılık eden fizyolojik dengesizliğin bir işareti olarak görev görür [3].

Glikoproteinler polipeptid yapılarına kovalent olarak bağlı oligosakkarit (glikan) zincirlerini içeren proteinlerdir. Glikoproteinler bakteri, hayvan, insan vb. pek çok canlıda bulunmakta ve farklı işlevlere sahip olduğu vurgulanmaktadır. Kısa glikan zincirlerine sahip bu proteinler birçok hücrel olayda hormonlar, virüsler ve başka maddeler aracılığıyla hücre yüzeyinin tanınmasında görev alırlar. Yapılarda içerdikleri bağ türüne göre glikoproteinler, Şekil 2.2’de gösterildiği gibi *N*-glikozidik bağlı (*N*-bağlı) ve Şekil 2.3’de gösterildiği gibi *O*-glikozidik bağlı (*O*-bağlı) olmak üzere iki sınıfa ayrılır. Glikanlar bir proteinin tüm yapısının değiştirebilir ve bundan dolayı da proteinlerin işlevlerine direkt etki gösterirler. [4]



Şekil 2.2. Proteinlerde *N*-bağlı glikolizasyon tipi [4].



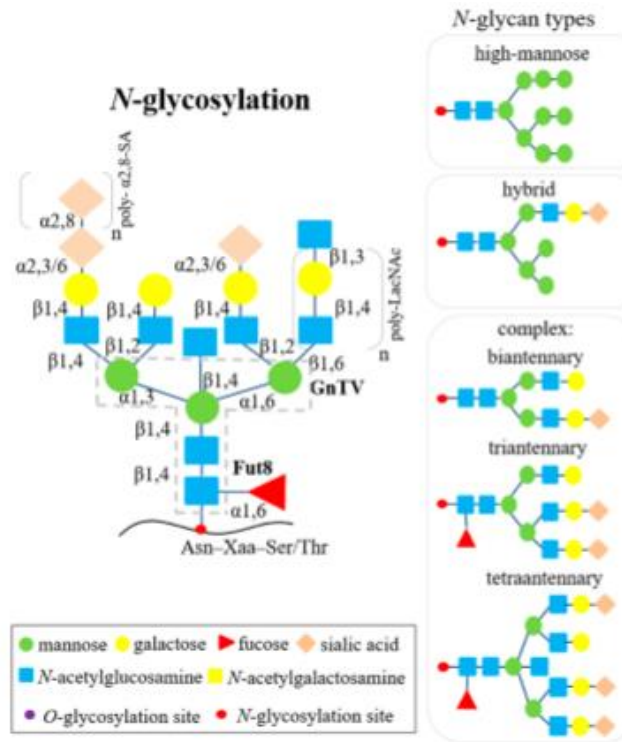
Şekil 2.3. Proteinlerde *O*-bağlı glikozilasyon tipi [4].

### 2.2.1. *N*-Bağlı Glikozilasyon

*N*-bağlı glikanlar, bir *N*-glikozidik bağ ile asparajin (Asn) amino asiti içeren protein kovalent bağlanır. Prokaryotlarda asparajine çeşitli şekerler eklenmesine karşın tüm ökaryotik *N*-glikanlar ilk olarak *N*-asetil glukoz amin içeren monosakkarit yapısı ile başlar. *N*-glikanların biyosentezi, memelilerde en kompleks olanıdır [5].



*N*-glikan sentezi, ökaryotlarda dolikol-fosfat (Dol-P) olarak isimlendirilen lipit benzeri poliizoprenoid molekülü üzerinde başlamaktadır. Yaklaşık 14 şeker içeren bir glikan sentezinden sonra, *N*-glikan içeren yapılar blok proteinine transfer edilir [5]. *N*-glikanlar, *N*-glikozidik bağlar vasıtasıyla Asn-Xaa-Ser / Thr konsensüs sekansında Asn'ye bağlanır. *N*-glikozilasyon'da üç tür *N*-glikan oluşturulur; aynı çekirdek yapıyı (GlcNAc2Man3) paylaşan oligomannoz (yüksek mannoz), hibrit tip ve kompleks tip (bi-, tri- veya tetraantennari) (Şekil 2.4) [6].

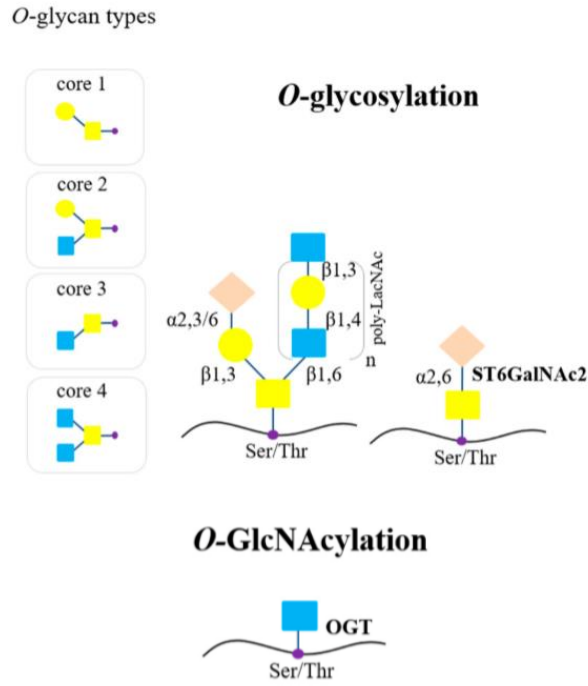


Şekil 2.4. *N*-glikan tipleri [6].

### 2.2.2. O-Bağlı Glikozilasyon

Bazı glikoproteinler, Ser veya Thr hidroksiline bağlı olan GalNAc ile başlatılan glikanları taşımaktadır. Müsin, çok fazla sayıda O-GalNAc glikan (*O*-glikan olarak da isimlendirilmekte) taşıyan glikoprotein sınıfıdır, fakat bu post-translasyonel modifikasyon bazı glikoproteinler arasında yaygındır [5].

Müsin tipi glikanlar (çoğunlukla müsinlerde ve ayrıca diğer glikoproteinlerde de bulunur) indirgeyici terminalde GalNAc'a sahiptir ve daha yüksek ökaryotlarda salgılanan ve membrana bağlı birçok glikoprotein üzerinde bulunur. N-bağlı glikozilasyonla ile kıyaslandığında, bir oligosakkaritin bir dolikol öncüsünden transferi ile başlamaz, fakat GalNAc olan tek bir monosakkarit eklenmesi ile başlar (Şekil 2.5) [7].



Şekil 2.5. *O*-glikan tipleri [6].

### 2.3. MİDE KANSERİ VE GÖRÜLME SIKLIĞI

Mide, kaslardan oluşan ve genişleyebilen bir sindirim sistemi organıdır. Yiyeceklerin geçici olarak büyük miktarlarda depolandığı organdır ve ek olarak yiyeceklerin kimyasal ve fiziksel olarak parçalandığı yerdir. Mide aynı zamanda salgı yapabilen bir organdır; iç duvarlarında bulunan hücreler ve bezler mukus vb. birçok önemli salgılar üretir (gastrin, somatostatin, mide asiti: hidroklorik asit vs.).

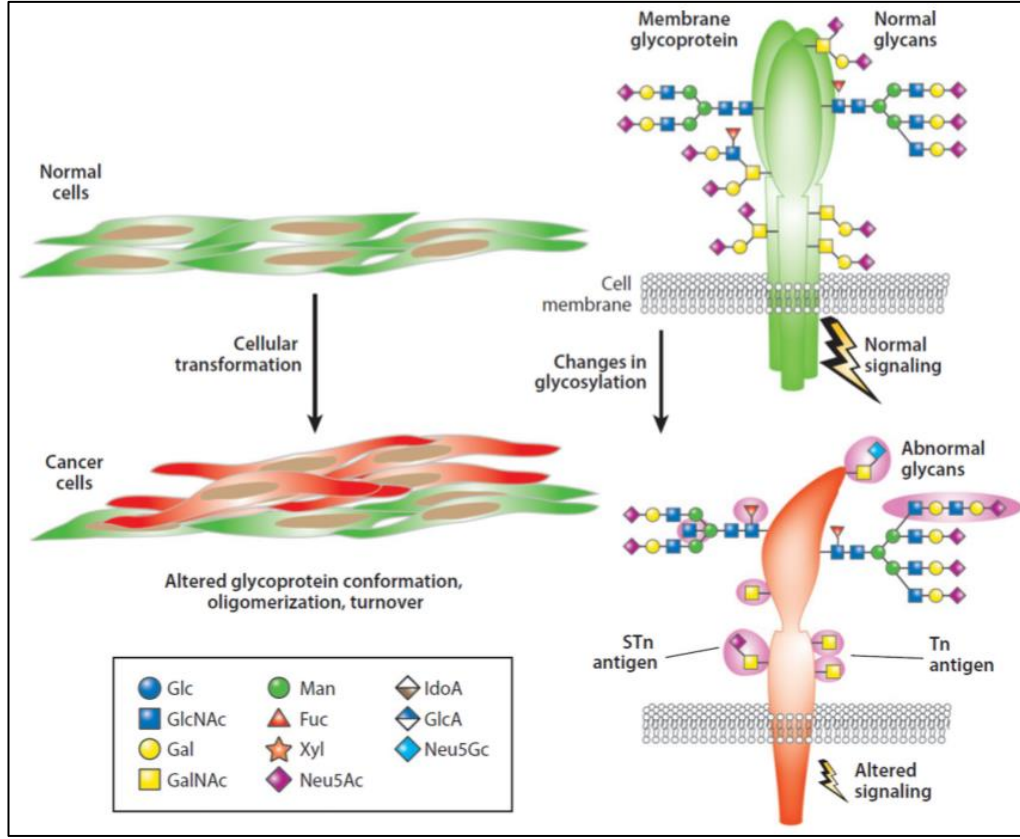
İnsan, yaşamı boyunca normal şekilde büyüyen ve gerektiği gibi bölünen trilyonlarca hücreden oluşur. Hücreler anormal bir yapıda olduğunda ya da yaşlandığında

çoğunlukla kendilerini imha ederler yani ölürler. Tam da bu süreçte bir şeyler ters gitmeye başladığında kanser başlar ve hücreler yeni hücreler yapmaya devam eder, eski ya da anormal olanlar ise gerektiği zaman ölmezler. Kanser hücreleri bu durumda kontrolden çıktıkça normal hücreler işlevlerini yerine getiremezler.

Mide kanseri dünya genelinde ölüm nedenleri arasında ilk sırada yer almaktadır. Amerikan Kanser Derneği (American Cancer Society)'nin yaptığı araştırmaya göre 2019 yılında ABD'de yaklaşık 27 bin mide kanseri vakası teşhis edilmiştir, bunların yaklaşık 17 bini erkek, 10 bini ise kadındır. Bu kanser türünden yaklaşık 11 bin kişi yaşamını yitirmiştir. Mide kanseri çoğunlukla da yaşlılarda görülmektedir, tanı konulanların ortalama yaşı 68'dir. Her yıl mide kanseri teşhisi konulan 10 kişiden yaklaşık 6'sı 65 yaş ve üzerindedir. Bir erkeğin yaşamı boyunca mide kanserine yakalanma riski 95'te 1, kadınlarda ise 154'te 1'dir. Adenokarsinom, lenfoma, gastrointestinal stromal tümör (GIST), karsinoid tümör gibi mide kanseri türleri bulunmaktadır.

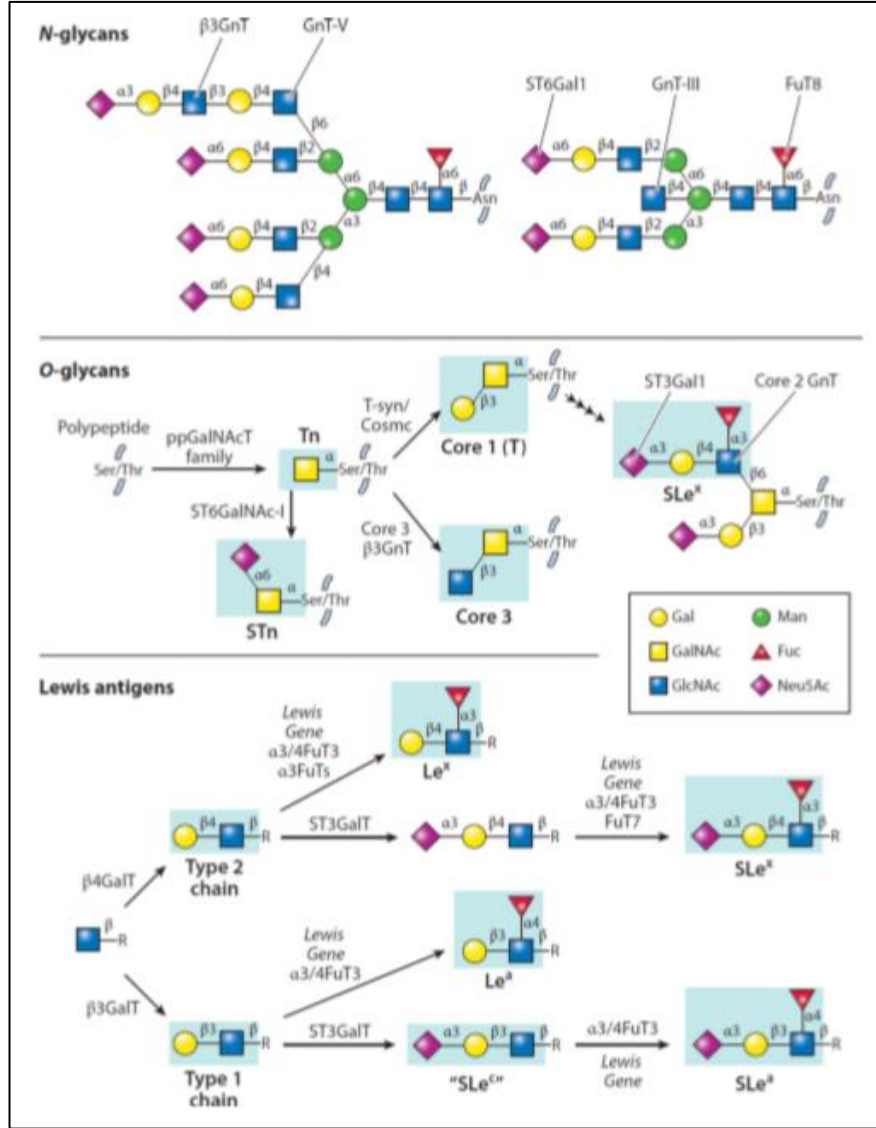
### **2.3.1. Kanser ve Glikozilasyon İlişkisi**

Post-translasyonel bir modifikasyon olan protein glikozilasyonu, biyobelirteç gibi kullanılmakla beraber kanserin gelişmesine ve ilerlemesine sebep olmaktadır. Hücresel dönüşümle ilgili protein glikozilasyonundaki esas değişiklikler arasında, kanserin ilerlemesinde ve metastazında erken ve geç ortaya çıkabilen *O*-glikanlardaki (GalNac-Ser/Thr) ve *N*-glikanlardaki değişimler yer alır (Şekil 2.6). Glikozilasyondaki bu değişiklikler, *O*- ve *N*-glikan çekirdek yapılarındaki spesifik değişikliklerle karakterize edilebilir. Ayrıca, transkriptlerin ve enzim aktivitesinin üretimi olarak tanımlanan glikozil-transferaz ekspresyonundaki farklılıklar sadece farklı çekirdek glikanların üretimini mühim ölçüde etkilemekle beraber, glikanı değiştirebilen çekirdek glikan dallanma derecesini de etkiler [8][9].



Şekil 2.6. Normal hücrelerde ve kanserli hücrelerdeki glikan değişimleri [8].

Çekirdek glikanlardaki değişimlere ilaveten, bu karbonhidratların her biri neoplastik transformasyondan sonra belirli değişikliklere uğrayabilen özgün uç glikan motifleri elde edilmek için değişebilmektedir. Örnek vermek gerekirse, Lewis antijenleri [ $\text{Lewis}^{a/b}$  ( $\text{Le}^{a/b}$ ) and  $\text{Lewis}^{x/y}$  ( $\text{Le}^{x/y}$ )] gibi yüksek derecede fukozylanmış glikanlar neoplastik dönüşümün ardından hücre yüzeyinde gelişebilir, zenginleşebilir (Şekil 2.7) [8].



Şekil 2.7. *N*- ve *O*- glikanların ve Lewis antijenlerinin özel yapıları [8].

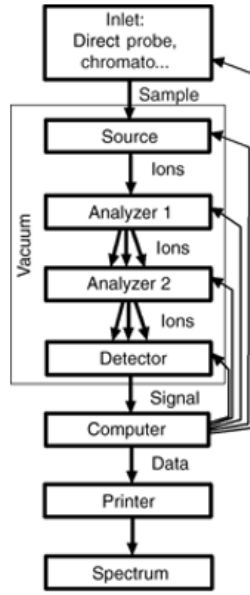
Ayrıca siyalizasyonda görülen artışlar, ( $\beta 1 \rightarrow 6$ ) GlcNAc bağlarının artması, bazı oligosakkaritlerin koparılması, sulfatlanmanın azalması ve zaman zaman yeni yapıların çıkması gibi çeşitli durumlar da kanser ve glikozilasyon arasındaki ilişkinin birer kanıtlarıdır [10].

## 2.4. KÜTLE SPEKTROMETRESİ

Kütle spektrometresi; atomların ve moleküllerin kütlelerini belirlemek amacıyla kullanılan analitik bir tekniktir. İyonların parçalanması yöntemine dayandığı için yapısal bir bilgi de vermektedir. Kütle spektrometresinin kullanımı geçtiğimiz yüzyılın başlarında başladı. İlk başlarda elementlerin atomik kütlelerini tespit etmek için kullanıldı. Bu ölçümler sayesinde elementlerin birkaç kütle gösterdiği anlaşıldı, bu sayede izotoplar keşfedildi ve kütle spektrometresi her bir elementin tüm izotoplarının belirlenmesinde kullanılmaya başlandı [11].

1960'lardan beri, yeni iyonizasyon yöntemleri proteinler, polisakkaritler vb. uçucu olmayan organik bileşiklerden gaz fazındaki iyonların üretilmesine izin vermiştir. Kütle analizörlerindeki bu gelişmeler ile elde edilen ağır iyonların kütlelerinin belirlenmesi kolaylaşmıştır. Ek olarak, gaz kromatografisi (GC)'nin geliştirilmesi, yüksek performanslı sıvı kromatografisi (HPLC) ve tandem kütle spektrometresindeki ilerlemeler kütle spektrometresini karışım analizi için epey güçlü bir teknik haline getirmiştir [12].

Günümüzde kullanılan modern kütle spektrometresi en az bir iyon kaynağı, bir kütle analizörü, bir dedektör ve yazıcıya sahip bir bilgisayar içeren bir cihazdır (Şekil 2.8). İyon kaynağı gaz fazında iyon üretmek amacıyla kullanılmaktadır, kütle analizörü iyonları  $m/z$  (kütle/yük) oranlarına göre ayırır, dedektör her bir  $m/z$  oranı için iyonları sayar. Bilgisayar ise analizörden ve dedektörden gelen verileri dönüştürerek bunları bir grafiğe dökmektedir. Ayrıca, kütle spektrometresi numuneyi bir ayırma cihazı aracılığıyla tanıtmak için gaz ve sıvı kromatografileri veya kapiller elektroforez gibi cihazlara ihtiyaç duyarlar. Kütle spektrometresi MS/MS ya da MS deneylerinin de yapılmasına izin veren birkaç tane analiz cihazı içerebilmektedir [11].



Şekil 2.8. Kütle spektrometresi şeması [11].

Proteomik veya glikproteomik çalışmalarında sıklıkla farklı tipte kütle analizörleri kullanılmaktadır. Proteinlerde meydana gelen post-translasyonel modifikasyonların belirlenmesinde QTOF-MS (quadropole time of flight mass spectrometry, dört kutuplu uçuş zamanlı kütle spektrometresi) cihazlar yaygın olarak kullanılmaktadır.

#### 2.4.1. İyonlaştırma Teknikleri

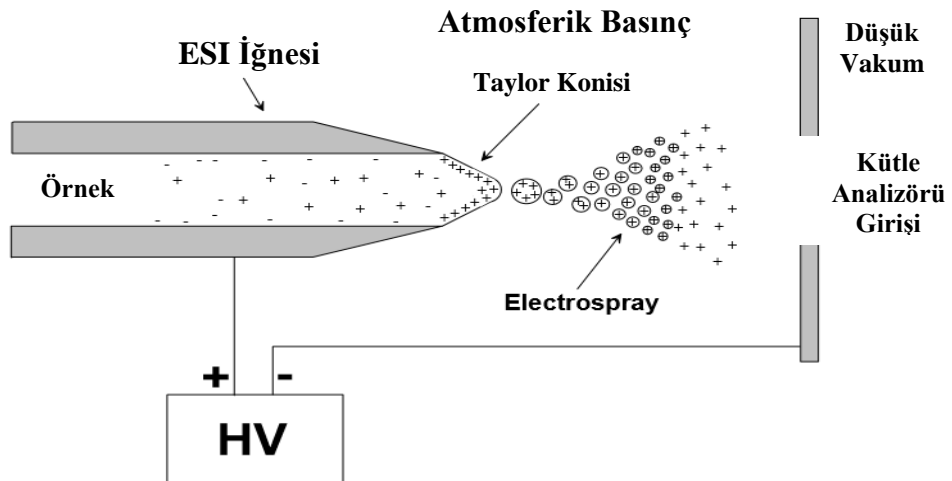
Analizi yapılacak olan numuneler kütle spektrometresindeki analizden önce iyon kaynaklarında iyonize edilir. Bu aşamada farklı iyonizasyon teknikleri kullanılır. İyonizasyon işlemi esnasında aktarılan enerji ve analitin iyonize olabilen fiziko-kimyasal özellikleri çok önem arz etmektedir. İyonizasyon tekniklerinden bazıları çok fazla enerjilidir ve parçalanmalara sebep olmaktadır. Geri kalan diğer teknikler ise biraz daha fazla yumuşaktır ve yalnızca moleküler türün iyonlarını üretir [13].

İyon kaynakları iki tiptir; birincisi sıvı-faz iyon kaynakları, ikincisi ise katı-hal iyon kaynaklarıdır. Sıvı-faz'da analit çözelti haldedir. Buradaki çözelti püskürtme yoluyla iyonların atmosferik basınçta üretildiği ve birtakım vakum pompalama evreleri ile kütle spektrometresine odaklandığı kaynağa damlacıklar şeklinde sokulmaktadır. Elektrosprey, atmosferik basınç kimyasal iyonizasyonu ve atmosferik basınç foto-

iyonizasyon kaynakları bu tipe karşılık gelmektedir. Katı-hal iyon kaynaklarında ise analit, uçucu olmayan bir tortu içerisinde bulunmaktadır. Genellikle katı ya da yapışkan olan sıvı bir matrisin sokulması prosedürü vb. çeşitli hazırlama metotları ile elde edilmektedir. Bu tortu daha sonrasında tortu yüzeyinin üzerinden dışarı salınan enerjik parçacıklar ya da fotonlar aracılığıyla ışınlanır. Bu iyonlar bir elektrik alanı ile çıkarılır ve analizöre odaklanır. MALDI'de iyon üretmek için bu yöntemi kullanır [14].

#### 2.4.1.1. Elektrosprey İyonlaştırma

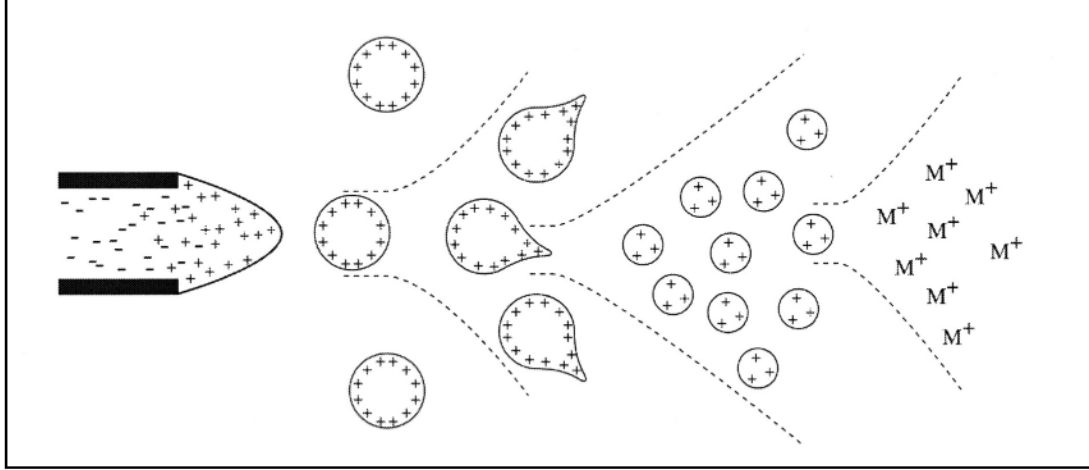
ESI (Elektrosprey İyonlaştırma), iyonların kütle spektrometrik analizi yapılmadan önce çözeltilerden gaz halindeki faza transferine yardımcı olmak için elektrik enerjisi kullanır (Şekil 2.9). Çözeltideki iyonik türler bu sayede ESI-MS ile yüksek hassasiyette analiz edilebilir. Nötr bileşikler ayrıca protonlama veya katyonizasyon (örneğin; metal katyonizasyonu) ile çözelti veya gaz fazında iyonik forma dönüştürülebilir ve bu sayede ESI-MS ile incelenebilir [15]. Elektrosprey, analitin başlangıçta çözelti içinde olduğu atmosferik basınç altında gerçekleştirilen bir iyonizasyon işlemidir. Bir elektrik potansiyelinin etkisi altında numune içi boş bir iğneden püskürtülür ve bu işlem sırasında analit, çözelti buharlaştıkça sıvıdan gaz fazına getirilir. İyonizasyon, gaz fazı geçişinden önce sıvı halde gerçekleşmektedir [16].



Şekil 2.9. Elektrosprey İyonizasyon işlemi şeması [16].



İyonik türlerin çözeltilerden ESI aracılığıyla gaz fazına iletimi üç adım içerir; birincisi, ince bir şarj damlacık spreyinin dağıtılması. İkincisi, çözücü buharlaşması ve üçüncüsü ise yüklü damlacıklardan iyon oluşumu ile kütle analizörüne doğru hızlandırılmasıdır (Şekil 2.10).



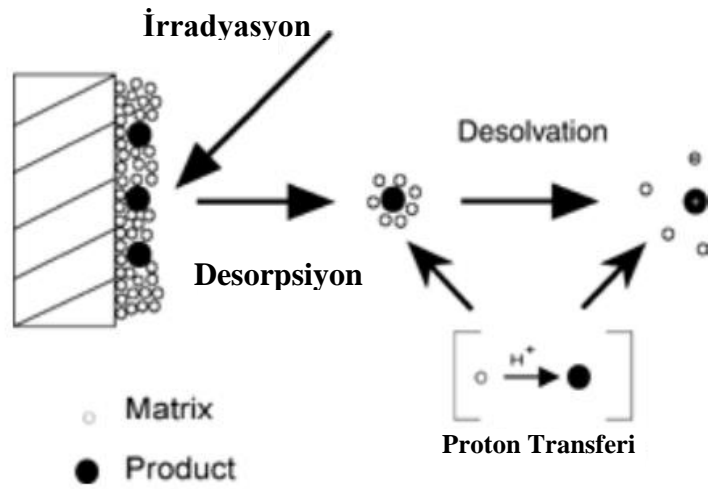
Şekil 2.10. Elektrosprey İyonlaştırma mekanizması [15].

ESI ilk başlarda protein analizine adanmış bir iyonizasyon kaynağı olarak kabul edilmiştir. Daha sonra kullanımı sadece diğer polimerlere ve biyopolimerlere değil, aynı zamanda küçük polar moleküllerin analizinde de kullanılmıştır. ESI çok yüksek hassasiyetlere kadar çıkabilen bir iyonlaştırma türüdür. Yüksek performanslı sıvı kromatografisi (HPLC- $\mu$ HPLC) ya da kılcal elektroforez ile birleştirilmesi oldukça basittir [13].

#### 2.4.1.2. Matriks Destekli Lazer Desorpsiyonu / İyonizasyonu (MALDI)

MALDI, temel olarak 1988'de Karas ve Hillenkamp tarafından ortaya atılmıştır [17]. O yıllardan beri proteinler, oligonükleotidler, sentetik polimerler ve büyük inorganik bileşikler gibi geniş, uçucu olmayan ve termal olarak kararsız bileşiklerden sağlam gaz fazı iyonlarının üretimi için yaygın ve güçlü bir kaynak haline gelmiştir. Desorpsiyon ve iyonizasyonu sağlayan bir MALDI matriksinin kullanılması, bu iyonizasyon yönteminin başarısı için oldukça önemli bir faktördür [13].

MALDI’de, bir numune başka bir çözülmüş katı, matriks içeren bir çözücü içinde çözülür. Çözünün buharlaştırılmasıyla matristeki analitin katı bir çözeltisini üretir. Bu katı çözelti, dalga boyu matriksin güçlü bir emilim bandına karşılık gelen bir lazer ışınının bir darbesi ile vakum altında ışınlanır. Bu sayede, matrikse ve numuneye enerji sağlanır. Buradaki enerji, örneğin gaz fazında iyonlar olarak dışarı atılmasına sebep olur (Şekil 2.11). Bu iyonlar sonra bir kütle analizörü ile analiz edilir. Darbeli bir lazerin kullanımıyla iyonlar çok kısa bir sürede üretilir ve çok kısa bir sürede yani birkaç milisaniyede birlikte analiz edilmelidir [11].



Şekil 2.11. MALDI kaynağının şematik gösterimi [11].

MALDI kütle spektrometresi hem sentetik polimerler hem de biyopolimerler için güçlü bir analitik araçtır. MALDI diğer lazer iyonizasyon tekniklerinden daha duyarlıdır. Aslında, matriks moleküllerinin sayısı analit moleküllerini büyük ölçüde aşar, bu da analit moleküllerini ayırır ve böylece moleküllerin görünümünü engelleyen örnek kümelerinin oluşumunu önler. MALDI, 1.000.000 Da’nı aşan çok yüksek moleküler kütleye sahip analitlerin desorpsiyonuna ve iyonizasyonuna imkan tanımaktadır [13].

#### 2.4.2. Kütle Analizörleri

Kütle spektrumları ve kütle analizörleri tipik olarak kütle hassasiyeti, kütle çözünürlüğü, algılama hassasiyeti ve sinyal-gürültü oranı (SNR) olan bir dizi parametre ile karakterize edilir. Kütle doğruluğu genellikle bir deney içindeki en

önemli verileri tanımladığından dolayı en önemli parametrelerden biridir. Kütle doğruluğu çoğunlukla milyonda bir parça (ppm) olarak sunulur. Örneğin; 20 ppm, 1000 Da olarak ölçülen bir iyonun gerçek kütlelerinin  $\pm 20$  mDa'lık bir tahmini hata payına sahip olduğu anlamına gelmektedir. Kütle doğruluğu, kalibrasyonun kalitesi ve zirvenin merkezinin ne kadar doğru belirlenebileceği ile sınırlıdır [18].

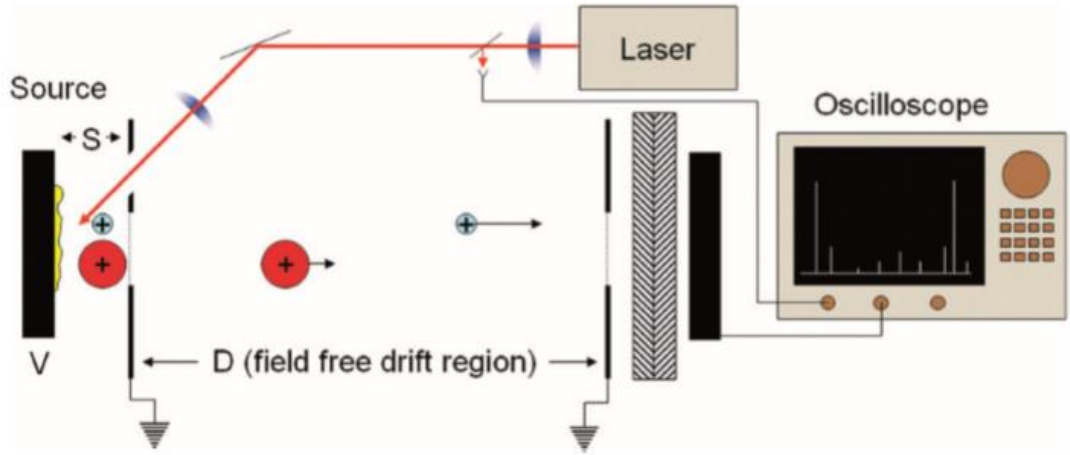
Pek çok çeşitli kaynaklar olduğu için çeşitli tipte kütle analizörleri geliştirilmiştir. Gerçekten de iyonların kütle/yük oranlarına göre ayrılması farklı prensiplere dayandırılabilir (Şekil 2.12). Tüm kütle analizörleri, tek başına ya da kombine edilebilen statik veya dinamik elektrikli ve manyetik alanlar kullanır [13].

| <b>Analizör Tipi</b>                      | <b>Sembol</b> | <b>Ayrırma Prensibi</b>      |
|---|---------------|------------------------------|
| Electric sector                           | E or ESA      | Kinetic energy               |
| Magnetic sector                           | B             | Momentum                     |
| Quadrupole                                | Q             | $m/z$ (trajectory stability) |
| Ion trap                                  | IT            | $m/z$ (resonance frequency)  |
| Time-of-flight                            | TOF           | Velocity (flight time)       |
| Fourier transform ion cyclotron resonance | FTICR         | $m/z$ (resonance frequency)  |
| Fourier transform orbitrap                | FT-OT         | $m/z$ (resonance frequency)  |

Şekil 2.12. Kütle spektrometresinde kullanılan analizör çeşitleri [13].

#### 2.4.2.1 Uçuş Zamanlı Kütle Analizörü (TOF–Time-of-Flight)

MALDI uygulamaları için en yaygın kütle analizörü, çalışma prensibi şekil 2.13'te verilen aksel (axial) TOF (uçuş zamanlı) kütle spektrometresidir. Aksel TOF enstrümanları MALDI kaynakları için idealdir, çünkü daha sonra uçuş tüpünden aşağı uçup çoklu-kanal plaka (MCP) gibi bir dedektöre çarpan çok kısa iyon atımlarına ihtiyaç duyarlar. İyon akım sinyali, bir dijital depolama osiloskopu kullanılarak dedektöre varış süresine kadar karşı basitçe çizilir [18].



Şekil 2.13. MALDI uçuş zamanlı kütle analizörü çalışma prensibi [18].

Prensip olarak TOF cihazının üst kütle aralığının sınırı yoktur, bu da onu özellikle yumuşak iyonizasyon teknikleri için kullanışlı uygun kılar. Örnek olarak, kütleleri 300 kDa'nın üzerinde olan numuneler MALDI-TOF tarafından gözlenmiştir. TOF cihazının diğer bir avantajı ise çok yüksek hassasiyetli yüksek iletim verimliliğidir. Örneğin,  $10^{-15}$  mol gramisidin spektrumu ve 100-200 attomol miktarının çeşitli proteinlerin (sitokrom C, ribonükleaz A, lizozim ve miyoglobin) saptanması TOF analizörleri ile elde edilmiştir [19].

## 2.5. MAKİNE ÖĞRENMESİ

Makine öğrenmesi, yapay zekanın bir alt dalıdır ve hesaplama sistemlerinde yapay zeka konseptlerini uygulamak için algortimalar kullanır, ayrıca makine öğrenmesinde çeşitli matematiksel ve istatistiksel işlemler kullanılarak veriler üzerinden çıkarımlar yapılması sağlanır. Makine öğrenmesi, tıp alanında özellikle de tıp mühendisliği ve biyomedikal mühendisliği uygulamalarında sıkça kullanılmaktadır. Birçok hastalıkta bireylerin risk faktörlerinin hesaplanmasında, kanser teşhislerinin yapılmasında, görüntü işleme uygulamalarında, ses tanıma, nesne tanıma vb. alanlarda sıklıkla makine öğrenmesi algoritmaları kullanılmaktadır [20].

Makine öğrenmesinde;

- Gözetimli (Supervised)
- Gözetimsiz (Unsupervised)
- Takviyeli (Reinforcement)

olmak üzere üç çeşit öğrenme türü vardır (Şekil 2.14).

### **2.5.1. Gözetimli Öğrenme**

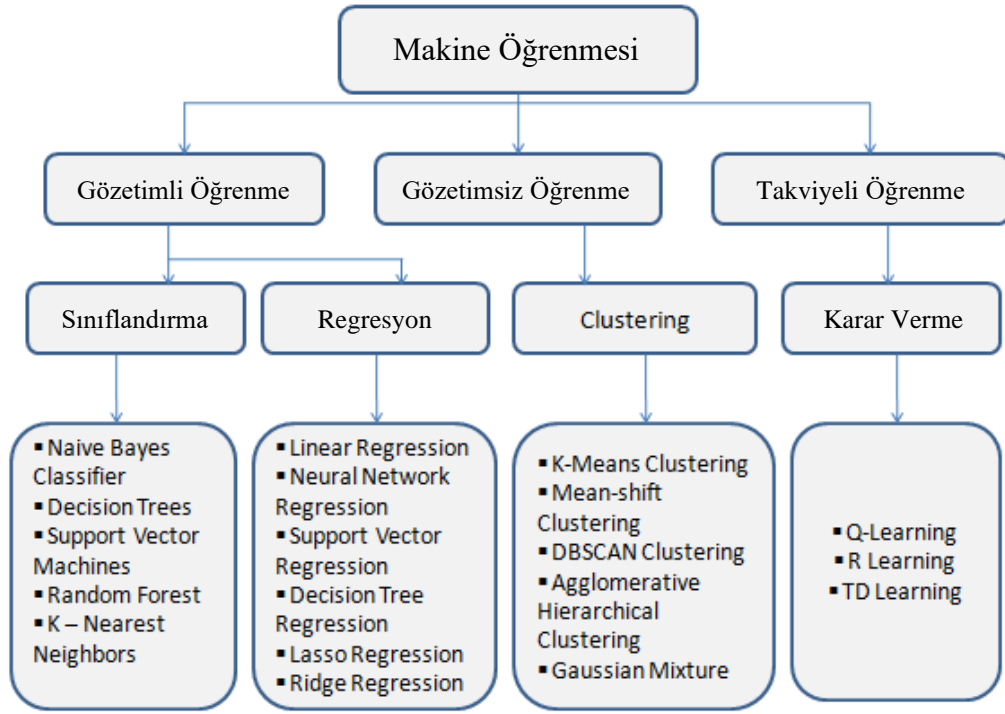
İşaretlenmiş veriler ile çıkış değerleri arasında eşleşme yapan fonksiyonlar oluşturulur. Buradaki eğitim verisi hem girdilerden hem de çıktılarından oluşur, veri setindeki çıkışlar kategorik ise sınıflandırma (classification), nümerik ise regresyon (regression) algoritmaları kullanılır.

### **2.5.2. Gözetimsiz Öğrenme**

Etiketlenmemiş veriler üzerinden bilinmeyen datayı tahmin etmek için gözetimsiz öğrenme algoritmaları kullanılır. Veri kümesindeki tüm gözlem birimleri etiketsizdir ve algoritmalar girdi verilerinden doğal yapıyı öğrenir. Gözetimsiz öğrenme modellerinde büyük miktarda sınıflandırılmamış veriyle çalışmak gerekir ve istenilen sonuçları elde edebilmek için büyük bir veri setine ihtiyaç duydukları için hesaplama açısından oldukça karmaşıktır.

### **2.5.3. Takviyeli Öğrenme**

RL (Reinforcement Learning), bir aracının kendi eylemlerinden ve deneyimlerinden gelen geri bildirimleri kullanarak deneme yanılma yoluyla etkileşimli bir ortamda öğrenmesini sağlayan makine öğrenmesi türüdür. Gözetimli ve gözetimsiz öğrenmenin aksine burada olumlu ve olumsuz davranış için sinyaller olarak ödülleri ve cezaları kullanır.



Şekil 2.14. Gözetimli Öğrenme Algoritmaları.

Bu çalışmada dokuların kanser/kontrol olarak sınıflandırılması için gözetimli öğrenme grubuna ait sınıflandırma algoritmaları kullanılmıştır.

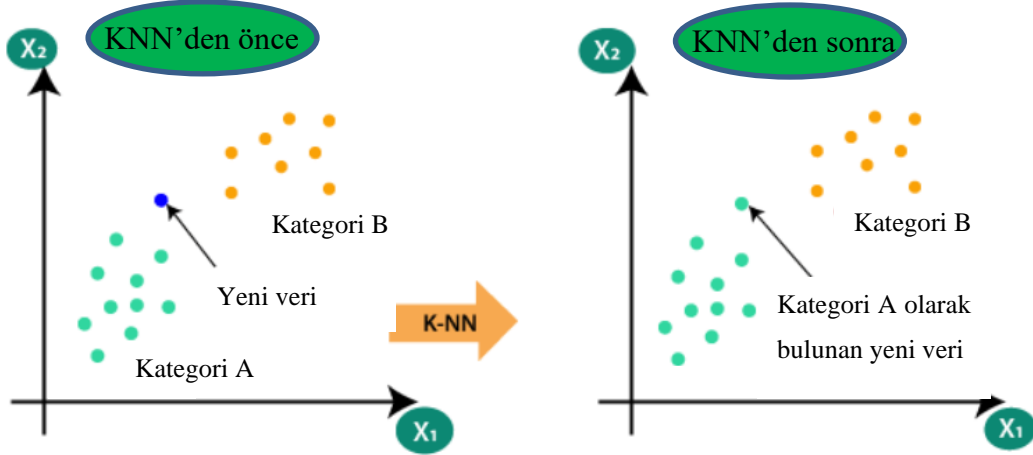
## 2.6. K-EN YAKIN KOMŞU ALGORİTMASI (KNN, K-NEAREST NEIGHBORS ALGORITHM)

Denetimli bir makine öğrenmesi algoritması olan K-en yakın komşu algoritması sınıflandırılmamış bir veri noktasının ne olarak sınıflandırılacağını bize gösterir. KNN algoritmasının basit mekanizması önceden tanımlanmış sınıflar ve değişen her örnek arasındaki öklid mesafesi fonksiyonunu (euclidean distance function) hesaplamaktır. Daha sonrasında KNN algoritması her kategoriye göre minimum en yakın komşu seçimini gerçekleştirir. Örnekler en yakın k komşuya göre sınıflara ayrılır. KNN algoritması sınıflandırmanın yanı sıra regresyon modellemeleri için de kullanılabilir ancak daha çok sınıflandırma problemleri için kullanılmaktadır [21].

KNN’de en çok kullanılan denklem eşitlik 2.1’deki öklid mesafesi denklemdir.

$$d = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2} \quad (2.1)$$

Burada k, her bir örnek vektöründeki değerlerin sayısı,  $x_1$  ve  $x_2$  ise girdi örnekleridir.



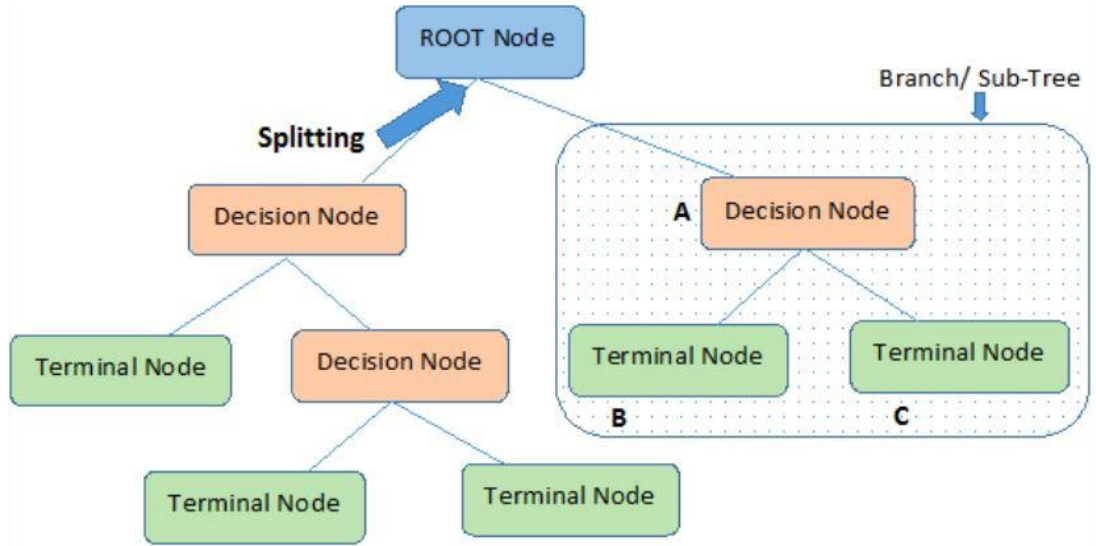
Şekil 2.15. K-En Yakın Komşu Algoritması [22].

Şekil 2.15’te gösterilen kategori A ve B olarak iki sınıfa göre  $x_1$  yeni veri noktasının (mavi renkli) hangi sınıfta yer aldığı için KNN algoritması kullanılmıştır. İlk olarak komşu sayısı ( $k$ ) seçilir, öklid uzaklığı hesaplanır, hesaplanan öklid mesafesine göre en yakın komşu  $k$  alınır. Burada  $k$  komşu arasında her kategorideki veri noktalarının sayısı sayılır ve sınıflandırılacak yeni veri, komşu sayısının maksimum olduğu kategoriye atanır.

## 2.7. DECISION TREE (KARAR AĞAÇLARI)

Karar ağaçları, bir denetimli makine öğrenmesi algoritmasıdır. Diğer denetimli öğrenme algoritmalarından farklı olarak karar ağaçları sınıflandırma ve regresyon problemlerini çözmek için de kullanılabilir. Karar ağacının kullanım amacı, önceki verilerden yani eğitim setlerinden çıkarılan basit karar kurallarını öğrenerek hedef değişkenin sınıfını veya değerini tahmin edebilmek için kullanılacak bir eğitim (training) modeli oluşturmaktır. Karar ağaçlarında bir datanın sınıf etiketini tahmin

edebilmek için ağacın kökünden başlanır. Kök özneliğinin değerlerini verininki ile kıyaslanır. Karşılaştırmadan sonra o değere karşılık gelen dalı takip edip bir sonraki düğüme atlanır (Şekil 2.16) [23].



Şekil 2.16. Karar ağaçları şeması [24].

**Root node (kök düğüm):** Tüm örneği temsil eder, ayrıca iki veya daha fazla homojen kümeye bölünür.

**Splitting (Bölme):** Bir düğümün iki veya daha fazla alt düğüme bölünmesi işlemidir.

**Decision Node (Karar düğümü):** Bir alt düğümün başka alt düğümlere bölünmesidir.

**Leaf / Terminal Node (Yaprak / Terminal Düğüm):** Bölünmeyen düğümlerdir.

**Pruning (Budama):** Bir karar düğümünün alt düğümleri kaldırıldığında bu işleme budama denir ve aynı zamanda bu işleme bölme işleminin tersi denilebilir.

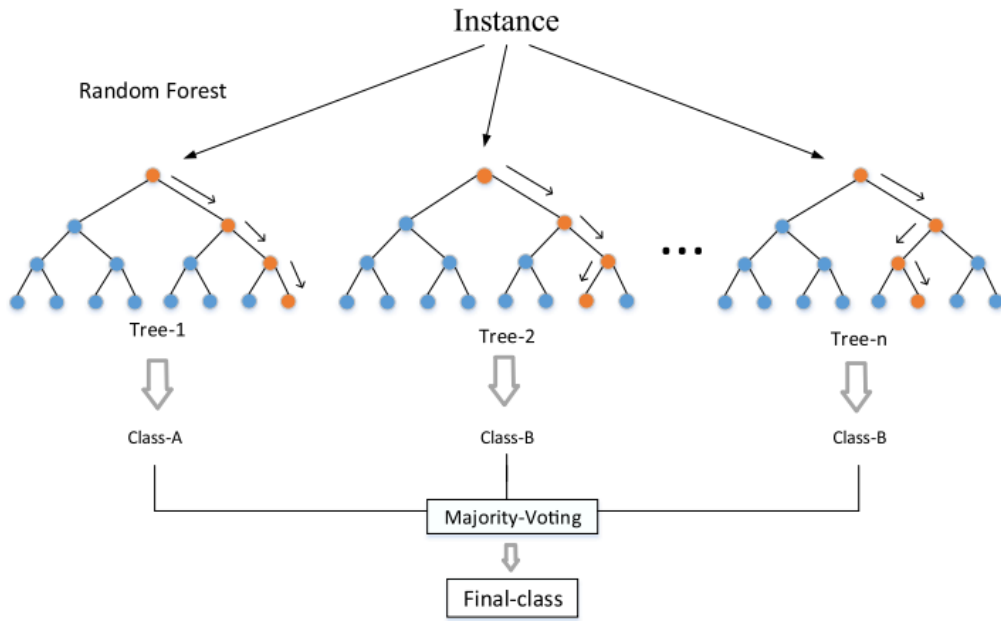
**Branch / Sub-tree (Dal / Alt ağaç):** Alt ağaç olarak isimlendirilir.

**Parent and Child Node (Ana ve Alt düğüm):** Alt düğümlere bölünmüş bir düğüm, alt düğümlerin ana düğümü olarak adlandırılır, alt düğümler ise bir ana düğümün çocuklarıdır.



## 2.8. RANDOM FORESTS (RASSAL - RASTGELE ORMANLAR)

Breiman (2001), regresyon ve sınıflandırma için Random Forest algoritmasını önermiştir. Rastgele Orman, denetimli bir sınıflandırma algoritmasıdır. Algoritma içerisindeki ağaç sayısı ve elde edilen sonuçlar arasında doğrudan bir ilişki bulunmaktadır. RF adından da anlaşılacağı üzere bir orman oluşturur ve bunu rastgele bir şekilde yapmaktadır. Kurulan bu orman “bagging” yöntemiyle eğitilen karar ağaçlarıdır. Bootstrap Aggregation olarak da bilinen Bagging, rassal orman tarafından kullanılan topluluk tekniğidir. Bagging, veri setinden rastgele örnek seçer ve bundan dolayı da her model satır örnekleme (bootstrap samples) ile orijinal veri tarafından örneklerden oluşturulmaktadır.



Şekil 2.17. RF'nin çalışma prensibi [25].

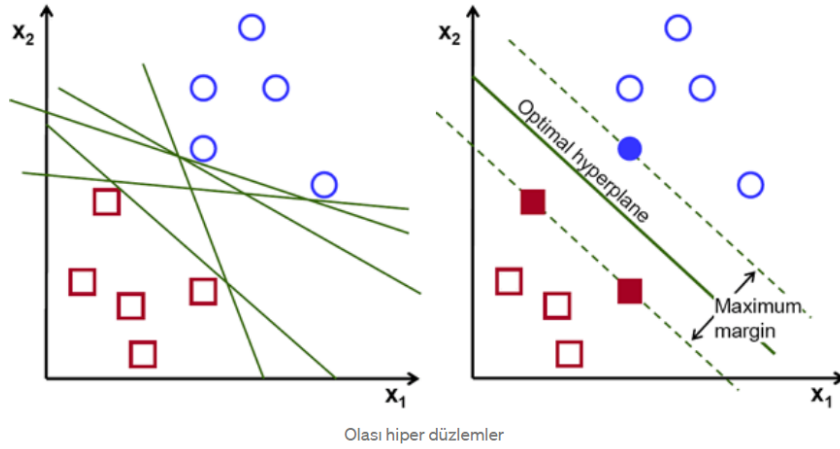
RF birkaç regresyon ağacından oluşmaktadır ve değişkenin bulunduğu düğüme kök düğüm denir. Her ağacın altındaki düğümlere yaprak düğümler veya terminal düğümler denir (Şekil 2.17). RF algoritmasında kullanıcı tarafından iki parametre tanımlanır. Bu parametreler ağaç sayısı (N) ve her bir düğüme kullanılan değişken sayı (m) (Breiman, 2001) ve ID3 algoritması kriterine göre entropi ile belirsizlik ölçülür ve eşitlik 2.2'de gösterildiği gibi formülize edilmektedir.

$$H(T) = - \sum_{i=1}^n p(x_i) \log_p(x_i) \quad (2.2)$$

H entropi,  $p(x_i)$  belirli bir sınıfa ait grubun yüzdesini belirtmektedir. Burada entropinin yüksek olması belirsizliğin de o kadar yüksek olduğunu göstermektedir. Entropi sıfıra eşit olana kadar ağaçların dallanması sürer ve ağaç derinliği de belli bir seviyede seçilerek ağacın budaması yapılmaktadır [25].

## 2.9. SUPPORT VECTOR MACHINES (SVM)

Sınıflandırma, regresyon ve aykırı değerlerin tespit edilmesi için kullanılan denetimli öğrenme yöntemidir. Bu algoritmanın amacı veri noktalarını belirgin bir şekilde sınıflandıran N boyutlu bir uzayda (N-öznitelik) bir hiperdüzlem bulmaktır.



Şekil 2.18. SVM [26].

İki veri noktası sınıfını ayırmak için seçilebilecek birçok olası hiperdüzlem vardır (Şekil 2.18). Amaç, maksimum marjine sahip yani her iki sınıfın veri noktaları arasındaki maksimum mesafeye sahip bir düzlem bulmaktır. Marj mesafesini en üst düzeye çıkarmak gelecekteki veri noktalarının daha güvenle sınıflandırılabilmesi için bir miktar takviye sağlamaktadır.

SVM algoritmasında, veri noktaları ile hiperdüzlem arasındaki marjı maksimize edilmeye çalışılır. Marjı maksimize etmeye yardımcı olan loss fonksiyonu 'hinge loss'tur ve eşitlik 2.3'te gösterilmiştir.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \quad (2.3)$$

Tahmin değeri ile gerçek değer aynı işaretli ise maliyet 0'dır. Eğer değillerse kayıp değeri hesaplanır. Maliyet fonksiyonuna bir düzenleme parametresi eklenmektedir. Düzenleme parametresi maksimum marj ve kaybı dengelemeye yardımcı olmaktadır. Düzenleme parametresini ekledikten sonra cost fonksiyonu eşitlik 2.4'te verilmiştir [27].

$$\min_w \lambda \|w\|^2 + \sum_{i=1}^n (1 - y_i(x_i, \omega)) \quad (2.4)$$

## 2.10. LOJİSTİK REGRESYON

Logistic regresyon bağımlı değişken ikili olduğunda yapılacak en uygun regresyon analizidir. Diğer tüm regresyon analizleri gibi logistic regresyon da tahmine dayalı bir analiz yapmaktadır. LR modeli genelleştirilmiş doğrusal modellerden biridir. Logistik regresyon modeli eşitlik 2.5'te gösterilmiştir.

$$p_i = f(y|x) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_i)\}} \quad (2.5)$$

Formülde  $p_i$  başarı olasılığıdır ve  $P(y_i = 1) = p$ ,  $P(y_i = 0) = q_i = 1 - p_i$ ,  $0 \leq p_i \leq 1$ .

Ayrıca  $\beta_0$  ve  $\beta_1$  modelin parametreleridir,  $x_i$  bağımsız değişken  $\exp$  ise Euler sayısı olarak adlandırılan ve yaklaşık olarak 2.78'e eşit olan matematiksel sabittir [28].

Lojistik regresyon, yirminci yüzyılın başlarında biyoloji bilimlerinde kullanılmıştır. Daha sonra birçok sosyal bilim uygulamasında da kullanılmıştır. Lojistik regresyon bağımlı değişken kategorik olduğunda kullanılır.

Örnek olarak;

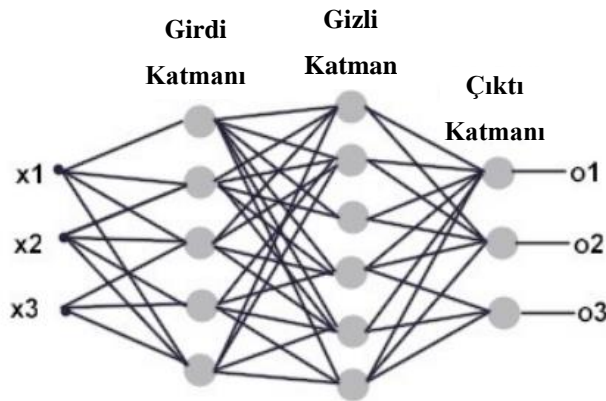
- 1) Bir e-postanın spam olup olmadığını tahmin etmek için (1 veya 0)
- 2) Bir mide kanseri dokusunun kötü huylu (malignant) olup olmadığını (1 veya 0) tespit edebilmek için

Örneklerden de anlaşılacağı üzere lineer regresyon ile kıyaslandığında lojistik regresyon sınıflandırma problemlerinde kullanılmaktadır.

## 2.11. MLP

Çok katmanlı algılayıcı, en popüler ve çok yönlü yapay sinir ağı (YSA) modellerinden biridir. Karmaşık karar bölgelerini haritalama yeteneğinden dolayı (Lippmann 1987) MLP, model sınıflandırılması için yaygın olarak kullanılmıştır. Bu YSA algoritması çeşitli sınıfların bir dizi örneğinin tekrar tekrar sunulması yoluyla istenen sınıflandırma görevini gerçekleştirmek için eğitilmiştir [29].

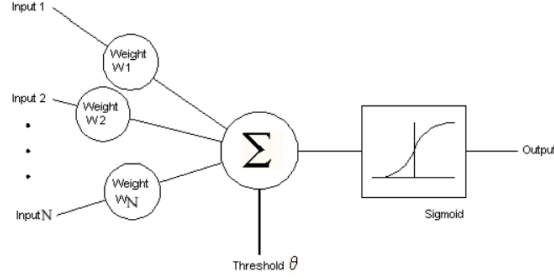
MLP sınıflandırma problemlerinde çok iyi çalışmaktadır. Çok katmanlı ağların yapısı şekil 2.19'daki gibidir.



Şekil 2.19. Çok katmanlı ağlar [30].

Görüldüğü gibi tek katmanlı algılayıcıdan farklı olarak arada gizli katman bulunmaktadır. Giriş katmanı verileri ara katmana iletir. Gelen bilgiler bir sonraki

katmana aktarılır. Ara katman sayısı ihtiyaca göre ayarlanmakla beraber probleme göre en az bir olarak ayarlanır. Her nöron bir sonraki katmanda bulunan diğer tüm nöronlara bağlıdır. Çıkış katmanı önceki nöronlardan gelen verileri işler. Sistemin çıkış ise çıkış katmanında bulunan eleman sayısına eşittir [30].



Şekil 2.20. Aktivasyon fonksiyonu (sigmoid) [30].

Sigmoid, linear, tang, threshold ve hard limiter en sık kullanılan fonksiyonlardır.

## 2.12. MODELİN DEĞERLENDİRİLMESİ (SINIFLANDIRILMANIN DEĞERLENDİRİLMESİ)

### 2.12.1 Karışıklık Matrisi (Confusion Matrix)

Tıp sektöründe özellikle tıbbi işlemlerde, biyomedikal uygulamalarda en çok kullanılan gösterimlerden biridir. Makine öğrenmesi sınıflandırma problemlerinde için performans ölçüm aracıdır.

Karışıklık matrisi, tahmini ve gerçek değerlerin 4 farklı kombinasyonunu içerir (Şekil 2.21). TP: kanserli dokunun gerçekte kanser olması, FN (tip 2 hata): kanserli dokulara kontrol demek, FP (tip 1 hata): kontrol dokularına kanser demek, TN: kontrol dokusunun gerçekte kontrol olması anlamlarına gelmektedir.

| GERCEK            |                   |         |        |
|-------------------|-------------------|---------|--------|
| POZİTİF           | NEGATİF           |         |        |
| True<br>Positive  | False<br>Positive | POZİTİF | TAHMİN |
| False<br>Negative | True<br>Negative  | NEGATİF |        |

Şekil 2.21. Confusion Matrix [31].

**Recall (Duyarlılık):** Kanser dokularını doğru tespit etme oranını ifade etmektedir ve formülü eşitlik 2.6’da gösterilmiştir.

$$Recall = \frac{TP}{TP + FN} \quad (2.6)$$

**Precision (Kesinlik):** Kanser olarak tahmin edilen dokuların gerçekte ne kadar doğru olduğunu ifade etmektedir ve formülü eşitlik 2.7’de gösterilmiştir.

$$Precision = \frac{TP}{TP + FP} \quad (2.7)$$

**F1 skoru:** Kesinlik ve duyarlılık değerlerinin harmonic ortalamasını göstermektedir. Formül eşitlik 2.8’de gösterilmiştir.

$$F1 Skor = \frac{2 * Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \quad (2.8)$$

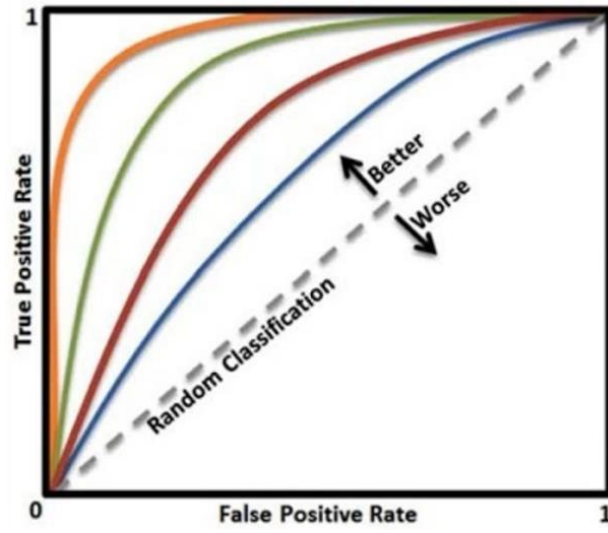
### 2.12.2. ROC ve AUC

ROC eğrisi biyomedikal uygulamalarında sınıflandırma problemleri için önemli bir performans ölçümüdür. ROC bir olasılık eğrisidir ve ROC eğrisi Şekil 2.22’deki turuncu renkli eğri gibi olması model performansının çok iyi olduğu anlamına gelmektedir. ROC eğrisinde x ekseninde FPR ve y ekseninde TPR bulunmaktadır.

AUC modelin sınıfları ne kadar başarılı ayırt edebildiğini anlatır. AUC arttıkça model 0'ları ve 1'leri tahmin etmede daha iyidir.

TPR, duyarlılık değeridir. Yani kanser olan dokuları tespit edebilme oranıdır. FPR ise kanser olmayan dokular için hatalı tahmin oranını ifade etmektedir.

Farklı thresh değerlerini deneyerek en iyi f1 skoru elde edildiğinde en iyi ROC eğrisi de tespit edilmektedir.



Şekil 2.22. ROC eğrisi.

## BÖLÜM 3

### 3.1. KULLANILAN MALZEMELER

Çalışmada kullanılan FFPE (formalinle sabitlenmiş parafilm ile kaplanmış) mide kanserli dokular Karabük Üniversitesi Eğitim ve Araştırma Hastanesi Patoloji biriminden temin edilmiştir. Çalışma retrospektif olarak dokuların toplanması ile gerçekleştirilmiştir. Çalışma girişimsel olmayan etik kurulu tarafından onaylanmıştır (Etik Kurul Karar No:2019/117, Ek1). Metanol, asetonitril, etanol, ksilen Sigma-Aldrich firmasından temin edilmiştir. 1,4-ditiyotreitol (DTT), trifloro asetik asit (TFA), DMSO, asetik asit (CH<sub>3</sub>COOH) Merck firmasından alınmıştır. 2-amino benzoik asit (2-AA), sodium dodesil sülfat (SDS) ve sodyum klorür (NaCl) Sigma-Aldrich'den alınmıştır. Peptit-N-glikozidaz F (PNGase F) enzimi Promega firmasından, 2,5-dihidroksi benzoik asit (DHB) Bruker Daltonics'den temin edilmiştir. Çalışmanın tamamında kullanılan dH<sub>2</sub>O (deiyonize su) ise Expe-Ultrapure su arıtma sistemleri ile elde edilmiş ve çalışmalarda kullanılmıştır.

### 3.2. DOKULARDAN PARAFİNİN UZAKLAŞTIRILMASI

Karabük Eğitim ve Araştırma Hastanesi Patoloji biriminden mide kanseri dokuları ependorfların içerisinde temin edilmiştir. Dokular 60 derecede 30 dakika inkube edildi. Sonrasında dokular sırasıyla 1 mL ksilen ile 2 kez yıkandı (5 dakika 600 RPM'de inkube edilip sıvılar uzaklaştırılmıştır), 1 mL %100 etanol ile 2 kez yıkanmıştır. Yıkanan dokulardan sıvılar uzaklaştırıldı ve dokular 45 derecede 5 dakika boyunca hızlı vakum konsantratör cihazı ile kurutulmuştur. Dokuların pH seviyesini belli bir aralıkta tutmak için 0.1 M Tris-HCl hazırlanmıştır. Aynı zamanda proteinleri primer yapıya kavuşturmak için 0.1 M DTT hazırlanmıştır. Hazırlanan karışımdan 150 µL alınıp her örneğe ilave edilmiştir ve proteinlerin verimli ekstraksiyonları için sonikasyon işlemi gerçekleştirilmiştir (1.5 saniye yap, 3 saniye bekle, 30 saniye toplam süre).



### **3.3. PROTEİNLERİN EKSTRAKSİYONU (KLOROFORM/METHANOL EKSTRAKSİYON YÖNTEMİ)**

Sonikasyon sonrası dokulara 60 µL %16 SDS ilave edilmiştir. 99 derecede 1 saat 600 rpm'de örnekler inkube edilmiştir. Dokular 2000 rcf'de 20 dk santrifüj edilmiştir. Santrifüj sonrasında her örnek için üstteki fazdan 150 µL alınıp ayrı ependorflara koyulmuştur ve üzerine 600 uL methanol, 150 µL kloroform ve 450 µL deionize su ilave edilmiştir. 1-2 saniye kadar vortekslenip ve 14000 rcf'de 5 dk santrifüjlenmiştir. Santrifüj sonrasında ortadaki pelete zarar vermeden üstteki faz uzaklaştırılmıştır. 450 µL metanol ilave edilip yine 1-2 saniye kadar vortekslenmiştir ve 14000 rcf'de 10 dk santrifüjlenmiştir. Ardından örneklerin içerisindeki sıvılar uzaklaştırılmıştır. Dipte kalan proteinler hızlı vakum konsantratör ile 45 °C'de kurutulmuştur ve -20 derece saklanmıştır.

### **3.4. N-GLİKANLARIN SALINIMI**

Kurutulan örnekler 25 µL %1 SDS ilave edilmiştir. 15-20 saniye kadar vortekslenmiştir. 60 °C 600 rpm'de 5 dakika inkube edilmiştir. %2 Igepal-CA360 hazırlanmıştır ve her bir örneğe 12.5 µL eklenmiştir. 5X PBS hazırlanmıştır ve tüm örnekler 12.5 µL eklenmiştir. Her örneğe 1 µL kadar 1 U/µL konsantrasyona sahip PNGase F eklenmiştir. Gece boyunca (16 saat) inkube edilmiştir.

### **3.5. 2-AA İLE N-GLİKANLARIN ETİKETLENMESİ**

48 mg 2-AA ve 63 mg NaCNBH<sub>3</sub> tartıldıktan sonra kimyasallar 10:3, v/v oranında DMSO ve Glacial Asetik Asit içerisinde çözülmüştür, hacimce 1/1, v/v oranında karıştırılmıştır ve sırasıyla örnekler 25 µL eklenmiştir. 65 °C'de örneklerin 2 saat inkubasyonu yapılmıştır.

### **3.6. SELÜLOZ İÇEREN KATI FAZ KARTUŞLAR İLE 2-AA ETİKETLİ N-GLİKANLARIN SAFLAŞTIRILMASI**

1 mL suda 100 mg selüloz olacak şekilde stok çözeltisi hazırlanmıştır. Ependorflara 0.2 mL stok çözeltisi aktarılmıştır ve sırasıyla 0.5 mL su ve %85 ACN ile 2'şer kez yıkanmıştır. 2-AA ile etiketlenmiş örneklerden 75 µL alınıp selüloz içeren ependorflara aktarılmıştır ve üzerlerine 425 µL %100 ACN eklenmiştir. Hazırlanan örnekler 600 rpm'de 20 dakika inkube edilmiştir. İlk olarak kontaminasyonu engellemek için kartuşlar 1 mL %100 etanol ile yıkanmıştır ve sonrasında inkube edilmiş örnekler selüloz içeren kartuşlara aktarılmıştır. 1 mL %85 ACN %1 TFA ile 3 kez ve 1 mL %85 ACN ile 3 kez yıkanmıştır. Son olarak 0.75 mL deionize su ile elüsyonları yapılmıştır. Örnekler daha sonradan tekrardan saflaştırılmak üzere -20 derecede saklanmıştır.

### **3.7. GÖZENEKLİ GRAFİTİZE KARBON KATI FAZ EKSTRAKSİYONU İLE SAFLAŞTIRMA (PGC- SPE)**

MALDI-MS analizinden önce örneklerdeki tuz vb. maddelerin arındırılması için gözenekli grafitize karbon içeren katı faz kartuşlar ile saflaştırılma yapılmıştır. 0.1 g/mL konsantrasyonunda PGC malzemesi ilk olarak ACN içerisinde hazırlanmıştır. Her bir pack'e 500 µL eklenmiştir. PGC malzemesi 2 kez 1 mL ACN/MQ (80:20 v/v), 3 kez 1 mL %0.1 TFA karışımı ile yıkanmıştır. Örnekler 0.5 µL TFA eklenmiştir. Örnekler katı faz kartuşlara yüklenmiştir. 3 kez %0.1 TFA ile yıkanmıştır. Son aşama olarak da yeni ependorflar koyularak 0.75 mL %0.1 TFA içeren %80 ACN ile elüsyonları yapılmıştır. Analizlerden önce örnekler hızlı vakum konsantratör ile kurutulmuştur ve MALDI-MS ile analizleri öncesi 20 µL deiyonize su ile örnekler yeniden çözülmüştür.

### **3.8. MALDI-MS ANALİZLERİ**

SmartBeam 3D lazer teknolojisi içeren bir RapiFlex MALDI-TOF/TOF-MS/MS (Bruker Daltonik GmbH, Bremen, Almanya) kullanılarak 2-AA etiketli N-glikanların MALDI-MS analizleri gerçekleştirilmiştir. Saflaştırılmış numune çözeltilerinden 1 µL

alınarak doğrudan MALDI hedef plakası üzerinde damlatılmıştır ve kurumaya bırakılmıştır. Kuruyan örnekler %50 Su ve % 50 ACN kullanarak hazırlanan 5 mg mL<sup>-1</sup> DHB [2,5-dihidroksibenzoik asit (2,5-DHB)] matriksinden 1 µL ilave edilmiştir. Spektrumlar, negatif iyonlaştırma modunda 20 kV hızlanma potansiyelinde ve refletron modu kullanılarak 1000–4000 Da kütle aralığında kaydedilmiştir. 2000 Hz frekansta toplam 8000 lazer atışı toplanarak spektrumlar elde edilmiştir. Her bir örneğe ait 4 farklı spektrum 4 farklı spottan elde edilmiş ve analizlerde kullanılmıştır.

### **3.9. MALDI-MS VERİ ANALİZLERİ**

Elde edilen MS ve MS/MS spektrumları Protein Scape V4 (Bruker Daltonik GmbH, Bremen, Almanya) yazılımına aktarıldıktan sonra GlycoQuest yazılımı kullanılarak bileşimleri ve yapıları belirlenmiştir. Tespit edilen *N*-glikanların bir listesi oluşturulmuştur. Bu listede yer alan glikanların alan bilgileri MassyTools yazılımı kullanılarak ekstrakte edilmiş ve toplam alan normalizasyon yaklaşımı kullanılarak her tespit edilen *N*-glikanın göreceli bolluğu ve analit alanı hesaplanmıştır. Ayrıca yazılım kullanılarak her bir glikana ait kütle intensite ve alan bilgileri elde edilmiştir.

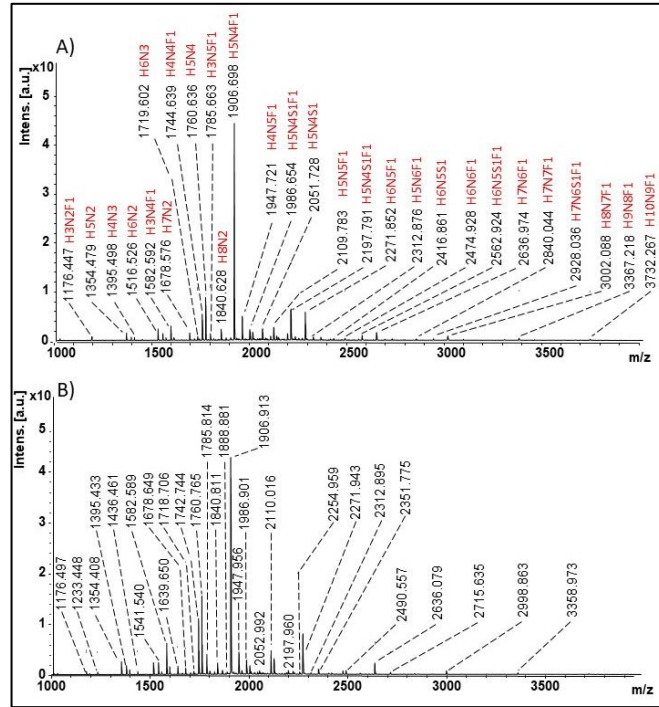
### **3.10. MAKİNE ÖĞRENMESİ**

130 kanser (33 kanserli doku), 122 kontrol (31 kontrol doku) dokularının verileri Python programlama dilinde makine öğrenmesi algoritmalarında kullanılmak üzere deneysel analizler sonucunda elde edilmiştir. Kanserli ve komşu kontrol dokuları ile bunların *N*-glikan türlerinin (tüm glikanlar ve nötr glikanlar olarak 2 farklı veri seti) göreceli alan ve analit alanlarının değerlerine göre 4 farklı veri seti oluşturulmuştur. Elde edilen veri setleri neticesinde python programlama dilinde KNN, RF, DT, SVM, LR ve MLPC algoritmaları kullanılarak algoritmaların kanserli ya da kontrol dokuyu tahmin etmedeki başarı oranları tespit edilmiştir. Bu bağlamda tüm *N*-glikanlar ve nötr glikanlarından hangisinin başarı oranının daha fazla olduğu tespit edilmiştir.

## BÖLÜM 4

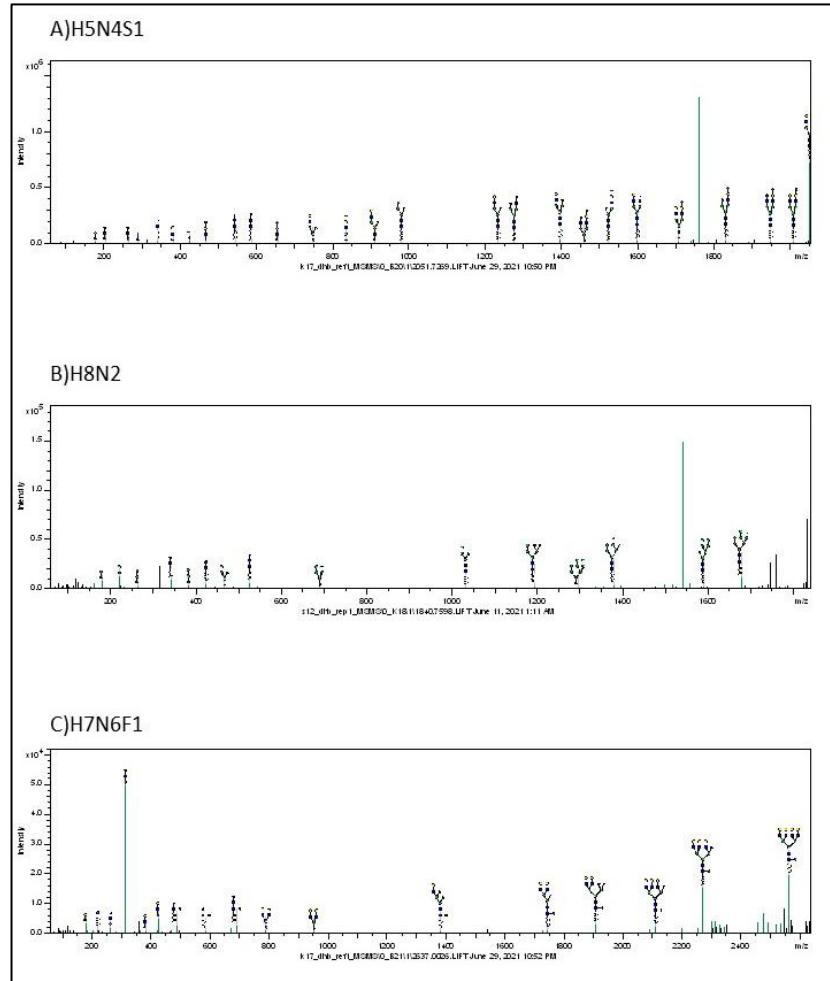
### 4.1. DOKU PROTEİNLERİNE AİT N-GLİKANLARIN MALDI-MS İLE ANALİZLERİ

Karabük Üniversitesinde Girişimsel Olmayan Klinik Araştırmalar Etik Kurulundan etik kurul onayı alınmıştır. Onay alındıktan sonra Karabük Eğitim ve Araştırma Hastanesi Patoloji biriminden mide dokuları patoloji uzmanının onayıyla (kanser/sağlıklı) temin edilmiştir. Sonrasında ilk aşamada dokulardan proteinlerin ekstraksiyonu metot kısmında anlatıldığı gibi gerçekleştirilmiştir. Bu işlem sonrasında glikanların proteinler üzerinden salınımının gerçekleştirilmesi için proteinler PNGase-F enzimine maruz bırakılmıştır. Glikanlar 2-AA florosans etiketiyle etiketlendikten sonra negatif iyonlaşma modunda MALDI-MS ile analizleri gerçekleştirilmiştir (Şekil 4.1).



Şekil 4.1. Kanser (A) ve Kontrol (B) dokularına ait MALDI-MS Spektrumu.

Şekil 4.1’de kanserli daku ve kanserli dokuya komşu kanser hücresi içermeyen kontrol dokusuna ait bir MALDI-MS spektrumu görülmektedir. Analizler sonucunda 59 tane *N*-glikan tespit edilmiştir. Bunların 45 tanesi nötr glikan, 14 tanesi ise asidik glikan olarak bulunmuştur. Nötr ve asidik glikanlar makinesi öğrenmesi çalışmamızda kanser-kontrol ayırımında ayrı kullanılmıştır. MALDI-MS analizlerinde tespit edilen *N*-glikanların yapısal karakterizasyonları da MALDI MS/MS sisteminde yapılmıştır. 37 adet *N*-glikana ait MS/MS spektrumları elde edilmiş olup bunların sonucunda elde edilen bu MS/MS parçalanma ürünleri ProteinScape glycoquest yazılımı ile tarandıktan sonra glikanların yapısı doğrulanmıştır. Şekil 4.2’de yazılım aracılığı ile yapıları doğrulanmış *N*-glikanlara ait spektrumlar görülmektedir.



Şekil 4.2. MALDI MS/MS analizleri sonucu yapıları doğrulanmış glikanlar (A) H5N4S1, (B) H8N2, (C) H7N6F1.

Analizler sonucunda tespit edilen 59 tane glikan Çizelge 4.1’de listelenmiştir. Bunlara ait m/z, skor değerleri ve parçalanma yüzdeleri verilmiştir. MassyTools yazılımı kullanılarak *N*-glikanlara ait kütle spektrometrik bilgiler ekstrakte edilmiş ve makine öğrenmesi yaklaşımı analizlerinde kullanılmıştır.

Çizelge 4.1. Mide kanseri dokularına ait tespit edilen *N*-glikanlar.

| Sıra | Kompozisyon | m/z ölçülen | z  | m/z hesaplanan | Mr hesaplanan | $\Delta$ MH+ [Da] | $\Delta$ m/z [ppm] | $\Delta$ MH+ [ppm] | Skor | FragCov. [%] |
|------|-------------|-------------|----|----------------|---------------|-------------------|--------------------|--------------------|------|--------------|
| 1    | H3N2-AA     | 1030.3994   | -1 | 1030.37327     | 1031.3811     | 0.0262            | 25.41              | 25.36              | 43.1 | 39           |
| 2    | H3N2F1-AA   | 1176.4467   | -1 | 1176.43118     | 1177.4390     | 0.0155            | 13.20              | 13.18              | 43.1 | 39           |
| 3    | H4N2-AA     | 1192.4276   | -1 | 1192.42609     | 1193.4339     | 0.0016            | 1.30               | 1.30               |      |              |
| 4    | H3N3-AA     | 1233.4540   | -1 | 1233.45264     | 1234.4605     | 0.0013            | 1.07               | 1.07               | 44.4 | 32           |
| 5    | H4N2F1-AA   | 1338.4687   | -1 | 1338.48400     | 1339.4918     | -0.0153           | -11.45             | -11.43             | 21.2 | 9            |
| 6    | H5N2-AA     | 1354.4786   | -1 | 1354.47891     | 1355.4867     | -0.0003           | -0.23              | -0.23              | 48.7 | 37           |
| 7    | H3N3F1-AA   | 1379.5098   | -1 | 1379.51055     | 1380.5184     | -0.0007           | -0.52              | -0.52              | 49.6 | 31           |
| 8    | H4N3-AA     | 1395.4976   | -1 | 1395.50546     | 1396.5133     | -0.0079           | -5.66              | -5.65              | 44.0 | 40           |
| 9    | H3N4-AA     | 1436.5066   | -1 | 1436.53201     | 1437.5398     | -0.0254           | -17.71             | -17.68             | 38.5 | 40           |
| 10   | H6N2-AA     | 1516.5262   | -1 | 1516.53174     | 1517.5396     | -0.0055           | -3.63              | -3.62              | 61.8 | 56           |
| 11   | H4N3F1-AA   | 1541.5584   | -1 | 1541.56337     | 1542.5712     | -0.0049           | -3.20              | -3.19              | 46.0 | 30           |
| 12   | H5N3-AA     | 1557.5462   | -1 | 1557.55829     | 1558.5661     | -0.0121           | -7.75              | -7.74              |      |              |
| 13   | H3N4F1-AA   | 1582.5920   | -1 | 1582.58992     | 1583.5977     | 0.0021            | 1.30               | 1.30               | 50.0 | 37           |
| 14   | H4N4-AA     | 1598.6109   | -1 | 1598.58484     | 1599.5927     | 0.0261            | 16.33              | 16.31              | 44.0 | 27           |

|    |             |           |    |            |           |         |        |        |      |    |
|----|-------------|-----------|----|------------|-----------|---------|--------|--------|------|----|
| 15 | H3N5-AA     | 1639.5985 | -1 | 1639.61139 | 1640.6192 | -0.0129 | -7.86  | -7.85  | 33.3 | 26 |
| 16 | H7N2-AA     | 1678.5765 | -1 | 1678.58456 | 1679.5924 | -0.0081 | -4.83  | -4.83  | 54.2 | 46 |
| 17 | H4N3S1-AA   | 1686.5937 | -1 | 1686.60088 | 1687.6087 | -0.0072 | -4.24  | -4.24  | 39.2 | 26 |
| 18 | H5N3F1-AA   | 1703.5892 | -1 | 1703.61620 | 1704.6240 | -0.0270 | -15.83 | -15.81 |      |    |
| 19 | H6N3-AA     | 1719.6021 | -1 | 1719.61111 | 1720.6189 | -0.0090 | -5.21  | -5.21  |      |    |
| 20 | H3N4S1-AA   | 1728.6382 | -1 | 1728.63609 | 1729.6439 | 0.0021  | 1.24   | 1.24   |      |    |
| 21 | H4N4F1-AA   | 1744.6390 | -1 | 1744.64274 | 1745.6506 | -0.0037 | -2.15  | -2.14  | 53.0 | 38 |
| 22 | H5N4-AA     | 1760.6362 | -1 | 1760.63766 | 1761.6455 | -0.0014 | -0.82  | -0.82  | 47.9 | 40 |
| 23 | H3N5F1-AA   | 1785.6627 | -1 | 1785.66929 | 1786.6771 | -0.0066 | -3.67  | -3.67  | 53.2 | 42 |
| 24 | H4N5-AA     | 1801.6579 | -1 | 1801.66421 | 1802.6720 | -0.0063 | -3.50  | -3.50  |      |    |
| 25 | H8N2-AA     | 1840.6275 | -1 | 1840.63738 | 1841.6452 | -0.0099 | -5.35  | -5.35  | 69.6 | 74 |
| 26 | H5N3S1-AA   | 1848.6408 | -1 | 1848.65370 | 1849.6615 | -0.0129 | -6.96  | -6.95  |      |    |
| 27 | H5N4F1-AA   | 1906.6979 | -1 | 1906.69557 | 1907.7034 | 0.0023  | 1.21   | 1.21   | 56.3 | 46 |
| 28 | H4N5F1-AA   | 1947.7214 | -1 | 1947.72212 | 1948.7299 | -0.0007 | -0.35  | -0.35  | 53.8 | 39 |
| 29 | H5N5-AA     | 1963.7043 | -1 | 1963.71703 | 1964.7249 | -0.0127 | -6.46  | -6.46  |      |    |
| 30 | H5N4S1F1-AA | 1986.6544 | -1 | 1986.65238 | 1987.6602 | 0.0020  | 1.01   | 1.00   | 15.4 | 7  |
| 31 | H9N2-AA     | 2002.6912 | -1 | 2002.69021 | 2003.6980 | 0.0010  | 0.52   | 0.52   | 40.6 | 24 |
| 32 | H4N5S1F1-AA | 2027.7221 | -1 | 2027.67893 | 2028.6868 | 0.0431  | 21.26  | 21.24  |      |    |
| 33 | H5N4S1-AA   | 2051.7284 | -1 | 2051.73308 | 2052.7409 | -0.0047 | -2.28  | -2.28  | 40.4 | 21 |

|    |             |           |    |            |           |         |       |       |      |    |
|----|-------------|-----------|----|------------|-----------|---------|-------|-------|------|----|
| 34 | H4N5F2-AA   | 2093.7914 | -1 | 2093.78003 | 2094.7879 | 0.0114  | 5.45  | 5.44  | 33.9 | 23 |
| 35 | H5N5F1-AA   | 2109.7826 | -1 | 2109.77494 | 2110.7828 | 0.0077  | 3.64  | 3.64  | 50.4 | 34 |
| 36 | H6N5-AA     | 2125.7809 | -1 | 2125.76986 | 2126.7777 | 0.0110  | 5.20  | 5.19  |      |    |
| 37 | H4N6F1-AA   | 2150.8280 | -1 | 2150.80149 | 2151.8093 | 0.0265  | 12.33 | 12.32 |      |    |
| 38 | H10N2-AA    | 2164.7813 | -1 | 2164.74303 | 2165.7509 | 0.0383  | 17.70 | 17.68 |      |    |
| 39 | H5N6-AA     | 2166.8113 | -1 | 2166.79640 | 2167.8042 | 0.0148  | 6.85  | 6.85  |      |    |
| 40 | H5N4S1F1-AA | 2197.7907 | -1 | 2197.79098 | 2198.7988 | -0.0003 | -0.15 | -0.15 | 40.2 | 18 |
| 41 | H5N5S1-AA   | 2255.8192 | -1 | 2255.82111 | 2256.8289 | -0.0019 | -0.83 | -0.83 | 22.3 | 9  |
| 42 | H6N5F1-AA   | 2271.8519 | -1 | 2271.82776 | 2272.8356 | 0.0242  | 10.63 | 10.62 | 49.6 | 37 |
| 43 | H3N6S1F1-AA | 2280.8762 | -1 | 2280.85275 | 2281.8606 | 0.0235  | 10.30 | 10.29 |      |    |
| 44 | H5N6F1-AA   | 2312.8765 | -1 | 2312.85431 | 2313.8621 | 0.0222  | 9.58  | 9.57  | 25.1 | 14 |
| 45 | H5N5S1F1-AA | 2400.8871 | -1 | 2400.87036 | 2401.8782 | 0.0167  | 6.96  | 6.96  |      |    |
| 46 | H6N5S1-AA   | 2416.8611 | -1 | 2416.86527 | 2417.8731 | -0.0042 | -1.73 | -1.72 | 28.6 | 13 |
| 47 | H5N6F2-AA   | 2458.9281 | -1 | 2458.91222 | 2459.9200 | 0.0158  | 6.44  | 6.43  |      |    |
| 48 | H6N6F1-AA   | 2474.9279 | -1 | 2474.90714 | 2475.9150 | 0.0208  | 8.41  | 8.40  | 55.1 | 39 |
| 49 | H7N6-AA     | 2490.9302 | -1 | 2490.90205 | 2491.9099 | 0.0282  | 11.30 | 11.29 |      |    |
| 50 | H6N5S1F1-AA | 2562.9243 | -1 | 2562.92318 | 2563.9310 | 0.0011  | 0.43  | 0.43  | 33.1 | 23 |
| 51 | H7N6F1-AA   | 2636.9745 | -1 | 2636.95996 | 2637.9678 | 0.0145  | 5.51  | 5.51  | 46.1 | 42 |
| 52 | H6N7F1-AA   | 2677.9976 | -1 | 2677.98651 | 2678.9943 | 0.0111  | 4.16  | 4.16  |      |    |
| 53 | H7N6F2-AA   | 2783.0241 | -1 | 2783.01787 | 2784.0257 | 0.0062  | 2.24  | 2.24  |      |    |
| 54 | H7N7F1-AA   | 2840.0437 | -1 | 2840.03933 | 2841.0472 | 0.0044  | 1.55  | 1.55  |      |    |
| 55 | H7N6S1F1-AA | 2928.0358 | -1 | 2928.05538 | 2929.0632 | -0.0196 | -6.68 | -6.68 | 12.1 | 3  |



|    |            |           |    |            |           |         |        |        |      |    |
|----|------------|-----------|----|------------|-----------|---------|--------|--------|------|----|
| 56 | H6N5S3-AA  | 2999.0027 | -1 | 2999.05610 | 3000.0639 | -0.0534 | -17.80 | -17.79 |      |    |
| 57 | H8N7F1-AA  | 3002.0879 | -1 | 3002.09216 | 3003.1000 | -0.0042 | -1.41  | -1.41  | 13.7 | 15 |
| 58 | H9N8F1-AA  | 3367.2185 | -1 | 3367.22435 | 3368.2322 | -0.0059 | -1.75  | -1.74  | 14.2 | 11 |
| 59 | H10N9F1-AA | 3732.2674 | -1 | 3732.35655 | 3733.3644 | -0.0891 | -23.88 | -23.87 |      |    |

#### 4.2. N-GLİKAN VERİLERİ ÜZERİNE MAKİNE ÖĞRENMESİ ALGORİTMALARININ TEST EDİLMESİ

*N*-glikanlara ait kütle spektrometrik veriler ekstrakte edildikten sonra elde edilen veriler kullanılarak makine öğrenmesi algoritmaları test edilmiştir. İlk olarak veriler nötr glikanlar (siyalik asit içermeyen) ve tüm *N*-glikanları içerecek şekilde ikiye ayrılmıştır. Elde edilen 2 veri seti üzerinde uygulamalar göreceli alan ve kütle alanı verileri kullanılarak gerçekleştirilmiştir. Her bir kanser ve kontrol örneğine ait maksimum 4 teknik tekrardan elde edilen veriler (toplamda 250 adet veri) analizlere dahil edilmiş olup tüm veri setlerindeki her bir algoritma için StratifiedShuffleSplit fonksiyonu kullanılmıştır. Bu fonksiyon ile veri seti test ve train olarak tabakalı bir şekilde parçalara ayrılmıştır. Veri setlerindeki 250 adet örneğin %20'si test olacak şekilde fonksiyonda tanımlanmış ve çapraz doğrulama katsayısı '5' olarak girilerek veriler her defasında farklı farklı test verilerini seçerek 5 kez algoritmaların sonucu elde edilmiştir ve bunların ortalaması analizlerde tolerans değerleriyle verilmiştir.

Makine öğrenmesi algoritmaları nötr ve tüm *N*-glikanlar üzerinde gerçekleştirilmiştir. Tüm *N*-glikanlar (relative ve analyte area) ve nötr *N*-glikanlar (relative ve analyte area) olmak üzere 4 farklı veri seti hazırlanmıştır. Çizelge 4.2 ve 4.3'te de gösterildiği gibi her bir doku örneği 1 (kanser) ve 0 (kontrol) olarak veri setinde belirtilmiştir. Anaconda Navigator Jupiter Notebook aracılığıyla makine öğrenmesi analizleri gerçekleştirilmiştir.

Çizelge 4.2. Tüm *N*-glikanlar analit alanı, “1” olarak işaretlenmiş kanserli dokular.

| Y | H3N2     | H3N2F1   | H4N2     | H3N3     | H4N2F1   | H5N2     | H3N3F1   |
|---|----------|----------|----------|----------|----------|----------|----------|
| 1 | 3163.357 | 4885.691 | 3670.848 | 3771.858 | 4772.232 | 8013.156 | 6065.159 |
| 1 | 2682.01  | 3283.915 | 2969.172 | 3009.11  | 3950     | 4904.764 | 4395.516 |
| 1 | 1011.699 | 2374.803 | 1685.587 | 1415.959 | 1734.229 | 4777.688 | 3575.582 |
| 1 | 2029.286 | 5595.329 | 2722.395 | 2742.439 | 3021.56  | 10516.33 | 8007.436 |
| 1 | 5171.1   | 14595.63 | 7797.227 | 6086.685 | 6317.381 | 34071.86 | 12907.86 |
| 1 | 6565.704 | 21391.56 | 10666    | 8118.562 | 7792.907 | 50190.53 | 19401.26 |
| 1 | 5688.255 | 18332.49 | 7915.281 | 6231.158 | 5162.146 | 63526.72 | 27781.83 |
| 1 | 6868.364 | 23212.44 | 9069.151 | 6363.016 | 5007.002 | 71640.18 | 28813    |
| 1 | 8064.862 | 10884.9  | 8984.042 | 10391.28 | 9667.907 | 35081.06 | 23056.62 |
| 1 | 10065.13 | 16738.21 | 12172.72 | 15880.92 | 11034.59 | 46085.96 | 37611.22 |
| 1 | 7547.198 | 10659.51 | 8166.127 | 10306.88 | 7321.29  | 27243.16 | 22616.73 |
| 1 | 6333.603 | 9030.857 | 7004.337 | 8722.978 | 6337.704 | 23701.51 | 19585.78 |
| 1 | 5590.865 | 11371.61 | 7498.85  | 6895.312 | 7041.481 | 34575.81 | 19715.58 |
| 1 | 6875.952 | 13860.58 | 8376.416 | 8115.543 | 8346.397 | 48360.44 | 25721.62 |

Çizelge 4.3. Tüm *N*-glikanlar analit alanı, “0” olarak işaretlenmiş kontrol dokular.

| Y | H3N2     | H3N2F1   | H4N2     | H3N3     | H4N2F1   | H5N2     | H3N3F1   |
|---|----------|----------|----------|----------|----------|----------|----------|
| 0 | 2877.194 | 4002.39  | 3598.584 | 4133.32  | 4342.009 | 9314.789 | 8449.51  |
| 0 | 2778.378 | 3813.889 | 3363.141 | 3976.488 | 4287.322 | 8581.87  | 7853.708 |
| 0 | 2253.941 | 4796.524 | 3496.138 | 6152.853 | 3649.158 | 29909.24 | 26303.11 |
| 0 | 1577.293 | 3426.975 | 2494.356 | 4345.374 | 2413.778 | 21819.54 | 18556.22 |
| 0 | 5328.989 | 8707.266 | 7413.137 | 6801.946 | 6919.254 | 35604.06 | 14704.84 |
| 0 | 8535.564 | 15117.65 | 12519.1  | 11068.89 | 9335.722 | 68124.56 | 24402.76 |
| 0 | 3689.18  | 6272.803 | 5187.03  | 4830.172 | 3966.55  | 29303.15 | 11725.71 |
| 0 | 6233.956 | 12329.66 | 8896.367 | 8630.773 | 5933.454 | 57432.3  | 28100.63 |
| 0 | 18933.33 | 54930.41 | 22724.11 | 18818.7  | 13026.16 | 138576.1 | 58163.54 |
| 0 | 4818.414 | 9582.07  | 5655.092 | 5234.391 | 5657.592 | 28698.54 | 15956.04 |
| 0 | 9521.97  | 26581.98 | 10512.85 | 9311.921 | 6657.85  | 62770.46 | 33628.21 |
| 0 | 7690.882 | 18850.86 | 7712.1   | 6726.304 | 4967.123 | 40125.16 | 21627.47 |
| 0 | 3823.447 | 6354.75  | 4511.324 | 4998.508 | 4806.575 | 19287.44 | 12494.24 |
| 0 | 3101.64  | 4771.137 | 3527.307 | 3849.955 | 3948.686 | 13318.19 | 8358.758 |
| 0 | 1293.128 | 9633.918 | 5087.82  | 8808.122 | 5110.091 | 60357.26 | 41941.24 |

Jupyter notebook'ta python programlama dili kullanılarak gerçekleştirilen makine öğrenmesi analizlerinde veri setleri sisteme tanıtılmıştır ve her veri seti için KNN, DT, SVM, RF, LR ve MPLClassifier skorları elde edilmiştir. Daha sonra her veri seti için model doğrulaması yapılmıştır ve bu bağlamda doğruluk (accuracy), karışıklık matrisi

(confusion matrix) -TP, FP, FN, TN, TPR, FPR, FNR, TNR-, f1 skor, ROC eğrisi ve ROC eğrisinin altında kalan alan olan AUC değerleri bulunmuştur.

Tüm veri setlerinde kütüphaneler import edildikten sonra veri setleri okutuldu, bunlar üzerinde gerekli ekleme-çıkarma işlemleri yapıldı ve normalizasyon işlemi gerçekleştirilmiştir. KNN, DT, SVM, RF, LR ve MPLClassifier algoritmaları stratifiedShuffleSplit fonksiyonu ile kullanılmıştır. Test size parametresi 0.2 alınarak örneklerin %20'si yani 50 tanesi test için ayrılmıştır. Özellikle birçok parametre içerdiği için MLPC'ye Grid Search hiperparametresi uygulanmıştır ve bu sayede en iyi skoru elde edebileceğimiz uygun parametreler bulunmuştur.

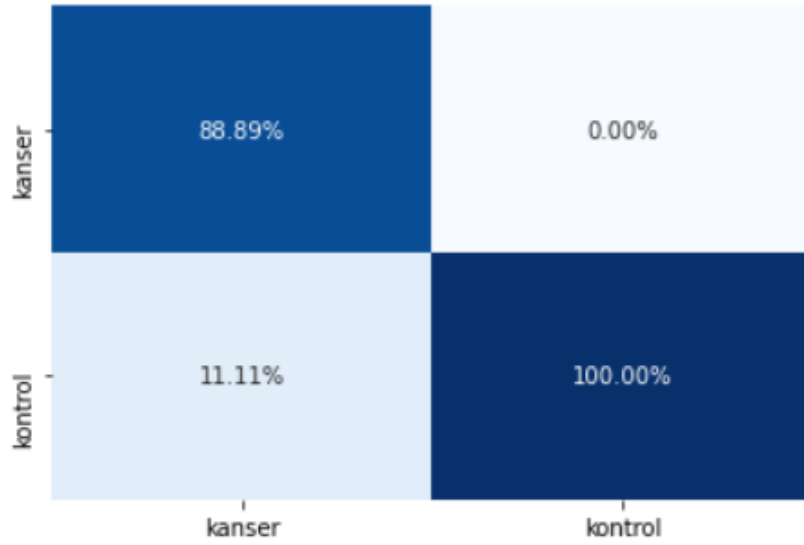
#### 4.2.1. Tüm N-Glikanlar Analit Alan Veri Setinin Makine Öğrenmesi Analizleri

Bu bağlamda ilk veri seti olan "Tüm N-Glikanlar Analyte Area (all\_a\_rev)" için en iyi skor 92.8+-3.49 ile MLPC algoritmasında bulunmuştur. Daha sonra MLPC için model değerlendirilmesi kısmına geçilmiştir. Model değerlendirmesinde doğruluk %94 olarak bulunmuştur. Model değerlendirilmesi sonucu elde edilen karışıklık matrisi Çizelge 4.4 ve Şekil 4.3'te verilmiştir.

Çizelge 4.4. Karışıklık Matrisi (TP,FN,FP,TN).

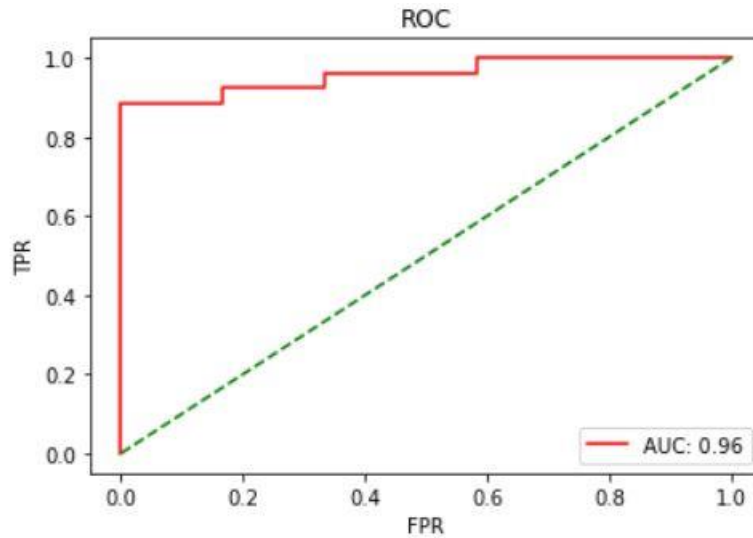
| <i>all_a_cm</i> |                |                 |
|-----------------|----------------|-----------------|
|                 | Doğru - Kanser | Doğru - Kontrol |
| Kanser Tahmini  | 24             | 0               |
| Kontrol Tahmini | 3              | 23              |

Şekil 4.3'te verildiği gibi kanserli olarak bulunan dokulardan yüzde kaçının gerçekte kanser olduğu TPR ile bulunur ve burada TPR %88.89 olarak bulunmuştur. TN/CN (Condition negative) oranı ile elde edilen TNR kontrol olan dokulardan %100'ü gerçekte de kontrol dokular olduğunun bilgisini vermektedir. Kontrol olduğu halde kanserli olarak bulunan dokular %11.11 (FNR). Pozitif olduğu halde negatif bulunan dokular yoktur yani FPR değeri %0'dır. F1 skor 0.94 olarak bulunmuştur.



Şekil 4.3. Karışıklık Matrisi Oranları (TPR, TNR, FNR ve FPR).

Şekil 4.4'te ilk veri setine ait ROC eğrisi ve AUC değeri görülmektedir. ROC eğrisi, sınıflandırma algoritmasının performansını kontrol etmek için en önemli değerlendirme ölçütlerinden biridir. Burada grafikteki kırmızı eğri sol üst tarafa yakın olduğu için kullanılan modelin başarılı olduğunu göstermektedir. Ayrıca kırmızı çizginin altında kalan alan (AUC) 0.96 olarak bulunmuştur, bu da başarı oranının yüksek olduğunu göstermektedir.



Şekil 4.4. ROC Analizi.

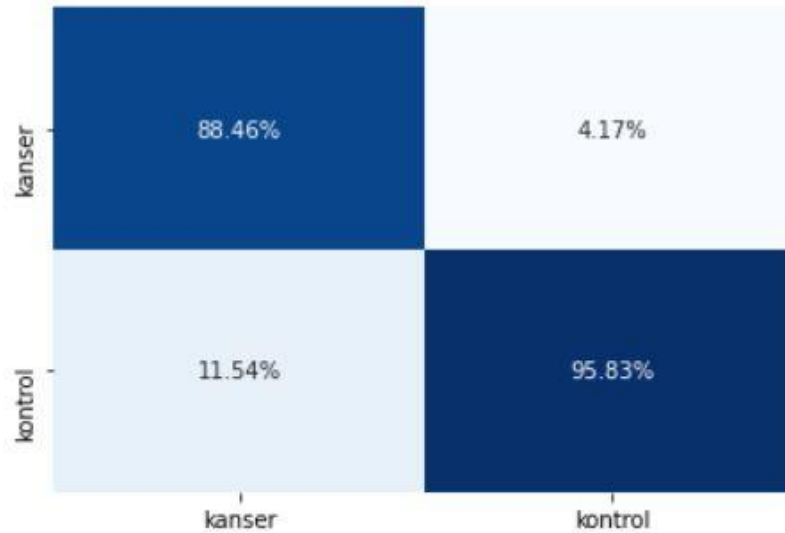
#### 4.2.2. Tüm N-Glikanlar Göreceli Alan Veri Setinin Makine Öğrenmesi Analizleri

İkinci veri setinde “Tüm N-Glikanlar Relative Area (all\_r\_rev)” en yüksek skorlar SVM’de %91.6+3.2, RF’de %90.8+-1.6 ve %92.8+-1.6 ile en iyi skor MLPC algoritmasından elde edilmiştir. Model değerlendirmesinde doğruluk %92 olarak bulunmuştur ve karışıklık matrisi Çizelge 4.5’te verilmiştir, TP:23, TN:23, FP:1, FN:3 olarak bulunmuştur.

Çizelge 4.5. Karışıklık Matrisi (TP,FN,FP,TN).

| <i>all_r_cm</i> |                |                 |
|-----------------|----------------|-----------------|
|                 | Doğru - Kanser | Doğru - Kontrol |
| Kanser Tahmini  | 23             | 1               |
| Kontrol Tahmini | 3              | 23              |

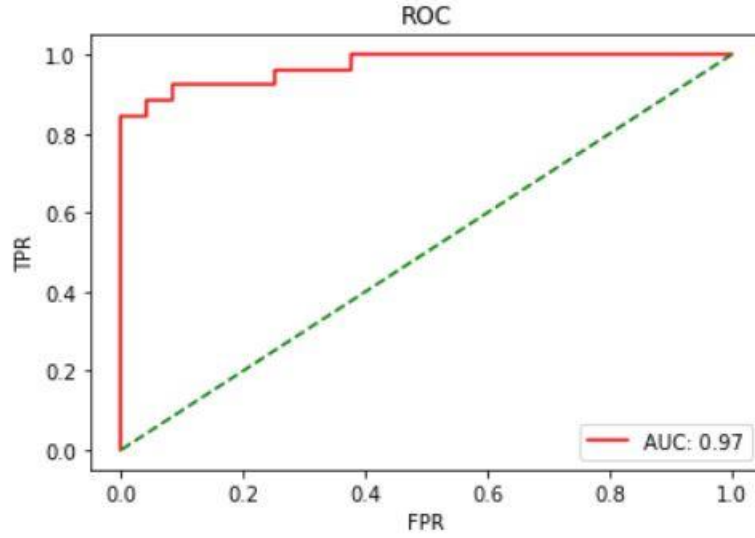
Şekil 4.5’te da gösterildiği üzere TPR %88.46, TNR %95.83, FPR %4.17 ve FNR %11.54 olarak elde edilmiştir.



Şekil 4.5. Karışıklık Matrisi Oranları.

Şekil 4.6’da ROC analizi verilmiştir, grafikten de görülebileceği üzere ilk veri setindeki gibi ROC analizi anlamı sonuç üretmiştir ve kullanılan modelin başarılı olduğunu göstermektedir, ayrıca eğri altında kalan alan 0.97 olarak bulunmuştur. Bu

da veri setinde kullanılmış olan algoritma ve modelin başarılı olduğunu göstermektedir.



Şekil 4.6. ROC Analizi.

#### 4.2.3. Nötr N-Glikanlar Analit Alan Veri Setinin Makine Öğrenmesi Analizleri

Üçüncü veri seti olan “Nötr N-Glikanlar Analite Area (notr\_a\_rev)”da bazı algoritmaların sonuçları; KNN: %81+-5, LR: %81.57+-5.9 ve en iyi sonuç MLPC: %89.02+-3.84 olarak bulunmuştur.

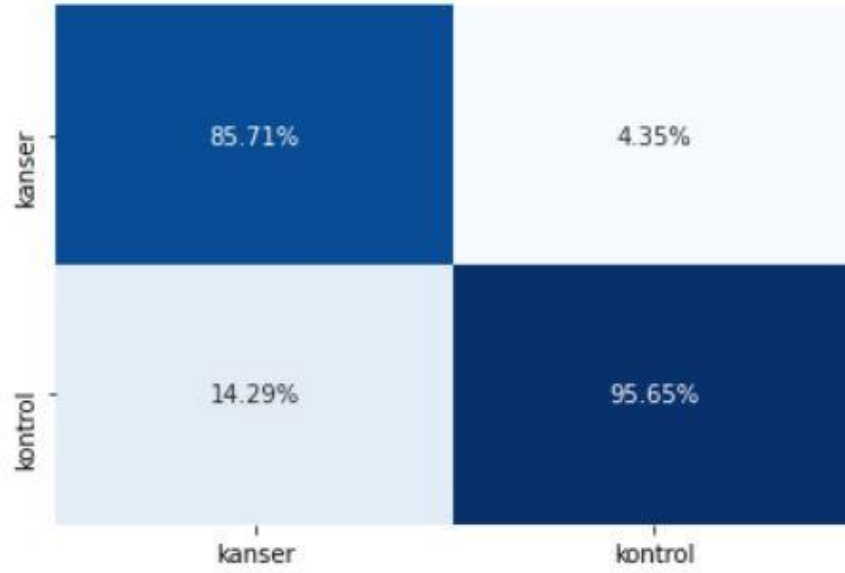
Çizelge 4.6’da üçüncü veri setindeki model değerlemesinde elde edilen karışıklık matrisi verilmiştir. Değerlendirmede karışıklık matrisi TP:24, TN:22, FN:4, FP:1 olarak elde edilmiştir.

Çizelge 4.6. Karışıklık Matrisi (TP,FN,FP,TN).

| <i>notr_a_cm</i> |                |                 |
|------------------|----------------|-----------------|
|                  | Doğru - Kanser | Doğru - Kontrol |
| Kanser Tahmini   | 24             | 1               |
| Kontrol Tahmini  | 4              | 22              |

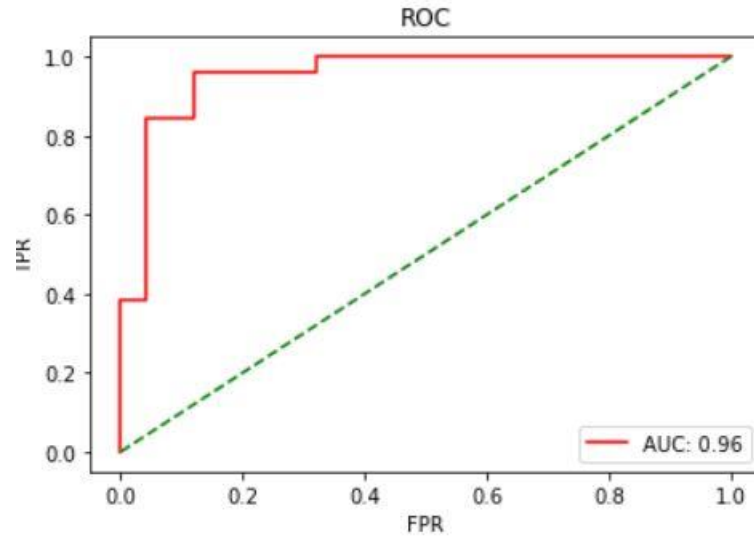
Nötr N-glikanlar Analit Alan veri setinde TPR değeri %85.71, TNR %95.65, FPR %4.35 ve FNR %14.29 olarak Şekil 4.7’deki gibi elde edilmiştir. Sonuç olarak bulunan TPR değeri ile sınıflandırma algoritmasının ne kadar gerçek pozitif değeri doğru

tahmin ettiğini yani kanserli dokuları doğru etme oranını göstermektedir. TNR ise gerçekte kontrol dokusu olan seçilen dokuların kaçının doğru olarak seçildiğinin oranını belirtmektedir.



Şekil 4.7. Karışıklık Matrisi Oranları.

Diğer veri setlerindeki ROC analizleri gibi bu veri setinde de ROC eğrisi ve altında kalan alan olan AUC değeri anlamlı sonuçlar üretmiştir. Hem AUC değerine bakarak hem de grafiği incelendiğinde kullanılan algoritma ve modellerin başarılı olduğu görülmektedir. Şekil 4.8’de verildiği üzere ROC eğrisi altında kalan alan 0.96 olarak bulunmuştur.



Şekil 4.8. ROC Analizi.

#### 4.2.4. Nötr N-Glikanlar Göreceli Alan Veri Setinin Makine Öğrenmesi Analizleri

Dördüncü ve son veri setinde “Nötr N-Glikanlar Relative Area (notr\_r\_rev)” en iyi sonucu üreten algoritma %93.6+/-0.8 ile MLPC algoritması olmuştur.

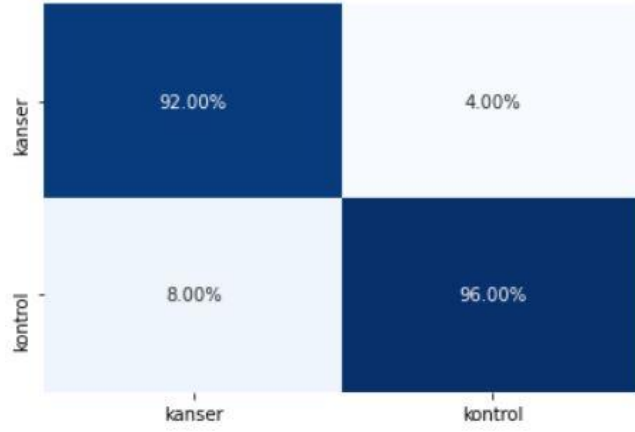
Model değerlendirmesi sonucu elde edilen confusion matrix değerleri Tablo 4.7’de verilmiştir. Burada TP:23, TN:24, FP:1, FN:2 olarak elde edilmiştir.

Çizelge 4.7. Karışıklık Matrisi (TP, FN, FP, TN).

| <i>notr_r_cm</i> |                |                 |
|------------------|----------------|-----------------|
|                  | Doğru - Kanser | Doğru - Kontrol |
| Kanser Tahmini   | 23             | 1               |
| Kontrol Tahmini  | 2              | 24              |

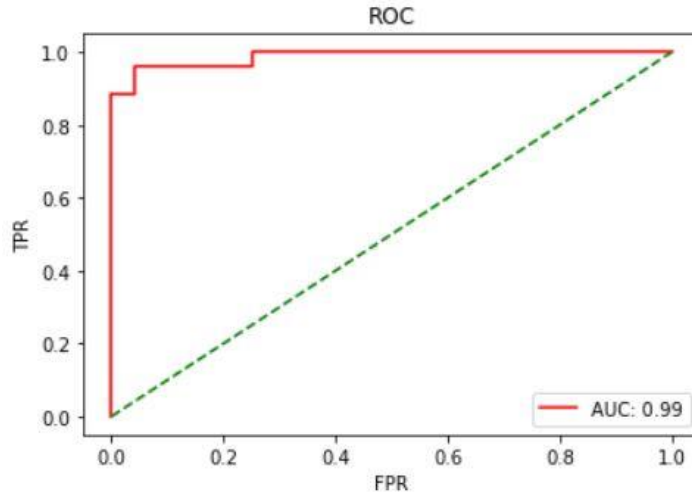
Nötr N-glikanlar Göreceli Alan veri setinde TPR, TNR, FPR ve FNR şekil 4.9’daki gibi elde edilmiştir. Sonuçlar sırasıyla %92, %96, %4 ve %8 olarak bulunmuştur. Sonuç olarak bakıldığı zaman hem yukarıdaki tablo 4.5’te hem de TPR ve TNR değerleri gerçekte kanser ve kontrol olan dokuların büyük oranda doğru tahmin edilildiğini göstermektedir.





Şekil 4.9. Karışıklık Matrisi Oranları.

Şekil 4.10'da gösterilen ROC analizinde AUC değeri 0.99 bulunmuştur. Diğer veri setleri ile kıyaslandığı zaman nötr N-glikanların göreceli alan veri setinin ROC analizi diğerlerinden daha yüksek bulunmuştur.



Şekil 4.10. ROC Analizi.

## BÖLÜM 5

### SONUÇLAR VE ÖNERİLER

Bu yüksek lisans tezinde mide kanser ve kontrol dokuları analizleri sonucunda *N*-glikan profili MALDI-MS ile başarılmıştır. Mide dokularına ait 59 adet *N*-glikan yapısı belirlenmiş olup bunların 37 adetinin yapısı MS/MS düzeyinde doğrulanmıştır. Glikanlara ait kütle spektrometrik veriler elde edildikten makine öğrenmesi uygulamaları nötr ve tüm *N*-glikanlar üzerinde gerçekleştirilmiştir.

DeneySEL analizler sonucu elde edilen 4 farklı veri seti makine öğrenmesi analizlerinde kullanılmıştır. Analizlerde KNN, SVM, DT, RF, LR ve MLPC algoritmaları kullanılmıştır. Sonuç olarak:

1. “Tüm *N*-Glikanlar Analit Area” veri setinde en iyi sonuçlar sırasıyla MLPC, SVM ve RF algoritmalarından elde edilmiştir.
2. “Tüm *N*-Glikanlar Relative Area” veri setinde en iyi sonuçlar sırasıyla MLPC, SVM ve RF algoritmalarından elde edilmiştir.
3. “Nötr *N*-Glikanlar Analit Area” veri setinde en iyi sonuçlar sırasıyla MLPC, SVM ve LR algoritmalarından elde edilmiştir.
4. “Nötr *N*-Glikanlar Relative Area” veri setinde en iyi sonuçlar sırasıyla MLPC, SVM ve RF algoritmalarından elde edilmiştir.

Yapılan deneyler ve makine öğrenmesi analizleri sonucunda en iyi ayrımlar nötr *N*-glikanlar göreceli alan verilerinde ve tüm *N*-glikanların analit alan veri setlerinde gözlemlenmiştir. Tüm veri setlerinde en yüksek doğruluk skoru MLPC algoritması ile elde edilmiştir. ROC analizi sonucunda en iyi AUC değeri 0.99 ile Nötr *N*-glikan Göreceli Alan veri setinden elde edilmiştir. En yüksek MLPC doğruluk skoruna sahip veri seti ise yine Nötr *N*-glikan Göreceli Alan veri seti bu tez kapsamında belirlenmiştir.

Bu alıřma ile mide kanseri dokularına ynelik ktle spektrometresi temelli glikan analizleri ve makine đrenmesi kullanılarak kanser/kontrol ayrımı bařarılı bir řekilde gerekleřtirmiřtir. İlerleyen alıřmalarda geliřtirilen biyoanalitik yntem kullanılarak farklı kanser dokuları zerine alıřma geniřletilebilir ve kanser trlerinin erken tanısında test edilen makine đrenmesi algoritmaları kullanılabilir.

## KAYNAKLAR

1. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S., "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence", *Proteomics*, 4 (6): 1633–1649 (2004).
2. Kayili, H. M., "Glikopeptitlerin kütle spektrometrik analizleri için yeni bir biyoanalitik yöntem geliştirilmesi ve insan c1-inhibitör proteininin ayrıntılı bölgeye özgü glikozilasyon analizinin gerçekleştirilmesi", *Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Kimya Anabilim Dalı, Ankara*, (2016).
3. Li, Z., Yuan, B., Lin, X., Meng, X., Wen, X., Guo, Q., Li, L., Jiang, H., and Wang, K., "Intramolecular trigger remodeling-induced HCR for amplified detection of protein-specific glycosylation", *Talanta*, 215 (August 2020): Makale: 120889 (2020).
4. Uslupehlivan, M., "Transkripsiyon faktörü tñib'nin glikozilasyon profilinin biyoinformatik ve deneysel olarak belirlenmesi", *Ege Üniversitesi Fen Bilimleri Enstitüsü Biyoloji Anabilim Dalı, İzmir*, (2017).
5. Stanley P, Taniguchi N, Aebi M. N-Glycans. 2017. In: Varki A, Cummings RD, E. J., "Essentials of Glycobiology [Internet]. 3rd Edition", Bölüm 7-48 (2017).
6. Zabczyńska, M., Kozłowska, K., and Pocheć, E., "Glycosylation in the thyroid gland: Vital aspects of glycoprotein function in thyrocyte physiology and thyroid disorders", *International Journal Of Molecular Sciences*, 19 (9): 2792 (2018).
7. Van Den Steen, P., Rudd, P. M., Dwek, R. A., and Opdenakker, G., "Concepts and principles of O-linked glycosylation", *Critical Reviews In Biochemistry And Molecular Biology*, 33 (3): 151–208 (1998).
8. Wang, H., Ramakrishnan, A., Fletcher, S., Prochownik, E. V., and Genetics, M., "Protein glycosylation in cancer", *Annu Rev Pathol*, 2 (2): 473–510 (2015).
9. The American Cancer Society medical and editorial content team, "About Stomach Cancer", *The American Cancer Society*, 1–14 (2018).
10. Alçay, N., "Sağlık ve hastalıkta glikoproteinlerin rolü", *Ankara Üniversitesi*

*Sağlık Bilimleri Enstitüsü Biyokimya Anabilim Dalı, Ankara, (2006).*

11. De Hoffmann, E., "Mass spectrometry", *Kirk-Othmer Encyclopedia Of Chemical Technology*, 1–20 (2005).
12. Nie, Z. X., Tzeng, Y. K., Chang, H. C., Chiu, C. C., Chang, C. Y., Chang, C. M., and Tao, M. H., "Microscopy-based mass measurement of a single whole virus in a cylindrical ion trap", *Angewandte Chemie - International Edition*, 45 (48): 8131–8134 (2006).
13. Hoffmann, E. de and Stroobant, V., "Mass Spectrometry", *Wiley*, Chichester, 1–489 (2007).
14. Berkel, G. J. Van, Pasilis, S. P., and Ovchinnikova, O., "Established and emerging atmospheric pressure surface sampling/ionization techniques for mass spectrometry", *Journal Of Mass Spectrometry*, 43 (9): 1161–1180 (2008).
15. Ho, C., Lam, C., Chan, M., Cheung, R., Law, L., Lit, L., Ng, K., Suen, M., and Tai, H., "Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications", *The Clinical Biochemist Reviews*, 24 (1): 3–12 (2003).
16. Sjö, J., "Miniaturized Techniques for Protein Analysis", *KTH Royal Institute of Technology, Division of Analytical Chemistry, Stockholm*, (2004).
17. Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F., "Matrix-assisted ultraviolet laser desorption of non-volatile compounds", *International Journal Of Mass Spectrometry And Ion Processes*, 78 (C): 53–68 (1987).
18. Editörler: Hillenkamp, F. and Peter-Katilinic, J., "MALDI MS - A Practical Guide to Instrumentation, Methods and Applications", Second. Ed., *Wiley*, Weinheim, 1–477 (2014).
19. Soler, L., Uzbekova, S., Blesbois, E., Druart, X., and Labas, V., "Intact cell MALDI-TOF mass spectrometry, a promising proteomic profiling method in farm animal clinical and reproduction research", *Theriogenology*, 150: 113–121 (2020).
20. Internet: Şafak, H. İ., "Makine Öğrenmesi Nedir", <https://medium.com/türkiye/makine-öğrenmesi-nedir-20dee450b56e> (2017).

21. Ali, A., Alrubei, M., Hassan, L. F. M., Al-Ja'afari, M., and Abdulwahed, S., "Diabetes classification based on KNN", *IJUM Engineering Journal*, 21 (1): 175–181 (2020).
22. Internet: JavaTpoint, "K-Nearest Neighbor(KNN) Algorithm for Machine Learning", <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning> (2021).
23. Ramasamy, U. and Sundar, S., "An Illustration of Rheumatoid Arthritis Disease Using Decision Tree Algorithm", *Informatika*, 46 (1): 107–119 (2022).
24. Internet: Chauhan, N. S., "Decision Tree Algorithm, Explained", <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html#> (2022).
25. Fidan, H., "Random Forest (Rastgele Orman) Algoritması Temelli Süreç İzleme Yönteminin Ambulatuvar Kan Basıncı İzlemede Hipertansiyonun Erken Tanısı İçin Kullanımı", (2020).
26. Internet: Gandhi, R., "Support Vector Machine — Introduction to Machine Learning Algorithms", <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> (2018).
27. Guenther, N. and Schonlau, M., "Support vector machines", *Stata Journal*, 16 (4): 917–937 (2016).
28. G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, and Mensinkal, K., "Logistic Regression Technique for Prediction of Cardiovascular Disease", *Global Transitions Proceedings*, 3 (April): 127–130 (2022).
29. Castellani, M., "Competitive co-evolution of multi-layer perceptron classifiers", *Soft Computing*, 22 (10): 3417–3432 (2018).
30. Internet: Işıkhani, E., "Multi Layer Perceptron (MLP) Nedir?", <https://isikhanelif.medium.com/multi-layer-perceptron-mlp-nedir-4758285a7f15> (2020).
31. Internet: Simsek, H. K., "Makine Öğrenmesi Dersleri 10: Sınıflandırma Modellerinde Başarı Kriterleri", <https://medium.com/data-science-tr/sınıflandırma-modellerinde-başarı-kriterleri-2d86488799c6> (2018).

**EK AÇIKLAMALAR A.**

**ETİK KURUL İZİNİ**



**T.C.  
KARABÜK ÜNİVERSİTESİ  
GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR  
ETİK KURULU**

27/2/2020

Karar No: 2020/192

Sayın Dr. Öğr. Üyesi Hacı Mehmet KAYILI

Girişimsel Olmayan Klinik Araştırmalar Etik Kurulumuza sunmuş olduğunuz "Kütle Spektrometresi Temelli Kliniksel Glikomik Çalışmalarına Yönelik Bir Yapay Zekâ Uygulaması" başlıklı araştırma projeniz amaç, gerekçe, yaklaşım ve yöntemle ilgili açıklamaları açısından Girişimsel Olmayan Etik Kurulu yönergesine göre incelenmiştir. Etik açıdan bir sakınca olmadığına oy çokluğu ile karar verilmiş ve uygun görülmüştür. Bilgilerinize rica ederim.

Dr. Öğr. Üyesi Zafer LİMAN  
Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu Başkanı

Ekler:

1-Kurul Üyeleri Değerlendirme Formları

Şekil Ek A.1. Etik Kurul İzni.



## ÖZGEÇMİŞ

Deniz Baran DEMİRHAN ilkokulu 2009 yılında Arif Eminoglu İlköğretim Okulu'nda tamamladı. Selaheddini Eyyubi Anadolu Lise'sinden 2013 yılında mezun oldu. 2013 yılında Karabük Üniversitesi Tıp Mühendisliği bölümünü kazandı ve aynı üniversitede 1 sene İngilizce hazırlık eğitimi gördükten sonra 2014 yılında lisans eğitimine başladı ve 2018 yılında Onur öğrencisi olarak mezun oldu. Yüksek lisans eğitimini Karabük Üniversitesi Lisansüstü Eğitim Enstitüsü Biyomedikal Mühendisliği Anabilim Dalı'nda "Kütle Spektrometresi Temelli Kliniksel Glikomik Çalışmalarına Yönelik Bir Makine Öğrenmesi Uygulaması" adlı tezini savunarak tamamladı ve mezun oldu.