

**T.C.**  
**FIRAT ÜNİVERSİTESİ**  
**FEN BİLİMLERİ ENSTİTÜSÜ**  
**YAZILIM MÜHENDİSLİĞİ ANABİLİM DALI**

**DNA DİZİLİMLERİNDEN HASTALIK TANILANMASI İÇİN İŞARET İŞLEME**  
**TEMELLİ YENİ YAKLAŞIMLARIN GELİŞTİRİLMESİ**

**DOKTORA TEZİ**

**Bihter DAŞ**  
**(122137204)**

**Tezin Enstitüye Verildiği Tarih: 03/07/2018**

**Tezin Savunulduğu Tarih: 06/08/2018**

**Danışman : Prof. Dr. İbrahim TÜRKOĞLU (Fırat Ü.) .....**

**Jüri Üyeleri : Prof. Dr. Mehmet KAYA (Fırat Ü.) .....**

**Prof. Dr. M. Sıraç ÖZERDEM (Dicle Ü.) .....**

**Doç. Dr. Bilal ALATAŞ (Fırat Ü.) .....**

**Dr. Öğr. Üyesi Mahmut SİNECEN (A.D.Ü) .....**

**AĞUSTOS - 2018**

## ÖNSÖZ

Doktora eğitimimin başladığı günden beri değerli fikirleri ve tecrübeleri ile akademik çalışmalarına ışık tutan saygıdeğer danışman hocam Sayın *Prof. Dr. İbrahim TÜRKOĞLU*'na teşekkürlerimi sunarım. Ayrıca, bu tez çalışmasında Intergen Genetik Hastalıklar Tanı Merkezinden Beta talasemi DNA dizilimlerinin elde edilmesinde *Doç. Dr. Serdar CEYLANER*'e, DNA verilerinin yorumlanması ve genetik alanında her türlü bilgiyi almada benden yardımını esirgemeyen *Dr. Öğr. Üyesi Aşkın Şen*'e katkılarından dolayı teşekkür ederim.

Çalışmalarım süresince desteklerini esirgemeyen aileme, çalışmalarımda beni yüreklendiren ve yaşamımda her konuda hep yanımda olan eşim *Doç. Dr. Resul DAŞ*'a ve biricik kızım *Zeynep Eyşan DAŞ*'a teşekkür ederim.

**Bihter DAŞ**  
**Elazığ - 2018**

# İÇİNDEKİLER

## Sayfa No

|          |                                                                               |           |
|----------|-------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>GİRİŞ.....</b>                                                             | <b>1</b>  |
| 1.1      | Doktora Çalışmasının Amacı ve Kapsamı.....                                    | 1         |
| 1.2      | Literatür Taraması ve Değerlendirmesi .....                                   | 1         |
| 1.3      | Tezin Organizasyonu.....                                                      | 5         |
| <b>2</b> | <b>DNA’NIN YAPISI VE DİZİLİMLERİ.....</b>                                     | <b>7</b>  |
| 2.1      | DNA’nın Yapısı .....                                                          | 7         |
| 2.2      | DNA Dizilimleri.....                                                          | 11        |
| <b>3</b> | <b>DNA DİZİLİMİNİN SAYISAL BİÇİME DÖNÜŞTÜRÜLME TEKNİKLERİ.....</b>            | <b>13</b> |
| 3.1      | Sabit Haritalama Teknikleri.....                                              | 13        |
| 3.1.1    | Voss Haritalama Tekniği.....                                                  | 13        |
| 3.1.2    | Tetrahedron Haritalama Tekniği.....                                           | 14        |
| 3.1.3    | Karmaşık Haritalama Tekniği.....                                              | 15        |
| 3.1.4    | Tamsayı Haritalama Tekniği .....                                              | 16        |
| 3.1.5    | Reel Haritalama Tekniği.....                                                  | 16        |
| 3.1.6    | Kuaterniyon Haritalama Tekniği.....                                           | 17        |
| 3.1.7    | Moleküler Kütle Haritalama Tekniği .....                                      | 18        |
| 3.2      | Fiziko-Kimyasal Özellik Tabanlı Haritalama Teknikleri .....                   | 18        |
| 3.2.1    | EIIP Haritalama Tekniği.....                                                  | 18        |
| 3.2.2    | Atomik Sayı Tekniği.....                                                      | 18        |
| 3.2.3    | Eşleştirilmiş Sayısal Haritalama Tekniği.....                                 | 18        |
| 3.2.4    | DNA-Walk Haritalama Tekniği.....                                              | 19        |
| 3.2.5    | Z-Eğrisi Haritalama Tekniği.....                                              | 19        |
| <b>4</b> | <b>SAYISAL HARİTALAMA TEKNİKLERİNİN BAŞARIMLARININ DEĞERLENDİRİLMESİ.....</b> | <b>21</b> |
| 4.1      | Kullanılan Sınıflandırma Yöntemleri .....                                     | 21        |
| 4.1.1    | Ayrık Fourier Dönüşümü Yöntemi .....                                          | 21        |
| 4.1.2    | Destek Vektör Makineleri.....                                                 | 23        |
| 4.1.3    | K-En Yakın Komşu Algoritması.....                                             | 24        |

|          |                                                                                      |           |
|----------|--------------------------------------------------------------------------------------|-----------|
| 4.2      | Önerilen Uygulama Süreci.....                                                        | 24        |
| 4.3      | DNA Verilerinin Elde Edilmesi.....                                                   | 25        |
| 4.4      | Sayısal Haritalama Teknikleri ile DNA Diziliminin Sayısallaştırılması .....          | 26        |
| 4.5      | DNA Dizilimlerindeki Ekzon/Intron Bölgelerinin Sınıflandırılması .....               | 27        |
| 4.5.1    | Ayrık Fourier Dönüşümü Yöntemi ile Sınıflandırma.....                                | 27        |
| 4.5.2    | Destek Vektör Makineleri Yöntemi ile Sınıflandırma .....                             | 28        |
| 4.5.3    | $k$ -En Yakın Komşu Algoritması ile Sınıflandırma.....                               | 28        |
| 4.6      | Sınıflandırma Başarımlarının Karşılaştırılması.....                                  | 28        |
| 4.6.1    | Sınıflandırma Başarımlarının ROC Eğrileriyle Analizi .....                           | 30        |
| 4.7      | Uygulama Sonuçlarının Değerlendirilmesi.....                                         | 32        |
| <b>5</b> | <b>DNA DİZİLİMLERİ İÇİN ENTROPİ TABANLI YENİ BİR SAYISAL HARİTALAMA TEKNİĞİ.....</b> | <b>33</b> |
| 5.1      | Entropiye Dayalı Yeni Sayısal Haritalama Tekniği .....                               | 33        |
| 5.2      | Ekzon ve Intron Bölgelerinin Belirlemesi İçin Kullanılan Yöntemler .....             | 35        |
| 5.2.1    | Tekli Değer Ayırışım Yöntemi .....                                                   | 36        |
| 5.2.2    | Kısa Zamanlı Fourier Dönüşümü.....                                                   | 36        |
| 5.3      | Önerilen Entropiye Dayalı Sayısal Haritalama Tekniğinin Uygulaması.....              | 36        |
| 5.4      | Uygulama Sonuçları .....                                                             | 39        |
| 5.4.1    | Tekli Değer Ayırışımı Yönteminin Sınıflandırma Sonuçları.....                        | 39        |
| 5.4.2    | Ayrık Fourier Dönüşümü Yönteminin Sınıflandırma Sonuçları.....                       | 41        |
| 5.4.3    | Kısa Zamanlı Fourier Dönüşümü Yönteminin Sınıflandırma Sonuçları.....                | 42        |
| 5.5      | Uygulama Sonuçlarının Değerlendirilmesi.....                                         | 43        |
| <b>6</b> | <b>DNA DİZİLİMLERİNDEN BETA TALASEMİ HASTALIĞININ TANILANMASI... 44</b>              |           |
| 6.1      | Materyal ve Yöntem.....                                                              | 45        |
| 6.1.1    | Veri Kaynağı.....                                                                    | 45        |
| 6.1.2    | Gen Diziliminden Özellik Çıkarımı .....                                              | 46        |
| 6.1.3    | Yapay Sinir Ağı ile Sınıflandırma.....                                               | 48        |
| 6.2      | Sınıflandırma Sonuçlarının Yorumlanması .....                                        | 50        |
| <b>7</b> | <b>SONUÇ.....</b>                                                                    | <b>56</b> |
| 7.1      | Sonuçların İrdelenmesi .....                                                         | 56        |

|     |                        |    |
|-----|------------------------|----|
| 7.2 | Öneriler.....          | 58 |
| 7.3 | Yayınlar ve Proje..... | 58 |



## ŞEKİLLER LİSTESİ

### Sayfa No

|                                                                                                                                |    |
|--------------------------------------------------------------------------------------------------------------------------------|----|
| Şekil 2.1 DNA'nın çift sarmal yapısı [32] .....                                                                                | 8  |
| Şekil 2.2 Bir DNA'nın kimyasal yapısı [34] .....                                                                               | 8  |
| Şekil 2.3 RNA polimerazın DNA'dan bir mRNA üretimi [31] .....                                                                  | 9  |
| Şekil 2.4 DNA molekülünü meydana getiren nükleotidlerin dizilişi [47] .....                                                    | 11 |
| Şekil 2.5 RNA'nın yapısı [34] .....                                                                                            | 12 |
| Şekil 3.1 Tetrahedron tekniği [56] .....                                                                                       | 14 |
| Şekil 3.2 Bazların sol ve sağ düzlemde gösterimi .....                                                                         | 16 |
| Şekil 3.3 Kuaterniyon tekniği [64] .....                                                                                       | 17 |
| Şekil 4.1 Destek vektör makine algoritması [102] .....                                                                         | 24 |
| Şekil 4.2 Ekzon ve intron sınıflandırma için önerilen yaklaşımın uygulama adımları .....                                       | 25 |
| Şekil 4.3 Ensembl veri tabanındaki ekzon görüntüsünden bir kesit .....                                                         | 25 |
| Şekil 4.4 Ensembl veri tabanındaki intron görüntüsünden bir kesit .....                                                        | 26 |
| Şekil 4.5 Ekzon ve intron sınıflandırmada 3 haritalama tekniğinin ROC eğrileriyle başarımları .....                            | 31 |
| Şekil 5.1 AFD, TDA ve KZFD yöntemlerini kullanılarak önerilen haritalama tekniği ile ekzon bölgelerinin tahmini .....          | 37 |
| Şekil 5.2 NCBI veri tabanından elde edilen AJ229040 geninden bir kesit .....                                                   | 37 |
| Şekil 5.3 AJ229040 genine ait genel bilgiler .....                                                                             | 38 |
| Şekil 5.4 AJ229040 genine ait DNA diziliminin ekzon bölgelerinin temsili .....                                                 | 38 |
| Şekil 5.5 ROC eğrisinde kullanılan terimler [131] .....                                                                        | 39 |
| Şekil 5.6 Ekzon bölgelerinin taranmasında farklı sayısal haritalama tekniklerinin karşılaştırılması .....                      | 40 |
| Şekil 5.7 Baz pozisyonlarına göre güç spektrumunu ifade eden ekzonlar .....                                                    | 41 |
| Şekil 5.8 KZFD yönteminde baz pozisyonlarına göre ekzon bölgeleri .....                                                        | 42 |
| Şekil 6.1 Önerilen Entropi tabanlı haritalama tekniğinin 3 katlı çaprazlama doğrulamasına göre sınıflandırma başarımları ..... | 54 |

## TABLÖLAR LİSTESİ

### Sayfa No

|                                                                                                                                                                    |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Tablo 2.1 Uluslararası saf ve uygulamalı kimya birliğinin kuralları.....                                                                                           | 12 |
| Tablo 3.1 Voss haritalama tekniğine göre sayısallaştırılmış DNA dizilimi .....                                                                                     | 13 |
| Tablo 4.1 DNA sayısal haritalama teknikleri .....                                                                                                                  | 27 |
| Tablo 4.2 Pencere fonksiyonlarına göre sınıflandırma başarımları .....                                                                                             | 29 |
| Tablo 4.3 Sayısal haritalama teknikleri için Fourier tabanlı yöntemin sınıflandırma başarımının<br>diğer makine öğrenme tabanlı yöntemlerle karşılaştırılması..... | 30 |
| Tablo 5.1 Gen bankasından elde edilen DNA dizilimindeki gerçek ekzon pozisyonları ile yöntem<br>sonucu tahmin edilen ekzon pozisyonlarının karşılaştırılması ..... | 42 |
| Tablo 5.2 Eşik değere göre sayısal haritalama tekniklerinin doğruluk oranlarının karşılaştırılması                                                                 | 42 |
| Tablo 5.3 KZFD yönteminde ortalama başarıya göre sayısal haritalama tekniklerinin başarımlarının<br>karşılaştırılması .....                                        | 43 |
| Tablo 6.1 EIIP haritalama tekniğine göre normalizasyon sonucu özellik değerlerinden bazıları....                                                                   | 47 |
| Tablo 6.2 Integer haritalama tekniğine göre normalizasyon sonucu özellik değerlerinden bazıları                                                                    | 48 |
| Tablo 6.3 Önerilen sayısal haritalama tekniğine göre normalizasyon sonucu özellik değerlerinden<br>bazıları .....                                                  | 48 |

## KISALTMALAR LİSTESİ

|             |                                                                           |
|-------------|---------------------------------------------------------------------------|
| <b>DNA</b>  | : Deoksiribo Nükleik Asit                                                 |
| <b>RNA</b>  | : Ribo Nükleik Asit                                                       |
| <b>A</b>    | : Adenin                                                                  |
| <b>G</b>    | : Guanin                                                                  |
| <b>C</b>    | : Sitozin                                                                 |
| <b>T</b>    | : Timin                                                                   |
| <b>EIIP</b> | : Electron Ion Intraction Potential – Elektron İyon Etkileşim Potansiyeli |
| <b>CIS</b>  | : Complex Indicator Sequence                                              |
| <b>SVM</b>  | : Support Vector Machine                                                  |
| <b>DVM</b>  | : Destek Vektör Makinesi                                                  |
| <b>YSA</b>  | : Yapay Sinir Ağı                                                         |
| <b>ANN</b>  | : Artifical Neural Network                                                |
| <b>k-NN</b> | : <i>k</i> - Nearest Neighbors – <i>k</i> -En Yakın Komşu                 |
| <b>RBF</b>  | : Radyal Tabanlı Fonksiyon                                                |
| <b>KZFD</b> | : Ayrık Zamanlı Fourier Dönüşümü                                          |
| <b>AFD</b>  | : Ayrık Fourier Dönüşümü                                                  |
| <b>DFT</b>  | : Discrete Fourier Transform                                              |
| <b>SVD</b>  | : Singular Value Decomposition                                            |
| <b>TDA</b>  | : Tekli Değer Ayrışımı                                                    |

## SİMGELER VE SEMBOLLER LİSTESİ

|              |                                                                 |
|--------------|-----------------------------------------------------------------|
| $X_n$        | : Bir DNA gen dizilimi                                          |
| $X_e(n)$     | : Bir DNA dizilimi boyunca serbest elektron enerji dağılımı     |
| $x_r(n)$     | : Tetrahedron tekniğinde red (kırmızı) rengi gösteren x vektörü |
| $x_g(n)$     | : Tetrahedron tekniğinde green (yeşil) rengi gösteren x vektörü |
| $x_b(n)$     | : Tetrahedron tekniğinde blue (mavi) rengi gösteren x vektörü   |
| $X_A(n)$     | : Bir DNA diziliminde sadece Adenin gösterilmesi                |
| $X_T(n)$     | : Bir DNA diziliminde sadece Timinin gösterilmesi               |
| $X_G(n)$     | : Bir DNA diziliminde sadece Guaninin gösterilmesi              |
| $X_C(n)$     | : Bir DNA diziliminde sadece Sitozinin gösterilmesi             |
| $P_n$        | : Z-Curve tekniğinde Z-eğrisi üzerindeki düğüm sayısı           |
| $L$          | : DNA diziliminin baz uzunluğu                                  |
| $N$          | : $X(n)$ dizisinin üzerinde bulunan ayırık nokta sayısı         |
| $W_n$        | : x dizisi ile AFD olan X dizisi arasındaki dönüşüm matrisi     |
| $X_p(n)$     | : N periyotlu $X(n)$ dizisi                                     |
| $X_m(k)$     | : Pencerelemiş her bir dizilim                                  |
| $X_T(k)$     | : DFT spektrumunun normalize edilmiş toplamı                    |
| $X_m(k)$     | : Pencerelemiş her bir dizilim                                  |
| $S(k)$       | : DNA diziliminin güç spektrumu                                 |
| $P_3$        | : 3 periyotlu spektral bileşeni                                 |
| $T_3$        | : Eşik değeri                                                   |
| $sdP_{3e}$   | : 3 periyotlu ekzonların standart sapması                       |
| $meanP_{3i}$ | : 3 periyotlu intronların ortalaması                            |
| $meanP_{3e}$ | : 3 periyotlu ekzonların ortalaması                             |
| $sdP_{3i}$   | : 3 periyotlu intronların standart sapması                      |
| $P_{3t}$     | : 3 periyotlu test dizisi                                       |
| $S$          | : Shannon entropisi                                             |

## DİL DÖNÜŞÜM TABLOSU

|                       |                                                                                                   |
|-----------------------|---------------------------------------------------------------------------------------------------|
| <b>Sekans</b>         | : DNA Dizilimi                                                                                    |
| <b>Ekzon</b>          | : İfade edilebilen bölgeler, genlerin kodlanan bölümü (Expressed regions)                         |
| <b>Intron</b>         | : İntragenik bölgeler, genlerin kodlanmayan bölümü (Intragenic regions)                           |
| <b>Pürin</b>          | : Adenin ve Guanin organik bazları                                                                |
| <b>Pirimidin</b>      | : Timin ve Sitozin organik bazları                                                                |
| <b>Kodon</b>          | : 3' lü baz                                                                                       |
| <b>Ribozom</b>        | : Hücrenin protein sentez yerlerine verilen ad                                                    |
| <b>Nükleik</b>        | : Nükleotid birimlerinden oluşmuş polimerler                                                      |
| <b>Polimer</b>        | : Küçük moleküllerin birbirlerine tekrarlar halinde eklenmesiyle oluşan çok uzun zincirli molekül |
| <b>Nükleotid</b>      | : 1 fosfat, 5 karbonlu şeker (deoksiriboz) ve 1 organik baz birleşiminden oluşan kimyasal bileşik |
| <b>Beta talasemi:</b> | Akdeniz Anemisi                                                                                   |
| <b>HBB geni</b>       | : Akdeniz Anemisi hastalığı HBB genindeki mutasyon sonucu meydana gelir.                          |
| <b>Baz</b>            | : Adenin(A), Guanin (G), Sitozin (C), Timin(T)'dir.                                               |

# **DNA DİZİLİMLERİNDEN HASTALIK TANILANMASI İÇİN İŞARET İŞLEME TEMELLİ YENİ YAKLAŞIMLARIN GELİŞTİRİLMESİ**

**Bihter DAŞ**

**Fırat Üniversitesi, Fen Bilimleri Enstitüsü**

**Yazılım Mühendisliği Anabilim Dalı**

**Doktora Tezi, Ağustos 2018**

**Tez Danışmanı: Prof. Dr. İbrahim TÜRKOĞLU**

## **ÖZET**

Son yıllarda, biyoenformatik alanında DNA ile ilgili yapılan araştırmalar, DNA yapısının anlaşılması ve canlılardaki genlerin etkilerinin tespit edilmesi açısından büyük önem taşımaktadır. Özellikle, DNA dizilimlerindeki ekzon bölgelerinin belirlenmesi, genetik hastalıkların tanınması, genomik özelliklerin karşılaştırılması ile ilgili çalışmalar yapılmaktadır.

Bu tez çalışmasında, DNA dizilimlerinden genetik hastalıkların tanınmasına yönelik optimum çözüm önerileri sunabilecek yeni yaklaşımlar geliştirilmiştir. Bu amaç doğrultusunda, üç aşamalı bir çalışma yapılmıştır:

- a) Birinci aşamada, DNA dizilimlerinin sayısallaştırılması için mevcut haritalama tekniklerinin başarımı karşılaştırmalı olarak incelenmiş, ekzon bölgelerinin belirlenmesine yönelik uygulamalar sonucunda en yüksek başarıma sahip haritalama teknikleri belirlenmiştir.
- b) İkinci aşamada, DNA dizilimlerini sayısallaştırmak için yeni bir sayısal haritalama tekniği geliştirilmiş ve bu tekniğin başarımı literatürdeki en yüksek başarıma sahip olan mevcut sayısal haritalama teknikleriyle farklı uygulamalarda karşılaştırılmıştır.
- c) Üçüncü aşamada ise, Beta talasemi hastaları ile sağlıklı bireylere ait DNA dizilimleri, önerilen sayısal haritalama tekniği ile sayısallaştırılmıştır. Bu verilerden özellik çıkarım kümesi elde etmek için, istatistiksel metrik tabanlı özellik seçim

teknikleri kullanılmıştır. Çıkarılan bu bilgilerden oluşturulan özellik vektörleri YSA ile sınıflandırılmış ve %96 tanıma doğruluğu başarımları elde edilmiştir.

Tez çalışmasında ortaya konan entropi tabanlı yeni sayısal haritalama tekniği ve işaret işleme ve yapay zekâyâ dayalı olarak geliştirilen, DNA dizilimlerinden hastalık tanımlanması yönteminin, gerçekleştirilen uygulamalardaki başarımları bu alanda yapılacak yeni çalışmalara ışık tutacaktır.

**Anahtar Kelimeler:** DNA Dizilimi, Ekzon, İntron, Sayısal Haritalama Tekniği, İşaret İşleme, Özellik Çıkarımı, Yapay Zekâ, Sınıflandırma



# **DEVELOPMENT OF NEW APPROACHES BASED ON SIGNAL PROCESSING FOR DISEASE DIAGNOSIS FROM DNA SEQUENCES**

**Bihter DAŞ**

**Firat University, Graduate School of Natural and Applied Sciences**

**Department of Software Engineering**

**Ph. D. Thesis, August 2018**

**Thesis Supervisor: Prof. Dr. İbrahim TÜRKOĞLU**

## **ABSTRACT**

Recently, studies with DNA in bioinformatics fields has been quite important in terms of understanding of DNA structure and determining of genes' effects in living organisms. Especially, several studies are carried out related to determining exon regions in DNA sequences, detection of genetic diseases, and comparison of genomic features.

In this thesis study, new approaches have been developed that will able to offer optimum solution suggestions intended for diagnosis of genetic diseases from DNA sequences. For this purpose, a comprehensive 3 stage study was completed.

- a) In the first stage, for the digitization of DNA sequences the performance of the existing mapping techniques was examined comparatively, the highest performing mapping techniques were identified in the result of applications intended for determining exon regions.
- b) In the second stage, a new numerical mapping technique was developed to digitize DNA sequences, and the performance of this technique was compared with other existing numerical mapping techniques in literature in the different applications.
- c) In the third stage, DNA sequences that belong to Beta talasemi patients and healthy individuals were digitized by the proposed numerical mapping technique. Feature selection techniques based on statistical metric were used to obtain the feature exaction set from this data. The feature vectors that generated from this extracted information were classified by YSA, and the diagnosis accuracy with 96% was obtained.

In the thesis study, the novel entropy based mapping technique which has been put forward, and the performance in the implemented applications of the disease diagnosis method from DNA sequences which has been improved based on signal processing and artificial intelligence will guide to new studies to do in this field.

**Keywords:** DNA Sequences, Exon, Intron, Numerical Mapping Technique, Signal Processing, Feature Extraction, Artificial Intelligence, Classification



# 1 GİRİŞ

## 1.1 Doktora Çalışmasının Amacı ve Kapsamı

Bu tezin amacı, DNA dizilimlerinden genetik hastalıkların sınıflandırılması veya tanınmasına yönelik optimum çözüm önerileri sunabilmek için yeni yaklaşımlar geliştirmektir. Bu yaklaşımlar için hedeflenmiş bileşenler şunlardır;

- a) DNA dizilimlerin sayısallaştırılması için yeni bir sayısal haritalama tekniği yaklaşımının geliştirilmesidir. DNA dizilimlerinden protein kod bölgelerinin(ekzon) belirlenmesine yönelik genomik uygulamalarda mevcut haritalama tekniklerinin başarımı incelenmiş, bu başarımı daha yüksek değerlere çıkarabilmek için yeni bir haritalama tekniği geliştirilmiştir.
- b) Ekzon bölgelerinin belirlenmesinden sonra, geliştirilen yeni sayısal haritalama tekniği kullanılarak sayısallaştırılmış DNA dizilimlerinden hastalık tanınması yapabilmek için işaret işleme ve bilgi ölçümüne (entropi, enerji) dayalı tekniklerle etkili anahtar özellik seçimi gerçekleştirilmiştir.
- c) DNA dizilimlerinden elde edilen bu özelliklerin sınıflandırılmasında YSA yöntemi kullanılarak Beta talasemi (Akdeniz Anemisi) genetik hastalığının tanınmasına yönelik bir uygulama süreci geliştirilmiş ve hedeflenen başarımların üstünde %96 tanımlama doğruluğu başarımları elde edilmiştir.

DNA dizilimlerinden genetik hastalık tanınması ile ilgili literatürde sadece birkaç çalışma mevcuttur. Var olan sayısal haritalama tekniklerinin haricinde yeni sayısallaştırma tekniğinin önerilmesi ve gerçeğe en yakın sonuçların elde edilebilmesi için gerçek Beta talasemi hastalarına ait DNA verilerinden bu genetik hastalığın tanınması, tez çalışmasının özgün yanlarını oluşturmaktadır.

## 1.2 Literatür Taraması ve Değerlendirmesi

Günümüzde DNA hakkında yapılan araştırmalar, DNA yapısının anlaşılması ve genlerin canlılardaki etkilerinin tespit edilmesi konusunda oldukça önemlidir. Literatürde

DNA dizilimlerindeki ekzonların (protein kodlayan bölgelerin) ve intronların tanınması, sayısal haritalama teknikleri, DNA dizilimlerinden özellik çıkarımı ve DNA dizilimlerindeki genetik hastalıkların tanınmasına yönelik birtakım araştırma ve deneysel çalışmalar bulunmaktadır. Yapılan literatür araştırmalarında bu akademik çalışmalar incelenmiş ve analiz edilmiştir. Tezin kapsamına ilişkin literatür çalışmaları 4 gruba ayrılarak irdelenmiştir:

***A. DNA dizilimlerinden protein kod bölgelerini (ekzon) tanıma ile ilgili çalışmalar;***

Malaya K. Hota ve diğ. yaptıkları çalışmada protein kod bölgelerini bulmada Kısa Zamanlı Ayırık Fourier Dönüşümü yönteminde 3 farklı Antinotch filtresi kullanarak farklı sayısal haritalama tekniklerinin performansları karşılaştırmalı incelenmiştir [1].

Qingda Zhou ve diğ. yaptıkları çalışmada, DNA dizilimlerinde bulunan nükleotid çiftlerinin frekans değerleri kullanarak 4 farklı bakteri türünü yüksek doğrulukla sınıflandırmışlardır [2].

Sathish Kumar S ve diğ. yaptıkları çalışmada örüntü madenciliği ve yapay sinir ağına dayalı bir sınıflandırma tekniği ile DNA dizilimlerinden bazı türlerin tanısını gerçekleştirmişlerdir [3].

Robertas Damasevicius yaptığı çalışmada, Destek Vektör Makinesi yöntemi ile DNA dizilimlerini sınıflandırmada ve promotör tanımda 11 farklı binary haritalama tekniğinin analizi yapılmıştır [4].

Zhang ve diğ. yaptıkları çalışmada, DNA dizilimlerin yapısal özelliklerini kullanarak ampirik mod ayrışma ve Fourier dönüşümüne dayalı bir yöntemle ekzon tahminini gerçekleştirmişlerdir [5].

S. D. Sharma ve diğ. yaptıkları çalışmada Gabor dalgacık dönüşümü yöntemi ile ekzon taramada sayısal haritalama tekniklerini değerlendirmişlerdir [6].

Ismail M. El-Badawy ve diğ. yaptıkları çalışmada Ayırık Fourier Dönüşümü yöntemi ile EIIP ve CIS(Complex Indicator Sequence) haritalama tekniklerinin DNA dizilimlerinde ekzon tahmini üzerindeki başarımını incelemişlerdir [7].

Baoshan Ma ve diğ. yaptıkları çalışmada Z-Curve haritalama tekniğini kullanarak protein kod bölgelerini tahmin etmede Fourier yönteminin farklı pencere uzunluklarındaki başarımlarını karşılaştırmışlardır [8].

Sitanshu Sekhar Sahu ve diğ. yaptıkları çalışmada Zaman-Frekans Filtreleme yaklaşımı kullanarak DNA dizilimlerindeki protein kod bölgelerin tanınmasını gerçekleştirmişlerdir [9].

Mahmood Akhtar ve diğ. yaptıkları çalışmada genomik dizilimleri protein kodlayan ve kodlanmayan olarak Gaussian mixture modele dayalı bir sınıflandırma yapmışlardır [10].

Ravi Gupta ve diğ. yaptıkları çalışmada ekzon ve intronları belirlemede bir zaman serisi yaklaşımı kullanmışlardır [11].

Lun Huang ve diğ. yaptıkları çalışmada DNA dizilimlerindeki 3 periyotlu protein kod bölgelerini tarayan yeni bir algoritma sunmuşlardır. Bu bölgeleri tahmin etmede sayısal bir korelasyon ve filtreleme yaklaşımı sunulmaktadır [12].

Sajid A. Marhon ve diğ. yaptıkları çalışmada protein kod bölgelerini tanımada yeni bir teknik önermişlerdir. Adaptif sayısal temsil tekniği ile DNA dizilimlerini sayısallaştırıp sinyal işleme yöntemleri ile ekzon bölgelerini tahmin etmişlerdir [13].

Lopamudra Das ve diğ. yaptıkları çalışmada Kaiser Pencereleme ve Ayrık Fourier Donuşumu yöntemlerini kullanarak yeni bir yaklaşım geliştirmişler ve ekzon bölgelerini tahmin etmeye çalışmışlardır [14].

### ***B. Sayısal haritalama teknikleri ile ilgili çalışmalar;***

Imen Messaoudi ve diğ. yaptıkları çalışmada DNA analizi uygulamalarında bir boyutlu ve iki boyutlu sayısal haritalama tekniklerinin başarımlarını karşılaştırmaktadır [15].

Daniel Ashlock ve diğ. yaptıkları çalışmada DNA sınıflandırılması için GC baz içeriği ve DNA üzerindeki motifler gibi DNA' nın kritik uzunluğu ile eşleşen blokları bularak DNA dizilimleri için değer dizisi üretir. DWP (Do What's Possible) diye adlandırılan bir temsil kullanılmaktadır.

Jin-yu Zhou ve diğ. yaptıkları çalışmada DNA dizilimlerini sayısala dönüştürme tekniği olarak yeni 3 boyutlu grafiksel bir sayısal haritalama tekniği geliştirmişler ve 11

farklı türün protein kod bölgelerinin tespit edilmesine yönelik bir uygulama yapmışlardır [17].

Mohammed Abo-Zahhad ve diğ. yaptıkları çalışmada DNA dizilimlerindeki ekzon ve intronları tanımda Dairesel haritalama tekniği diye adlandırılan yeni bir sayısal haritalama tekniği geliştirmişlerdir ve bu tekniğin başarımını varolan diğer sayısal haritalama teknikleriyle kıyaslamışlardır. Dairesel tekniğe göre her baz kodonlardaki pozisyon ve nükleotid tipine göre karmaşık sayı değerleriyle temsil edilmektedir [18].

Durga Ganesh Grandhi ve diğ. yaptıkları çalışmada DNA dizilimlerindeki protein kod bölgelerinin tespiti için Simplex haritalama tekniği geliştirmişlerdir [19].

Parameswaran Ramachandran ve diğ. yaptıkları çalışmada Quasi-Newton algoritmasını kullanarak DNA dizilimlerindeki ekzon bölgeleri için sayısal haritalama tekniği optimize etmişlerdir [20].

Hiroshi Dozono ve diğ. yaptıkları çalışmada Hidden Markov Modelini kullanarak DNA dizilimlerini sayısala dönüştürerek haritalamışlardır [21].

### ***C. DNA dizilimlerinden özellik çıkarma ve seçimi ile ilgili çalışmalar;***

Witold Kinsner ve diğ. yaptıkları çalışmada DNA sekansları için multifraktal ölçümlerin ekstraksiyonunu ve tahminlerini sunmaktadır ve DNA sekansı içinde yer alan bilgileri kullanarak biyolojik işlevselliği tanımlamak için bir korelasyon tekniği geliştirmişlerdir [22].

Qingda Zhou ve diğ. yaptıkları çalışmada çeşitli bakteri türlerine ait DNA dizilimlerin özellik kümesi çıkarılmış ve Destek Vektör Makinesi yöntemi ile bu özellikler sınıflandırılmıştır[23].

Lan Tao ve diğ yaptıkları çalışmada GA-PSO-SVM isimli DNA dizilimlerinden yeni bir promotor tanıma tekniği geliştirmişlerdir. Kromozomlara göre özellik kümesi elde edilmiş ve geniş ölçekli DNA dizilimlerinde yöntem test edilmiştir [24].

Alan Wee-Chung Liew ve diğ. yaptıkları çalışmada DNA dizilimlerinden ekzon ve intron bölgelerini sınıflandırmak için istatistiksel tekniklere dayanan özellik kümesi oluşturulmuştur ve bunun için korelasyon analizi yöntemi kullanılmıştır [25].

Muhammad Javed Iqbal ve diğ. yaptıkları çalışmada değişken uzunluktaki protein dizisini temsil etmek için bir mesafe bazlı özellik kodlama tekniği önerilmiştir. Bu tekniği

kullanarak, farklı amino asitlerden her amino asidin bir farklı seviyedeki dekompozisyon düzeyindeki bir dizide bulunmasıyla özellik kümesi elde edilmiştir [26].

#### ***D. DNA dizilimlerinden genetik hastalık tanılanması ile ilgili çalışmalar;***

Safaa M. Naeem ve diğ. yaptıkları çalışmada farklı güç spektrum yöntemlerini kullanarak geliştirilen bir algoritma ile göğüs kanseri hastalığının genetik çeşitliliğini taramışlardır. Bu algoritma NCBI veri tabanından elde edilen 6 sağlıklı ve 6 göğüs kanseri hastalığına sebep olan gen üzerinde test edilmiştir [27].

Shilpi Chakraborty ve diğ. yaptıkları çalışmada EIIP sayısal haritalama tekniğini ve Ayrık Dalgacık Dönüşümü Güç Spektrum Yöntemini kullanarak protein kod bölgelerindeki anormallikleri tahmin etmeye çalışmışlardır. Bu çalışmanın amacı, hastalığı sınıflandırmak için kullanılabilecek genlerin familyasını keşfetmektir [28].

KwongSak Leung ve diğ. yaptıkları çalışmada moleküler evrim analizi, kümeleme, özellik seçimi ve sınıflandırmayı içeren bir veri madenciliği yapısı tanıtılmıştır. Hepatit B virüsündeki genetik işaretçileri belirlemek için 200'den fazla hastadan, ya B veya C genotip HBV DNA dizileri toplanmıştır [29].

Literatürde DNA dizilimlerinin sayısallaştırılması ve ekzon bölgelerinin tanılanmasına yönelik birçok akademik çalışma bulunurken, DNA dizilimlerinden genetik hastalıkların tanılanmasına yönelik çalışmalar yeterince mevcut değildir. Bu tez çalışmasında, ekzon bölgelerini tanımada sayısal haritalama tekniklerinin başarımını ortalama %95'in üstüne çıkarmak, DNA dizilimlerinden etkili özellik kümesi oluşturarak Beta talasemi (Akdeniz Anemisi) hastalığının tanılanması amaçlanmıştır.

### **1.3 Tezin Organizasyonu**

Tez bölümlerinin organizasyonu ve tezdeki orijinal katkılar aşağıda açıklanmıştır:

**Bölüm 1'de**, teze genel bir bakış açısı kazandırmaya yönelik olarak temel bilgiler verilmiş, literatür irdelemesi yapılarak tez çalışmasının amacı ve kapsamı ortaya konulmuştur.

**Bölüm 2'de**; DNA' nın detaylı yapısı ve fonksiyonu, DNA'da bulunan genler ve arasındaki bağlantılar, baz dizilimleri, protein kodlayan bölgelerin (ekzon) ve protein

kodlamayan bölgelerin (intron) ne anlama geldiği açıklanmaktadır. Kısaca, DNA terminolojisi hakkında alt yapı oluşturmak için detaylı bilgiler verilmektedir.

**Bölüm 3'de;** işaret işleme uygulamalarında DNA dizilim verilerinin kullanılabilmesi için literatürde en çok tercih edilen sayısal haritalama teknikleri incelenmekte ve bu tekniklerin DNA dizilimlerini nasıl sayısallaştırıldığı hakkında detaylı bilgiler sunulmaktadır.

**Bölüm 4'de;** DNA dizilimlerindeki ekzon ve intron bölgelerinin sınıflandırılması ve bu sınıflandırmada mevcut sayısal haritalama tekniklerinin başarımlarının incelenmesi üzerine gerçekleştirilen bir uygulama anlatılmaktadır. Bu uygulama da literatürde en çok kullanılan sayısal haritalama tekniklerinin DNA dizilimlerini sayısallaştırmadaki başarımları incelenmiş ve daha sonra Ayrık Fourier Dönüşümü, DVM ve k-NN gibi sınıflandırma yöntemleri ile bu haritalama tekniklerinin DNA dizilimlerini ekzon ve intron olarak sınıflandırmadaki başarımları karşılaştırılmıştır. Uygulama sonucunda en yüksek sınıflandırma başarıma sahip sayısal haritalama teknikleri belirlenmiştir.

**Bölüm 5'de;** DNA dizilimindeki ekzon bölgelerinin tespiti ve mevcut haritalama tekniklerinin göstermiş olduğu başarımın üstünde bir başarımlar elde etmek için entropiye dayalı yeni bir sayısal haritalama tekniği önerilmiş ve bu önerilen haritalama tekniğinin başarımlarını var olan diğer sayısal haritalama teknikleriyle karşılaştırılmıştır. Uygulamada önerilen tekniğin ekzon bölgelerini tanımadaki başarımlarını Ayrık Fourier Dönüşümü, Tekli Ayrışım Yöntemi ve Kısa Zamanlı Fourier Yöntemlerinde de ayrı ayrı test edilmiştir.

**Bölüm 6'da;** DNA dizilimlerinin ekzon bölgelerindeki bazlarda meydana gelen mutasyon sonucu gözlemlenebilen Beta talasemi (Akdeniz Anemisi) genetik hastalığının tanımlanmasına yönelik bir uygulama gerçekleştirilmiştir. Bunun için önerilen haritalama tekniği ile sayısallaştırılan Beta talasemi hastalarına ve sağlıklı bireylere ait DNA dizilimlerinden etkili bir özellik kümesinin elde edilmesi ve Beta talasemi hastalığının tanımlanmasına yönelik geliştirilen bir uygulama süreci sunulmaktadır.

**Bölüm 7'de** tezin sonuçları irdelenmiş ve özgün katkıları vurgulanmıştır. Ayrıca ileriye dönük uygulama alanları ve öneriler tartışılmıştır.

## 2 DNA’NIN YAPISI VE DİZİLİMLERİ

Genetik bilgileri depolayan DNA parçaları gen olarak adlandırılırlar. DNA dizilimleri ise protein kodlayan genler ve canlı türleri ile ilgili bilgi edinmede kullanılmaktadır. Tez çalışmasının bu bölümünde; DNA’ nın yapısı, DNA dizilimleri, DNA’daki ekzon ve intron bölgeleri, genler, kodonlar hakkında bilgiler verilmiş ve terminolojiye yönelik alt yapı oluşturulmuştur.

### 2.1 DNA’nın Yapısı

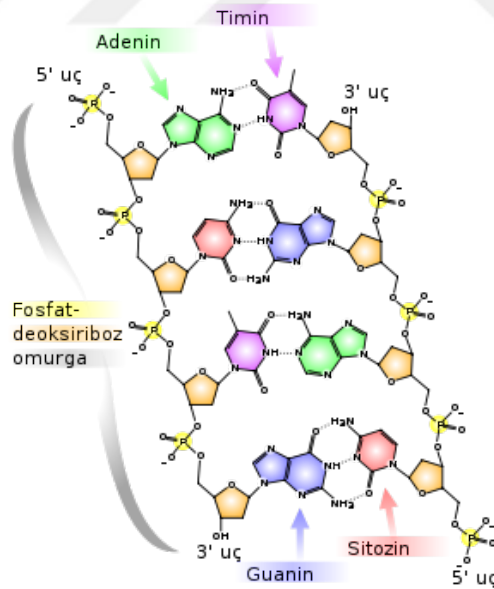
Deoksiribo Nükleik Asit (DNA), tüm organizmalar ve bazı virüslerin canlılık işlevleri ve biyolojik gelişmeleri için gerekli olan genetik talimatları taşıyan bir nükleik asittir. DNA'nın başlıca rolü bilginin uzun süreli saklanmasıdır. Protein ve RNA gibi hücrenin diğer bileşenlerinin inşası için gerekli olan bilgileri içermesinden dolayı DNA, bir kalıp, şablon veya reçeteye benzetilir [30,31].

DNA’daki genetik bilgiler hücre çekirdeğindeki kromozomlarla taşınır. Kromozomlar DNA ve özel proteinlerin birleşmesinden oluşur. DNA’ lar nükleotid olarak adlandırılan iki uzun polimerden oluşur. Bu polimerlerin omurgaları, ester bağları ile birbirine bağlanmış şeker ve fosfat gruplarından meydana gelir. Şekil 2.1’de DNA’nın çift sarmal yapısı görülmektedir [32]. Bu iki iplik birbirlerine ters yönde uzanırlar. Her bir şeker grubuna baz olarak adlandırılan dört tip molekülden biri bağlıdır. Bu bazlar Adenin (A), Guanin (G), Sitozin (C) ve Timin (T)’dir [33,34]. Bazlar halka yapılarına göre iki tip olarak sınıflandırılırlar: Adenin ve Guanin, Pürin grubudur. Bunlar beş ve altı üyeli halkalardan oluşmuş bileşiklerdir. Sitozin ve Timin ise Pirimidin grubudur. Bunlar da altı üyeli bir halkadan oluşmaktadırlar. Bir diğer baz olan urasil (U), Sitozinin yıkımı sonucu seyrek olarak DNA'da bulunabilir. Kimyasal olarak DNA'ya benzeyen RNA'da Timin yerine Urasil bulunur.



**Şekil 2.1** DNA'nın çift sarmal yapısı [32]

DNA' nın yarısı dişi bireyden yarısı da erkek bireyden gelir. Canlılarda DNA genelde tek bir molekül değil, birbirine sıkıca sarılı bir çift molekülden oluşur. Bu iki uzun iplik sarmaşık gibi birbirine sarılarak bir çift sarmal oluşturur. Şeker ve fosfat DNA molekülünün omurgasını oluşturur, baz ise çifte sarmaldaki öbür DNA ipliği ile etkileşir. Genel olarak bir şekerle bağlı baza nükleozit, bir şeker ve bir veya daha çok fosfata bağlı baza ise nükleotid denir [31]. Birden çok nükleotidin birbirine bağlı haline polinükleotid denir. Şekil 2.2'de bir DNA'nın kimyasal yapısı gösterilmektedir.

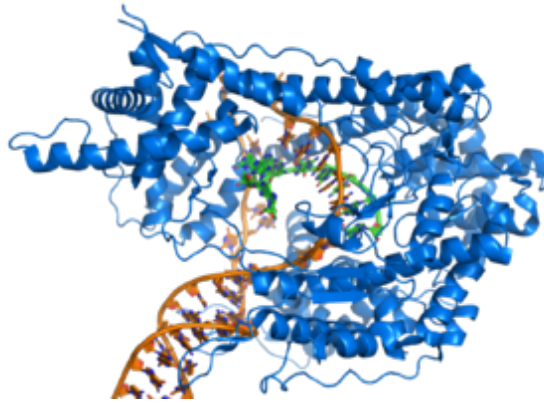


**Şekil 2.2** Bir DNA'nın kimyasal yapısı [34]

DNA' da bulunan şeker 2-deoksiribozdur, bu bir pentozdur. Pentoz ise 5 karbonlu bir şekerdir. Bitişik iki şekerden birinin 3 numaralı karbonu ile öbürünün 5 numaralı

karbon atomu arasındaki fosfat grubu, bir fosfodiester bağı oluşturarak şekerleri birbirine bağlar. Fosfodiester bağı asimetrik olması nedeniyle DNA ipliğinin bir yönü vardır [31]. Çifte sarmalda bir iplikteki nükleotidlerin birbirine bağlanma yönü, öbür ipliktekilerin yönünün tersidir. DNA ipliklerinin bu düzenine anti-paralel denir. DNA ipliklerinin asimetrik uçları 5' (beş üssü) ve 3' (üç üssü) olarak adlandırılır, 5' uç bir fosfat grubu, 3' uç ise bir hidroksil grubu taşır. DNA ve RNA arasındaki başlıca farklardan biri, içerdikleri şekerdir, RNA'da 2-deoksiriboz yerine başka bir pentoz şeker olan riboz bulunur [34].

DNA'daki bazların oluşturduğu dizi, genetik bilgiyi kodlar. Protein kodlayan genler durumunda DNA dizisi bir mesajcı RNA dizisini tanımlar, mRNA olarak adlandırılan bu dizi bir veya birkaç proteinin dizisini belirler. Genlerdeki DNA dizisi ile proteinlerdeki amino asit dizisi arasındaki ilişki, biyolojik çevrim (translasyon) kuralları tarafından belirlenir, bunlar topluca genetik kod ile özetlenir. Genetik kod, üç nükleotidlik dizilere karşılık gelen, üç harfli 'kelimelerden' oluşur (örneğin, ACT, CAG, TTT), bu üçlüler *kodon* olarak adlandırılır [31,32]. Transkripsiyonda, protein kodlayan bir genin kodonları önce RNA polimeraz tarafından bir mesajcı RNA şeklinde kopyalanır. Bu RNA kopya, ardından bir ribozom tarafından deşifre edilir; ribozom, mesajcı RNA ile amino asit taşıyan taşıyıcı RNA'lar arasında baz eşleşmesi yaparak onu okur. Dört bazın 3'lü kombinasyonları olabildiği için 64 olası kodon vardır.



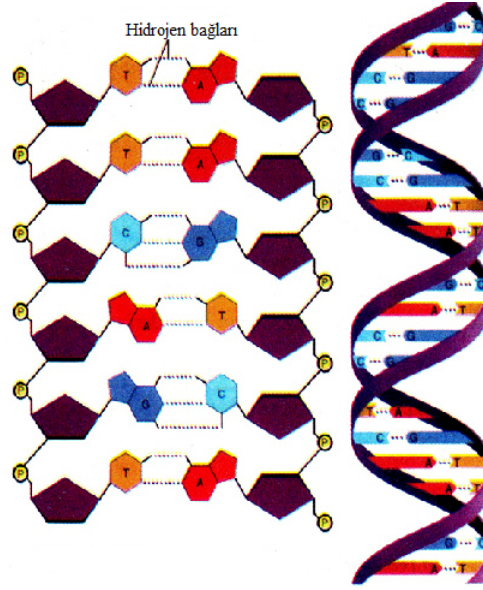
Şekil 2.3 RNA polimerazın DNA'dan bir mRNA üretimi [31]

Şekil 2.3 de RNA polimerazın DNA'dan bir mRNA üretimi gösterilmektedir [33,34]. Bunlar yirmi standart amino asidi kodlarlar, böylece çoğu amino asite birden çok kodon düşer. Ayrıca, protein kodlayıcı bölgenin sonuna işaret eden üç tane de 'stop' veya anlamsız (*nonsense*) kodon vardır, bunlar TAA, TGA ve TAG kodonlarıdır [35-37]. Protein sentezi

sırasında bu bilgi, genetik kod aracılığıyla okununca proteinlerin amino asit dizisini belirler. Bu esnada DNA'daki bilgi, DNA'ya benzer yapıya sahip başka bir nükleik asit olan RNA'ya kopyalanır [38,39].

Protein ve enzimler üretilirken DNA üzerindeki bazların dizilimleri örnek alınarak bu genlere karşılık gelen RNA kopya dizilimleri çıkarılır [40,41]. mRNA'lar çıkartılırken bir genin DNA dizilimleri baştan sona tümüyle okunmaz. DNA'nın okunmadan atlanan, mRNA ve protein kodlamasına katılmayan bu bölümlerine intron, kodlanan kısımlarına ise ekzon adı verilir [42]. Eskiden intronların bir işe yaramadığı düşünülürken, günümüzde intronlar olmasaydı yanlış sentezlemeler olacağı düşünülmektedir. Bir gene ait olan DNA diziliminde o dizilimi ekzon ve intron olarak sınıflandırmak, bir DNA dizi analizinde büyük önem taşımaktadır. [43,44]. Bir genin dizi analizinde, benzerlik (homoloji) araştırması, yeni bulunan bir DNA diziliminin diğer tüm dizilimlerle karşılaştırılmasıdır. Karşılaştırma sonucunda benzerlerdeki veri tabanında ya da literatürde tanımlanmış bazı biyolojik işlevler, yeni bulunan dizilime yakıştırılmaktadır. Bu yöntemle benzerlikler ve protein kodlayan ekzonlar araştırılır ve bir genin mutasyona uğrayıp uğramadığı belirlenebilir [46, 47]. Hangi proteinin nerede, nasıl, ne kadar ve ne zaman kodlanacağını, büyüme ve gelişmenin nerede ve nasıl düzenleneceğinin bilgisi ekzon ve intron bölgelerinin tespiti ile mümkündür. Benzer şekilde kök hücrelerin nerede hangi hücre, doku ve organlara dönüşeceğinin, hücrelerin hangi koşullarda çoğaltılıp ya da öldürüleceğinin, ne zaman hastalık geliştirileceğinin araştırılması ekzon ve intronların sınıflandırılmasının önemini arttırmaktadır [41,46].

Bir DNA diziliminde Adenin bazı daima Timin bazı ile karşılıklı bağlanırken, Guanin bazı ise daima Sitozin bazı ile karşılıklı olarak bağlanır. Şekil 2.4'de DNA molekülünü meydana getiren nükleotidlerin dizilişi verilmektedir. Adenin Timin arasında ikili, Guanin ile Sitozin arasında üçlü hidrojen bağı kurulur. DNA' da hidrojen bağlarının fazla olması DNA'nın daha kuvvetli olmasını sağlar [47,48]. Eşit sayıda oluşan DNA'lardan G/C oranı fazla olan daha sağlamdır. Bir DNA molekülünde Adenin sayısı Timine, Guanin sayısı ise Sitozin sayısına eşittir. Bir DNA molekülündeki pürin sayısı, pirimidin sayısına eşittir. Yani  $A+G=C+T$  dir [49,50].



Şekil 2.4 DNA molekülünü meydana getiren nükleotidlerin dizilişi [47]

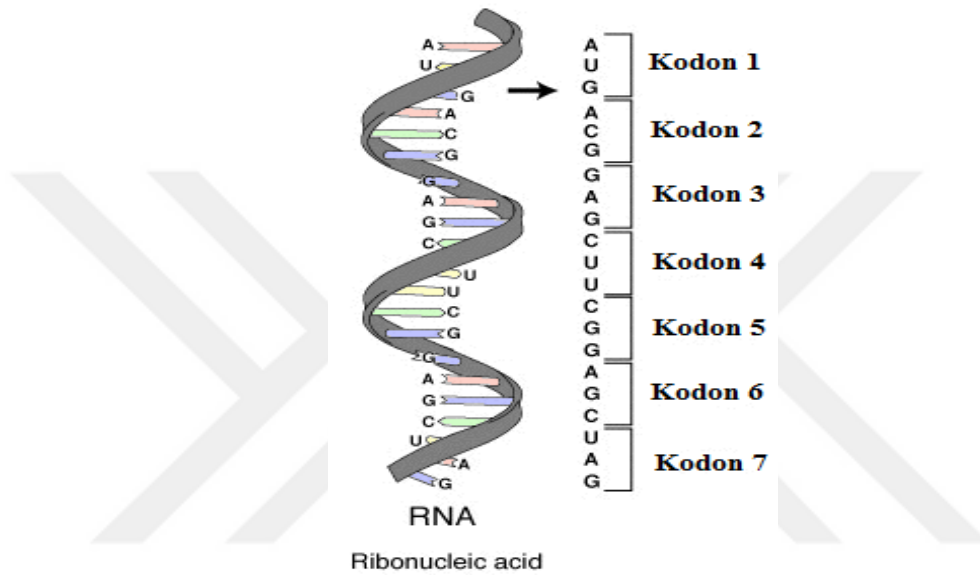
## 2.2 DNA Dizilimleri

DNA dizilimlerinin incelenmesi, dizisi bilinmeyen bir DNA bölgesindeki nükleotid bazlarının (Adenin, Guanin, Sitozin ve Timin) sırasının belirlenmesi işlemidir. DNA dizilimlerinin bilinmesi; temel biyoloji, biyoteknoloji, adli bilim, tıbbi tanı koyma gibi pek çok alanda önem kazanmıştır. DNA dizilemesi, biyolojik araştırma ve keşifleri çok hızlandırmıştır. Modern DNA dizileme teknolojilerin sayesinde; insan, pek çok hayvan, bitki ve mikrop genomunun tam dizisi üretilmiştir [50-52].

DNA dizilimi üzerindeki bilgiyi taşıyan DNA parçası bir geni oluşturur. Genlerde meydana gelen bazı nükleotid değişimler mutasyon olarak adlandırılır ve çeşitli kalıtsal hastalıklara neden olur. Hastalığa sebep olan gen bölgesi, Sekans Analizi tekniği ile taranarak mutasyonun gen üzerindeki yeri tespit edilir. Sekans Analizi aletlerinin geliştirilmesi ile birlikte sekanslama işleminin süresi kısalmıştır. İlk DNA dizileri 1970' lerin başlarında üniversite araştırmacıları tarafından iki-boyutlu kromatografiye dayanan zahmetli yöntemlerle elde edilmiştir. Otomatik analizle çalışan boya-tabanlı dizileme yöntemlerinin gelişimiyle DNA dizilemesi çok daha kolaylaşmıştır [50-52].

DNA dizilimleri genel olarak 5'-3' şeklinde sıralanırlar. Çünkü nükleotidler lineer(dallanmamış) polimerlerdir. Bir sekans bilgiyi temsil eden kapasiteye sahiptir. Biyolojik DNA, canlının fonksiyonlarını yöneten bilgileri temsil eder. Bu nedenle nesneler

için genetik sekans sıkça kullanılmaktadır. Sekanslar DNA dizileme yöntemleri ile biyolojik ham maddeden okunabilmektedirler [52]. DNA sekanslarında taşınan genetik bilginin proteine çevirisinde ise RNA önemli rol oynar. Mesajcı RNA, DNA'daki bilgiyi protein sentez yeri olan ribozomlara taşır, ribozomal RNA ribozomun en önemli kısımlarını oluşturur, taşıyıcı RNA ise protein sentezinde kullanılmak üzere kullanılacak aminoasitlerin taşınmasında gereklidir. Ayrıca çeşitli RNA tipleri genlerin ne derece aktif olduğunu düzenlemeye yarar. Şekil 2.5' de bir RNA yapısı görülmektedir.



Şekil 2.5 RNA'nın yapısı [34]

DNA diziliminde bulunan A, G, C, T bir pozisyonda o nükleotidleri temsil ederken bazen bir arada bulunarak başka bir harf ile temsil edilebilirler. Uluslararası saf ve uygulamalı kimya birliğinin belirlediği kurallar Tablo 2.1'de belirtilmiştir.

**Tablo 2.1** Uluslararası saf ve uygulamalı kimya birliğinin kuralları

| Kısaltma | Açıklama        | Kısaltma | Açıklama                |
|----------|-----------------|----------|-------------------------|
| <b>A</b> | Adenin          | <b>S</b> | (G C) Güçlü Bağ         |
| <b>C</b> | Sitozin         | <b>W</b> | (A T) Zayıf bağ         |
| <b>G</b> | Guanin          | <b>B</b> | (G T C) A dışında hepsi |
| <b>T</b> | Timin           | <b>D</b> | (G A T) C dışında hepsi |
| <b>R</b> | (G A) purin     | <b>H</b> | (A C T) G dışında hepsi |
| <b>Y</b> | (T C) pirimidin | <b>V</b> | (G C A) T dışında hepsi |
| <b>K</b> | (G T) Keto      | <b>N</b> | (A G C T) herhangi biri |
| <b>M</b> | (A C) Amino     |          |                         |

### 3 DNA DİZİLİMİNİN SAYISAL BİÇİME DÖNÜŞTÜRÜLME TEKNİKLERİ

Bir DNA dizilimini sayısal hale dönüştürürken çeşitli sayısal haritalama teknikleri kullanılır. Bu teknikler sabit haritalama teknikleri ve fiziko-kimyasal özellik tabanlı haritalama teknikleri olarak 2 grupta toplanır [51]. Bu kapsamda, DNA dizilimlerinin sayısal biçime dönüştürmek için literatürde en çok kullanılan haritalama teknikleri aşağıda verilmiştir:

#### 3.1 Sabit Haritalama Teknikleri

Sabit haritalama tekniklerinde DNA dizilimlerinde bulunan bazlar keyfi sayısal değerlere dönüştürülürler. Bu teknikler iki kısımda incelenmektedir [51]. Birincisi A,G,T,C bazlarının bir dizilimde var olup olmadığının 0 ve 1 sayıları ile temsil edilmesidir. İkincisi ise A,G,C,T bazlarına anlamlı reel ve karmaşık sayıların atanmasıdır. Genel olarak kullanılan sabit haritalama teknikleri: Voss tekniği, Tetrahedron tekniği, Karmaşık tekniği, Tamsayı tekniği, Reel tekniği ve Kuaterniyon tekniği ve Moleküler Kütle tekniğidir [53,54].

##### 3.1.1 Voss Haritalama Tekniği

En çok kullanılan haritalama tekniklerinden biridir [53]. Bir DNA diziliminde bulunan A, G, C, T bazları  $X_A(n)$ ,  $X_G(n)$ ,  $X_C(n)$ ,  $X_T(n)$  şeklinde 4 boyutlu vektör şeklinde gösterilir [55]. N bir DNA diziliminin uzunluğunu ifade eder.

$X_A(n)$  satırında bir DNA diziliminde A yerine 1, diğer bazlar yerine ise 0 yazılır.  $X_G(n)$  satırında G yerine 1 diğer bazlar yerine 0 yazılır.  $X_C(n)$  ve  $X_T(n)$  satırları da aynı şekilde ifade edilir [56]. Tablo 3.1’de Voss haritalama tekniğine göre sayısallaştırılmış örnek bir DNA dizilimi gösterilmektedir.

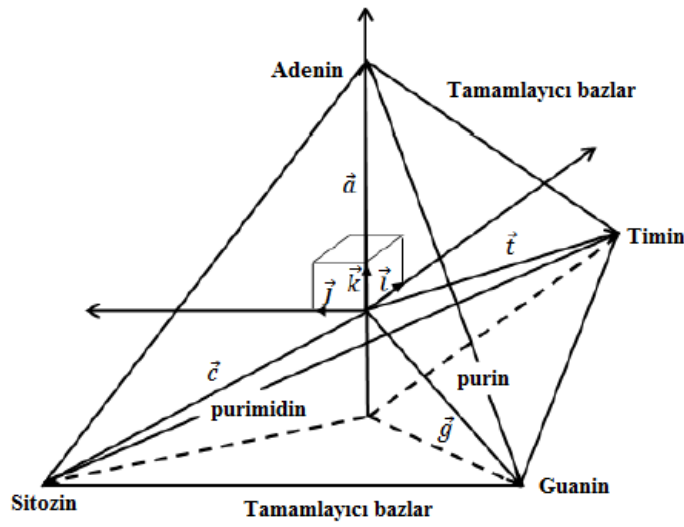
**Tablo 3.1** Voss haritalama tekniğine göre sayısallaştırılmış DNA dizilimi

| DNA Dizilimi | .....A G G T C T A T G C A G G A A C T C A C G G..... |
|--------------|-------------------------------------------------------|
| $X_A(n)$     | .....1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 0..... |
| $X_G(n)$     | .....0 1 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 1..... |
| $X_C(n)$     | .....0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0..... |
| $X_T(n)$     | .....0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0..... |

Voss haritalama tekniğinde bazlar arasındaki matematiksel ilişki önceden tanımlanmaz, sadece bazların frekansı işaret edilir. Bu teknik DNA dizilimlerinin spektral analizi için etkin bir yöntemdir [57].

### 3.1.2 Tetrahedron Haritalama Tekniği

Bu teknikte Voss haritalama tekniğinin aksine işaret dizilerinin sayısı 4 boyut yerine 3 boyut kullanılır [56]. Şekil 3.1'deki gibi A, G, C, T gibi bazları temsil eden 4 tane 3 boyutlu vektörler merkezden köşelere doğru işaret edilmektedir [58].



Şekil 3.1 Tetrahedron tekniği [56]

$$(a_r, a_g, a_b) = (0, 0, 1) \quad (3.1)$$

$$(t_r, t_g, t_b) = \left(\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}\right) \quad (3.2)$$

$$(g_r, g_g, g_b) = \left(-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}\right) \quad (3.3)$$

$$(c_r, c_g, c_b) = \left(-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}\right) \quad (3.4)$$

Yukarıdaki denklemlerden aşağıdaki 3 sayısal vektör elde edilir. Bunlar

$$x_r(n) = \frac{\sqrt{2}}{3} (2X_T(n) - X_C(n) - X_G(n)) \quad (3.5)$$

$$x_g(n) = \frac{\sqrt{6}}{3} (X_C(n) - X_G(n)) \quad (3.6)$$

$$x_b(n) = \frac{1}{3}(3X_A(n) - X_T(n) - X_C(n) - X_G(n)) \quad (3.7)$$

r, g ,b kırmızı, yeşil ve mavi renk belirteci olarak kullanılır.

$X(n)=[CGAT]$  şeklinde verilen bir DNA diziliminin sayısallaştırılmış temsili:

$$x_r(n) = \frac{\sqrt{2}}{3}[-1 -1 0 2] \quad x_g(n) = \frac{\sqrt{6}}{3}[1 -1 0 0] \quad x_b(n) = \frac{1}{3}[-1 -1 3 -1] \text{ şeklinde olacaktır.}$$

Biomoleküler dizilerin DNA spektrumunun elde edilmesi tetrahedron tekniği dönüşümlerinin esas uygulamalarındandır. Elde edilen bu spektrumlar A, G, C, T bazları için lokal frekans bilgisini sağlar [56, 59].

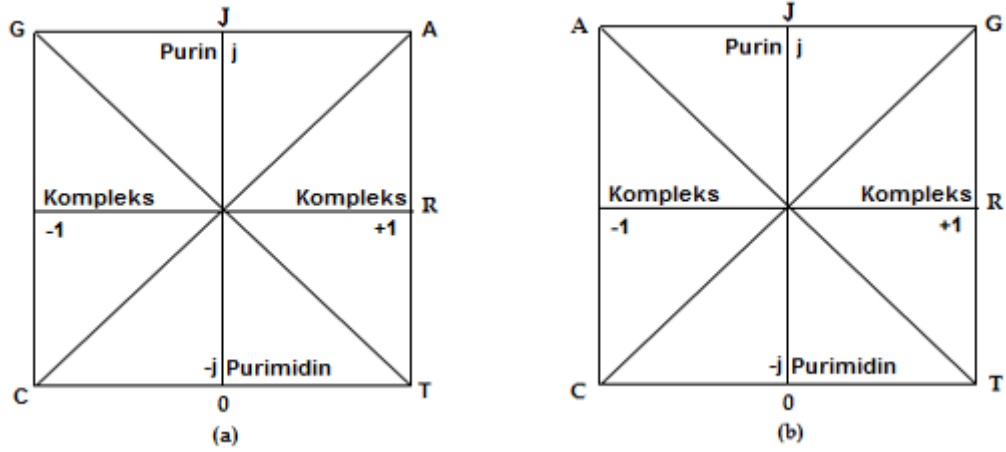
### 3.1.3 Karmaşık Haritalama Tekniği

Farklı düzlemlerde tetrahedron tekniğinin yansıtılması ile tetrahedron sayısal haritalama tekniğinin boyutu 2'e dönüştürülebilir [58]. Böyle düzlemler sayısal dönüşümlerin simetrisinin korunması ve biyolojik özelliklere karşılık gelen matematiksel özelliklerin yansıtılması yolları ile seçilebilir. Düzlemler bir çift koordinat eksenini ile tanımlanır. Diğer taraftan bu düzlemler karmaşık(complex) düzleme dönüştürülür ve böylece A,G,C,T bazlarının karmaşık temsilleri elde edilir. Bir  $x(n)$  diziliminde bazların karmaşık temsili Denklem 3.8'de gösterilmiştir [60,61].

$$x(n) = Ax_A(n) + Cx_C(n) + Tx_T(n) + Gx_G(n) \quad (3.8)$$

Bu denklemde  $A = 1 + j, T = 1 - j, C = -1 - j, G = -1 + j$  şeklindedir.

Karmaşık teknik, bazların bazı özelliklerini matematiksel özelliklere dönüştürme bakımından sinyal işlemede çok avantajlıdır. A-T ve G-C baz çiftleri karmaşık eşlenikler olarak ifade edilir. Pürinler ve pirimidinler eşit sanal parça ve zıt işaretli gerçek parça olarak ifade edilirken onların gösterimleri, karmaşık eşlenikler olarak ifade edilir. Şekil 3.2(a)'da bazların sağ düzlemde Şekil 3.2(b) ise sol düzlemde gösterimi verilmiştir [62,63].



Şekil 3.2 Bazların sol ve sağ düzlemde gösterimi

### 3.1.4 Tamsayı Haritalama Tekniği

Tamsayı (integer) teknik DNA bazlarının 1 boyutlu haritalama tekniğidir. Bir DNA diziliminde *eğer pürin(A,G) > pirimidin(C,T)* 4 baza sırayla  $T=0, C=1, A=2, G=3$  şeklinde değer verilir [59,61].

$T > A$  ve  $G > C$  olması durumunda bazlara  $A=0, C=1, T=2$  ve  $G=3$  değerleri verilir. Tamsayı tekniğinde bazlara keyfi olarak atanan bu değerlerin matematiksel özelliklerle eşleştirilmesi zordur. Bundan dolayı işaret işlemede tamsayı tekniği sınırlı bir şekilde önerilmektedir.

### 3.1.5 Reel Haritalama Tekniği

Bu teknikte bir DNA dizilimde  $A=-1.5, T=1.5, C=0.5, G=-0.5$  değerlerini alır.  $TGGAAC$  şeklinde verilen bir dizilimin reel (real) sayı temsilleri  $1.5 -0.5 -0.5 -1.5 -1.5 0.5$  şeklinde ifade edilir [62,64].

Karmaşık, reel ve tamsayı teknikleri takımyıldızı diyagramları olarak görülebilir. Reel ve tamsayı haritalama teknikleri, reel tekniğin doğal tamamlayıcı özelliği ve tamsayı tekniğinin tamamlayıcı olmayan özelliğinden dolayı farklı dizilim eşleşmelerini sağlar.

$S(n)=5''-TACGTG-3''$  şeklinde verilen bir dizi reel teknik ile sayısal dizilimlere dönüştürülürken aynı otoregresif parametrelere sahip olabilir. Aşağıda 3 tip otoregresif yapı görülmektedir.

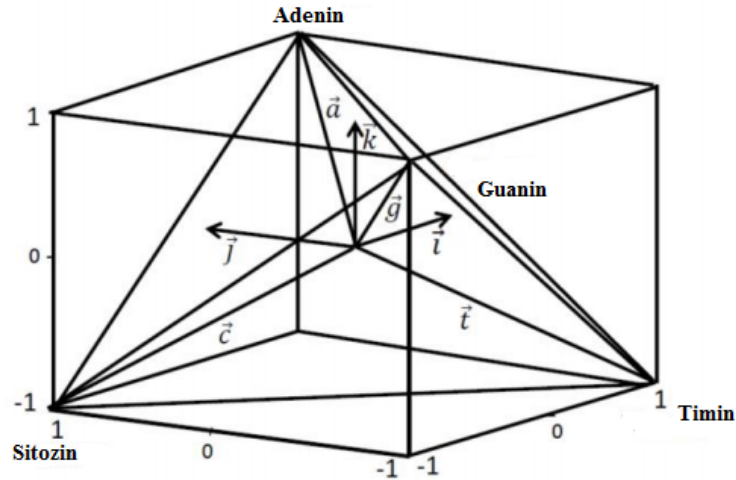
- a) Tersine tamamlayıcı dizilim:  $5''-1.5 1.5 -0.5 0.5 -1.5 0.5 -3''=5''-ATGCAC-3''$

- b) Ters dizilim:  $5'' - 0.5 - 0.5 \ 1.5 - 0.5 \ 0.5 - 1.5 - 3'' = 5'' - CGTGCA - 3''$   
c) Tamamlayıcı dizilim:  $5'' - -0.5 \ 0.5 - 1.5 \ 0.5 - 0.5 \ 1.5 - 3'' = 5'' - GCACGT - 3''$

Bu dizilim yapılarından dolayı ters işaretli sayısal diziler ve gerçek sayısal dizilimler aynı lineer bağımlılığa sahiptir ve bu yüzden aynı otoregresif parametreleri içerirler. Aynı zamanda ileri ve geri beslemeli lineer tahmin hataları da teorik olarak aynı otoregresif modelleri vermektedirler.

### 3.1.6 Kuaterniyon Haritalama Tekniği

Kuaterniyon (Quaternion) tekniğinde A, G, C, T bazlarına kuaterniyon değerleri geometrik gösterim ile atanır [63]. Şekil 3.3'de kuaterniyon değerlerinin geometrik temsili gösterilmektedir [64]. Küpün köşelerinde bulunan A, G, C, T bazlarına atanan kuaterniyon değerleri merkezden köşelere doğru uzanan vektörlerle belirlenir. Küpün kenar uzunluklarının 2 birim olduğu, orijinin küpün merkezinde yer aldığı ve küpün köşe koordinatlarının  $(\pm 1, \pm 1, \pm 1)$  olduğu farz edilirse; A, G, C, T vektörleri Denklem 3.9'daki gibi tanımlanır.



Şekil 3.3 Kuaterniyon tekniği [64]

$$(A = i + j + k, G = i - j + k, C = i - j - k, T = -i + j - k) \quad (3.9)$$

Bu haritalama tekniği DNA dizilimlerinin periyodik dönüşümünün hesaplanmasında kullanılır.

### 3.1.7 Moleküler Kütle Haritalama Tekniği

Moleküler Kütle (Molecular Mass) haritalama tekniğinde DNA dizilimindeki bazlara sırasıyla A=134, G=150, C=110 ve T=125 değerleri verilir [62-64].

$X(n)=[ATTGCGC]$  şeklinde verilen bir DNA diziliminin sayısal temsili:

$X(n)=[134\ 125\ 125\ 150\ 110\ 150\ 110]$  şeklinde olacaktır.

### 3.2 Fiziko-Kimyasal Özellik Tabanlı Haritalama Teknikleri

Bu haritalama teknikleri DNA biyomoleküllerin biyofiziksel ve biyokimyasal özelliklerinin haritalanmasında kullanılır [65]. Genel olarak kullanılan fiziksel-kimyasal özellik tabanlı haritalama teknikleri: EIIP (Electron-Ion Interaction Potential) tekniği, Atomik Sayı tekniği, Eşleştirilmiş Sayısal tekniği, DNA-Walk ve Z-Eğrisi tekniğidir [66].

#### 3.2.1 EIIP Haritalama Tekniği

Bu teknikte DNA dizilimindeki her bir nükleotid EIIP temsilindeki yarı değerlik sayısı ile eşleştirilir. A= 0.1260, G=0.0806, C=0.1340, T=0.1335 değerleri verilir [65,66]. Eğer  $X(n)$  dizisinde A, G, C, T için EIIP değerleri yerine koyulursa, oluşan yeni sayısal dizilim  $X_e(n)$  bir DNA dizilimi boyunca serbest elektron enerji dağılımlarını temsil eder [67,68].

$X(n)=[A\ T\ T\ G\ C\ A\ T\ G\ C]$  iken  $X(n)$  diziliminin sayısal temsili  $X_e(n)$  olarak gösterilmiştir.

$X_e(n)=[0.1260\ 0.1335\ 0.1335\ 0.0806\ 0.1340\ 0.1260\ 0.1335\ 0.0806\ 0.1340]$  dir.

#### 3.2.2 Atomik Sayı Tekniği

Atomik Sayı tekniğinde (Atomic Number), tek atom numaralı işaret dizisi, her bir nükleotide A=70, G=78, C=58, T=66 değerlerini atayarak elde edilir [67].

$X(n)=[GCCATG]$  şeklinde verilen bir DNA diziliminin sayısal temsili:

$X(n)=[78\ 58\ 58\ 70\ 66\ 78]$  şeklinde olacaktır.

#### 3.2.3 Eşleştirilmiş Sayısal Haritalama Tekniği

Eşleştirilmiş Sayısal (Paired Numerical) teknikte eşleştirilmiş olan A-T ve G-C bazları +1 ve -1 değerlerini alırlar [68,69]. Bir ya da iki işaret dizilimleri olarak temsil edilebilirler. Bu sayısal dönüşüm DNA'nın karmaşıklığını azaltarak DNA'nın yapısal özellikleri birleştirilir. Nükleotid çiftlerini bu teknikle sayısal hale dönüştürmek için 7 kural vardır. Bunlar;

1. *Pürin-pirimidin kuralı (RY kuralı):* Eğer  $n_i$  nükleotid dizisi pürin ise ( $A$  yada  $G$ )  $u_i=1$  dir, eğer  $n_i$  pirimidin ise ( $C$  yada  $T$ )  $u_i=-1$  dir.
2.  $A\bar{A}$  kuralı: Eğer  $n_i=A$  ise  $u_i=1$  dir. Diğer tüm durumlarda  $u_i=-1$  dir.
3.  $T\bar{T}$  kuralı: Eğer  $n_i=T$  ise  $u_i=1$  dir. Diğer tüm durumlarda  $u_i=-1$  dir.
4.  $G\bar{G}$  kuralı: Eğer  $n_i=G$  ise  $u_i=1$  dir. Diğer tüm durumlarda  $u_i=-1$  dir.
5.  $C\bar{C}$  kuralı: Eğer  $n_i=C$  ise  $u_i=1$  dir. Diğer tüm durumlarda  $u_i=-1$  dir.
6. *Hidrojen Bağ Enerji kuralı(SW Kuralı):* Kuvvetle bağlanmış çiftler için ( $G$  veya  $C$ )  $u_i=1$ , zayıfça bağlanmış çiftler için ( $A$  veya  $T$ )  $u_i=-1$  dir
7. *Hibrid kuralı (KM kuralı):*  $A$  yada  $C$  için  $u_i=1$ ,  $T$  yada  $G$  için  $u_i=-1$  dir.

Yukarıdaki kurallardan en çok kullanılan RY kuralıdır. Eşleştirilmiş sayısal teknik en çok gen ve ekzon tahmininde kullanılır. Bu teknik ekzon ve intron tahmininde, bazların ekzon ve intronda bulunma sıklığından faydalanırlar. Yani intronlar  $A$  ve  $T$  bakımından zengin, ekzonlar ise  $G$  ve  $C$  bakımından zengindir. Ekzon ve intronların bu ayrılma özelliğinden faydalanmak için  $A-T$  ve  $G-C$  birbiriyle eşleştirilir.  $A-T$  çiftleri  $+1$ ,  $G-C$  çiftleri ise  $-1$  değerini alırlar [70,71].

“ $AATGCTGCCATTA$ ” şeklinde verilen bir DNA diziliminin sayısal temsili  $X_n$  ile gösterilir.

$$X_n = [+1 +1 +1 -1 -1 +1 -1 -1 -1 +1 +1 +1 +1] \text{ olur [72].}$$

### 3.2.4 DNA-Walk Haritalama Tekniği

DNA dizilimlerinin sabit ölçekli uzun menzilli korelasyonu ile çalışmak için geliştirilmiş bir tekniktir [70]. Geleneksel tek boyutlu DNA-Walk modeli için bir yürüteç her bir yürüyüşün  $i$  adımı için hem  $[u(i)=+1]$  yukarı hem de  $[u(i)=-1]$  de aşağı doğru hareket eder. Bu teknikte bir baz çifti seçilir. Bu baz çiftine göre değerler verilir. ( $C$  yada  $T$ )  $=1$  iken ( $A$  yada  $G$ )  $=-1$  değerini alır. Ancak burada seçilen baz çifti önemlidir. Örneğin  $C$  ve  $A$  baz çifti seçilmişse DNA diziliminde  $C=1$ ,  $A=-1$ , diğer bazlar ise  $0$  değerini alır. Ters durumda  $T$  ve  $G$  baz çifti seçilmişse  $T=1$ ,  $G=-1$  ve diğer bazlar  $0$  değerini alır [73,74].

### 3.2.5 Z-Eğrisi Haritalama Tekniği

Z-Eğrisi (Z-Curve) bir DNA dizilimini destekleyen 3 boyutlu bir eğridir. DNA diziliminin taşıdığı tüm bilgileri Z eğrisi içerir [75]. Sonuç eğrisi zig zag şeklinde olduğu için Z-Eğrisi denilmektedir. Z-eğrisinin avantajlarından biri de sezgisel olmasıdır. Genomun

Z-Eğrisi bir ekran üzerinde yada kağıt üzerinde izlenebilir. Z-Eğrisi  $P_0, P_1, P_2 \dots L$  düğümlerinden ve  $x_n, y_n, z_n$  koordinatlarından oluşmaktadır. Burada  $n=0, 1, 2 \dots L$  'dir.  $L$  ise DNA diziliminin baz uzunluğudur.

$$X_n(A_n + G_n) - (G_n + T_n) = R_n - Y_n \quad (3.10)$$

$$Y_n(A_n + C_n) - (G_n + T_n) = M_n - K_n \quad (3.11)$$

$$Z_n = (A_n + T_n) - (G_n + C_n) = W_n - S_n \quad (3.12)$$

Burada  $A_n, G_n, T_n, C_n$  bir DNA diziliminde bazların kaçınıcı sırada olduğunu gösterir.

$A_0 = C_0 = G_0 = T_0 = 0$  olduğunda  $X_0 = Y_0 = Z_0 = 0$  dır.

$R, Y, M, K, W$  ve  $S$  sırasıyla pürin, pirimidin, amino, keto, zayıf hidrojen (H) ve güçlü ve kalın hidrojen (H), temsil eder.

Örneğin  $\{T, A, G, C, G, A\}$  diziliminde 4 baz işaretçisi için:

$\{0, 1, 0, 0, 0, 1\}$  (A),

$\{0, 0, 1, 0, 1, 0\}$  (G),

$\{1, 0, 0, 0, 0, 0\}$  (T)

$\{0, 0, 0, 1, 0, 0\}$  (C)

$S_n = \{-1, 1, 1, -1, 1, 1\}$  dir.

## **4 SAYISAL HARİTALAMA TEKNİKLERİNİN BAŞARIMLARININ DEĞERLENDİRİLMESİ**

Bölüm 2’de belirtildiği üzere protein üretimi sırasında mRNA çıkarılırken, bir genin DNA dizilimi baştan sona okunmamaktadır. DNA’nın okunmadan atlanan bu bölümüne intron, kodlanan kısımlarına ise ekzon denilmektedir. Bir DNA dizilimindeki ekzon ve intronların sınıflandırılması, DNA dizilimlerindeki genetik hastalıkların tespiti, DNA üzerindeki baz mutasyonların yerinin belirlenmesine yönelik genomik çalışmalarda oldukça önem taşımaktadır [76-78]. Sayısal işaret işleme teknikleri sembolik işaretlere uygulanamadığından, DNA dizilimlerinin sayısal haritalama teknikleri ile sayısal işaretlere dönüştürülmesi gerekmektedir [79-81].

Tezin bu bölümünde, literatürde var olan ve işaret işleme tekniklerinde en çok kullanılan sayısal haritalama tekniklerinin ekzon ve intron bölgelerini ayırmadaki sınıflandırma başarımları incelenmiştir. Bu uygulamanın amacı en yüksek başarımlara sahip haritalama tekniklerini belirlemektir. Bunun için insan türünün MEFV genine ait DNA baz dizilimleri sayısal haritalama teknikleri ile sayısal işaretlere dönüştürülmüştür. Daha sonra Ayırık Fourier Dönüşümü (AFD), Destek Vektör Makinesi (DVM), K-En yakın Komşu Algoritması (k-NN) gibi sınıflandırma yöntemleri ile bu tekniklerin DNA dizilimlerini ekzon ve intron olarak sınıflandırmadaki başarımlarını karşılaştırılmıştır.

### **4.1 Kullanılan Sınıflandırma Yöntemleri**

Ekzon ve intron bölgelerinin sınıflandırılmasında literatürde genomik uygulamalarda en çok tercih edilen AFD yöntemi kullanılırken, makine öğrenmesi metotlarından DVM ve k-NN yöntemleri ile de bir sınıflandırma yapıp sistemin başarımının test edilmesi amaçlanmıştır.

#### **4.1.1 Ayırık Fourier Dönüşümü Yöntemi**

Ayrık zamanlı Fourier Dönüşümü, ayrık zamanlı işaret işleme algoritma ve sistemlerinin analizi, tasarımı, gerçekleştirilmesi ile doğrusal filtreleme, korelasyon analizi ve spektrum analizi gibi işaret işleme uygulamalarında önemli bir rol oynar[82,83]. İşaretler sınırlı sayıda nokta için değerlendirilir [84,85]. Bu uygulamada AFD, DNA dizilimlerinin 3-

periyot değerlerinin çıkarılması amacıyla kullanılmaktadır [86-88]. 3-periyotlu olmasının sebebi aminoasit üreten kodonların 3 baz uzunluğunda olmasıdır. AFD, Fourier dönüşümünün eşit aralıklı frekanslardaki örneklerine özdeştir [89,90].

N-noktalı bir AFD'nin hesaplanması Denklem 4.1'de verilmiştir [87].

$$X[k] = \frac{1}{\sqrt{N}} \sum_{n=1}^N x[n] W_N^{(k-1)(n-1)} \quad 1 \leq k \leq N \quad W_N = e^{\frac{-j2\pi}{N}} \quad (4.1)$$

Burada,  $x[n]$  sayısal biçime dönüştürülmüş DNA dizilimidir.  $N$ , DNA dizilimdeki toplam baz sayısıdır.

AFD için pencereleme yapılırken  $L$  baz uzunluğunda ve 2 komşu pencere arasındaki  $L-3$  baz genişliğinde örtüşme ile pencere yaklaşımı kullanılmaktadır [90].  $L$ , kullanılan penceredeki baz uzunluğudur. AFD spektrumunun normalizasyon toplamı Denklem 4.2'de verilmiştir [91,92].

$$X_T[k] = \frac{1}{N_W} \sum_{m=1}^{N_W} X_m[k] \quad (4.2)$$

$X_m[k]$ , pencerelenmiş her bir dizilimdir.

$X_T[k]$ , AFD spektrumunun normalize edilmiş toplamıdır.

DNA diziliminin güç spektrumu için Denklem 4.3 kullanılmaktadır.

$$S[k] = |X_T[k]|^2 \quad (4.3)$$

Sayısal biçime dönüştürülmüş dizilimin spektral içerik ölçümünden elde edilen 3 periyotlu spektral bileşeni Denklem 4.4'de verilmiştir. Bu denklemdeki,  $k$  değeri  $N/3+1$  olup, bu ise  $2\pi/3$  periyotuna karşılık gelmektedir. Ekzonlara karşılık gelen DNA segmentleri  $2\pi/3$  frekansında nispeten güçlü bir bileşen sergilemektedirler, oysa intronlar  $2\pi/3$  frekansında zayıf bir frekans oluşturmaktadır.

$$P_3 = S[\frac{N}{3} + 1] \quad (4.4)$$

Ekzon ve intron olarak sınıflandırmak için kullanılan eşik değeri  $T_3$  olarak isimlendirilir [28].  $T_3$  değeri için Denklem 4.5 kullanılmaktadır.

$$T_3 = \frac{sdP_{3e} * meanP_{3i} + sdP_{3i} * meanP_{3e}}{sdP_{3e} + sdP_{3i}} \quad (4.5)$$

Ekzon ve intronların 3 periyotlu standart sapma ve ortalama değerlerini bulmak için ekzon ve intron dizilimlerinin eğitim dizilerimleri kullanılmaktadır. Ekzon ve intron test

dizilerimleri için 3 periyotlu  $P_{3t}$  değeri,  $T_3$  eşik değerden büyük ya da eşit olursa o test dizisi ekzon olarak,  $T_3$  eşik değerden küçük olursa ise o test dizisi intron olarak sınıflandırılır [87]. AFD yönteminde, eşik değer sınıflandırıcının bir parametresidir.

*Pencereleme:* İşaret işleme uygulamalarında sonsuz uzunluktaki bir işaret dizisi ile çalışmak imkânsız olduğunda bütün işaret analizlerinde pencereleme yapılması gerekmektedir [92,93]. Orijinal veriyi pencerelemek için işaretin bir bölümü seçilir. En basit pencereleme tekniğinde verilen işaretin incelenecek kısmı 1 ile dışarıda kalan kısmı ise 0 ile çarpılır. İşaretler işlenmeden önce belli sayıda örnek içeren parçalara ayrılır. İşte bu parçaların her birine pencere adı verilir. Bu uygulamada kullanılan pencere fonksiyonları Dikdörtgen Pencere Fonksiyonu (Rectangular), Hamming Pencere Fonksiyonu (Hamming), Gaussion Pencere Fonksiyonu ve Blackman Pencere Fonksiyonu'dur [94,95].

*Dikdörtgen(Rectangular) penceresi:*  $\omega(n)=1$ ,  $B=1$

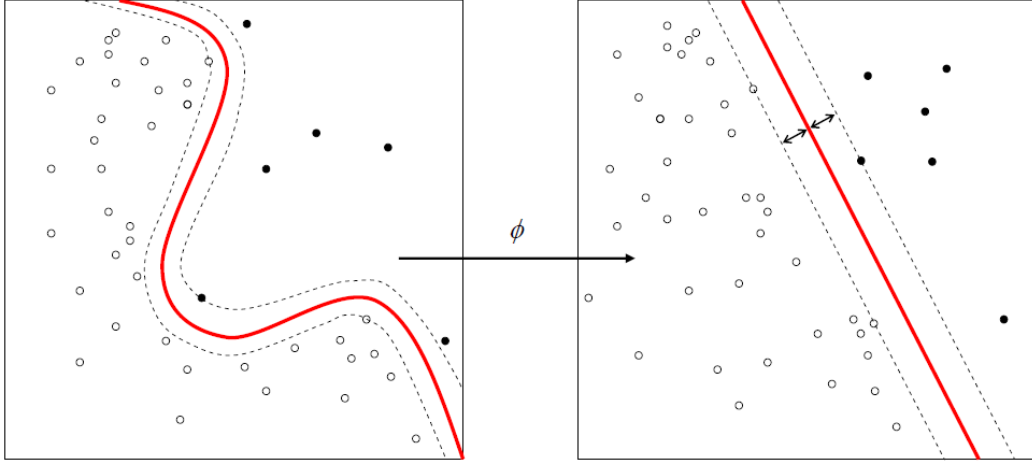
*Hamming penceresi:*  $\omega(n)=0,54-0,46*\cos(2*\pi*n/(N-1))$ ,  $B=1,37$

*Blackman penceresi:*  $\omega(n)=a_0 - a_1*\cos(2*\pi*n/(N-1)) + a_2*\cos(4*\pi*n/(N-1))$ ,  $a_0=(1-\alpha)/2$ ,  $a_1=1/2$ ,  $a_2=\alpha/2$  ve genelde kullanılan  $\alpha$  değeri: 0,16 ve  $B=1,73$

*Gaussion Penceresi:*  $\omega(n)=\exp\left[\frac{-1}{2}\left(\alpha \frac{n}{N/2}\right)^2\right]$ ,  $0 \leq |n| \leq N/2$

#### 4.1.2 Destek Vektör Makineleri

Son yıllarda, sınıflandırma problemlerinin çözümü için geliştirilmiş en başarılı makine öğrenimi algoritmalarından biri Destek Vektör Makineleri'dir [95-97]. Destek Vektör Makineleri, değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki birçok sınıflandırma probleminin çözümünde başarıyla uygulanmış, başarımı yüksek ve etkin makine öğrenimi algoritmalarından biri olarak veri madenciliği uygulamalarındaki yerini almıştır [98-100]. Bu yöntem, sınıflandırmayı bir doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yerine getirir [101,102]. Doğrusal olmayan dönüşümlerde kernel fonksiyonu kullanılmakta ve verilerin daha yüksek boyutta doğrusal olarak ayırımına imkân sağlanmaktadır ve Şekil 4.1'de destek vektör algoritmasının genel yapısı görülmektedir [103-107].



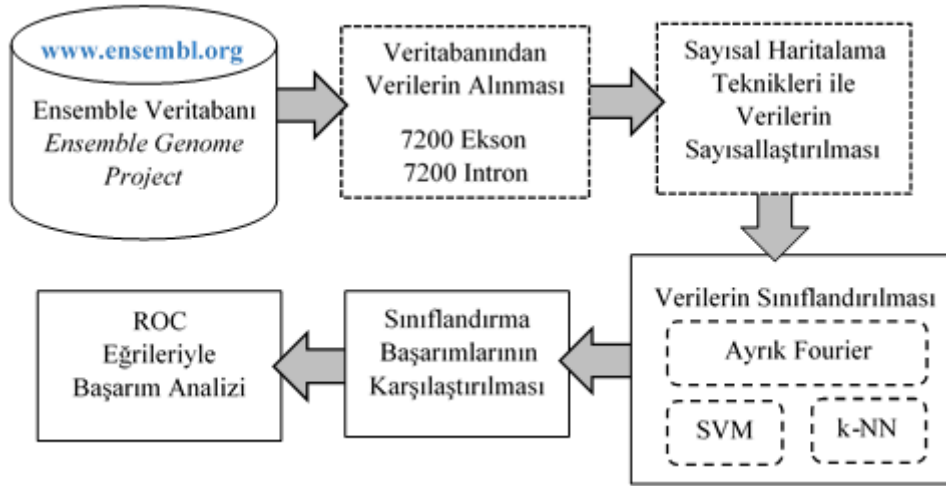
Şekil 4.1 Destek vektör makine algoritması [102]

#### 4.1.3 K-En Yakın Komşu Algoritması

K-En Yakın Komşu Algoritması (k-NN), yeni bir veri geldiğinde var olan öğrenme verisi üzerinde sınıflandırma yapan eğitilmiş öğrenme algoritmasıdır [108,109]. Algoritma, yeni bir veri geldiğinde, onun en yakın k komşusuna bakarak bu verinin sınıfına karar verir [110]. Her sınıfın özelliklerinin önceden belirlenmiş olması çok önemlidir. Yeni gelen verinin daha önceki verilerden k tanesine yakınlığına bakılır. Bu iki veri arasındaki mesafe çeşitli uzaklık fonksiyonları kullanılarak hesaplanır [111,112]. Manhattan Uzaklık Fonksiyonu, Minkowski Uzaklık Fonksiyonu, Öklid Uzaklık Fonksiyonu içerisinde en çok tercih edilen fonksiyon Öklid uzaklık fonksiyonudur. En yakın mesafe neresi ise yeni veri o sınıfa atanır [112].

#### 4.2 Önerilen Uygulama Süreci

Bu uygulama çalışmasında [114], bir DNA diziliminin ekzon ve intron olarak sınıflandırılması için Ayırık Fourier tabanlı bir yaklaşım önerilmektedir. Ensembl veri tabanından elde edilen DNA verilerinin sayısal haritalama tekniklerine göre sınıflandırılması ve analizi için uygun olan DVM ve k-NN sınıflandırma yöntemleri de ayrıca kullanılarak, karşılaştırma analizi yapılmıştır. Sınıflandırma için önerilen yaklaşımın uygulama adımları Şekil 4.2’de verilmiştir.



Şekil 4.2 Ekzon ve intron sınıflandırma için önerilen yaklaşımın uygulama adımları

### 4.3 DNA Verilerinin Elde Edilmesi

Bu uygulama için Ensembl veri tabanındaki Mediterranean fever [Source: HGNC Symbol; Acc:HGNC:6998] Human GRCh38 Chromosome:16p13.3 Erişim Numarası: ENSG00000103313 olan MEFV geni kullanılmıştır [115]. Uygulamada 4800 ekzona ait baz dizisi eğitim amaçlı, 2400 ekzon baz dizisi ise test amaçlı kullanılmıştır. Aynı şekilde 4800 introna ait baz dizisi eğitim amaçlı, 2400 intron baz dizisi test amaçlı kullanılmıştır. Şekil 4.3’de Ensembl veritabanından alınan ekzonlardan bir kesit gösterim biçimi verilmiştir. Ekzonlar koyu ve kırmızı renkli olarak gösterilmektedirler.

| Exons                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | MEFV exons | All exons in this region |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------------------------|
| <pre> &gt;chromosome:GRCh38:16:3241428:3257227:-1 GGCTGGAAGAACCAGTCAACTGGAACCGGATCAACAGGGGTGATGGCATGGCAAGAGTTA TCTCCTGGCAGTGCCCTTCTGGCCTCACTTGCCTTCTTGGGCCAGGAAAGGCAAAGCTCA CAGGACTGTATTAGTGCACCCCTTCCCGTCTGTGCCATTGGCTCTGGAAGGTCC CTGAAACCCGAGTCTGGAGGAGAAGTTCAGGAGCAGGGCGGGCCCTCAGCATAGTCC GAAGCCAGACAGCTGGCTCGAGCCTCTCTGCTCAGCACCATTGGCTAAGACCCCTAGTGA CCATCTGCTGTCCACCCTGGAGGAGCTGGTGCCCTATGACTTCGAGAAGTTCAAGTTCAA GCTGCAGAACACCAGTGTGCAGAAGGAGCACTCCAGGATCCCCGGAGCCAGATCCAGAG AGCCAGGCCCGTGAAGATGGCCACTCTGCTGGTCACCTACTATGGGGAAGAGTACGCCGT GCAGCTCACCTGCAGGTCTGCGGGCCATCAACCAGCGCTGCTGGCCGAGGAGCTCCA CAGGGCAGCCATTAGGTAAGCGGGGCCAGGCCTCTCCTCATCCAGTGCTGAGTGCTG GCTGCTTTGTGGAAAGGGGACCAGGAGCTCAGAGCAGCTCACTCTGACCTGGGGATTGG GAGTCTCAGGTCTACCAAAATCCAGATGACTTTAGTTCAGGAACGTCCCTTTCTTCACTC </pre> |            |                          |

Şekil 4.3 Ensembl veri tabanındaki ekzon görüntüsünden bir kesit

Şekil 4.4’de Ensembl veritabanındaki intronların gösterim biçimi küçük harfle ve mavi olarak gösterilmektedir.

```
GTCGGCGCGACCCGCGACGTCGCCACGCGCCGACAATGGCGGCTGCGTCGGCCTGAGCAG
GGCTTAGTTTAGAAGTAAATTCCTGACGTTGCCGAGGGAGCCGAGTCGCCTCAGATCCG
GTCGGCGCGCGCGCGGTGGCGCGCGCGCAGCGCGCTGCGCGCTCCCGCAGCGCCCTG
GACCTAGCGGCGGTGCCGAGGCCGCGGAGCAAGCCAG

gtggggcgggcgcgcgcgagctcagtaagtgggttggttgacctccctttctctgctcag

CTCCAGCGTCATTCGGCCTCTTAGTTCCTCTGAACCCTGCTCCTGAGCTAGGTAGGAAA
CATGAGCGGCACCAACTTGGATGGGAACGATGAGTTTGATGAGCAGTTGCGAATGCAAGA
ATTGTACGGAGACGGCAAGGATGGTGACACCCAGACCGATGCCGGCGGAGAACCCGATTC
TCTCGGGCAGCAGCCGACGGAACCTCCCTACGAGTGGGACCTGGACAAAAAGGCTTGTT
CCCCAAG

tagggagagtgccacggcgccacttatattttgtcggtctgttgtaattgcttgtag

ATTACTGAAGATTTCAITGCTACATATCAGGCCAATTATGGCTTCTCTAACGATGGCGCA
TCTAGTTCTACCGCAAATGTTGAAGATGTCCATGCTAGGACTGCAGAGGAACCTCCACAA
GAAAAAGCCCCGGAACCCACTGATGCCAGAAAGAAGGGAGAAAAAGAAAGGCTGAGTCA
```

Şekil 4.4 Ensembl veri tabanındaki intron görüntüsünden bir kesit

#### 4.4 Sayısal Haritalama Teknikleri ile DNA Diziliminin Sayısallaştırılması

Bu uygulamada öncelikle Ensembl veri tabanından elde edilen ekzon ve intron dizilimleri, literatürde yaygın kullanılan 9 adet haritalama tekniklerine göre sayısal biçime dönüştürülmüştür. Bu teknikler tezin 2. Bölümünde açıklanan, Tamsayı Haritalama Tekniği, Reel Haritalama Tekniği, Moleküler Kütle Haritalama Tekniği, DNA-Walk Haritalama Tekniği, Eşleştirilmiş Sayısal Haritalama Tekniği, Karmaşık 1 Haritalama Tekniği, Karmaşık 2 Haritalama Tekniği ve EIIP Haritalama Tekniği yöntemleridir. Bu 9 farklı sayısal haritalama tekniklerinin örnek bir DNA diziliminde uygulaması Tablo 4.1’de gösterilmektedir.

**Tablo 4.1** DNA sayısal haritalama teknikleri

| Teknik                             | Sayısal Temsili                                                                                        | $X(n)=[G T G T A C C A C A C T T T C T T T A T C C A G]$                                                                                                                      |
|------------------------------------|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Tamsayı Tekniği</b>             | pürin(A,G) > pirimidin(C,T) ise<br>T=0, C=1, A=2, G=3<br>T>A ve G>C ise A=0, C=1, T=2 ve G=3           | [3 2 3 2 0 1 1 0 1 0 1 2 2 2 1 2 2 0 2 1 1 0 3 2]                                                                                                                             |
| <b>Reel Tekniği</b>                | A=-1,5 T=1,5 C=0,5 G=-0,5                                                                              | -0,5 1,5 -0,5 1,5 -1,5 0,5 0,5 -1,5 0,5 -1,5 0,5 1,5 1,5 1,5 0,5 1,5 1,5 1,5<br>-1,5 1,5 0,5 0,5 -1,5 -0,5                                                                    |
| <b>Atomik Tekniği</b>              | A=70, G=78, C=58, T=66                                                                                 | 78 66 78 66 70 58 58 70 58 70 58 66 66 66 58 66 66 66 70 66 58 58 70<br>78 66 58                                                                                              |
| <b>Moleküler<br/>Kütle Tekniği</b> | A=134, G=150, C=110, T=125                                                                             | 150 125 150 125 134 110 110 134 110 134 110 125 125 125 110 125<br>125 125 134 125 110 110 134 150 125                                                                        |
| <b>DNA-Walk Tekniği</b>            | A=1, G=-1, T=j ve C=-j                                                                                 | 0 0 0 0 -1 1 1 -1 1 -1 1 0 0 0 1 0 0 0 -1 0 1 1 -1 0 0 1                                                                                                                      |
| <b>Eşleştirilmiş<br/>Tekniği</b>   | Sayısal x(n)pürin ise (A ya da G) $u_i=1$ dir,<br>eğer x(n) purimidin ise (C ya da T)<br>$u_i=-1$ dir. | -1 1 -1 1 1 -1 -1 1 -1 1 -1 1 1 1 -1 1 1 1 1 -1 -1 1 -1 -1                                                                                                                    |
| <b>Karmaşık 1 Tekniği</b>          | $a=1+j$ , $t=1-j$ , $c=-1-j$ ve $g=-1+j$ dir.                                                          | -1-j 1-j -1-j 1-j 1+j -1+j -1+j 1+j -1+j 1+j -1-j 1-j -1-j 1-j 1+j -1-j<br>1+j -1+j 1+j -1-j                                                                                  |
| <b>Karmaşık 2 Tekniği</b>          | A=1, T=j, C=-j ve G=-1                                                                                 | -1 j -1 j 1 -j -j 1 -j 1 -j j j j -j j j j 1 j -j -j 1 -1 j -j                                                                                                                |
| <b>EIIP Tekniği</b>                | A= 0,1260 G=0,0806 C=0,1340<br>T=0,1335                                                                | 0,0806 0,1335 0,0806 0,1335 0,1260 0,1340 0,1340 0,1260 0,1340<br>0,1260 0,1340 0,1335 0,1335 0,1335 0,1340 0,1335 0,1335 0,1335<br>0,1260 0,1335 0,1340 0,1340 0,1260 0,0806 |

#### 4.5 DNA Dizilimlerindeki Ekzon/Intron Bölgelerinin Sınıflandırılması

DNA dizilimlerindeki ekzon ve intron bölgelerinin sınıflandırılması için üç farklı yöntem kullanılmış ve başarımları karşılaştırılmıştır.

##### 4.5.1 Ayırık Fourier Dönüşümü Yöntemi ile Sınıflandırma

Sayısal işaretlere dönüşmüş DNA dizilimlerine Ayırık Fourier Dönüşümü metodu uygulanmıştır. Denklem 4.1, 4.2, 4.3 uygulanarak ekzon ve intron eğitim dizilimlerinden güç spektrumları bulunmuştur. Aynı şekilde  $T_3$  değeri Denklem 5'e göre ekzon ve intron eğitim dizilimlerinden elde edilmiştir. Daha sonra ekzon ve intron test dizilerinden elde edilen  $P_{3t}$  değeri  $T_3$  değeri ile karşılaştırılmıştır. Elden edilen  $P_{3t}$  değeri  $T_3$  değerinden büyük veya eşit olma durumunda ekzon sınıfı olarak nitelendirilir, küçük olması durumunda ise intron sınıfı olarak nitelendirilmektedir. Ayırık Fourier dönüşüm metodunda DNA diziliminin AFD spektrumun bulunmasında 4 farklı pencere (Dikdörtgen, Hamming, Gaussian, Blackman) kullanılmış ve bu pencerelerinin sınıflandırma başarımına etkisi incelenmiştir. 9 haritalama tekniğinin sınıflandırma başarımı, pencereleme fonksiyonlarının

etkisine göre Tablo 4.2’de verilmiştir. Ekzon ve intron sınıflandırma başarımları Denklem 4.6 ve Denklem 4.7’de göre hesaplanmıştır. Doğruluk oranı için Denklem 4.8 kullanılmaktadır.

$$\text{Ekzon Sınıflandırma} = \% \frac{\text{Doğru Bulunan Ekzon Sayısı}}{\text{Toplam Ekzon Sayısı}} * 100 \quad (4.6)$$

$$\text{Intron Sınıflandırma} = \% \frac{\text{Doğru Bulunan Intron Sayısı}}{\text{Toplam Intron Sayısı}} * 100 \quad (4.7)$$

$$\text{Doğruluk} = \% \frac{\text{Doğru Bulunan Ekzon Sayısı} + \text{Doğru Bulunan Intron Sayısı}}{\text{Ekzon ve Intronların Toplam Sayısı}} * 100 \quad (4.8)$$

#### 4.5.2 Destek Vektör Makineleri Yöntemi ile Sınıflandırma

DNA dizilimlerinin 9 farklı haritalama tekniği ile sayısal işaretlere dönüştürülmüş dokuz farklı DNA dizilimlerinin her birine destek vektör makineleri yöntemi uygulanmıştır. Destek vektör makineleri (DVM) ile gerçekleştirilecek sınıflandırma için radyal tabanlı fonksiyon (RBF) kerneli kullanılmıştır. Uygulamada kernel parametresi (RBF kerneli için band genişliği değeri) 2, düzenleme parametresi (C) ise 100000 olarak seçilmiştir.

#### 4.5.3 k-En Yakın Komşu Algoritması ile Sınıflandırma

Haritalama teknikleri ile sayısal işaretlere dönüşmüş DNA dizilimlerine K-En Yakın Komşu algoritması (k-NN) uygulanmıştır. Bu çalışmada mesafe hesaplama işleminde öklid uzaklığı kullanılmış, k değeri ise 10 olarak seçilmiştir.

#### 4.6 Sınıflandırma Başarımlarının Karşılaştırılması

Önerilen yaklaşıma göre, gerçekleştirilen uygulamalarda, pencereleme fonksiyonlarının da etkisi alınarak sayısal haritalama tekniklerinin sınıflandırılma başarımlarının karşılaştırılma sonuçları Tablo 4.2’de verilmiştir. Tablo 4.2’de görüldüğü üzere 4 farklı pencereleme fonksiyonuna göre %96,2 sınıflandırma başarımları ile Tamsayı Tekniği ve %95,58 sınıflandırma başarımları ile Eşleştirilmiş Sayısal Tekniği en yüksek doğruluk oranlarına sahiptir. Ancak, bu iki tekniğin doğruluk oranları, pencereleme fonksiyonlarına göre farklılık göstermektedir. Kullanılan Hamming penceresi fonksiyonunda bu iki teknik en yüksek sınıflandırma başarımları göstermektedir. Ayrıca, DNA-

Walk Tekniği %90,82 sınıflandırma başarımı ile Blackman pencere fonksiyonunda en yüksek doğruluk oranına sahipken, diğer 3 pencereleme fonksiyonda da yüksek doğruluk oranıyla sınıflandırma başarısı göstermektedir. Ayrıca AFD sınıflandırma yöntemi ile elde edilen sonuçlar, makine öğrenme tabanlı Destek Vektör Makineleri ve K-En Yakın Komşu algoritması yöntemleriyle elde edilen sonuçlarla karşılaştırılmıştır. Tablo 4.2’de görüldüğü üzere, en yüksek sınıflandırma başarımına sahip Hamming pencereleme fonksiyonu için %96,21 başarımla Tamsayı tekniği, %87,49 başarımla DNA-Walk Tekniği ve %95,58 başarımla Eşleştirilmiş Sayısal tekniğinin sınıflandırma başarımları diğer sayısal haritalama tekniklerine göre daha yüksek çıkmıştır.

**Tablo 4.2** Pencere fonksiyonlarına göre sınıflandırma başarımları

| Pencereleme Fonk.             | Dikdörtgen (Rectangular) Pencere (%) |       |        |              | Hamming Pencere (%)    |       |        |              |
|-------------------------------|--------------------------------------|-------|--------|--------------|------------------------|-------|--------|--------------|
| Sayısal Haritalama Tekn.      | Eşik Değer                           | Ekzon | Intron | Doğruluk     | Eşik Değer             | Ekzon | Intron | Doğruluk     |
| Tamsayı Tekniği               | 1.3778e+007                          | 87,10 | 88,25  | <b>87,68</b> | 4.6856e+005            | 92,75 | 99,66  | <b>96,21</b> |
| Reel Tekniği                  | 2.0559e+006                          | 17,70 | 14,79  | 16,24        | 6.1854e+005            | 41,60 | 35,50  | 38,55        |
| Atomik Tekniği                | 1.4429e+013                          | 21,40 | 75,08  | 48,24        | 2.8925e+009            | 11,30 | 95,04  | 53,17        |
| Moleküler Kütle Tekniği       | 1.9188e+014                          | 63,25 | 60,66  | 61,95        | 4.2838e+010            | 47,45 | 52,70  | 50,07        |
| DNA-Walk Tekniği              | 2.9003e+005                          | 73,20 | 75,19  | 74,19        | 7.8216e+004            | 87,87 | 87,12  | <b>87,49</b> |
| Eşleştirilmiş Sayısal Tekn    | 2.9106e+005                          | 87,80 | 88,02  | <b>87,91</b> | 7.8517e+004            | 96,29 | 94,87  | <b>95,58</b> |
| Karmaşık 1 Tekniği            | 1.2601e+006                          | 54,54 | 67,75  | 61,14        | 1.4822e+005            | 25,55 | 31,38  | 28,46        |
| Karmaşık 2 Tekniği            | 7.6681e+005                          | 24,35 | 35,35  | 29,85        | 1.5100e+005            | 25,97 | 30,00  | 27,98        |
| EIIP Tekniği                  | 157.8496                             | 78,86 | 76,20  | 77,53        | 0.1106                 | 51,08 | 49,66  | 50,37        |
| Pencereleme Fonk.             | Gaussian Penceresi (%)               |       |        |              | Blackman Penceresi (%) |       |        |              |
| Sayısal Haritalama Tekn.      | Eşik Değer                           | Ekzon | Intron | Doğruluk     | Eşik Değer             | Ekzon | Intron | Doğruluk     |
| Tamsayı Tekniği               | 0.4583(1.0e+006)                     | 86,12 | 86,38  | <b>86,25</b> | 4.6323e+005            | 85,25 | 87,12  | <b>86,18</b> |
| Reel Tekniği                  | 6.3082e+005                          | 54,37 | 46,01  | 50,19        | 6.6045e+005            | 52,08 | 43,91  | 47,99        |
| Atomik Tekniği                | 1.1192e+009                          | 26,25 | 99,95  | 63,1         | 4.8616e+008            | 26,54 | 96,00  | 61,27        |
| Moleküler Kütle Tekniği       | 1.7505e+010                          | 47,40 | 60,83  | 54,11        | 8.1113e+009            | 46,00 | 71,16  | 58,58        |
| DNA-Walk Tekniği              | 8.0278e+004                          | 88,66 | 89,58  | <b>89,12</b> | 8.3740e+004            | 89,75 | 91,90  | <b>90,82</b> |
| Eşleştirilmiş Sayısal Tekniği | 8.0288e+004                          | 92,92 | 91,60  | <b>92,26</b> | 3.2685e+005            | 89,00 | 86,58  | <b>87,79</b> |
| Karmaşık 1 Tekniği            | 1.4493e+005                          | 24,44 | 29,44  | 26,94        | 1.4116e+005            | 23,05 | 25,37  | 24,21        |
| Karmaşık 2 Tekniği            | 1.4817e+005                          | 25,55 | 28,47  | 27,01        | 1.4458e+005            | 24,86 | 26,66  | 25,76        |
| EIIP Tekniği                  | 0.0806                               | 49,66 | 50,41  | 50,03        | 0.0746                 | 49,66 | 50,41  | 50,03        |

Tablo 4.3’de görüldüğü üzere, 3 yöntemde de aynı veri kümesi kullanılmasına rağmen DVM ve k-NN gibi makine öğrenme tabanlı yöntemlerin ekzon ve intron sınıflandırma başarımlarının düşük olduğu, buna rağmen Ayrılcı Fourier Dönüşümü yönteminin ise diğer iki yöntemle göre DNA dizilimi sınıflandırmada güçlü bir metot olduğu açıkça görülmektedir. Bunun nedeni, AFD sınıflandırma yönteminde, eşik değeri parametresinin

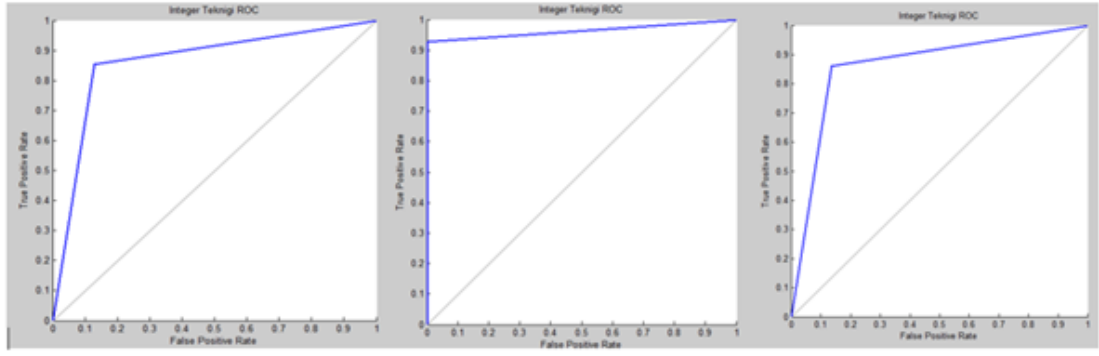
ekzon ve intron bölgelerini ayırmada oldukça etkili olmasıdır. DNA dizilimindeki ekzon bölgelerinin güç spektrumları intron bölgelerine göre her zaman yüksektir. Dolayısıyla eşik değeri tanımlayan denklemde de eşik değere eşit veya eşik değerden büyük güç spektrumu ekzon bölgesi olarak, eşik değerden küçük güç spektrumları ise intron bölgesi olarak kabul edilir. AFD yönteminde, eşik değeri olarak etkili bir parametre kullanıldığı için sınıflandırma başarımı diğer yöntemlerden daha yüksek çıkmıştır.

**Tablo 4.3** Sayısal haritalama teknikleri için Fourier tabanlı yöntemin sınıflandırma başarımının diğer makine öğrenme tabanlı yöntemlerle karşılaştırılması

| Sayısal Haritalama Teknikleri         | AFD (%) | DVM (%) | k-NN (%) |
|---------------------------------------|---------|---------|----------|
| Tamsayı Tekniği [9]                   | 96,21   | 63,2653 | 67,44    |
| Reel Tekniği [10]                     | 38,55   | 61,2245 | 54,07    |
| Atomik Tekniği [16]                   | 53,17   | 53,0612 | 58,14    |
| Moleküler Kütle Tekniği [17]          | 50,075  | 54,0816 | 59,88    |
| DNA-Walk Tekniği [18]                 | 87,495  | 66,013  | 69,56    |
| Eşleştirilmiş Sayısal Tekniği [19-20] | 95,58   | 71,6326 | 75,68    |
| Karmaşık 1 Tekniği [11]               | 28,465  | 52,33   | 56,13    |
| Karmaşık 2 Tekniği [15]               | 27,985  | 50,66   | 54,17    |
| EIIP Tekniği [21-22]                  | 50,37   | 48,23   | 50,58    |

#### 4.6.1 Sınıflandırma Başarımlarının ROC Eğrileriyle Analizi

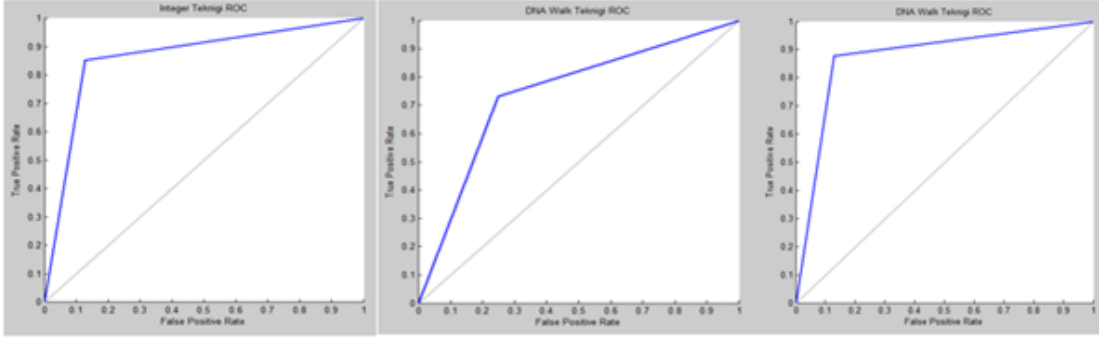
AFD sınıflandırma yönteminin uygulanması sonucuna göre en yüksek başarımlara sahip Tamsayı Tekniği, Eşleştirilmiş Sayısal Tekniği ve DNA-Walk Tekniğinin ROC eğrileri Şekil 4.5’de gösterilmiştir. Şekil 4.5’deki ROC eğrilerinden görüldüğü üzere Tamsayı tekniğinin Hamming penceresi fonksiyonunda, DNA-Walk tekniğinin Blackman penceresi fonksiyonunda, Eşleştirilmiş Sayısal tekniğinin ise Hamming penceresi fonksiyonunda eğrinin altında kalan alan diğer pencereleme fonksiyonlarına göre daha büyüktür. Alanın büyük olması daha yüksek bir doğruluk başarımına sahip olduğunu göstermektedir.



a) Tamsayı tekniği ile Dikdörtgen pencere

b) Tamsayı tekniği ile Hamming pencere

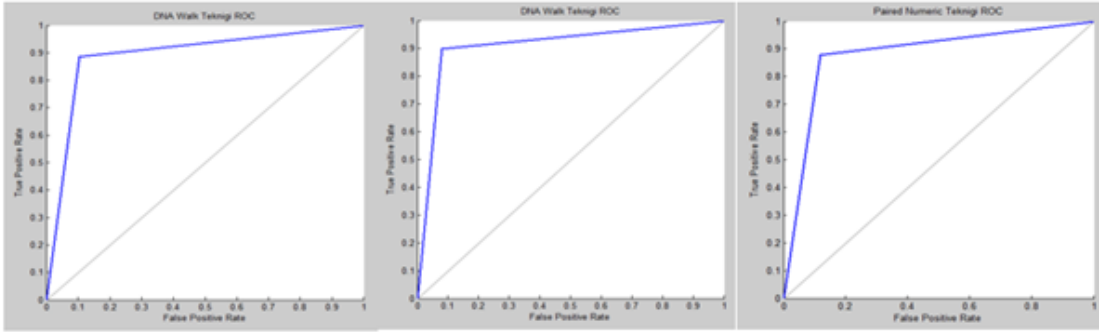
c) Tamsayı tekniği ile Gaussion pencere



d) Tamsayı tekniği ile Blackman pencere

e) DNA-Walk tekniği ile Dikdörtgen pencere

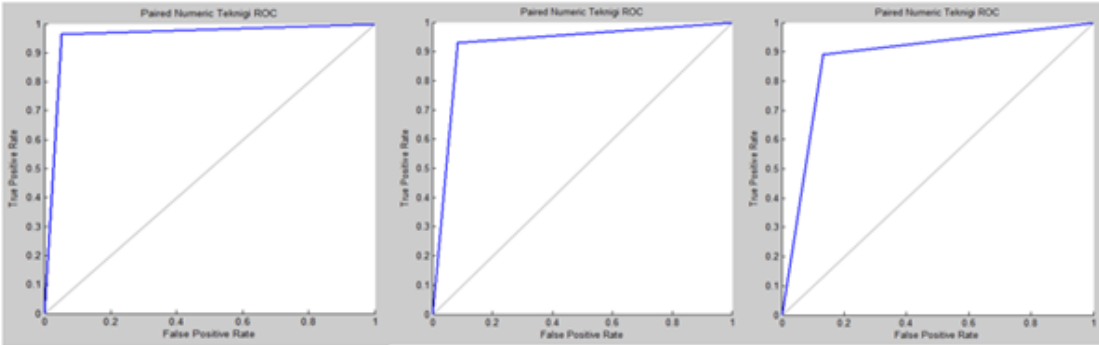
f) DNA-Walk tekniği ile Hamming pencere



g) DNA-Walk tekniği ile Gaussion pencere

h) DNA-Walk tekniği ile Blackman pencere

ı) Eşleştirilmiş tekniği ile Dikdörtgen pencere



j) Eşleştirilmiş tekniği ile Hamming pencere

k) Eşleştirilmiş tekniği ile Gaussion pencere

l) Eşleştirilmiş tekniği ile Blackman pencere

Şekil 4.5 Ekzon ve intron sınıflandırmada 3 haritalama tekniğinin ROC eğrileriyle başarımları

#### 4.7 Uygulama Sonuçlarının Değerlendirilmesi

Gerçekleştirilen bu uygulamada, bir DNA diziliminin ekzon ve intron olarak sınıflandırılmasında farklı sayısal haritalama tekniklerinin başarımları karşılaştırılmıştır. Karşılaştırılma sonucunda insan türünün MEFV genine ait DNA dizilimleri, 9 farklı haritalama tekniği ile sayısal dizilere dönüştürülmüştür. Dönüştürülen bu sayısal dizileri ekzon ve intron olarak sınıflandırmak için AFD yöntemi kullanılmıştır. Ayrık Fourier Dönüşümü yönteminde 4 farklı pencereleme fonksiyonu kullanılmış ve bu pencereleme fonksiyonlarının da sınıflandırma başarımı üzerindeki etkileri karşılaştırılmıştır. Uygulama sonucunda Tamsayı Tekniği ve Eşleştirilmiş Sayısal Tekniği dört pencereleme fonksiyonunda diğer sayısal tekniklere göre daha yüksek sınıflandırma başarımı göstermiştir. DNA-Walk Tekniği ise Rectangular pencereleme fonksiyonu dışında diğer 3 pencereleme fonksiyonu için yüksek başarımlar göstermiştir. Hamming Pencereleme Fonksiyonunun sınıflandırma başarısı diğer pencereleme fonksiyonlarından daha yüksek olduğu gözlemlenmiştir. Ayrıca Fourier tabanlı yöntemle elde edilen sonuçlar makine öğrenmesi tabanlı Destek Vektör Makineleri ve K-En Yakın Komşu Algoritması yöntemleri ile de karşılaştırılmıştır. Fourier tabanlı yöntemin diğer iki yöntemle göre sınıflandırma başarımı oldukça yüksek çıkmıştır. Paired Numerik tekniği, DNA'nın yapısal özelliğini iyi yansıtmaması ve Ayrık Fourier tekniğinde karmaşıklığı azaltmasından dolayı protein kod bölgelerini bulmada diğer tekniklere göre yüksek başarımlar vermiştir. Tamsayı tekniği ise DNA dizilimini 0,1 şeklinde basit olarak sayısallaştırdığı için bu hesaplamanın etkili olmasını sağlamıştır. DNA Walk tekniği ise uygulamada kullanılan gen dizilimi üzerinde geniş aralıklı korelasyon bilgisi sağladığından, bu dizilimde bazı kalıpların periyodik aralıklarla tekrarlanmasından ve bazı dizilimlerin kombinasyonlarının değişiminden dolayı (AG-GA veya AGA-GAG şeklinde) diğer sayısal haritalama tekniklerine göre daha yüksek başarımlar sağlamıştır.

## 5 DNA DİZİLİMLERİ İÇİN ENTROPİ TABANLI YENİ BİR SAYISAL HARİTALAMA TEKNİĞİ

Protein sentezindeki her amino asit üç nükleotid ile kodlanır. DNA'daki bu trinükleotid, kodlanmış amino asitler “Kod” olarak adlandırılır. RNA’da ise bu üçlü nükleotidlere “Kodon” denir. Burada kodon tek bir amino asite karşılık gelir. Kodlar, DNA'daki mRNA sentezi sırasında kodonlara dönüştürülür. Kod ve kodon tamamlayıcıdır. RNA'da 64 çeşit kodon vardır [116,117] [61].

Tezin bu bölümünde mevcut haritalama tekniklerine alternatif olarak, ekzon ve intron sınıflandırmadaki başarımların oranını dahada arttırmak ve ayrıca ekzon bölgelerini işaret üzerinde gösterebilmek için entropi tabanlı yeni bir haritalama tekniği geliştirilmiştir. Önerilen bu tekniğin başarımlarını en yüksek başarımla sahip mevcut haritalama tekniklerinin başarımlarıyla Tekli Değer Ayırımı, Ayrık Fourier Dönüşümü ve Kısa Zamanlı Fourier Dönüşümü yöntemlerinde karşılaştırılmıştır. Bu uygulamanın yazılımı, MATLAB yazılım ortamında gerçekleştirilmiştir.

### 5.1 Entropiye Dayalı Yeni Sayısal Haritalama Tekniği

Shannon entropisi bilgi içeriği ve karmaşıklığın bir ölçümüdür. İlk olarak 1948 yılında Shannon tarafından tanıtılmıştır. İletişimde 0 ve 1 olmak üzere iki kararlı durum vardır. Herhangi bir zamanda 1 bitlik bilgi taşıyan bir aygıt söz konusu ise, N adet aygıt toplam  $2^N$  durum için N bit’lik bilgi taşıyacaktır. Dolayısıyla  $\log_2 2^N = N$  olarak yazılabilir. Buradan işaret sayısının logaritmasının zamanla doğrusal olarak arttığı görülmektedir. Shannon meydana gelen olasılıkları  $p_1, p_2, p_3, \dots, p_n$ , olan ve bu olasılıklardan hangisinin önce meydana geldiği bilinmeyen bir olaylar kümesini ele almıştır. S belirsizlik ölçüsü olarak tanımlanırsa bu belirsizlik ölçüsü Denklem 5.1 ile yazılabilir [118,119]

$$S = - \sum_i p(x_i) \log(p(x_i)) \quad (5.1)$$

Denklem 5.1’deki ifade Shannon belirsizliği, Shannon entropisi ya da Bilgi Teorisi olarak adlandırılır. Denklem 5.1’ de olasılıklar ile  $x_i$  değerleri üzerinde rastgele bir değişken dağıtılır. Belli sayıda olası durum için, bu işlev, tüm durumların olasılıkları eşit olduğunda bir maksimuma sahiptir. Shannon entropisi zaman içerisinde topolojik, metrik, Kolmogorov-Sina ve Rényi entropisi olmak üzere birçok farklı hal almıştır. Bu entropiler, bir sistemi karmaşıklık veya basitlik ölçütü ile sınıflandırmak amacıyla tanımlanmıştır [120].

Entropi ve bilgi teorisi uygulamaları, genetik dizinin evrimi üzerine temel görüşleri sağlamada yararlı olduğu kanıtlanmıştır [121,122]. DNA gibi sembolik dizilimlerin düzensizliğini ölçmede entropi kullanılmaktadır [123]. Özellikle topolojik entropi, yüksek boyutluluk problemleri nedeniyle genetik uygulamalarında nadiren kullanılır [119]. Bu sorunlar, topolojik entropinin matematiksel kavramının sonsuz uzunluk dizileri üzerinde çalışılmaya başlandığı gerçeğinden kaynaklanmaktadır. Entropi uygulamasındaki en zor konunun, sınırlı örneklem etkisinden kaynaklanan yakınsama problemi olduğu genel olarak kabul edilmektedir [124,125]. Bu problemleri topolojik entropi ile önlemek ve sonlu uzunluk dizilerine uyarlamak için birkaç farklı yaklaşım daha önce denenmiştir. Örneğin, dilsel karmaşıklık (toplam albilginin toplam muhtemel albilginin fraksiyonu), sonlu örnek problemlerini atlamak için kullanılır [126]. Bu, intron bölgelerinin karmaşıklığının, ekzon bölgelerinin karmaşıklığından daha düşük olduğunu gözlemlemeye yol açar.

Moleküler düzeydeki bilgi DNA ve protein dizileri tarafından taşınır. Shannon'un entropi kavramı, genellikle sembolik dizilerin istatistiksel topluluklarına uygulandığında şeffaf bir rol oynamaktadır. Entropinin işlevselliğin derecesini bu sembolik dizilere özgüdür. Bir DNA sekansının işlevselliği, etkileşim özgünlüğü anlamına gelir; Dolayısıyla, yüksek sekans değişkenliği yani entropisi bu dizilerin gerçekleştirebileceği daha fazla spesifik etkileşimdir [127-129].

Shannon entropisi, nükleotid dizilimi analizinde entropi hesaplamaları, DNA dizilimindeki ekzon, intron ve aminoasitlerin dağılımını karakterize etme gibi alanlarda uygulanır [118,119]. Bilginin ölçüsü olarak bilinen entropinin farklı DNA dizilimlerinin uzunluklarını hesaba kattığı kabul edilmektedir. DNA dizilimleri, bu dizilimlerin periyodik olması ve bazı bazların veya altdizilimlerin sürekli tekrar edilmesinden dolayı düzenli, farklı baz kombinasyonları yani kodonları bulundurmasından ve bazı kombinasyonların eşit frekansa sahip olmasından dolayı da düzensiz olarak kabul edilir. Hem düzenli hemde düzensiz olan DNA dizilimlerini en iyi ölçmede Shannon entropi kullanılır.

Farklı türlerin kromozom bilgisini belirlemede [118], DNA dizilim analizi için karmaşıklık ölçümlerinde [123] entropi kullanılmasına rağmen, DNA dizilimlerinin sayısallaştırılması için Shannon entropi kullanılması ile ilgili bir çalışmaya rastlanılmamıştır.

Uygulamanın bu bölümünde, DNA dizisinin sayısal gösterimi için önerilen yeni bir sayısal haritalama tekniği sunulmaktadır. Dizi değerlerini sayısal değerlere dönüştürmek için

yeni Shannon entropi yöntemini kullanılmıştır. Shannon entropisi, uygulamada DNA dizilimlerinin sayısal gösterimi için yetersiz kabul edilir. Bu nedenle, dizilimi sayısal değerlere dönüştürmek için yeni bir entropi yönteminin geliştirilmesine gereksinim bulunmaktadır. Tez çalışmasında geliştirilen yöntem, Ali Karcı tarafından önerilen Shannon denkleminin kesirli bir türevi olan fraksiyonel Shannon entropi temellidir [130]. Shannon entropisi reel sayılarda, fraksiyonel entropi denklemi ise hem reel hemde karmaşık sayılarda 0.01 ile 1 arasında daha iyi sonuçlar vermektedir. Fraksiyonel entropi denklemini, DNA dizilimlerine yeni bir sayısal gösterim tekniği olarak uygularken,  $\alpha$  ve  $p(x_i)$  değerleri üzerinde DNA dizilimlerine göre uyarlamalı olarak belirlendi. Önerilen haritalama tekniğinde, DNA dizilimlerindeki kodon dağılımları için entropi hesaplanmaktadır. Her kodon için entropi, 64 olası durumun frekans dağılım değerlerini yansıtır. Bir DNA dizilimindeki AAA, ATT, TGC, GCC, vs. gibi kodonlar sayısı  $i$  ile gösterilmektedir.  $i$ 'nin değeri 64'tür. Önerilen yaklaşım, Denklem 5.2 ile tanımlanır.

$$Sf = -\sum_i [(-p(x_i))^{\alpha_i} p(x_i) \log(p(x_i))] \quad (5.2)$$

Denklem 5.2' deki  $p(x_i)$  değeri bir DNA dizilimindeki her bir kodonun tekrarlama sıklığıdır. Bu denklemde  $\alpha_i$  parametresi için Denklem 5.3'te verilen yeni bir tanımlama yapılmıştır.

$$\alpha_i = \frac{1}{\log(p(x_i))} \quad (5.3)$$

Buna göre ( $\alpha_i$ ) değerini sabit bir değer almak yerine, alfa doğrudan genom dizilim verisinden uyarlamalı olarak hesaplanmaktadır. Şöyleki; dizideki bir kodonun tekrarlama olasılığı ( $p(x_i)$ ) yüksek ise  $\alpha_i$  değeri büyük, tersi durumunda küçük olacaktır. Burada ( $p(x_i)$ ) ile  $\alpha_i$  arasında doğrusal bir ilişki kurması amaçlanmıştır. Bu sayede kodon olasılıkları arasındaki korelasyonlar çok iyi yansıtılmış ve DNA diziliminin sayısal temsili için daha geniş bir sayısal aralık sağlanmıştır.

## 5.2 Ekzon ve Intron Bölgelerinin Belirlemesi İçin Kullanılan Yöntemler

En yaygın kullanılan TDA, KZFD ve AFD yöntemleri kullanılmıştır. AFD Bölüm 4'de açıklanmış olup diğer 2 yöntem aşağıda açıklanmıştır.

### 5.2.1 Tekli Değer Ayırışım Yöntemi

Tekli Değer Ayırışım (TDA- Singular Value Decomposition) yöntemi lineer cebirde, karmaşık matrislerde kullanılan bir algoritmadır. Gen analizine, bilgi getirimi, bilgi çıkarımından boyut azaltma (temel bileşenler analizi gibi) ve veri sıkıştırmaya kadar uzanan geniş bir yelpazede kullanılan yöntemdir. Basitçe bir matrisi 3 parçaya ayırarak tutar ve bu üç parçayı kullanarak aynı matrisin yeniden elde edilmesini sağlar [131,132]. Gen analizinde kullanılan TDA yönteminin matematiksel gösterim biçimi Denklem 5.4’da verilmiştir.

$$A_{n \times p} = U_{n \times n} S_{n \times p} V_{p \times p}^T \quad (5.4)$$

Denklem 5.4’de U ve  $V^T$  ortonormal matris, S ise diyagonal matristir.

Denklem 5.4’de U’nun sütunları sol tekli vektörüdür (gen katsayı vektörleri).  $V^T$  sağ tekli vektörleri (ifade seviyesi vektörleri) olan satırlara, S ise A ile aynı boyutlu tekli değerlere sahip diyagonal bir matristir. TDA, orijinal verilerin, kovaryans matrisinin köşegen olduğu bir koordinat sisteminde genişlemeyi temsil eder [133].

### 5.2.2 Kısa Zamanlı Fourier Dönüşümü

Kısa Zamanlı Fourier dönüşümü (KZFD) işaret isleme uygulamalarında durağan olmayan işaretlerin analizi için iyi bilinen bir yöntemdir. KZFD yöntemi işareti dar zaman aralıklarına ayırmaktadır ve her aralığının da Fourier dönüşümünü almaktadır. Sonlu enerjiye sahip  $x(t)$  işareti alınırsa, denklemde  $x(t)$  işaretinin tüm t değerleri için tanımlı olduğu kabul edilmektedir [134,135].

$$\int_{-\infty}^{\infty} |x(t)|^2 dt < \infty \quad (5.5)$$

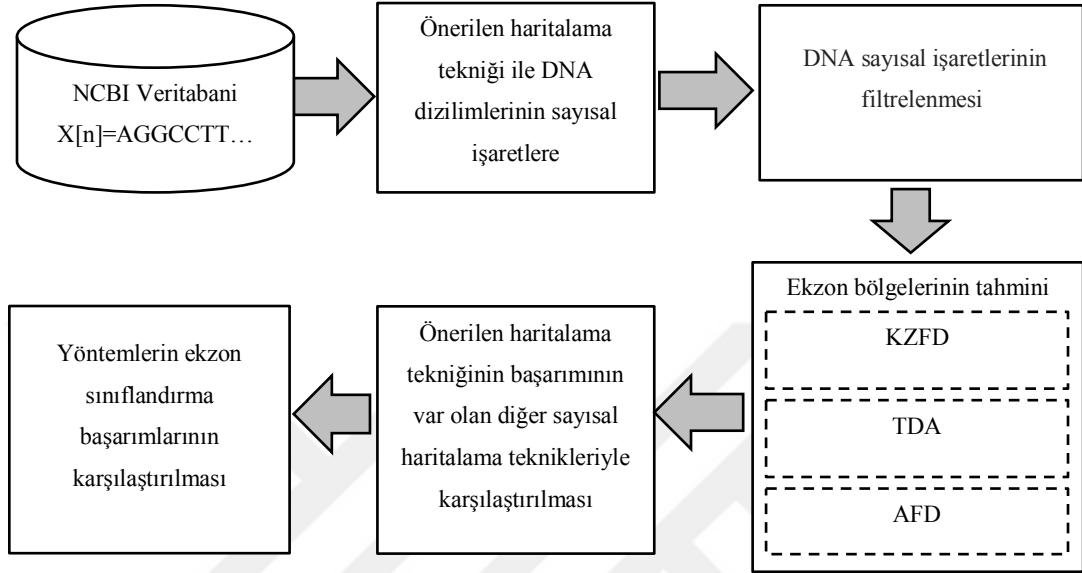
n zaman ve  $\omega_s$  frekans domaininde gösterilen  $x(n)$  işaretinin KZFD formülü Denklem 5.6’da gösterilmektedir[136-138].

$$X_n(e^{j\omega_s}) = \sum_k \omega(n - k)x(k)e^{-j\omega_s k} \quad (5.6)$$

### 5.3 Önerilen Entropiye Dayalı Sayısal Haritalama Tekniğinin Uygulaması

Önerilen sayısal haritalama tekniği [113,139,140] ekzon bölgelerinin tespiti için Ayrık Fourier Dönüşümü, Kısa Zamanlı Fourier Dönüşümü ve Tekli Değer Ayırışımı yöntemlerinde kullanılmıştır. Önerilen haritalama tekniği ekzon bölgelerinin

belirlenmesinde yüksek başarımlar göstermiştir. AFD, KZFD ve TDA yöntemlerinde önerilen sayısal haritalama tekniği ile ekzon bölgelerinin belirlenmesine yönelik uygulamanın blok diyagramı Şekil 5.1’de gösterilmektedir [113].



Şekil 5.1 AFD, TDA ve KZFD yöntemlerini kullanılarak önerilen haritalama tekniği ile ekzon bölgelerinin tahmini

Uygulamada test için 3500-10500 baz aralığındaki AJ229040 geni kullanılmıştır. Kullanılan 7000 baz uzunluğuna sahip DNA dizilimi NCBI veri tabanından alınmıştır [141]. NCBI gen bankasından elde edilen bu gene ait bir kesit Şekil 5.2’de gösterilmektedir.

```

>gi|41223386|emb|AJ229040.1| Homo sapiens 959 kb contig between AML1 and CBR1 on chromosome 21q22GATCACTGAGCAT.
TTTGACGGGTTGTTTCTCTGAGGGCTGTGAGGAGAATCTGTTCTGGGCCCTCTGACTTGACCTGAGATGGTCATCTTCTCTCTCTCTGTCCCCCTTCTCTCTC
CTGTACCGCTCAAGTCTGTGCCAAGACACCACCTGTGGAGCTTCTCACTCAGCTCTGGGCTTTCTGGCCCTCCATGAGCTCTAGTTAATTTCCATGTGGGCTCCAGCAAT
CCTGGCACATAAAGCACAATATGAATTTTTAGGTAAAAATAAAGATTGAAAGCTTTCTCAGTGCCGGGCACTGTGATAAGGACTTTAAACGCGGTACTATAATGATG
GTAACCTGGGATTGCTCTGGCCAATGGGGTGATGGCCACCCCAACTACAGTCTACCAAGGATGACAGAGCACTGGGTGCAGACCCAGGAGAGGTGGGGAGGAGTGGCAGAG
GCTTGCTCATGCTGTGTGTGTGTGTGCACACCAATGTGTATACAGCATGCATGGATCCCAATAGGTCTGACCCCGGCCCATCTCAGTGGCCATCTGGTTGGTGCCCTTT
AATATTTTTTAGTGGTGGTGGGGAGTAAAGGATTAACACACACACACACACACACACACACACAGATTCTTTGGATGTTAGGTGAGAAAAGTTCC
ACTGGCAATACGGGAAAATTTATAGTCAGCACATTTTAAATTAGCCAGACACGAACCTTAGCCGTTTCGCTGAATACATCGCACAGAGTGCTAGAACCTGCAAG
AGGGTCTCTGAAATTAGGATACATCAGAATTCGGAAGTATTTCTCATGTAGCATCTTTTAGATAAGTCCATGTCAGAGGTGGAAAACACAGGTTGTTCTGTTTTGC
AATTTCTGTGGCATATGGAAGCTATTAATTTCTCTTTTCTACTTACTCCCCAGCTACTTTCTTTTAAAGGCCAGGCCGATGTTACACAGTTGCTGTGGCTTCTT
TTCTCAAGAGGACTGGTTAGACATTTGTAGAATGTTCCCATTTGCATGTGTCTGATGTTTTCTTGATTAGACCAGGGTTTTGGGTTTTCTGGGAAGAATACCATGAA
AGGCTTATCTAACAGTCATATTGTGGGTTAGGGCTTCAACATATAAATTTTGCAGGGACACAATTCATCCCTAACAGAGCTTTTTTTTTTTTTTTCTTCTGAGAGT
  
```

Şekil 5.2 NCBI veri tabanından elde edilen AJ229040 geninden bir kesit

Şekil 5.3’de gen bankasından alınan verinin genomik bilgileri, Şekil 5.4’de ise AJ229040 genine ait DNA diziliminin ekzon bölgelerinin NCBI veritabanında gösterim biçimi yer almaktadır.

## Homo sapiens 959 kb contig between AML1 and CBR1 on chromosome 21q22

GenBank: AJ229040.1

[FASTA](#) [Graphics](#)

[Go to:](#) ☒

LOCUS AJ229040 958952 bp DNA linear CON 14-MAR-2015  
DEFINITION Homo sapiens 959 kb contig between AML1 and CBR1 on chromosome 21q22.  
ACCESSION AJ229040  
VERSION AJ229040.1  
DBLINK BioSample: [SAMEA3293435](#)  
KEYWORDS .  
SOURCE Homo sapiens (human)  
ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.  
REFERENCE 1 (bases 897811 to 898239)  
AUTHORS Miyoshi,H., Ohira,M., Shimizu,K., Mitani,K., Hirai,H., Imai,T., Yokoyama,K., Soeda,E. and Ohki,M.  
TITLE Alternative splicing and genomic structure of the AML1 gene involved in acute myeloid leukemia  
JOURNAL Nucleic Acids Res. 23 (14), 2762-2769 (1995)  
PUBMED [7651838](#)  
REFERENCE 2 (bases 897811 to 898239)  
AUTHORS Levanon,D., Bernstein,Y., Negreanu,V., Ghazi,M.C., Bar-Am,I., Aloya,R., Goldenberg,D., Lotem,J. and Groner,Y.

Şekil 5.3 AJ229040 genine ait genel bilgiler

```
repeat region 2878..2964
/inference="non-experimental evidence, no additional
details recorded"
/rpt_family="MIR"
repeat region complement(2981..3198)
/inference="non-experimental evidence, no additional
details recorded"
/rpt_family="MIR"
exon 3702..3821
/inference="non-experimental evidence, no additional
details recorded"
/note="GRAIL, score = 60.000%, comment = good"
exon 3770..3826
/inference="non-experimental evidence, no additional
details recorded"
/note="GenScan, score = 9.66%, comment = Initial_exon 57
bp frame: 1 phase: 0"
exon 4584..4601
/inference="non-experimental evidence, no additional
details recorded"
```

Şekil 5.4 AJ229040 genine ait DNA diziliminin ekzon bölgelerinin temsili

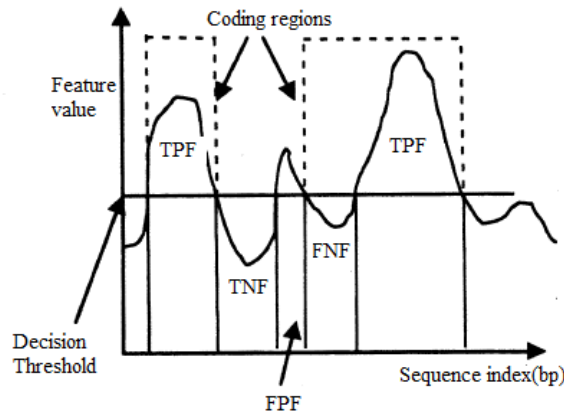
Önerilen sayısal haritalama tekniğinde, genin DNA dizilimindeki her bir kodunun tekrarlanma sıklığı Denklem 5.2 ve Denklem 5.3'e göre hesaplanmaktadır. Örneğin AAC kodunun DNA dizilimdeki tekrarlanma sıklığı 0.0649 olarak hesaplanmıştır. TTT kodunun değeri ise 0.3014 olarak bulunmuştur. Benzer şekilde, 64 çeşit kodonun değerleri Denklem 5.2 ve Denklem 5.3'deki formüllere göre hesaplanmıştır.

## 5.4 Uygulama Sonuçları

Önerilen sayısal haritalama tekniğinin deneysel sonuçları Tekli Değer Ayırışımı, Ayrık Fourier Dönüşümü ve Kısa Zamanlı Fourier Dönüşümü yöntemleri için ayrı ayrı gösterilmiştir. Önerilen sayısal haritalama tekniğinin ekzon bölgelerinin tahmin edilmesindeki başarımı 3 farklı yöntemde test edilmiştir.

### 5.4.1 Tekli Değer Ayırışımı Yönteminin Sınıflandırma Sonuçları

Önerilen sayısal haritalama tekniğinin protein kod bölgelerinin(ekzon) belirlenmesindeki başarımını göstermek için Receiver Operating Characteristic (ROC) eğrilerinin terimleri kullanılmıştır [142]. Bu terimler Doğru Pozitif Oranı (True Positive Rate, TPR, TPF), Yanlış Pozitif Oranı (Yanlış Pozitif Oranı, FPR, FPF), Doğru Negatif Oranı (True Negative Rate, TNO, TNF), Yanlış Negatif Oranı (False Negative Rate, FNO, FNF) olarak Şekil 5.5’ de gösterilmektedir [131,142].



Şekil 5.5 ROC eğrisinde kullanılan terimler [131]

TPF: Protein kod bölgelerin doğru bir şekilde tahmin edilmesi

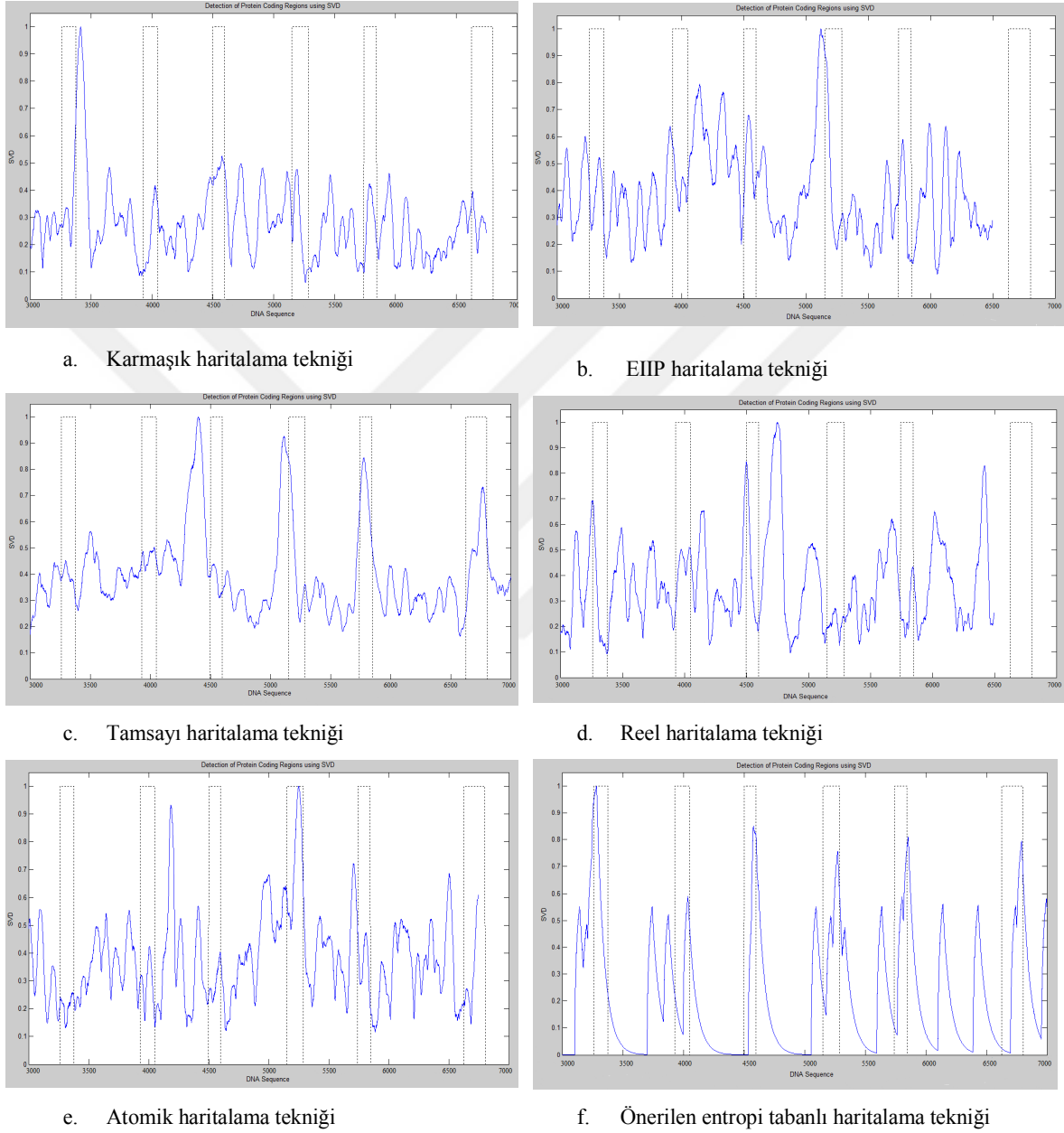
FPF: Protein kod bölgelerin yanlış bir şekilde tahmin edilmesi

FNF: Protein kodlamayan bölgelerin yanlış bir şekilde tahmin edilmesi

TNF: Protein kodlamayan bölgelerin doğru bir şekilde tahmin edilmesi

Protein kodlayan bölgeler için  $TPF + FNF = 1$  ve protein kodlamayan bölgeler için  $TNF + FPF = 1$ ’dir. Şekil 5.6’daki grafik için karar eşik değeri 0.6 seçilmiştir. Şekil 5.6’da ekzon bölgelerinin taranması için farklı haritalama tekniklerinin başarımları karşılaştırılmıştır. Ekzon bölgeleri noktalı dikdörtgen şeklinde belirtilmiştir. Dikdörtgen içinde kalan ve 0.6 eşik değerini aşan pik işaretleri DNA sayısal sinyalinin ekzon bölgeleri olarak kabul edilir.

Önerilen haritalama tekniği, DNA diziliminin belirli aralığındaki 6 ekzon bölgelerinin taranmasında diğer sayısal haritalama tekniklerine göre daha iyi bir başarımlı göstermektedir. Hedeflenen bu 6 ekzon bölgeleri AJ229040 genindeki 3770-3826, 4584-4601, 4671-4730, 4999-5277, 5730-5823 ve 6719-6898 baz aralıklarıdır.



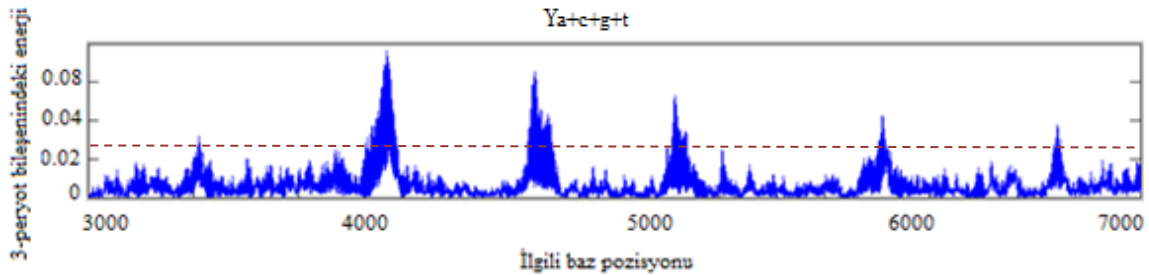
**Şekil 5.6** Ekzon bölgelerinin taranmasında farklı sayısal haritalama tekniklerinin karşılaştırılması

Şekil 5.6'da ekzon bölgeleri noktalı dikdörtgenlerle işaretlenmiştir. Şekil 5.6(f) grafiğine dikkatli bakıldığında DNA işaretinin noktalı dikdörtgen ile belirtilen 6 alan içinde yüksek oranda pik yapmaktadır. Önerilen sayısal haritalama tekniğinin 6 ekzon bölgesini de doğru

tahmin ettiği görülmektedir. Şekil 5.6 (a), (b), (c), (d), (e) deki diğer grafiklere bakıldığında işaretin noktalı dikdörtgen alanları içinde pik yapmadığı yani ekzon bölgelerinin yanlış tespit edildiği veya tam olarak 6 alanda da tespit edilmediği görülmektedir. Tamsayı haritalama tekniği ekzon bölgelerinin taranmasında ikinci en yüksek başarıma sahip olup, 4 alanda ekzon bölgesini doğru tahmin etmiştir.

#### 5.4.2 Ayırık Fourier Dönüşümü Yönteminin Sınıflandırma Sonuçları

Genomik uygulamalarda 3 periyotlu temel ekzon (kodon) bölgelerinin belirlenmesi zordur. Bu ekzon tarama uygulamasında Fourier tabanlı filtreleme yaklaşımı kullanılarak ekzonların baz pozisyonları tahmin edilmiştir. İlk olarak sembolik DNA dizilimleri önerilen sayısal haritalama tekniği ile sayısal işaretlere dönüştürülmüştür. Daha sonra Fourier dönüşümü yöntemi ile ekzon pozisyonları tahmin edilmiştir. Tahmin edilen pozisyonlar, gen bankasından alınan DNA dizilimlerindeki ekzonların gerçek pozisyonlarıyla karşılaştırılmış ve doğruluk oranları belirlenmiştir. Şekil 5.7 uygulama sonuçlarındaki ekzon pozisyonlarını ifade eden güç spektrumlarının pik işaretlerini göstermektedir. Literatür dikkate alınarak filtrelenmiş DNA diziliminin 99.persentil değeri eşik değeri olarak belirlenmiştir. Şekil 5.7'e göre 3770-3826, 4584-4601, 4671-4730, 4999-5277, 5730-5823 ve 6719-6898 aralıklarındaki baz pozisyonları 0.0275 eşik değerini aşmıştır. Ama yine de 3770-3826 aralığındaki ekzon tahmini zayıf kalmıştır.



Şekil 5.7 Baz pozisyonlarına göre güç spektrumunu ifade eden ekzonlar

Tablo 5.1'da gen bankasından elde edilen DNA dizilimlerinin gerçek ekzon pozisyonları, önerilen yöntem ile elde edilen ekzon pozisyonları ile karşılaştırılmıştır. Her ekzon pozisyonu için yöntemin doğruluk oranı ve ortalama başarıma yaklaşık olarak %96'dır.

**Tablo 5.1** Gen bankasından elde edilen DNA dizilimindeki gerçek ekzon pozisyonları ile yöntem sonucu tahmin edilen ekzon pozisyonlarının karşılaştırılması

| NCBI Veritabanı<br>Baz Pozisyonları(bp) | AFD Yöntemi        | Ekzon Bölgeleri<br>Doğruluk Oranı (%) |
|-----------------------------------------|--------------------|---------------------------------------|
| 867-1065 (198 bp)                       | 867-1065 (164 bp)  | 82,85                                 |
| 2505-2915 (410 bp)                      | 2505-2915 (410 bp) | 100                                   |
| 4038-4365 (328 bp)                      | 4038-4365 (328 bp) | 100                                   |
| 5404-5690 (287 bp)                      | 5404-5690 (287 bp) | 100                                   |
| 7376-7638 (262 bp)                      | 7376-7638 (256 bp) | 97,72                                 |

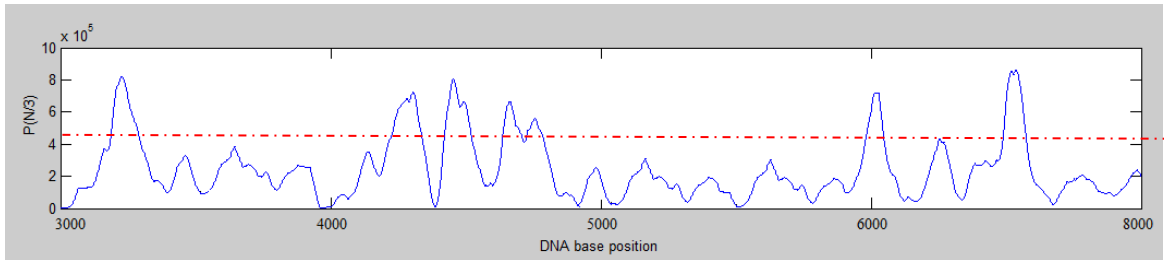
Tablo 5.2’de görüldüğü gibi, önerilen sayısal haritalama tekniğinin ekzon bölgelerinin belirlenmesindeki doğruluk oranları, mevcut diğer sayısal haritalama tekniklerinin başarımlarıyla karşılaştırılmıştır.

**Tablo 5.2** Eşik değere göre sayısal haritalama tekniklerinin doğruluk oranlarının karşılaştırılması

| Haritalama Tekniği            | Eşik değeri | Ekzon Bölgeleri Doğruluk Oranı (%) |
|-------------------------------|-------------|------------------------------------|
| Tamsayı Tekniği               | 0.0463      | 86,125                             |
| EIIP Tekniği                  | 0.000142    | 77,53                              |
| DNA-Walk Tekniği              | 0.0305      | 90,82                              |
| Eşleştirilmiş Sayısal Tekniği | 0.0290      | 87,79                              |
| Önerilen Haritalama Tekniği   | 0.0275      | <b>96,114</b>                      |

#### 5.4.3 Kısa Zamanlı Fourier Dönüşümü Yönteminin Sınıflandırma Sonuçları

DNA dizilimi önerilen sayısal haritalama tekniği ile sayısal bir işarete dönüştürülmüştür. Elde edilen sayısal işaretin ekzon bölgeleri KZFD yöntemi ile tespit edilmiştir. Bu uygulama için Hamming pencereleme fonksiyonu kullanılmıştır. 3770–3826, 4584–4601, 4671–4730, 4999–5277, 5730–5823 ve 6719–6898 aralıklarındaki baz pozisyonları  $4.972 \times 10^5$  olan eşik değerini aşmıştır. Ekzon bölgelerini ifade eden işaretin 6 tepesi de eşik değerini aşmıştır.



**Şekil 5.8** KZFD yönteminde baz pozisyonlarına göre ekzon bölgeleri

Şekil 5.8’de görüldüğü gibi KZFD yönteminde önerilen sayısal haritalama tekniği iyi bir başarımla göstermiştir. Bu sonuçlar ortalama başarımlar bakımından, var olan diğer sayısal haritalama tekniklerinin başarımlarıyla karşılaştırılmıştır ve karşılaştırma sonuçları Tablo 5.3’ de gösterilmiştir.

**Tablo 5.3** KZFD yönteminde ortalama başarıya göre sayısal haritalama tekniklerinin başarımlarının karşılaştırılması

| Haritalama Tekniği            | Eşik değeri          | Ekzon Bölgeleri Doğruluk Oranı (%) |
|-------------------------------|----------------------|------------------------------------|
| Tamsayı Tekniği               | $4,583 \times 10^5$  | 90,21                              |
| EIIP Tekniği                  | $0.1106 \times 10^5$ | 74,19                              |
| DNA-Walk Tekniği              | $7,81 \times 10^5$   | 80,495                             |
| Eşleştirilmiş Sayısal Tekniği | $7,8517 \times 10^5$ | 85,561                             |
| Önerilen Haritalama Tekniği   | $4,972 \times 10^5$  | <b>92,037</b>                      |

Tablo 5.3’de önerilen sayısal haritalama tekniği %92.037 ile en yüksek başarımla sahiptir.

## 5.5 Uygulama Sonuçlarının Değerlendirilmesi

Bu uygulamada DNA dizilimindeki protein kodlayan bölgelerin (ekzon) tespiti için entropiye dayalı yeni bir sayısal haritalama tekniği önerilmiştir. Bu teknik, dizilimdeki her bir kodunu Shannon denkleminin geliştirilmiş fraksiyonel türevi ve uyarlamalı parametreleri ile sayısallaştırmıştır. Önerilen sayısal haritalama teknik 6 ekzon bölgesinin tamamını doğru bulmada büyük bir başarı sağlamıştır. Diğer haritalama tekniklerinin DNA dizilimlerinin ekzon bölgelerini bulmadaki başarımları daha düşük kalmıştır. Önerilen haritalama tekniğinin başarımları TDA, AFD ve KZFD yöntemlerinde test edilmiştir. Diğer haritalama tekniklerinin başarımlarının üstünde bir başarı elde edilmiştir. Önerilen tekniğin AFD ve KZFD yöntemlerindeki başarımları sırayla %96.116, %92.037. Önerilen tekniğin yüksek başarımla göstermesinin sebebi ekzon bölgelerinin ayırım oranlarını derinleştirmesidir. Önerilen teknik DNA diziliminin karmaşık yapısını daha iyi yansıtmakta ve sayısallaştırmayı kodonların tekrarlama sıklığına göre yapmaktadır. Ayrıca bu haritalama tekniği gen dizisi üzerinde kodonların tekrarlama sıklık değerlerine göre geniş bir korelasyon bilgisi aralığı sunmaktadır. Önerilen haritalama yaklaşımı yalnızca ekzon ve intron bölgelerinin tanımlanmasıyla değil aynı zamanda pek çok uygulamada da uygulanabilir.

## 6 DNA DİZİLİMLERİNDEN BETA TALASEMİ HASTALIĞININ TANILANMASI

Mendel, genetik biliminin babası olarak bilinir. Genetik özelliklerin daha sonraki nesillere miras kaldığını keşfetmiştir. İnsan genomu yaklaşık 3 milyar baz çifti ve yaklaşık 35.000 gen içerir. Genler karmaşık bir DNA makro molekülünde kalıtsal taşıyıcıdır [143-144]. Genlerde meydana gelen mutasyonlar sonucunda bazı genetik hastalıklar tespit edilir. Beta talasemi de bu genetik hastalıklardan biridir.

Beta talasemi (Akdeniz Anemisi), ebeveynlerden çocuklara miras kalan önlenebilir bir kan bozukluğudur. Akdeniz ülkelerinde önemli bir halk sağlığı sorunudur. Taşıyıcıların tanımlanması, genetik danışmanlık ve prenatal tanı ile önlenebilir bir hastalık olmasına rağmen, talasemisi olan en az 365,000 hasta dünyaya her yıl doğar ve tedaviye ihtiyaç duyar [145-146].

Beta talasemi hastalığı, yaşam süresini önemli ölçüde kısaltan ve tedavinin gerektiği gibi yapılmaması durumunda yaşam kalitesini olumsuz yönde etkileyen ciddi bir hastalıktır. Hastalığın tedavisi zor ve maliyetlidir.

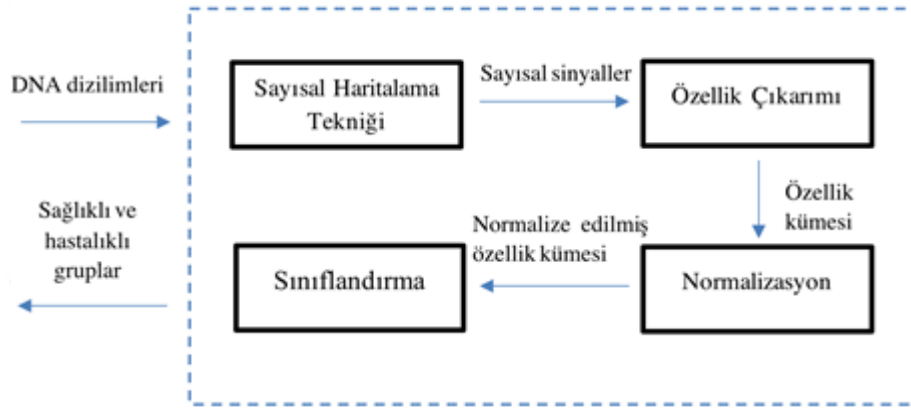
Beta talasemi, insan DNA diziliminin HBB genindeki bazı mutasyonların bir sonucu olarak ortaya çıkar. HBB genindeki mutasyonlar beta-globin üretimini azaltır veya sentezi durdurur. En sık görülen tipler alfa ( $\alpha$ ) ve beta ( $\beta$ ) talasemidir. Bugün beta-globin geninde 200'ün üzerinde mutasyon olduğu belirtilmektedir. Beta talasemi mutasyonları oldukça çeşitlidir ve bu hastalık dünyanın farklı bölgelerine yayılmıştır. HBB genindeki bazı mutasyonlar *IVS 1.110 (G> A)*, *IVS 2.745 (C> G)*, *c-31C> T*, *E7V (c.20A> T)*'dir. Beta talasemi majör, bir mutant beta-globin aleli için heterozigot veya homozigot olan bireylerde görülür. Hastalığın hafif bir formu olan beta talasemi minör bireyler heterozigot mutant allellere sahiptir [145-147].

Tezin bu bölümünde Ankara Intergen Genetik Hastalıklar Tanı Merkezinden alınan gerçek Beta talasemi hastaları ile sağlıklı bireylere ait DNA dizilimlerinden Beta talasemi hastalığının tanılanmasına yönelik örüntü analizi tekniğine dayalı bir uygulama süreci geliştirilmiştir. Burada, literatürde en çok kullanılan ve en yüksek başarıma sahip haritalama teknikleriyle önerilen sayısal haritalama tekniğinin Beta talasemi hastalığının tanılanması üzerindeki yorumlama başarımı test edilmektedir. Bu çalışma, Fırat Üniversitesi Girişimsel Olmayan Araştırmalar Etik Kurulu Başkanlığının (tarih: 31.05.2018, sayı:10/16 ) izni ile gerçekleştirilmiştir.

## 6.1 Materyal ve Yöntem

DNA dizilimlerinden Beta talasemi hastalığının tanılanmasına yönelik geliştirilen uygulama sürecinin blok diyagramı Şekil 6.1’de gösterilmektedir. Bu yapıda örüntü analizi tekniğinin özellik çıkarma kısmını sayısal haritalama tekniği oluştururken, sınıflandırma kısmını ise Yapay Sinir Ağı (YSA) oluşturmaktadır.

Önerilen yaklaşımda, veri setini hem Beta talasemi hastalarının hem de sağlıklı bireylerin DNA dizilimleri oluşturmaktadır. Özellik çıkarma sürecinde ilk olarak, alınan sembolik DNA dizileri, üç sayısal haritalama tekniği ile sayısal dizilere dönüştürülür. Akabinde, özellik çıkarma işlemi, DNA sayısal işaretlerinden istatistiksel teknikler kullanılarak özellik seti hesaplanıp normalizasyon işlemine tabi tutulur. Son aşamada ise, haritalama tekniklerinin başarımı YSA sınıflandırıcısı üzerinden karşılaştırılmaktadır.



Şekil 6.1 Önerilen yaklaşımın blok diyagramı

### 6.1.1 Veri Kaynağı

Sınıflandırma uygulamasında kullanılan veriler, Türkiye’de Intergen Genetik Hastalıklar Tanı Merkezi’nden elde edilmiştir. Beta talasemi hastalığının örüntü analizi için 53 HBB gen dizilimi kullanıldı. Bunlardan 20’si hastalıklı, 33’ü sağlıklı bireylere ait olmak üzere toplam 53 HBB gen dizilimi kullanıldı. Şekil 6.2’ de HBB gen diziliminin bir bölümü gösterilmektedir.

ATAATATCCCCAGTTTAGTAGTTGGACTTAGGGAACAAAGGAACCTTTAATAGAAATTGGACAGCAA  
GAAAGCGAGCTTAGTGATACTTGTGGGCCAGGGCATTAGCCACACCAGCCACCCTTTCTGATAGGCA  
GCCTGCACTGGTGGGTGAATTCTTTGCCAAAGTGATGGGCCAGCACACAGACCAGCACGTTGCCAG  
GAGCTGTGGGAGGAAGATAAGAGGTATGAACATGATTAGCAAAAGGGCCTAGCTTGGACTCAGAATAA  
TCCAGCCTTATCCCAACCATAAAATAAAAGCAGAATGGTAGCTGGATTGTAGCTGCTATTAGCAATAT  
GAAACCTCTTACATCAGTTACAATTTATATGCAGAAATATTTATATGCAGAAATATTGCTATTGCCTT  
AACCCAGAAATTATCACTGTTATTCTTTAGAATGGTGCAAAGAGGCATGATACATTGTATCATTATTG  
CCCTGAAAGAAAGAGATTAGGGAAAGTATTAGAAATAAGATAAAACAAAAAAGTATATTTAAAGAAAGAA  
AGCATTTTTTAAATTACAAATGCAAATTACCCTGATTTGGTCAATATGTGTACACATATTTAAACA

Şekil 6.2 HBB gen diziliminden bir kesit

### 6.1.2 Gen Diziliminden Özellik Çıkarımı

DNA dizilimlerinin özellik çıkarma Özelliği çıkartımı işlemi öncesinde sembolik DNA dizilimleri 5. Bölümde önerilen entropi tabanlı sayısal haritalama tekniği [113] ile literatürde en çok tercih edilen Integer ve EIIP teknikleri [148] kullanılarak sayısal işaretlere dönüştürülmüştür. Akabinde, bu sayısal işaretlerden aşağıda belirtilen bilgi ölçümü ile istatistiksel teknikler [149-151] kullanılarak özellikler hesaplanmıştır.

**Entropi:** Bir işaretin düzensizlik derecesini ölçmek için kullanılır [152]. Entropi, Denklem 6.1'de [153,154] gösterildiği gibi hesaplanır.

$$H(p) = -\sum_{i=1}^M p_i \log p_i \quad (6.1)$$

**Standart Sapma:** Verilerin ortalamaya göre yayılmasının bir ölçüsüdür. Verilerin standart sapma ile nasıl dağıldığı ve ortalamadan ne kadar uzakta olduğu öğrenilmiştir [155,156]. Standart sapma değeri Denklem 6.2 ile hesaplanır.

$$\text{Standart Sapma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x - x_i)^2} \quad (6.2)$$

**Ortalama Mutlak Değer:** Ortalama mutlak değer, işaret genliğinin mutlak değerlerinin ortalamasıdır [157-159]. Denklem 6.3 ile hesaplanır.

$$\text{Ortalama Mutlak Değer} = \frac{1}{N} = \sum_{i=1}^N |X_i| \quad (6.3)$$

**Basit Kare İntegrali:** İşaretin enerjisini belirlemek için kullanılır [158,159]. Denklem 6.4 ile hesaplanır.

$$\text{Basit Kare Integral} = \sum_{i=1}^N |X_i|^2 \quad (6.4)$$

**Dalga Boyu:** Dalga boyu, zaman dilimi boyunca dalganın kümülatif uzunluğudur [159-161]. Dalga boyu genlik, frekans ve zaman ile ilgilidir. Denklem 6.5 ile hesaplanır.

$$\text{Dalgaboyu} = \sum_{i=1}^{N-1} |X_{i+1} - X_i| \quad (6.5)$$

Yukardaki hesaplamalar sonucunda elde edilen özelliklerin, etkin ve doğru bir sınıflandırmada kullanılabilmesi için normalizasyona ihtiyaç duyulmuştur. YSA'da normalizasyon işlemi oldukça önemlidir [162,163]. Bunun için Denklem 6.6 kullanılarak hesaplanan değerler normalizasyona tabi tutularak ölçeklendirilmiş özellik vektörü elde edilmiştir.

$$X' = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \quad (6.6)$$

Tablo 6.1, 6.2, 6.3, bazı hasta ve sağlıklı gruplar için haritalama tekniklerine göre çıkarılan özelliklerin değerlerini göstermektedir.

**Tablo 6.1** EIP haritalama tekniğine göre normalizasyon sonucu özellik değerlerinden bazıları

| SN | Grup Tipi    | Entropi | Standart Sapma | Ortalama Mutlak Değer | Basit Kare İntegral | Dalga boyu |
|----|--------------|---------|----------------|-----------------------|---------------------|------------|
| 1  | Hastabt01    | 0       | 0.081          | 0.333                 | 0.160               | 0.810      |
| 2  | Hastabt02    | 0.666   | 0.363          | 0.666                 | 0.381               | 0.400      |
| 3  | Hastabt03    | 0.666   | 0.018          | 1.000                 | 0.588               | 0.212      |
| 4  | Hastabt04    | 1       | 0              | 1                     | 0.747               | 0.023      |
| 5  | Hastabt05    | 0       | 0.090          | 0                     | 0.215               | 0.855      |
| 6  | Sağlıklıbt01 | 0       | 0.818          | 0.333                 | 0.251               | 1          |
| 7  | Sağlıklıbt02 | 0.666   | 0.181          | 1                     | 0.748               | 0.265      |
| 8  | Sağlıklıbt03 | 0.333   | 0.636          | 0.333                 | 0.342               | 0.839      |
| 9  | Sağlıklıbt04 | 1       | 0              | 1                     | 0.971               | 0.013      |
| 10 | Sağlıklıbt05 | 0.666   | 0.363          | 0.666                 | 0.525               | 0.517      |

**Tablo 6.2** Integer haritalama tekniğine göre normalizasyon sonucu özellik değerlerinden bazıları

| SN | Grup Tipi    | Entropi | Standart Sapma | Ortalama Mutlak Değer | Basit Kare İntegral | Dalga boyu |
|----|--------------|---------|----------------|-----------------------|---------------------|------------|
| 1  | Hastabt01    | 0.880   | 0.044          | 0.327                 | 0.333               | 1          |
| 2  | Hastabt02    | 0.986   | 0.049          | 0.109                 | 0.185               | 0.400      |
| 3  | Hastabt03    | 0.230   | 0.011          | 0.812                 | 0.833               | 0.200      |
| 4  | Hastabt04    | 0.124   | 0.006          | 1                     | 0.981               | 0.600      |
| 5  | Hastabt05    | 1       | 0.050          | 0                     | 0                   | 0.800      |
| 6  | Sağlıklıbt01 | 0.828   | 0.615          | 0.066                 | 0.028               | 0.500      |
| 7  | Sağlıklıbt02 | 0.328   | 0.769          | 0.711                 | 0.778               | 0          |
| 8  | Sağlıklıbt03 | 0.167   | 0.385          | 0.855                 | 0.889               | 0.167      |
| 9  | Sağlıklıbt04 | 0.661   | 0.692          | 0.276                 | 0.278               | 1          |
| 10 | Sağlıklıbt05 | 0.005   | 0              | 1                     | 1                   | 0.500      |

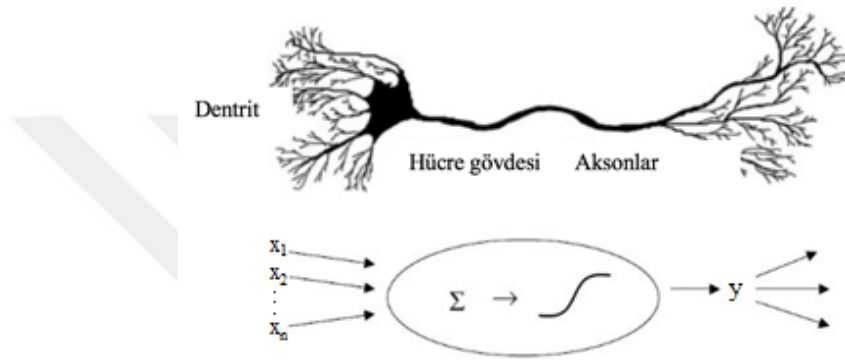
**Tablo 6.3** Önerilen sayısal haritalama tekniğine göre normalizasyon sonucu özellik değerlerinden bazıları

| SN | Grup Tipi    | Entropi | Standart Sapma | Ortalama Mutlak Değer | Basit Kare İntegral | Dalga boyu |
|----|--------------|---------|----------------|-----------------------|---------------------|------------|
| 1  | Hastabt01    | 0.034   | 0.144          | 0.044                 | 0.237               | 0.069      |
| 2  | Hastabt02    | 0.651   | 1              | 0.502                 | 0.386               | 0.034      |
| 3  | Hastabt03    | 0.074   | 0              | 1                     | 0.500               | 0.428      |
| 4  | Hastabt04    | 0.689   | 0.049          | 0.012                 | 1                   | 0.662      |
| 5  | Hastabt05    | 0       | 0.788          | 0.275                 | 0.035               | 0.255      |
| 6  | Sağlıklıbt01 | 0.046   | 0.188          | 0.157                 | 0.192               | 0          |
| 7  | Sağlıklıbt02 | 0.999   | 0              | 0                     | 0.263               | 0.199      |
| 8  | Sağlıklıbt03 | 0.709   | 1              | 1                     | 0.911               | 0.034      |
| 9  | Sağlıklıbt04 | 0       | 0.290          | 0.297                 | 0.659               | 0.210      |
| 10 | Sağlıklıbt05 | 1       | 0.054          | 0.043                 | 1                   | 0.641      |

### 6.1.3 Yapay Sinir Ağı ile Sınıflandırma

YSA, biyolojik sinir ağlarının yapı ve işlevlerine dayanan bir hesaplama modelidir. Girişe ve çıkışa dayalı bir sinir ağı olduğu için ağ üzerinden akan bilgiler YSA yapısını etkiler. YSA' lar, girdiler ve çıktılar arasındaki karmaşık ilişkilerin modellendiği veya kalıpların bulunduğu doğrusal olmayan istatistiksel veri modelleme araçları olarak kabul edilir [164,165]. Bir YSA, yapay nöronlar olarak adlandırılan bağlı ünitelerin veya düğümlerin bir koleksiyonuna dayanır. Yapay nöronlar canlı beyindeki biyolojik nöronların basitleştirilmiş bir versiyonu olarak kabul edilir. Yapay nöronlar arasındaki her bağlantı bir işareti diğerine iletir. İşareti alan yapay nöron onu işler ve ona bağlı yapay

nöronlara etki eder. Genel YSA uygulamalarında, yapay nöronlar arasındaki bir bağlantıdaki işaret gerçek bir sayıdır ve her bir yapay nöronun çıkışı, girdilerinin toplamının doğrusal olmayan bir fonksiyonu ile hesaplanır. Yapay nöronlar ve bağlantılar tipik olarak öğrenme ilerledikçe ayarlanan bir ağırlık parametresine sahiptir. Ağırlıklar, bir bağlantıdaki işaretin gücünü artırır veya azaltır. Tipik olarak, yapay nöronlar katmanlarda organize edilir. Farklı katmanlar, girdilerinde farklı türden dönüşümler yapabilir. İşaretler, katmanları geçtikten sonra, ilk (giriş) katmanından son (çıkış) katmanına akarlar. Şekil 6.3'te bir yapay sinir ağı modeli gösterilmektedir [166,167].



Şekil 6.3 Bir YSA yapısı

Çalışmada ileri beslemeli YSA yapısı kullanılmış olup, ağırlık öğrenmesi için danışmanlı öğrenme algoritmalarından birisi olan Levenberg-Marquardt (LM) öğrenme algoritması tercih edilmiştir. LM algoritması maksimum komşuluk fikri üzerine kurulmuş en az kareler hesaplama metodudur [168,169]. LM, Gauss-Newton ve Gradyant Azalan algoritmalarının en iyi özelliklerini kullanır. Geriye yayılım algoritmaları arasında en hızlı olanıdır. Problemin türüne uygun olması, optimum sonuca en hızlı ulaşan algoritma olması ve öğrenme hatasının düşük olması öğrenme algoritmasının tercih edilme sebepleridir. Uygulamada kullanılan YSA modelinin yapısı ve parametreleri Tablo 6.4' de gösterilmiştir.

Tablo 6.4 Kullanılan YSA modelinin yapısı ve eğitim parametreleri

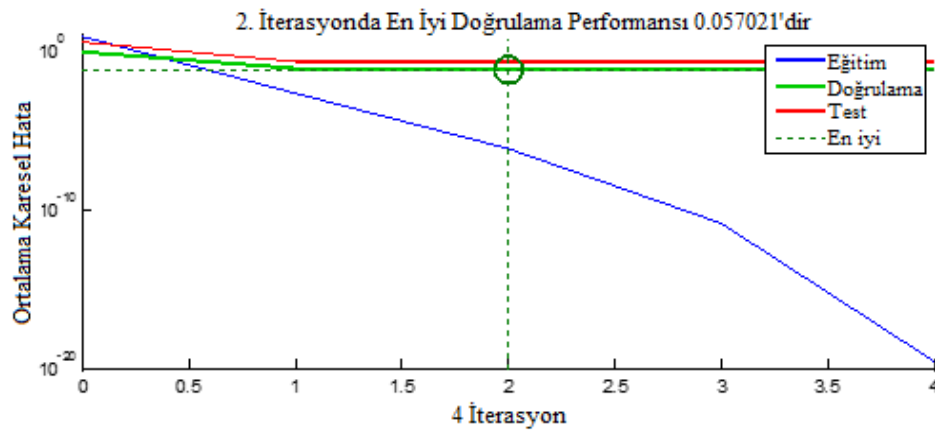
| YSA Yapısı                 |                                                                                                        |
|----------------------------|--------------------------------------------------------------------------------------------------------|
| YSA modeli                 | Çok katmanlı ileri beslemeli                                                                           |
| Katman sayısı              | 3                                                                                                      |
| Katmanlardaki nöron sayısı | Giriş: Özellik vektörünün bileşen sayısı<br>Gizli: Özellik çıkarım yöntemine göre değişken<br>Çıkış: 1 |
| Başlangıç ağırlıkları      | Rasgele                                                                                                |
| Aktivasyon Fonksiyonları   | Logaritmik Sigmoid                                                                                     |

| YSA Eğitim Parametreleri        |                                              |
|---------------------------------|----------------------------------------------|
| Öğrenme Algoritması             | Levenberg-Marquardt geri yayılım algoritması |
| $\lambda$ Marquardt parametresi | 0.001                                        |
| $\lambda$ Artış oranı           | 10                                           |
| $\lambda$ Azalış oranı          | 0.1                                          |

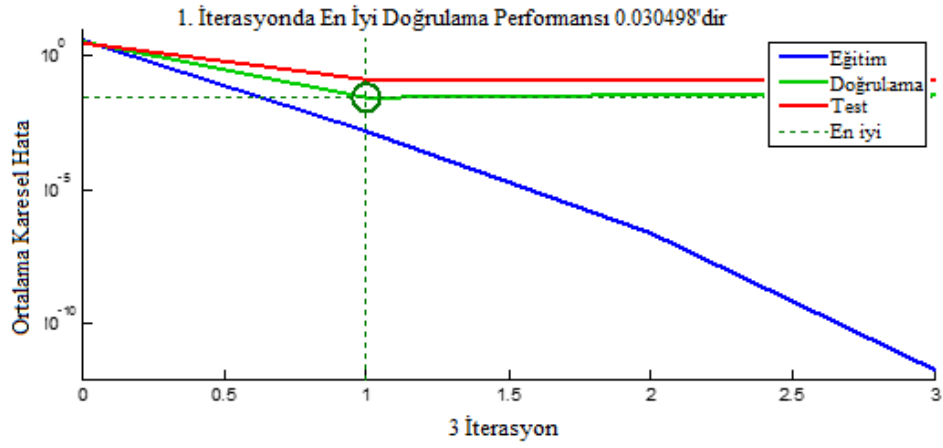
YSA ile sınıflandırma işlemi için 20 Beta talasemi hastası ve 33 sağlıklı kişiden oluşan veri seti çaprazlama doğruluk testine göre 3 gruba ayrılmıştır. Buna göre 1. Grup: 7 hasta-11 sağlıklı, II.Grup: 7 hasta-11 sağlıklı, III.Grup: 6 hasta-11 sağlıklı şeklinde ayrıştırılarak 3 katlı çaprazlama doğruluk tekniğine göre YSA’da eğitim ve test amaçlıyla giriş ve test veri seti olarak kullanılmıştır. Böylece veri setinin tamamı hem test hemde eğitim amaçlı kullanılmıştır.

## 6.2 Sınıflandırma Sonuçlarının Yorumlanması

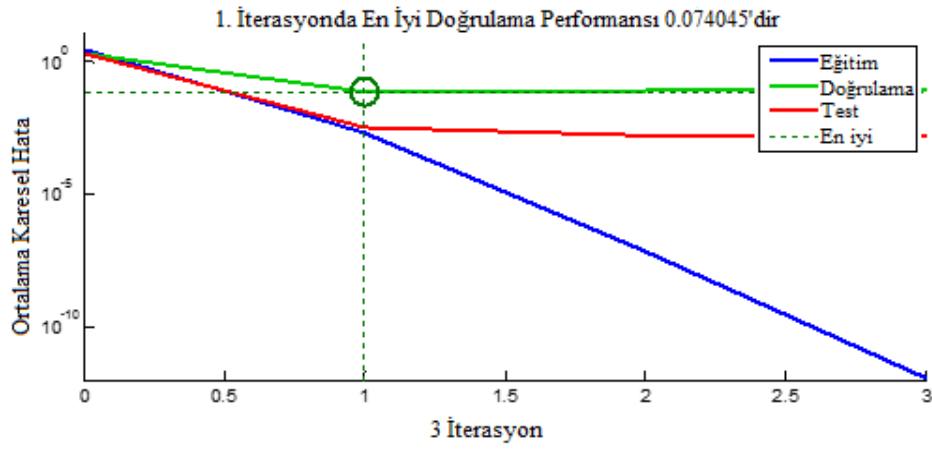
Sınıflandırmada önerilen haritalama tekniğinin başarımını göstermek üzere, entropi tabanlı sayısal haritalama tekniği ve literatürde sınıflandırmada başarımı yüksek olan EIIP ve Integer haritalama teknikleri HBB gen veri setlerine uygulandı ve yorumlama başarımları karşılaştırıldı. Şekil 6.4, 3 farklı eğitim ve test grubunda EIIP haritalama tekniğinin sınıflandırma başarımlarını göstermektedir. EIIP tekniği ortalama % 86,80 doğruluk başarımına sahiptir.



a)Eğitim Test Eğitim



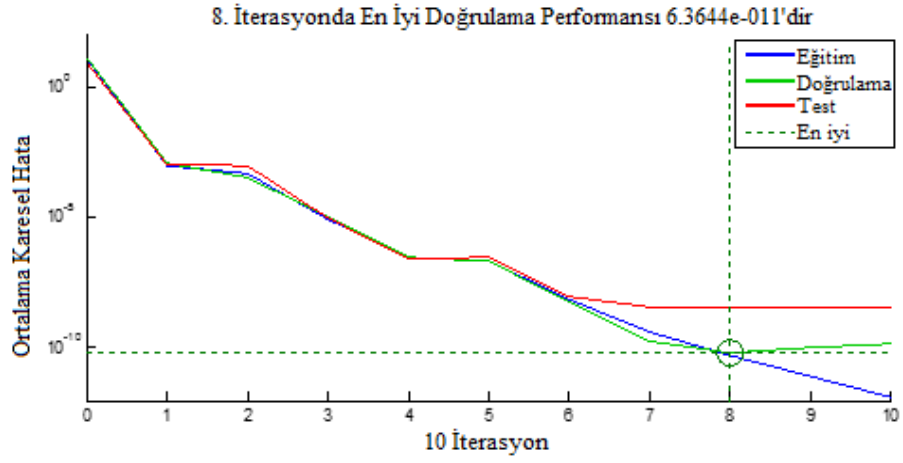
b)Eğitim Eğitim Test



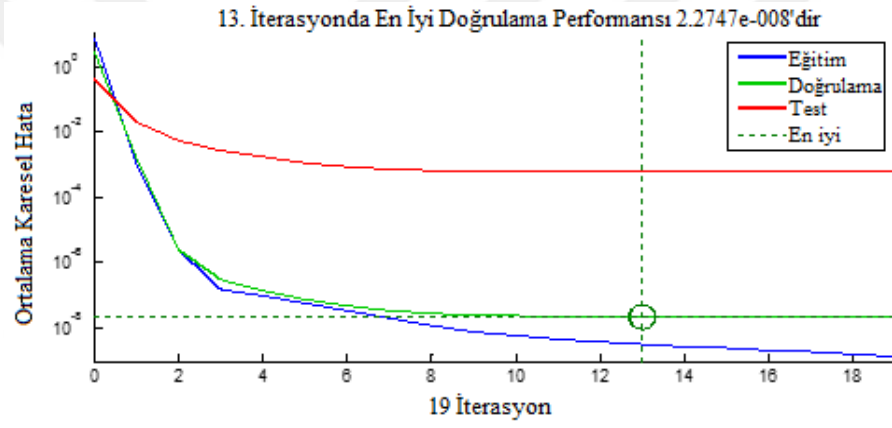
c)Test Eğitim Eğitim

Şekil 6.4 EIIP haritalama tekniğinin 3 katlı çaprazlama doğrulamasına göre sınıflandırma başarımı

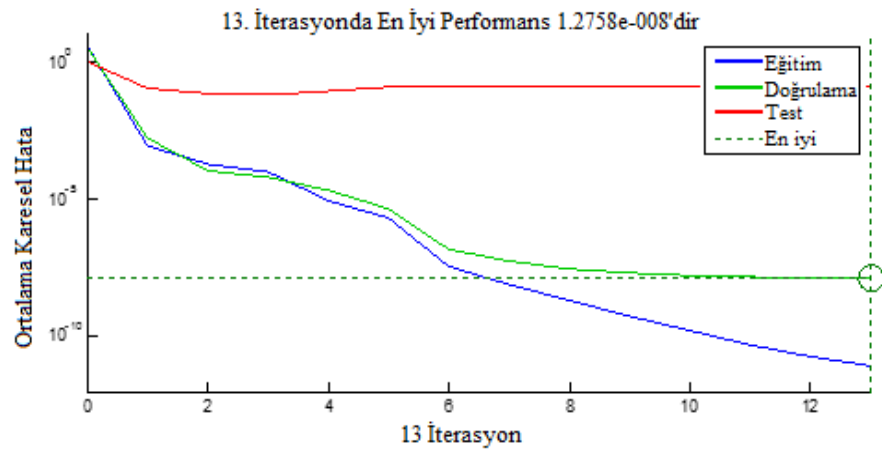
Şekil 6.5, 3 farklı eğitim ve test grupları üzerinde Integer haritalama tekniğinin sınıflandırma başarımlarını göstermektedir. Integer tekniği ortalama % 85,83 doğruluk başarımına sahiptir.



a) Eğitim-Eğitim-Test veri seti



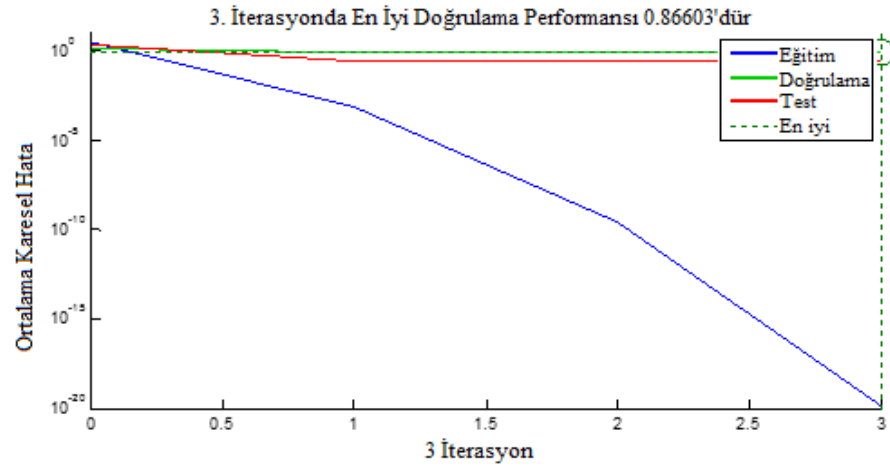
b) Test-Eğitim-Eğitim



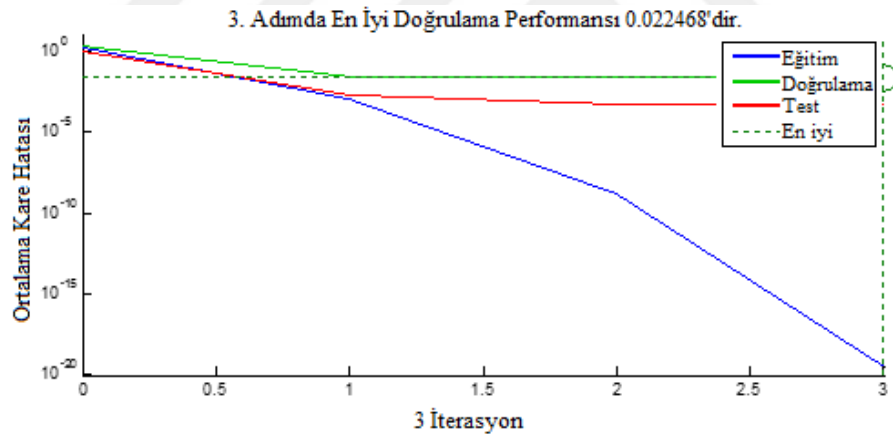
c) Eğitim-Test-Eğitim

Şekil 6.5 Integer haritalama tekniğinin 3 katlı çaprazlama doğrulamasına göre sınıflandırma başarımı

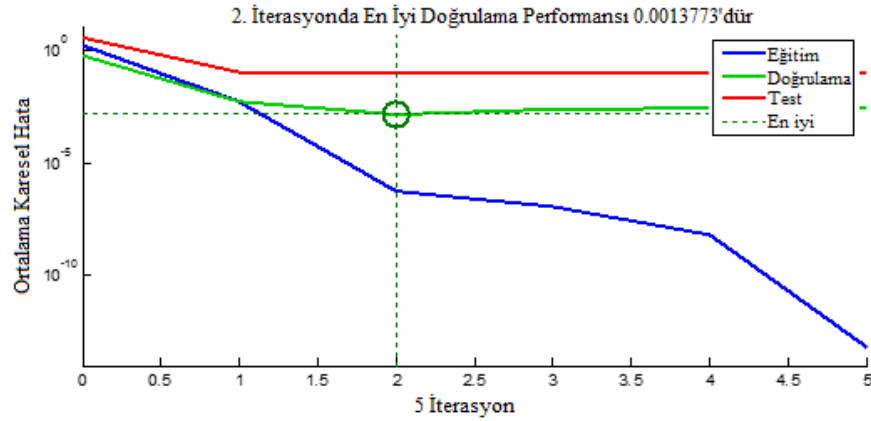
Şekil 6.6'da, 3 farklı eğitim ve test grubunda önerilen entropi haritalama tekniğinin sınıflandırma başarımlarını gösterir. Entropiye dayalı haritalama tekniği ortalama %96,49 doğruluk başarımına sahiptir.



a)Eğitim-Eğitim-Test



b)Test-Eğitim-Eğitim



c)Eğitim Test Eğitim

Şekil 6.1 Önerilen Entropi tabanlı haritalama tekniğinin 3 katlı çaprazlama doğrulamasına göre sınıflandırma başarımı

Tablo 6.5, Üç haritalama tekniğine ait hassasiyet, özgülük ve doğruluk değerlerini göstermektedir. Önerilen haritalama tekniği,% 96.49 ile en yüksek doğruluk sınıflandırma başarımlarına sahiptir.

Tablo 6.5 Üç haritalama tekniğine göre sınıflandırma sonuçları

|                        | Veri Seti                       |         |           | Duyarlılık(%) | Özgüllük(%) | Doğruluk(%)  |
|------------------------|---------------------------------|---------|-----------|---------------|-------------|--------------|
|                        | I.Grup                          | II.Grup | III. Grup |               |             |              |
| EHP Tekniği            | Eğitim                          | Eğitim  | Test      | 54.55         | 100         | 72.22        |
|                        | Test                            | Eğitim  | Eğitim    | 88.89         | 100         | 93.75        |
|                        | Eğitim                          | Test    | Eğitim    | 90            | 100         | 94.44        |
|                        | Ortalama Sınıflandırma Başarımı |         |           |               |             | 86.80        |
| Tamsayı Tekniği        | Eğitim                          | Eğitim  | Test      | 71.43         | 100         | 80           |
|                        | Test                            | Eğitim  | Eğitim    | 57.14         | 100         | 81.25        |
|                        | Eğitim                          | Test    | Eğitim    | 85.71         | 100         | 94.74        |
|                        | Ortalama Sınıflandırma Başarımı |         |           |               |             | 85.33        |
| Entropi Tabanlı Teknik | Eğitim                          | Eğitim  | Test      | 100           | 100         | 100          |
|                        | Test                            | Eğitim  | Eğitim    | 100           | 100         | 100          |
|                        | Eğitim                          | Test    | Eğitim    | 83.33         | 100         | 89.47        |
|                        | Ortalama Sınıflandırma Başarımı |         |           |               |             | <b>96.49</b> |

Bu çalışma, önerilen entropi tabanlı haritalama tekniği ile sayısallaştırılmış DNA dizilerinden elde edilen özellik kümesinin YSA ile sınıflandırmasını içeren, Beta talasemi hastalığını tanımlamaya yönelik bir uygulama süreci sunmaktadır. Örüntü analizi çalışmalarında doğru sınıflandırmanın başarımları anahtar özelliklerin seçiminden

geçmektedir. Bu çalışmada önerilen haritalama tekniği ile elde edilen sayısal diziye, çeşitli istatistiksel ve bilgi ölçüm teknikleri uygulanmış olup bu şekilde elde edilen özellikler doğru sınıflandırma başarımını arttırmak üzere normalizasyon işlemine tabi tutulmuştur. Yapay Sinir Ağı ile sınıflandırma sürecinde özellik veri seti, 3 katlı çapraz doğrulama yöntemine göre gruplandırılarak kullanıldı. DNA dizilimlerinin, önerilen entropi tabanlı sayısal haritalama tekniği ile birlikte 3 farklı sayısal gösterimi üzerine yapılan uygulamalar sonucunda, önerilen entropi tabanlı haritalama tekniğinin % 96.49 ile en yüksek Beta talasemi hastalığı yorumlama başarımına sahip olduğu görülmektedir.



## 7 SONUÇ

Bu tez çalışmasında, DNA dizilimlerinden gen yapısal bozukluklarının ortaya çıkarılması, genetik hastalıklarının tanınması veya sınıflandırılmasına yönelik yeni yaklaşımlar geliştirilmiştir. Tezin hedefleri doğrultusunda önerilen yaklaşımlar:

- a) DNA dizilimlerinin sayısallaştırılması için, entropi tabanlı sayısal haritalama tekniğinin geliştirilmesi,
- b) Beta talasemi hastalığı için DNA dizilimlerinin sayısal işaretlerinden anahtar özelliklerin elde edilmesine yönelik olarak işaret işleme, bilgi ölçümü ve istatistiksel tekniklere dayalı yöntemlerin kullanılması ile bu özelliklerin Yapay Sinir Ağlarında sınıflandırılması sonucunda Beta talasemi hastalığının tanınması için örüntü analizi ve yapay zeka temelli bir uygulama sürecinin geliştirilmesi,

Yukarıdaki yaklaşımlar çeşitli uygulamalarla test edilmiş ve bir takım sonuçlar elde edilmiştir.

### 7.1 Sonuçların İrdelenmesi

- a) Literatürde en çok tercih edilen NCBI[141], Ensembl[115] gen bankalarından elde edilen DNA dizilimlerinin sayısal işaretlere dönüştürülmesi için çeşitli sayısal haritalama teknikleri bulunmaktadır[55-75]. Tezin hedefleri doğrultusunda geliştirilen yöntem öncesi varolan sayısal haritalama teknikleri üzerinde, başarımlarını değerlendirmesine yönelik bir uygulama gerçekleştirilmiştir [114]. Bu uygulamada DNA dizilimlerinden ekzon ve intron bölgelerinin sınıflandırılması için literatürdeki en çok kullanılan sayısal haritalama tekniklerinin başarımları Ayırık Fourier Dönüşümü, Destek Vektör Makinesi ve  $k$ -En Yakın Komşu algoritması yöntemlerinde karşılaştırılmış ve en yüksek başarımlara sahip sayısal haritalama teknikleri tespit edilmiştir. Uygulama sonucunda 4 farklı pencereleme fonksiyonuna göre Tamsayı Tekniği %96,2 sınıflandırma başarımları ile Gaussion pencere fonksiyonunda ve Eşleştirilmiş Sayısal Tekniği %95,58 sınıflandırma başarımları ile Hamming pencere fonksiyonunda en yüksek doğruluk oranlarına sahiptir. Ayrıca, DNA-Walk Tekniği %90,82 sınıflandırma başarımları ile Blackman pencere fonksiyonunda en yüksek doğruluk oranına sahiptir. DVM ve  $k$ -NN gibi makine öğrenme tabanlı yöntemlerin ekzon ve intron sınıflandırma başarımlarının düşük olduğu, buna rağmen Ayırık Fourier

Dönüşümü yönteminin ise diğer iki yönteme göre sınıflandırmada güçlü bir metot olduğu açıkça görülmüştür.

- b)** Önerilen birinci yaklaşımda DNA dizilimlerinin sayısallaştırılması için entropi tabanlı yeni bir sayısal haritalama tekniği geliştirilmiştir. Söz konusu tekniğin  $\alpha$  parametresi belirlenirken DNA diziliminin yapısından faydalanıldığı için geliştirilen bu yöntem Fiziko-kimyasal özellik tabanlı sayısal haritalama teknikleri grubunda olabileceği değerlendirilmektedir. Böylece Fiziko-kimyasal özellik tabanlı sayısal haritalama teknikleri kapsamında DNA dizilimlerinin sayısallaştırılmasına yönelik yeni bir tekniğin uygulanması önerilmiştir. Yapılan uygulamalarda bu tekniğin başarımının gerek ekzon ve intron bölgelerinin ayırımında, gerekse yapay zeka temelli hastalık tanılanmasında diğerlerinden daha yüksek doğrulukta bir yorumlama başarımına sahip olduğu gösterilmiştir. Bu sonuçlar neticesinde literatüre yeni bir sayısal haritalama tekniği kazandırılmıştır.
- c)** Önerilen entropi tabanlı sayısal haritalama tekniğinin başarımı literatürde en çok kullanılan diğer haritalama teknikleriyle karşılaştırılmıştır. DNA dizilimlerinden ekzon bölgelerinin tespitine yönelik bir uygulama gerçekleştirilmiştir [139]. Önerilen haritalama tekniğinin başarımı Tekli Değer Ayırımı (TDA), Ayrık Fourier Dönüşümü (AFD) ve Kısa Zamanlı Fourier Dönüşümü (KZFD) yöntemlerinde test edilmiştir. Diğer haritalama tekniklerinin başarımalarının üstünde bir başarı elde edilmiştir. Önerilen tekniğin AFD ve KZFD yöntemlerindeki başarı başarımları sırayla %96.116, % 92.037' dir.
- d)** Literatürde, DNA dizilimlerinden hastalık tespit edilmesine yönelik olarak kanser tanılanması için çalışmalar [27-29] mevcut iken, bu tezde ele alınan Beta talasemi hastalığının tanılanmasına yönelik bir çalışmaya rastlanılmamıştır. Tez çalışması ile bu boşluk doldurularak DNA dizilimlerinden Beta talasemi genetik hastalığın tanılanması için önerilen entropi tabanlı sayısal haritalama tekniği ile yapay zekâya dayalı bir örüntü tanıma analizi yaklaşımı önerilmiştir. Bu yaklaşımda % 96 tanılanma başarımı elde edilmiştir. Beta talasemi hastalığının tanılanma uygulamasında, doğru yorumlama başarımının yüksek olması önerilen entropi tabanlı sayısal haritalama tekniğinden kaynaklandığı değerlendirilmiştir. DNA dizilimlerindeki hastalık yapıcı kısımlarının doğru yorumlanmasına yönelik önerilen bu yaklaşım, DNA'nın yorumlanamayan bölgeleri içinde ışıktır.

## 7.2 Öneriler

- a) Önerilen hastalık tanılama yaklaşımı, DNA dizilimlerinden Beta talasemi genetik hastalığının tanılanması dışında farklı genetik hastalarının tanılanmasında da değerlendirilebilir.
- b) Beta talasemi hastalığının tanılanması için özellik çıkarma işleminde kullanılan istatistiksel ve bilgi ölçüm teknikleri dışında farklı özellik değerleri araştırılıp hastalık tanılama yönteminin etkinliği artırılabilir.
- c) HBB genini taşıyan DNA dizilimi örneği sayısı artırılarak, Akdeniz Anemisine yönelik bir tarama testi geliştirilebilir.
- d) Hasta ve sağlıklı kişilere ait özellik kümelerini sınıflandırmada YSA dışında başka bir yöntem kullanılabilir ve sınıflandırma yöntemlerinin başarımları kıyaslanabilir.
- e) Önerilen sayısal haritalama tekniğinin başarımı ekzon ve intron bölgelerinin bulunması, genetik hastalıkların sınıflandırılması dışında başka genetik araştırma konuları için de incelenebilir.
- f) Hesaplanan özellik veri setinin YSA sınıflandırıcısına verilmesinde  $k$  parçalı çapraz doğrulama tekniğinde  $k$  yerine farklı sayılar verilip sistemin performansı test edilebilir.
- g) Literatüre yeni kazandırılan entropi tabanlı sayısal haritalama tekniği ve yapay zekâya dayalı bir örüntü tanıma analizi yaklaşımı ile DNA dizilimlerindeki ekzonların dışında, promotor bölgesi, intron bölgelerinin ilk 3 baz aralıkları ve dizilimdeki diğer belirsizliklerin bulunduğu kısımlar incelenerek DNA ile ilgili yeni çıkarımlar elde edilebilir.

## 7.3 Yayınlar ve Proje

- a) Tez çalışmasının sonuçları, 2 adet uluslararası dergi makalesi (SCI-E) [113,114], 4 adet uluslararası konferans bildirisi [139,140,170,171] ve 1 adet ulusal konferans bildirisi [172] olmak üzere toplam 7 adet yayın ile değerlendirilmiştir.
- b) Tez çalışması kapsamında TÜBİTAK BİDEB 2214 proje desteği ile Kanada, Alberta Üniversitesi, Bilgisayar Bilimleri Bölümü, Biyoinformatik Laboratuvarında 05 Eylül 2017 - 1 Haziran 2018 tarihleri arasında uygulama çalışmaları yapılmıştır.

## KAYNAKLAR

- [1] M. K. Hota, V. K. Srivastava, "Identification of protein coding regions using antinotch filters", *Digital Signal Processing*, ss. 869-877, Vol. 22, 2012.
- [2] Q. Zhou, Q. Jiang, D. Wei, "A new method for classification in DNA sequence", *Computer Science & Education (ICCSE), IEEE 2011 6th International Conference on*, ss. 218-221, 3-5, 2011.
- [3] S. Sathish Kumar, M. E. Dr.N.Duraipandian, Ph.D, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, ss. 198-207, Vol. 5, 2008.
- [4] R. Damasevicius, "Analysis of binary feature mapping rules for promoter recognition in imbalanced DNA sequence datasets using Support Vector Machine", *Intelligent Systems '08. 4th International IEEE Conference (Volume:2)*, ss. 11-20 - 11-25, 2008.
- [5] W.F. Zhang H. Yan, "Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences", *Elsevier Science Inc. New York, NY, USA*, ss. 947-95, 2012.
- [6] S. D. Sharma K. Shakya S. N. Sharma, "Evaluation of DNA mapping schemes for exon detection", *International Conference on Computer, Communication and Electrical Technology (ICCCET)*, ss. 71 – 74, 2011.
- [7] I.M. El-Badawy, A. M. Aziz, S. Gasser, M.E. Khedr, "A new multiple classifiers soft decisions fusion approach for exons prediction in DNA sequences", *2013 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2013, ss. 281 – 286, 2013.
- [8] Z. Baoshan , Y. Sheng, Y. Chen, "An Improved Fourier Method for DNA Sequence Classification", *Bioinformatics and Biomedical Engineering ICBBE 2009. 3rd International Conference on* , ss. 1 – 4, 2009.
- [9] S. S. Sitanshu, "Ganapati Panda Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach' Volume 9, Elsevier, ss. 45–55, 2011.
- [10] M. A. Akhtar, E. Ambikairajah, J. Epps, "GMM-Based Classification of Genomic Sequences", *Digital Signal Processing 15th International Conference on*, ss. 103 – 106, 2007.
- [11] R. Gupta, A. Mittal, K. Singh, P.A. Bajpai, "Time Series Approach for Identification of Exons and Introns", *Information Technology 10th International Conference on*, ss. 91 – 93, 2007.
- [12] L. Huang, M. A. Bataineh, G. E. Atkin, S. Wang, ve W. Zhang, "A Novel gene detection method based on period-3 property", in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, ss. 3857–3860, 2009.
- [13] S. A. Marhon ve S. C. Kremer, "Protein coding region prediction based on the adaptive representation method", in *2011 24th Canadian Conference on Electrical and Computer Engineering(CCECE)*, ss. 000415–000418, 2011.
- [14] L. Das, J. K. Das, ve S. Nanda, "Identification of exon location applying kaiser window and DFT techniques", *2nd International Conference for Convergence in Technology (I2CT)*, ss. 211–216, 2017.
- [15] I. Messaoudi, A. E. Oueslati, ve Z. Lachiri, "Genomic data visualization", in *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, ss. 772–778, 2012
- [16] D. Ashlock, S. Gillis, ve W. Ashlock, "Infinite string block matching features for DNA classification", in *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2017, ss. 1–8.

- [17] J. Zhou, Q. J. Xiao, Z.G. Deng, Z.D. Ouyang, "A novel 3D graphical representation of DNA and its application", *2010 3rd International Conference on Computer Science and Information Technology*, 2010, ss. 1-4
- [18] M. Abo-Zahhai, S.M. Ahmed S.A. Abd-Elrahman K11. "A New Numerical Mapping Technique for Recognition of Exons and Introns in DNA Sequences", *Radio Science Conference (NRSC)*, 2013.
- [19] D. G. Grandhi, C. V. Kumar, "Simplex Mapping for Identifying the Protein Coding Regions in DNA" *TENCON-2007 TAIWAN*, 2007.
- [20] P. Ramachandran, W. S. Lu, A. Antoniou, "Optimized numerical mapping scheme for filter-based exon location in DNA using a quasi-Newton algorithm", *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium*, 2010, ss. 2231 – 2234.
- [21] H. Dozono, G. Niina, "Mapping of DNA sequences using hidden Markov model self-organizing maps", *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) IEEE Symposium on*, 2013, ss. 212 – 217.
- [22] W. Kinsner ve H. Zhang, "Multifractal analysis and feature extraction of DNA sequences", *8th IEEE International Conference on Cognitive Informatics*, 2009, ss. 29–37.
- [23] Q. Zhou, Q. Jiang, ve D. Wei, "A new method for classification in DNA sequence", *6th International Conference on Computer Science Education (ICCSE)*, 2011, ss. 218–221.
- [24] L. Tao, H. Chen, Y. Xu, ve Z. Zhu, "A New Promoter Recognition Method Based on Features Optimal Selection", in *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, 2011, ss. 1–4.
- [25] A. W.-C. Liew, Y. Wu, ve H. Yan, "Selection of statistical features based on mutual information for classification of human coding and non-coding DNA sequences", in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2004, c. 3, s. 766–769 Vol.3.
- [26] M. J. Iqbal, I. Faye, A. M. Said, ve B. B. Samir, "A distance-based feature-encoding technique for protein sequence classification in bioinformatics", in *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, 2013, ss. 1–5.
- [27] S. M. Naeem, M. S. Mabrouk, ve M. A. Eldosoky, "Detecting genetic variants of breast cancer using different power spectrum methods", in *2017 13th International Computer Engineering Conference (ICENCO)*, 2017, ss. 147–153.
- [28] S. Chakraborty ve V. Gupta, "DWT Based Cancer Identification Using EIIP", *Second International Conference on Computational Intelligence Communication Technology (CICT)*, 2016, ss. 718–723.
- [29] K. Leung vd., "Data Mining on DNA Sequences of Hepatitis B Virus", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2), ss. 428–440, Mar. 2011.
- [30] R. E. Dickerson, "Definitions and nomenclature of nucleic acid structure components", *Nucleic Acids Res.*, 1989, s. 1797-1803 Vol. 17.
- [31] Internet: <http://www.biyoklinikder.org/>, Erişim Tarihi: 25.01.2018.
- [32] J. D. Watson, *Molecular Biology of the Gene*, First Edition edition. W. A. Benjamin, Inc, 1965.
- [33] E. F. Meyer, "The first years of the Protein Data Bank", *Protein Sci.*, ss. 1591-1597 Vol. 6, 1997.
- [34] F. Crick, "Central Dogma of Molecular Biology," vol. 227, pp. 561–563, Aug. 1970.
- [35] S. Clancy, W. Brown, "Translation: DNA to mRNA to Protein", *Nature Education*, 2008, 1(1):101.

- [36] F. Dragon, "A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis", *Nature*, 2002, ss.967–970.
- [37] E. Kritikou, "Transcription elongation and termination: It ain't over until the polymerase falls off", *Nature Milestones in Gene Expression*, 2005.
- [38] J. Y. Lee, J. Y. Park, B. Tian, "Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags and trace", *Methods in Molecular Biology*, ss. 23–37, 2008.
- [39] S. Nabavi, R. N. Nazar, "Nonpolyadenylated RNA polymerase II termination is induced by transcript cleavage", *Journal of Biological Chemistry*, ss. 13601–13610, 2008.
- [40] C. Yin, "Representation of DNA sequences in genetic codon context with applications in exon and intron prediction", *J Bioinform Comput Biol*, 13(2), 1550004, Nisan 2015
- [41] J. Y. Y. Kwan, B. Y. M. Kwan, H. K. Kwan, "Spectral Analysis of Numerical Exon and Intron Sequences", *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 2010, ss. 876 – 877.
- [42] J. C. O. Guerra ve P. Licinio, "A method for extracting period~10 modulations from DNA sequence correlations applied to the Drosophila melanogaster genome", *Genet. Mol. Res.*, c. 11, sy 3, ss. 2835-2846, Ağu. 2012.
- [43] S. A. Marhon, S. J. Kremer, "A dynamic representation - based, denovomethod for protein - coding region prediction and biological information detection", *Elsevier, Digital Signal Processing* 46, 2015, ss. 10 – 18.
- [44] J. Zhang, C. Yang, "DNA Sequence Recognition Based on the Markov Model", *6th International Conference on Biomedical Engineering and Informatics (BMEI 2013)*, 2013.
- [45] S. B. Mandal, S. Saha, A. Mandal, M. Roy, "Prediction of Protein Coding Regions of a DNA Sequence through Spectral Analysis", *IEEE/OSA/IAPR International Conference on Informatics Electronics & Vision*, 2012.
- [46] J. Xia, D. Caragea, S. J. "Brown, Prediction of Alternatively Spliced Exons Using Support Vector Machines", *Int. J. Data Mining and Bioinformatics*, 2010, ss. 411 – 30.
- [47] Internet: <https://www.dna.gen.tr/>, Erişim Tarihi: 15.02.2018.
- [48] W. A. El-Zayat, W. A. El-Sayed, ve A. A.-H. Abdel-Rahman, "Anti-hepatitis B virus activity of new pyrimidine and adenine peptide nucleic acid analogues", *Z. Naturforsch., C, J. Biosci.*, c. 64, sy 1-2, ss. 6-10, Şub. 2009.
- [49] J. B. Stafford, R. L. Eoff, A. Kozekova, C. J. Rizzo, F. P. Guengerich, ve L. J. Marnett, "Translesion DNA synthesis by human DNA polymerase eta on templates containing a pyrimidopurinone deoxyguanosine adduct, 3-(2'-deoxy-beta-d-erythro-pentofuranosyl) pyrimido-[1,2-a]purin-10(3H)-one", *Biochemistry*, c. 48, sy 2, ss. 471-480, Ocak 2009.
- [50] N. Schnetz-Boutaud, J. S. Daniels, M. F. Hashim, P. Scholl, T. Burrus, ve L. J. Marnett, "Pyrimido[1,2-alpha]purin-10(3H)-one: a reactive electrophile in the genome", *Chem. Res. Toxicol.*, c. 13, sy 10, ss. 967-970, Eki. 2000.
- [51] H. K. Kwan, S. B. Arniker, "Numerical representation of DNA sequences" [C]. *IEEE Inter, Conf. on Electro/Information Technology*, 2009, ss. 307-310.
- [52] P. Ramachandran, A. Antoniou, "Genomic Digital Signal Processing". *Lecture notes*.
- [53] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", *Physical Review Letters*, 1992, ss. 3805–3808.
- [54] D.G. Grandhi, C. Vijaykumar, "2-Simplex Mapping for Identifying the Protein Coding Regions in DNA", [C]. *TENCON-2007*, 2007, ss. 530.

- [55] T. P. George, T. Thomas, "Discrete wavelet transform de-noising in eukaryotic gene splicing", *BMC Bioinformatics*, 2010.
- [56] B. D. Silverman, R. Linker, "A measure of DNA periodicity", [*J*]. *Theor. Biol.*, ss. 295-300.
- [57] M. Akhtar, J. Epps, E. Ambikairajah, "Signal Processing in Sequence Analysis Advances in Eukaryotic Gene Prediction", [*J*]. *IEEE Journal of selected topics in signal processing*, 2008, ss. 310-321.
- [58] D. Anastassiou, "Genomic signal processing", [*M*]. *IEEE Signal Processing Magazine*, 2001, ss. 8–20.
- [59] P. D. Cristea, "Genetic signal representation and analysis", [*C*]. in *Proc. SPIE Inter. Conf. on Biomedical Optics*, 2002, ss. 77–84.
- [60] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals", [*J*]. *Cell. Mol. Med*, 2002, ss. 279-303.
- [61] P. D. Cristea, "Representation and analysis of DNA sequences. in Genomic signal processing and statistics", *EURASIP Book Series in Signal Processing and Communications*, 2005, ss. 15-66.
- [62] N. Chakravarthy, A. Spanias, L. D. Lasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences", [*J*]. *EURASIP Journal of Genomic Signal Processing*, 2004, ss. 13-28.
- [63] M. Akhtar, J. Epps, E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction", in *Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, 2007, ss.1-4.
- [64] A. K. Brodzik, O. Peters, "Symbol-Balanced Quaternionic Periodicity Transform for Latent Pattern Detection in Dna Sequences", [*C*] *Proceedings of IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, 2005, ss. 5: 373-376.
- [65] S. N. Achuthsankar, S. Sreenadhan, A. Pillai, "Coding measure scheme employing electron-ion interaction pseudo potential (EIIP)", *Bio-information*, 2006, ss.1: 197-202.
- [66] I. Cosic, "Macromolecular Bioactivity: Is it resonant interaction between macromolecules? Theory and Applications", *IEEE Transactions on Biomedical Eng.*, 1994, ss. 41:1101-1114.
- [67] T. Holden, R. Subramaniam, R. Sullivan, E. Cheng, C. Sneider, G. Tremberger, Jr. A. Flamholz, D. H. Leiberman, T. D. Cheung, "ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes", in *Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE)*, 2007, ss. 669417-1 to 669417-10.
- [68] S. V. Buldyrev, A. L. Goilberger, S. Havlin, R. N. Mantegna, M. E. Masta, C.-K. Peng, M. Simons, H. E. Stanley, "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis", *Phy. Rev. E*, 1995, ss. 5084-5091.
- [69] M. Akhtar J. Epps and E. Mbikairajah "Aired Spectral Content Measure for Gene and Exon Prediction in Eukaryotes", [*C*]. *Inter. Conf. on Information and Emerging Technologies, ICIET'07*, 2007, ss. 1- 4.
- [70] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, "Analysis of DNA sequences using methods of statistical physics", *Physica A, Elsevier Science B.V*, 1998, ss. 430-438.
- [71] G. Mendizabal-Ruiz, I. Román-Godínez, S. Torres-Ramos, R. A. Salido-Ruiz, ve J. A. Morales, "On DNA numerical representations for genomic similarity computation", *PLoS ONE*, 2017.

- [72] T. Hoang, C. Yin, ve S. S.-T. Yau, "Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison", *Genomics*, 2016, ss. 134-142.
- [73] S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, ve J.-F. Gibrat, "Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis", *J. Comput. Biol.*, 2012, ss. 796-813.
- [74] J. A. Berger, S. K. Mitra, M. Carli, A. Neri, "Visualization and analysis of DNA sequences using DNA walks", *Journal of the Franklin Institute*, 2004, ss. 341:37-53.
- [75] R. Zhang ,C. T. Zhang, "Identification of replication origins in archaeal genomes based on the Z-curve method", *2005 Heron Publishing-Victoria, Canada, Archaea 1*,2004 ss.335–346.
- [76] M. A. Zahhad, S. M. Ahmed, S. A. A. Elrahman, "Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques", *I.J. Information Technology and Computer Science*, ss. 22-36.
- [77] N. F. Law, K. Cheng and W. Siu, "On relationship of Z-curve and Fourier approaches for DNA coding sequence classification", *Bioinformatics*, 2006, ss. 242-246.
- [78] B. Y. M. Kwan, J. Y. Y. Kwan, H. K. Kwan, "Spectral Methods for Classifying Short Exon and Intron Sequences", *2012 International Conference on Biomedical Engineering and Biotechnology*, 2012, ss. 527-530.
- [79] D Karolchik, AS Hinrichs, TS Furey, KM Roskin, CW Sugnet, D Haussler, WJ Kent, "The UCSC Table Browser data retrieval tool", *Nucleic Acids Res. 32(Database issue)*, 2004, ss. 493-496.
- [80] J Goecks, A Nekrutenko, J Taylor, "The Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences", *Genome Biology*, 2010, Vol. 11(8).
- [81] D Blankenberg, G Von Kuster, N Coraor, G Ananda, R Lazarus, M Mangan, A Nekrutenko, J Taylor, "Galaxy: a web-based genome analysis tool for experimentalists", *Curr. Protoc. Mol. Biol. Chapter 19*, 2010, Unit 19.10.1-21.
- [82] Y. Keung, H. Kwan, "Spectral Analysis of Numerical Exon ve Intron Sequences", *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 2010.
- [83] D. K. Shakya ,R. Saxena, S. N. Sharma, "An Adaptive Window Length Strategy for Eukaryotic CDS Prediction", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2013, ss.1241 – 1252.
- [84] S. Datta, A. Asif, 'A Fast DFT Based Gene Prediction Algorithm For Identification of Protein Coding Regions" , *ICASSP*, 2005, ss.653 – 656.
- [85] S. M. Naeem, M. S. Mabrouk, ve M. A. Eldosoky, "Detecting genetic variants of breast cancer using different power spectrum methods", *13th International Computer Engineering Conference (ICENCO)*, 2017, ss. 147-153.
- [86] H. Saberhari, M. Shamsi, M. Sedaaghi, ve F. Golabi, "Prediction of protein coding regions in DNA sequences using signal processing methods", *IEEE Symposium on Industrial Electronics and Applications*, 2012, ss. 355-360.
- [87] H. Saberhari, M. Shamsi, M. Sedaaghi, F. Golabi, "Prediction of protein coding regions in DNA sequences using signal processing methods", *Industrial Electronics and Applications (ISIEA)*, 2012.
- [88] P. Ramachandran, W. S. Lu, A. Antoniou, "Filter - Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA", *IEEE Transactions on Biomedical Engineering*, 2012, ss. 1598 – 1609.

- [89] M. Roy, S. Barman, S. Saha, ve A. Mandal, "Spectral analysis of DNA sequence using recursive weiner khinchine theorem: a comparative approach", in 2011 International Conference on Recent Trends in Information Systems, 2011, ss. 315-318.
- [90] P. Ramachandran, W.-S. Lu, ve A. Antoniou, "Location of exons in DNA sequences using digital filters", in 2009 IEEE International Symposium on Circuits and Systems, 2009, ss. 2337-2340.
- [91] M. K. Hota ve V. K. Srivastava, "DSP technique for gene and exon prediction taking complex indicator sequence", in TENCON 2008 - 2008 IEEE Region 10 Conference, 2008, ss. 1-6.
- [92] T. S. Gunawan, "On the optimal window shape for genomic signal processing", in 2008 International Conference on Computer and Communication Engineering, 2008, ss. 252-255.
- [93] M. Akhtar, E. Ambikairajah, ve J. Epps, "Optimizing period-3 methods for eukaryotic gene prediction", in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, ss. 621-624.
- [94] J. A. Berger, S. K. Mitra, ve J. Astola, "Power spectrum analysis for DNA sequences", in Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings., 2003, c. 2, ss. 29-32 c.2.
- [95] P. P. Vaidyanathan ve B.-J. Yoon, "Digital filters for gene prediction applications", in Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002., 2002, c. 1, ss. 306-310 c.1.
- [96] R. Damaevicius, "Splice Site Recognition in DNA Sequences Using K-mer Frequency Based Mapping for Support Vector Machine with Power Series Kernel", in 2008 International Conference on Complex, Intelligent and Software Intensive Systems, 2008, ss. 687-692.
- [97] R. Gupta, A. Mittal, K. Singh, P. Bajpai, ve S. Prakash, "A Time Series Approach for Identification of Exons and Introns", in 10th International Conference on Information Technology (ICIT 2007), 2007, ss. 91-93.
- [98] A. Stanescu, K. Tangirala, ve D. Caragea, "Study of transductive learning and unsupervised feature construction methods for biological sequence classification", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, ss. 999-1006.
- [99] G. Su, Y. F. Sun, ve J. Li, "The Identification of Human Cryptic Exons Based on SVM", 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009, ss. 1-4.
- [100] J. Xia, D. Caragea, ve S. Brown, "Exploring Alternative Splicing Features Using Support Vector Machines", in 2008 IEEE International Conference on Bioinformatics and Biomedicine, 2008, ss. 231-238.
- [101] N. F. Law, K. Cheng, W. Siu, "On Relationship of Z - Curve and Fourier Approaches for DNA Coding Sequence Classification", Bioinformation, 2006, ss.242 – 246.
- [102] M. Akhtar, J. Epps, E. Ambikairajah, "On DNA Numerical Representations for Period - 3 Based Exon Prediction", IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), 2007 .
- [103] Yüces oy E., Nabiev V., D etermination of a speaker's age and gender with an SVM classifier based on GMM supervectors , Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (3), 501 - 509, 2016.
- [104] Sengur A ., Multiclass Least - Squares Support Vector Machines for A nalog Modulation Classification , Expert Systems with A ppli cations, 36 (3), 6681 - 6685, 2009.

- [105] Yıldız O., Tez M., Bilge H. Ş., Akcayol M.A., Güler İ., Gene Selection for Breast Cancer Classification Based on Data Fusion and Genetic Algorithm, Journal of the Faculty of Engineering and Architecture of Gazi University, 27 (3), 659 - 668, 2012.
- [106] Kumar M., Gromiha M. M., Raghava G. P. S., Identification of DNA - Binding Proteins Using Support Vector Machines and Evolutionary Profiles, BMC Bioinformatics, 463 (8), 1471 - 2105, 2007.
- [107] K. He, Z. Yang, Y. Bai, J. Long, ve C. Li, "Intelligent Fault Diagnosis of Delta 3D Printers Using Attitude Sensors Based on Support Vector Machines", Sensors (Basel), c. 18, sy 4, Nisan 2018.
- [108] J. Jo, J. Seo, ve J. D. Fekete, "A progressive k-d tree for approximate k-nearest neighbors", in 2017 IEEE Workshop on Data Systems for Interactive Analysis (DSIA), 2017, ss. 1-5.
- [109] A. Dairi, F. Harrou, Y. Sun, ve M. Senouci, "Obstacle Detection for Intelligent Transportation Systems Using Deep Stacked Autoencoder and k-Nearest Neighbor Scheme", IEEE Sensors Journal, ss. 1-1, 2018.
- [110] S. Chen vd., "A Study of Cell-free DNA Fragmentation Pattern and Its Application in DNA Sample Type Classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics, ss. 1-1, 2018.
- [111] M. J. Iqbal, I. Faye, A. M. Said, ve B. B. Samir, "Optimized tree-classification algorithm for classification of protein sequences", in 2015 International Symposium on Mathematical Sciences and Computing Research (iSMSC), 2015, ss. 110-115.
- [112] D. Kolbe, Q. Zhu, ve S. Pramanik, "On k-Nearest Neighbor Searching in Non-Ordered Discrete Data Spaces", in 2007 IEEE 23rd International Conference on Data Engineering, 2007, ss. 426-435.
- [113] B. Das, I. Turkoglu, "A novel numerical mapping method based on entropy for digitizing DNA sequences", Neural Comput & Applic, DOI 10.1007/s00521-017- 2871-5, 2017.
- [114] B. Das, I. Türkoğlu, "Sayısal Haritalama Teknikleri ve Fourier Dönüşümü Kullanılarak DNADizilimlerinin Sınıflandırılması", Journal of the Faculty of Engineering and Architecture of Gazi University, 31(4), 2016 ss. 921-932, 2016. doi:10.17341/gazimmfd.278447.
- [115] Ensembl Genbankası veritabanı, online erişim: <http://www.ensembl.org>
- [116] E. V. Koonin, A. S. Novozhilov, "Origin and evolution of the genetic code: the universal enigma", IUBMB Life, 61(2),2009, ss. 99–111.
- [117] J. Tugan, A. Rushdi, "A DSP based approach for finding the codon bias in DNA sequences", IEEE Journal on Signal Processing, 2008, 2(3), ss. 343 – 356.
- [118] J A. D. Machado, "Shannon entropy analysis of the genome code", Mathematical Problems in Engineering, 2012, ss. 10.1155/2012/132625.
- [119] D. Koslicki, "Topological entropy of DNA sequences", Bioinformatics, 2011, 27(8) ss. 1061-1067.
- [120] X. Zhang, X. A. Zhou, Y. H. Yu, "Similarity Analysis of DNA Using Improved Approximate Entropy", 2012 International Conference on Biomedical Engineering and Biotechnology,2012, ss. 511-514.
- [121] B. Kozarzewski, "A method for nucleotide sequence analysis", Computational Methods in Science and Technology, 2012, ss. 18(1): 5-10.
- [122] Zhang, J. Su, D. Yu, Q. Wu, ve H. Yan, "EpiDiff: Entropy-based quantitative identification of differential epigenetic modification regions from epigenomes", 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, ss. 655-658, 2013.

- [123] R. E. Monge, J. L. Crespo, "Comparison of complexity measures for DNA sequence analysis", 3rd IEEE International Work-Conference on Bioinspired Intelligence, ss. 71-75, 2014.
- [124] S. Vinga ve J. S. Almeida, "Rényi continuous entropy of DNA sequences", J. Theor. Biol., c. 231, sy 3, ss. 377-388, Ara. 2004.
- [125] O. V. Kirillova, "Entropy concepts and DNA investigations", Physics Letters A, c. 274, sy 5, ss. 247-253, Eylül 2000.
- [126] R. Damasevicius, "Complexity Estimation of Genetic Sequences Using Information-Theoretic and Frequency Analysis Methods", Informatica, Lith. Acad. Sci., c. 21, ss. 13-30, Ocak 2010.
- [127] T. R. Wojcik ve D. Krapf, "Solid-State Nanopore Recognition and Measurement Using Shannon Entropy", IEEE Photonics Journal, c. 3, sy 3, ss. 337-343, Haz. 2011.
- [128] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, ve J. C. Mueller, "Gene mapping and marker clustering using Shannon's mutual information", IEEE/ACM Transactions on Computational Biology and Bioinformatics, c. 3, sy 1, ss. 47-56, Ocak 2006.
- [129] C. J. Langmead, C. R. McClung, ve B. R. Donald, "A maximum entropy algorithm for rhythmic analysis of genome-wide expression patterns", in Proceedings. IEEE Computer Society Bioinformatics Conference, 2002, ss. 237-245.
- [130] A. Karcı, "New kinds of entropy: fractional entropy", International Conference on Natural Science and Engineering (ICNASE'16), 2016.
- [131] M. Akhtar, E. Ambikairajah, ve J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition", in Proceedings of the IEEE Symposium on Emerging Technologies, 2005., 2005, ss. 13-17.
- [132] S. A. Aghili, O. D. Sahin, D. Agrawal, ve A. E. Abbadi, "Efficient filtration of sequence similarity search through singular value decomposition", in Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering, 2004, ss. 403-410.
- [133] T. J. Peters, R. Smolikova-Wachowiak, ve M. P. Wachowiak, "Microarray Image Compression Using a Variation of Singular Value Decomposition", 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, ss. 1176-1179, 2007.
- [134] G. Köktürk, "Kısa süreli Fourier dönüşümü için başarımlı karşılaştırma", Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, cilt:8 sayı:3 ss.11-18, 2006.
- [135] J. Majewski ve R. Wojtyna, "Evolutionary algorithm for transformation of short-time signal into frequency-domain description", Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), ss. 30-33, 2016.
- [136] G. Manhertz, D. Modok, ve Á. Bereczky, "Evaluation of short-time fourier-transformation spectrograms derived from the vibration measurement of internal-combustion engines", IEEE International Power Electronics and Motion Control Conference (PEMC), ss. 812-817, 2016.
- [137] M. C. Şahingil ve Y. Özkazanç, "Visualization of repetitive DNA sequence regions via Short Time Fourier Transform", in 2012 20th Signal Processing and Communications Applications Conference (SIU), 2012, ss. 1-4.
- [138] F. X. Socheleau, D. Pastor, ve M. Duret, "On Symmetric Alpha-Stable Noise After Short-Time Fourier Transformation", IEEE Signal Processing Letters, 20(5) ss. 455-458, May. 2013.
- [139] B. Das, I. Turkoglu, "A new numerical mapping approach for identification protein coding regions in DNA sequences by using SVD method", International Conference on Engineering and Natural Science. Sarajevo, 2016, ss. 2773-2777.

- [140] B. Das, I. Turkoglu, "A new mapping technique for separation of exons and introns by using DFT method", International Conference on Engineering and Natural Science. Sarajevo, 2016, ss. 2778-2784.
- [141] NCBI GenBank veritabani: <http://www.ncbi.nlm.nih.gov/Genbank> Erişim Tarihi: Ocak 2016.
- [142] G. H. Sendra, "Dynamic speckle algorithms comparison using receiver operating characteristic", Optical Engineering, Opt. Eng, 2008, 47(5), doi:10.1117/1.2920429.
- [143] Genetic Alliance Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals. NCBI Bookshelf, 2009.
- [144] M. C. Frith, S. Khan, "A survey of localized sequence rearrangements in human DNA", Nucleic Acids Research, 2017, doi: 10.1093/nar/gkx1266.
- [145] S. H. Orkin, D. G. Nathan, D. Ginsburg, A. T. Look, D. E. Fisher, S. E. Lux, "Nathan and Oski's Hematology of Infancy and Childhood", 8 th Edition, Philadelphia, USA: Elsevier Saunders, 2015, ss.1055-1061.
- [146] R. E. Behrman, R. M. Kliegman, H. B. Jenson, "Nelson Textbook of Pediatrics", 17th ed.Philadelphia: W.B. Saunders Company, 2004.
- [147] D. Canatan, Y. Aydınok, "Talasemi ve hemoglobinopatiler tanı ve tedavi kitabı", Talasemi Federasyonu, Antalya, 2007.
- [148] H. K. Kwan, S. B. Arniker, "Numerical representation of DNA sequences". IEEE Inter, Conf. On Electro/Information Technology, EIT &#39;09, Windsor,2009, ss. 307-310.
- [149] S. Toraman, "Feature extraction using infrared spectroscopy from blood samples related to colon cancer" Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Mart 2016.
- [150] A. Wee, C. Liew, Y. Wu, H. Yan, "Selection of Statistical Features Based on Mutual Information for Classification of Human Coding and Non-coding DNA Sequences", Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), 1051-4651/04, IEEE, 2006.
- [151] R. Saidi, M. Maddouri, E. M. Nguifo, "Protein sequences classification by means of feature extraction with substitution matrices", BMC Bioinformatics, 2010, 1471-2105/11/175.
- [152] H. Heidari Bafroui, A. Ohadi, "Application of wavelet energy and Shannon entropy for feature extraction in gearbox fault detection under varying speed conditions," Neurocomputing, 2014, ss. 437-445.
- [153] J. Li, "Recognition of the optical image based on the wavelet space feature spectrum entropy," Optik (Stuttg), 2015, ss. 3931-3935.
- [154] A. Siva Sankar, S. S. Nair, V. S. Dharan, and P. Sankaran, "Wavelet sub band entropy based feature extraction method for BCI," Procedia Comput. Sci., vol. 46, no. Icict 2014,2015, ss. 1476-1482.
- [155] C. Ley, O. Klein, P. Bernard, L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", Journal of Experimental Social Psychology, 2013, Vol 49(4), ss. 764-766.
- [156] H. Özbek and S. Keskin, "Standart Sapma mı Yoksa Standart Hata mı?," Van Tıp Derg., vol. 14, no. 2, ss. 64-67, 2007.
- [157] A. Phinyomark, A. Nuidod, P. Phukpattaranont, C. Limsakul, "Feature extraction and reduction of wavelet transform coefficients for EMG pattern classification", Elektron. ir Elektrotehnika, vol. 122, no. 6, ss. 27-32, 2012.
- [158] A. Phinyomark, S. Hirunviriya, C. Limsakul, P. Phukpattaranont, "Evaluation of EMG feature extraction for hand movement recognition based on Euclidean distance and standard

deviation”, Electr. Eng. Comput. Telecommun. Inf. Technol. (ECTICON) Conf., 2010, ss. 856–860.

- [159] A. Phinyomark, C. Limsakul, P. Phukpattaranont, “A Novel Feature Extraction for Robust EMG Pattern Recognition”, J. Comput., 2009, ss. 71–80.
- [160] C. Vigneshwari, V. Vimala, S. V. Vignesh, G. Sumithra, “Analysis of Finger Movements Using EEG Signal”, Int. J. Emerg. Technol. Adv. Eng., ss. 583–588, 2013.
- [161] P. F. Bach, “Myoelectric signal features for upper limb prostheses,” Norwegian University of Science and Technology, 2009.
- [162] İ. Türkoğlu, “Durağan olmayan işaretler için zaman-frekans entropilerine dayalı akıllı örüntü tanıma”, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2002.
- [163] S. Yavuz and M. Deveci, “İstatiksel Normalizasyon Tekniklerinin Yapay Sinir Ağın Performansına Etkisi ”, Erciyes Üniversitesi İktisadi ve İdari Bilim. Fakültesi Derg., vol. 40, 2012, ss. 167–187.
- [164] E. Guresen ve G. Kayakutlu, “Definition of artificial neural networks with comparison to other networks”, Procedia Computer Science, c. 3, ss. 426-433, Ocak 2011.
- [165] B. Muller, J. Reinhardt, M.T. Strickland, Neural Networks an Introduction, Springer-Verlag Berlin Heidelberg, 1995.
- [166] Q. Mingyue, L. Cheng, ve S. Yu, “Application of the Artificial Neural Network in Predicting the Direction of Stock Market Index”, in 2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), 2016, ss. 219-223.
- [167] K. C. Sriharipriya ve R. Sanju, “Artificial neural network based multi dimensional spectrum sensing in full duplex cognitive radio networks”, International Conference on Computing Methodologies and Communication (ICCMC), ss. 307-312, 2017.
- [168] K. Levenberg, “A Method for the Solution of Certain Problems in Last Squares”, Quart. Appl. Math. 2, ss. 164– 168, 1944.
- [169] D. Marquardt, “An Algorithm for Least Squares Estimation of Nonlinear Parameters”, SIAMJ. Appl. Math. 11, ss.431–441, 1963.
- [170] B. Das, I. Turkoglu, “Sayısal Haritalama Tekniklerinin Başarımlarının K-Ortalamlar Kümeleme Yöntemi ile Analizi”, International Conference on Computer Science and Engineering Tekirdağ, ss. 301-305, 2016.
- [171] B. Das, I. Turkoglu, “Fourier-based filtering approach for identification of protein-coding regions in DNA sequences”, 23th Signal Processing and Communications Applications Conference (SIU), ss. 2529-2532, 2015.
- [172] B. Das, I. Türkoğlu, “DNA Dizilimindeki Nükleotid Çiftlerinin Frekans Değerlerine Göre Farklı Sınıflandırma Yöntemleri ile Karşılaştırılması”, Tıp Teknolojileri Ulusal Kongresi, 2014.

## ÖZGEÇMİŞ

***Bihter DAŞ***

### KİŞİSEL BİLGİLER

**ADRES :** Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği 23119 – Elazığ

**TELEFON :** +90 (424) 247 91 92

**EPOSTA :** [bihterdas@gmail.com](mailto:bihterdas@gmail.com)

**DOĞUM  
YERİ/TARİHİ:** Elazığ, 1982

**MEDENİ HALİ:** Evli

### EĞİTİM BİLGİLERİ

**DOKTORA :** **2012 – 2018**, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yazılım Mühendisliği Anabilim Dalı, Elazığ

**YÜKSEK LİSANS :** **2005–2007**, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik ve Bilgisayar Eğitimi Anabilim Dalı, Elazığ

**LİSANS :** **2000–2004**, Fırat Üniversitesi, Teknik Eğitim Fakültesi, Bilgisayar Öğretmenliği, Elazığ (*Bölüm ve Fakülte Birincisi*)

**LİSE :** **1998–2000**, Mehmet Akif Ersoy Süper Lise (İngilizce hazırlık), Elazığ

### İŞ DENEYİMİ

**2014 – Halen** MEB, Bahçelievler Ortaokulu, Elazığ, (Bilgisayar Öğretmeni)

**2017 – 2018** TÜBİTAK BİDEB-2214 bursu ile Alberta Üniversitesi, Bilgisayar Bilimleri, Biyoinformatik Lab. Edmonton, Alberta/Kanada, Araştırmacı doktora öğrencisi.

**2005 – 2014** MEB, 24 Kasım İlköğretim Okulu, Elazığ, (Bilgisayar Öğretmeni)

**2004 – 2005** MEB, Ovaakça Endüstri Meslek Lisesi, Bursa, (Bilgisayar Öğretmeni)

### BİLDİĞİ YABANCI DİLLER

İngilizce (B2)

### ARAŞTIRMA DENEYİMİ

Biyoinformatik, Sayısal İşaret İşleme, Genom Analizi, Tip-2 Bulanık Mantık.

### İLGİLENDİĞİ DİĞER BİLİMLER

Tiyatro, Sinema