Journal of the Faculty of Engineering and Architecture of Gazi University 31:4 (2016) 921-932

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Sayısal haritalama teknikleri ve Fourier dönüşümü kullanılarak DNA dizilimlerinin sınıflandırılması

Bihter Daş *, İbrahim Türkoğlu
Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü, 23119, Elazığ, Türkiye

Ö N E Ç I K A N L A R

- Sayısal haritalama tekniklerinin performanslarının karşılaştırılması
- DNA dizilimlerinin ekson ve intron olarak sınıflandırılması
- Pencereleme fonksiyonlarının sınıflandırma performansına etkisi

Makale Bilgileri
Geliş: 12.06.2015
Kabul: 04.08.2016

DOI:
10.17341/gazimmfd.278447

Anahtar Kelimeler:
Deoksiribo nükleik asit sıralama,
intron ve ekson sınıflandırma,
sayısal haritalama teknikleri

ÖZET

Bir DNA dizilimindeki bazların oluşturdukları kombinasyonlar, o DNA dizilimindeki bir gene karşılık gelir, bu genlerden de RNA kopya dizilimleri çıkarılır. Kopyalanan bu RNA'lar oluşurken genin baz dizilimi baştan sona tümüyle okunmaz. Genlerin okunmayan ve kodlanmayan bölümüne *intron*, kodlanan kısımlarına ise *ekson* denir. Bir DNA dizilimindeki protein nerede, ne kadar kodlanır? Büyüme ve gelişme nerede düzenlenir? Kök hücreler nerede başka hücreye dönüştürülür? Tüm bu soruların cevabı ve kanser gibi genetik hastalıkların araştırılması DNA dizilimlerinin ekson ve intron olarak sınıflandırmasıyla mümkündür. Çalışmanın amacı, DNA diziliminin ekson ve intron olarak sınıflandırılmasında farklı sayısal haritalama tekniklerinin performanslarını karşılaştırmaktır. Bu amaç doğrultusunda insan türünün MEFV genine ait DNA dizilimleri, 9 farklı haritalama tekniği ile sayısal dizilere dönüştürülmüştür. Dönüştürülen sayısal dizilerin sınıflandırılmasında Ayrık Fourier Dönüşümü yöntemi kullanılmıştır. Bu yöntemde 4 farklı pencereleme fonksiyonu kullanılmış, sınıflandırma başarımları karşılaştırılmıştır. Ayrıca Fourier tabanlı yöntemle elde edilen sonuçlar, Destek Vektör Makineleri ve K-en yakın komşu algoritması gibi makine öğrenme tabanlı yöntemlerle karşılaştırılmıştır. İnteger haritalama tekniği Ayrık Fourier Dönüşümü yönteminde %96,2 ile diğer makine öğrenme yöntemlerine göre en yüksek sınıflandırma başarımı sağlamıştır. Hamming pencereleme fonksiyonunda sınıflandırma başarısı diğer pencereleme fonksiyonlarından daha yüksek çıkmıştır.

Classification of DNA sequences using numerical mapping techniques and Fourier transformation

H I G H L I G H T S

- Comparison of the performance of digital mapping techniques
- Classification of the DNA sequences as exon and intron.
- Effect on the classification performance of windowing functions

Article Info
Received: 12.06.2015
Accepted: 04.08.2016

DOI
10.17341/gazimmfd.278447

Keywords:
Deoxyribonucleic acid sequences,
intron and exon classification,
numerical mapping techniques

ABSTRACT

The combinations of bases in a DNA sequence correspond to a gene in that DNA sequence, RNA copy sequences are extracted from these genes. When these copied RNA's extracted, the base sequence of gene is not read from the beginning to the end completely. The uncoded and unreadable section of the gene is called 'intron' and the coded section of the gene is called 'exon'. Where is a protein coded? How much is encoded? Where are growth and development regulated? Where are stem cells converted to other cells? The answer to all of these questions and the investigation of genetic diseases, such as cancer, are possible by DNA sequences that can be classified as the exon and intron. The aim of this study is to compare the performance of different digital mapping techniques for classification of DNA sequence as the exon or intron. For this purpose, DNA sequences of the MEFV gene in human species are transformed to numeric sequences by nine different digital mapping techniques. The Discrete Fourier Transform Method (DFT) is used to classify these transformed sequences. Four different windowing functions are used and their classification performance are compared in this method. Also, the results obtained from the Fourier-based method have been compared using the Support Vector Machine and the K-Nearest Neighbor methods. Integer mapping technique achieved the highest classification performance with 96.2% in the DFT method than other machine learning methods. Classification performance in Hamming windowing function is higher than other windowing functions.

* Sorumlu Yazar/Corresponding author: bihterdas@gmail.com / Tel:+90 424 237 0000 - 4222

1. GİRİŞ (INTRODUCTION)

Genetik özellikler hücre çekirdeğindeki kromozomlarla taşınır. Kromozomlar DNA ve özel proteinlerin birleşmesinden oluşur. Bir DNA'nın yapı birimi nükleotidlerdir. Nükleotidler şeker ve fosfat ve organik bazlardan oluşur. Bu bazlar Adenin (A), Guanin (G), Timin (T), Sitozin (C) dir. Bir nükleotid hangi organik bazı içeriyorsa o bazın ismiyle nitelendirilirler. Şekil 1'de örnek bir nükleik asit dizilimi görülmektedir. Protein ve enzimler üretilirken DNA üzerindeki bazların dizilimlerini örnek alınarak bu genlere karşılık gelen RNA kopya dizilimleri çıkarılır. mRNA olarak isimlendirilen bu RNA'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 ekson adı verilir. Bir gene ait olan DNA diziliminde o dizilimi ekson ve intron olarak sınıflandırmak, bir DNA dizi analizinde oldukça önemlidir [1-5]. Bir genin dizi analizinde homoloji (benzerlik) araştırması, yeni bulunan bir DNA diziliminin diğer tüm dizilimlerle karşılaştırılması ve bunun sonucunda benzerlerdeki veri tabanında ya da literatürde tanımlanmış bazı biyolojik işlevlerin, yeni bulunan dizilime yakıştırılması olarak tanımlanabilir. Bu yöntemle benzerlikler ve protein kodlayan eksonlar araştırılır ve bir genin mutasyona uğrayıp uğramadığı belirlenebilir [6, 7]. Hangi proteinin nerede, nasıl, ne kadar ve ne zaman kodlanacağını, büyüme ve gelişmenin nerede nasıl düzenleneceğini, kök hücrelerin nerede hangi hücre, doku ve organlara dönüşeceğini, hücrelerin hangi koşullarda çoğaltılıp ya da öldürüleceğini, ne zaman kanser geliştirileceğinin araştırılması ekson ve intronların sınıflandırılmasının önemini arttırmaktadır [1-3]. Sayısal sinyal işleme teknikleri sembolik sinyallere uygulanamaz. Bu yüzden DNA dizilimlerinin sayısal haritalama teknikleri ile sayısal sinyallere dönüştürülmesi gerekir. Bu makalede insan türünün MEFV genine ait DNA baz dizilimleri sayısal haritalama teknikleri ile sayısal sinyallere dönüştürülmüş ve daha sonra Ayrık Fourier Dönüşümü (DFT), Destek Vektör Makinesi (SVM), K- En yakın Komşu Algoritması (KNN) gibi sınıflandırma yöntemleri ile bu tekniklerin DNA dizilimlerini ekson ve intron olarak sınıflandırmadaki başarımları karşılaştırılmıştır.

2. KULLANILAN DNA SAYISAL HARİTALAMA TEKNİKLERİ (USED DNA DIGITAL MAPPING TECHNIQUES)

Sinyal işleme uygulamalarında DNA dizilimleri üzerinde çalışabilmek için DNA dizilimlerinin sayısal sinyallere dönüştürülmesi gerekir. Bu dönüştürme işleminde kullanılan yöntemler sayısal haritalama teknikleri olarak adlandırılırlar. Bu teknikler sabit haritalama teknikleri ve fiziko-kimyasal özellik tabanlı haritalama teknikleri olarak 2 grupta toplanır [8].

Bu makale çalışmasında, sabit haritalama tekniklerinden Integer Haritalama Tekniği [9], Reel Haritalama Tekniği [10], Karmaşık1 ve Karmaşık2 Haritalama Tekniği [11-15], fiziko-kimyasal özellik tabanlı haritalama tekniklerinden Atomic Haritalama Tekniği [16], Molecular Haritalama Tekniği, DNA-Walk Haritalama Tekniği [17, 18], Paired Numeric Haritalama Tekniği [19, 20] ve EIIP Haritalama Tekniği [21, 22] kullanılmıştır.

2.1. Tamsayı Haritalama Tekniği (Integer Mapping Technique)

Integer teknik DNA bazlarının 1 boyutlu haritalama tekniğidir [6]. Bir DNA diziliminde eğer pürin (A,G) > purimidin (C, T) 4 baza sırayla T=0, C=1, A=2, G=3 şeklinde değer verilir. T>A ve G>C olması durumunda bazlara A=0, C=1, T=2 ve G=3 değerleri verilir. S(n)=[ACCGTATAGG] şeklinde verilen bir DNA diziliminin integer tekniği ile sayısal hale dönüştürülmüş biçimi şu şekildedir. Bu DNA diziliminde pürin (A,G) >purimidin (C, T) olduğundan S(n)=[2 1 1 3 0 2 0 2 3 3] şeklinde olacaktır [9].

2.2. Reel Haritalama Tekniği (Reel Mapping Technique)

Bu teknikte bir DNA dizilimde A=-1,5 T=1,5 C=0,5 G=-0,5 değerlerini alır. Reel ve integer haritalama teknikleri, reel tekniğin doğal tamamlayıcı özelliği ve integer tekniğinin tamamlayıcı olmayan özelliğinden dolayı farklı dizi eşleşmelerini sağlar. S(n)=5"-TGGAAC-3" şeklinde verilen bir dizi reel teknik ile sayısal diziye dönüştürülürken aynı otoregresif parametrelere sahip olabilir.

Şekil 1. Bir nükleik asit dizilim örneği (The example of a nucleic acid sequence)

1. 5”- 1,5 -0,5 -0,5 -1,5 -1,5 0,5 -3”= 5”-TGGAAC-3” (tersine tamamlayıcı dizilim)
2. 5”- -0,5 1,5 1,5 0,5 0,5 -0,5 -3” =5”- GTTCCG-3” (Ters dizilim)
3. 5”- 0,5 -1,5 -1,5 -0,5 -0,5 0,5 -3”=5”-CAAGGC-3” (Tamamlayıcı dizilim)

Bu dizilim yapılarından dolayı ters işaretli sayısal diziler ve gerçek sayısal diziler aynı lineer bağımlılığa sahiptir ve bu yüzden aynı otoregresif parametrelere sahiptir. Aynı zamanda ileri ve geri beslemeli lineer tahmin hataları da teorik olarak aynı otoregresif modelleri vermektedirler [10].

2.3. Atomik Haritalama Tekniği (Atomic Mapping Technique)

Fiziko-kimyasal özellik tabanlı tekniklerden biri olan atomik işaret dizisi her bir nükleotide A=70, G=78, C=58, T=66 değerlerini atayarak elde edilir. S(n)=[TGGAAC] şeklinden verilen bir DNA dizilimi S(n)=[66 78 78 70 70 58] şeklinde sayısal diziye dönüştürülür [16].

2.4. Moleküler Kütle Haritalama Tekniği (Molecular Mass Mapping Technique)

Fiziko-kimyasal özellik tabanlı tekniklerden biri olan molekül işaret dizisi her bir nükleotide A=134, G=150, C=110, T=125 değerlerini atayarak elde edilir. S(n)=[TGGAAC] şeklinde verilen bir DNA dizilimi S(n)=[125 150 150 134 134 110] şeklinde sayısal diziye dönüştürülür [17].

2.5. DNA-Walk Haritalama Tekniği (DNA Walk Mapping Technique)

DNA dizilerinin sabit ölçekli uzun menzilli korelasyonuyla çalışmak için geliştirilmiş bir tekniktir [18]. 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. Bir DNA diziliminde bulunan A, G, C, T bazları A=1, G=-1, T=j ve C=-j değerlerini alır. S(n)=[TGTCAC] şeklinde verilen bir DNA dizilimi S(n)=[j -1 j -j 1 1 -j] şeklinde sayısal diziye dönüştürülür [20-21].

2.6. Eşli Sayısal Haritalama Tekniği (Paired Numeric Mapping Technique)

Bu teknikte eşleştirilmiş olan A-T ve G-C bazları +1 ve -1 değerlerini alırlar. Bir ya da iki işaret olarak temsil edilebilirler. Bu sayısal dönüşüm DNA’nın karmaşıklığını azaltarak DNA yapısal özellikleri birleştirilir. Nükleotid çiftlerini bu teknikle sayısal hale dönüştürmek için 7 kural vardır [19-20]. Bunlar;

1. Purin-purimidin kuralı (RY kuralı): Eğer n_i nükleotid dizisi pürin ise (A ya da G) u_i=1 dir, eğer n_i purimidin ise (C ya da T) u_i=-1 dir.
2. $A\bar{A}$ kuralı: Eğer n_i=A u_i=1 dir. Diğer tüm durumlarda u_i=-1 dir.
3. $T\bar{T}$ kuralı: Eğer n_i=T u_i=1 dir. Diğer tüm durumlarda u_i=-1 dir.
4. $G\bar{G}$ kuralı: Eğer n_i=G u_i=1 dir. Diğer tüm durumlarda u_i=-1 dir.
5. $C\bar{C}$ kuralı: Eğer n_i=C 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 ya da C için u_i=1, T ya da G için u_i=-1 dir.

Yukarıdaki kurallardan en çok kullanılan RY kuralıdır. Eşleştirilmiş sayısal teknik en çok gen ve ekson tahmininde kullanılır. Bu teknik gen ve ekson tahmininde bazların ekson ve intronda bulunma sıklığından faydalanırlar. Yani Intronlar A ve T bakımından zengin, eksonlar ise G ve C bakımından zengindir. Ekson 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. “…AATGCTGCCATTA…” şeklinde verilen bir DNA diziliminde x_n=[+1 +1 +1 -1 -1 +1 -1 -1 -1 +1 +1 +1 +1] olur [20].

2.7. Karmaşık Sayısal Haritalama Tekniği (Complex Digital Mapping Technique)

Farklı düzlemlerde tetrahedron tekniğinin yansıtılması ile tetrahedron sayısal haritalama tekniğinin boyutu 2’e dönüştürülebilir [11-15]. Böyle düzlemler sayısal dönüşümlerin simetrisinin korunması ve biyolojik özelliklere karşılık gelen matematiksel özelliklerin yansıtılması yolları ile seçilebilir. Bu düzlemler karmaşık (complex) düzleme dönüştürülür ve böylece A, G, C, T bazlarının karmaşık temsilleri elde edilir. Bir x(n) diziliminde bazların karmaşık temsili aşağıda gösterilmiştir [11-15].

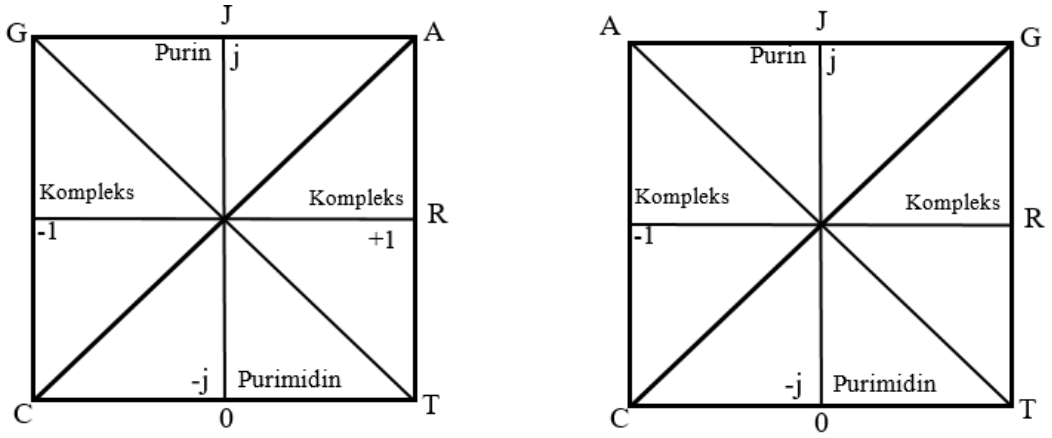
$$x(n)=ax_A(n)+ cx_C(n)+ tx_T(n)+ gx_G(n)$$

a=1+j, t=1-j, c=-1-j ve g=-1+j dir.

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. Purinler ve purimidinler eşit sanal parça ve zıt işaretli gerçek parça olarak ifade edilirken onların gösterimleri, karmaşık eşlenikler olarak ifade edilir. Şekil 2a da bazların sağ düzlemde Şekil 2b ise sol düzlemde gösterimi verilmiştir [32-38]. Karmaşık teknikler için kullanılan diğer bir dönüşümde ise A=1, T=j, C=-j ve G=-1 değerleri verilir. Bu sayısal dönüşüm bazların zayıf reel ve zayıf imajiner karmaşık temsilinde kullanılır [11-15].

2.8. EIIP Haritalama Tekniği (EIIP Mapping Technique)

Bu teknikte DNA dizilimindeki her bir nükleotid EIIP temsilindeki yarı değerlik sayısı ile eşleştirilir [18, 19]. A= 0,1260 G=0,0806 C=0,1340 T=0,1335 değerleri verilir.



Şekil 2. Bazların sol ve sağ düzlemde gösterilimi (Graphical representation of bases on the left and right plane)

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.

S(n)=[A T T G C A T G C] iken $S_e(n)$=[0,1260 0,1335 0,1335 0,0806 0,1340 0,1260 0,1335 0,0806 0,1340]'dir [21, 22].

3. KULLANILAN SINIFLANDIRMA YÖNTEMLERİ (USED CLASSIFICATION METHODS)

3.1. Ayrık Fourier Dönüşümü Yöntemi (Discrete Fourier Transform Method)

Ayrık zamanlı Fourier dönüşümü (DFT), ayrık zamanlı sinyal işleme algoritma ve sistemlerinin analizi, tasarımı, gerçekleştirilmesi ile doğrusal filtreleme, korelasyon analizi ve spektrum analizi gibi sinyal işleme uygulamalarında önemli bir rol oynar [23, 24]. Sinyaller sınırlı sayıda nokta için değerlendirilir [25-27]. Bu makalede DFT, DNA dizilimlerinin 3-periyot değerlerinin çıkarılması amacıyla kullanılmaktadır. 3-periyotlu olmasının sebebi aminoasit üreten kodonların 3 baz uzunluğunda olmasıdır. DFT, Fourier dönüşümünün eşit aralıklı frekanslardaki örneklerine özdeştir [39-42].

N -noktalı bir DFT'nin hesaplanması Eşitlik 1'de verilmiştir [28].

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

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

L baz uzunluğunda ve 2 komşu pencere arasındaki L-3 baz genişliğinde örtüşme ile pencere yaklaşımı kullanılmıştır. DFT spektrumunun normalizasyon toplamı Eşitlik 2'de verilmiştir [42-50].

$$X_T[k] = \frac{1}{N_W}\sum_{m=1}^{N_w} X_m[k] \quad (2)$$

$X_m[k]$, pencerelenmiş her bir dizilimdir. $X_T[k]$, DFT spektrumunun normalize edilmiş toplamıdır.

DNA diziliminin güç spektrumu için Eşitlik 3 kullanılmaktadır.

$$S[k]=|X_T[k]|^2 \quad (3)$$

Sayısal biçime dönüştürülmüş dizilimin spektral içerik ölçümünden elde edilen 3 periyotlu spektral bileşeni Eşitlik 4'de verilmiştir.

$$P_3=S[N/3+1] \quad (4)$$

Ekson ve intron olarak sınıflandırmak için kullanılan eşik değer T_3 olarak isimlendirilir [28]. T_3 değeri için Eşitlik 5 kullanılmaktadır.

$$T_3=\frac{sdP_{3e}*meanP_{3i}+sdP_{3i}*meanP_{3e}}{sdP_{3e}+sdP_{3i}} \quad (5)$$

Ekson ve intronların 3 periyotlu standart sapma ve ortalama değerlerini bulmak için ekson ve intron dizilimlerinin eğitim dizileri kullanılmaktadır. Ekson ve intron test dizileri için 3 periyotlu P_{3t} değeri, T_3 eşik değerden büyük ya da eşit olursa o test dizisi ekson olarak, T_3 eşik değerden küçük olursa ise o test dizisi intron olarak sınıflandırılır [28].

Pencereleme (Windowing): Sinyal işleme uygulamalarında sonsuz uzunluktaki bir işaret dizisi ile çalışmak imkansız olduğunda bütün işaret analizlerinde pencereleme yapılması gerekmektedir [29-35]. 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. Sinyaller işlenmeden önce belli sayıda örnek içeren parçalara ayrılır. İşte bu parçaların her birine pencere adı verilir. Bu makalede kullanılan pencere fonksiyonları Dikdörtgen Pencere Fonksiyonu (Rectangular Window Function), Hamming Pencere Fonksiyonu (Hamming Window Function), Gaussion Pencere Fonksiyonu (Gaussion Window Function) ve Blackman Pencere Fonksiyonu (Blackman Window Function)'dur [34-36].

Rectangular penceresi: ω(n)=1, B=1

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

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

Gaussion Penceresi: exp(−x T Ax+s T x)

3.2. Destek Vektör Makineleri (Support Vector Machines)

Son yıllarda, sınıflandırma problemlerinin çözümü için geliştirilmiş en başarılı makine öğrenimi algoritmalarından biri Destek Vektör Makineleri'dir. Destek Vektör Makineleri, değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki birçok sınıflandırma probleminin çözümünde başarıyla uygulanmış, performansı yüksek ve etkin makine öğrenimi algoritmalarından biri olarak veri madenciliği uygulamalarındaki yerini almıştır [51-52]. Bu yöntem, sınıflandırmayı bir doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yerine getirir. Doğrusal olmayan dönüşümlerde kernel fonksiyonu kullanılmakta ve verilerin daha yüksek boyutta doğrusal olarak ayrımına imkân sağlanmaktadır ve Şekil 3'de destek vektör algoritmasının genel yapısı görülmektedir [56-60].

Bu çalışmada destek vektör makineleri (DVM) ile gerçekleştirilecek iki sınıflı bir sınıflandırma için radyal tabanlı fonksiyon (RBF) kerneli kullanılmıştır. Uygulamada kernel parametresi (RBF kerneli için band genişliği değeri) 2, düzenleme parametresi (C) ise 100000 olarak belirlenmiştir.

3.3. K-En Yakın Komşu Algoritması (K-Nearest Neighbor Algrorithm)

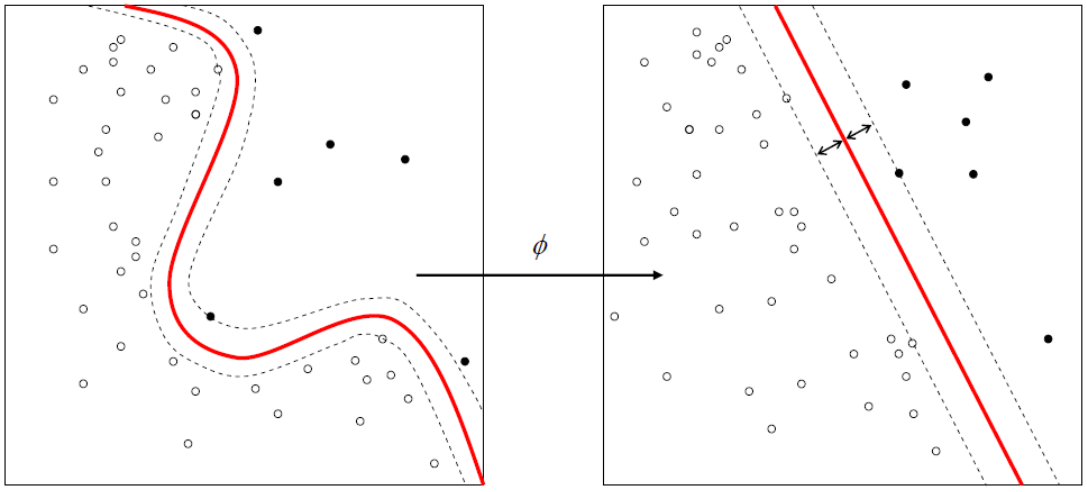
K-En Yakın Komşu Algoritması (K-Nearest Neighbor Algorithm), yeni bir veri geldiğinde var olan öğrenme verisi üzerinde sınıflandırma yapan eğitimli öğrenme algoritmasıdır. Algoritma, yeni bir veri geldiğinde, onun en yakın K komşusuna bakarak bu verinin sınıfına karar verir. Her sınıfın özelliklerinin 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. Manhattan Uzaklık Fonksiyonu, Minkowski Uzaklık Fonksiyonu, Öklid Uzaklık Fonksiyonu içerisinden en çok tercih edilen fonksiyon Öklid uzaklık fonksiyonudur. En yakın mesafe neresi ise yeni veri o sınıfa atanır. Bu çalışmada kullanılan KNN algoritmasında k değeri 10 olarak alınmıştır [50].

4. ÖNERİLEN YAKLAŞIM VE UYGULAMASI (THE PROPOSED APPROACH AND IMPLEMENTATION)

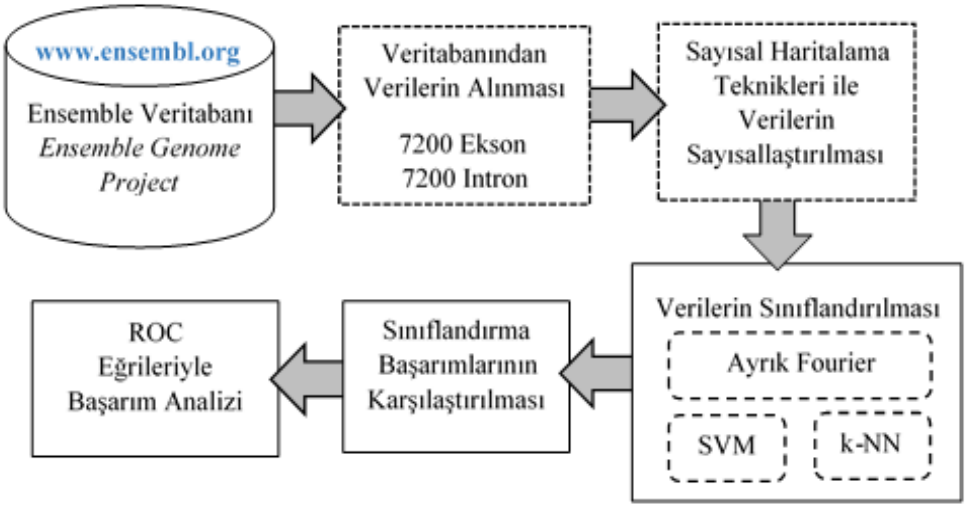
Bu makale çalışmasında, bir DNA diziliminin ekson ve intron olarak sınıflandırılması için Ayrık Fourier tabanlı bir yaklaşım önerilmektedir. Ensembl veri tabanından elde edilen DNA verilerinin sayısal haritalama tekniklerine göre sınıflandırılması ve analizi için uygun olan DVM ve k-NN sınıflandırma yöntemleri de ayrıca kullanılarak, karşılaştırma analizi yapılmıştır. Sınıflandırma için önerilen yaklaşımın uygulama adımları Şekil 4'de verilmiştir.

4.1. Verilerin Elde Edilmesi (Data Acquisition)

Bu makaledeki 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 [55]. Kullanılan veri kümesiyle ilgili literatür taraması yapılmış olup ekson ve intron sınıflandırmaya yönelik bu veri kümesinin kullanıldığı bir makale çalışmasına rastlanılmamıştır. Ayrıca diğer veri tabanlarında bu veri kümesine Blast algoritması uygulanmış ve önemli bir benzerlik bulunmamıştır. Uygulamada 4800 eksona ait baz dizisi eğitim amaçlı, 2400 ekson baz dizisi ise test amaçlı kullanılmıştır. Aynı şekilde 4800 introna ait baz dizisi eğitim amaçlı, 2400 intron baz dizisi test amaçlı kullanılmıştır. Şekil 5'de Ensembl veritabanından alınan



Şekil 3. Destek vektör makine algoritması (Support vector machine algorithm) [52]



Şekil 4. Ekson ve intron sınıflandırma için önerilen yaklaşımın uygulama adımları
(Application steps of the proposed approach for exon and ıntron classification)

eksonlardan bir kesit gösterim biçimi verilmiştir. Eksonlar koyu ve kırmızı renkli olarak gösterilmektedirler. Şekil 6'da ensembl veritabanındaki intronların gösterim biçimi küçük harfle ve mavi olarak gösterilmektedir.

4.2. Sayısal Haritalama Teknikleri ile Verilerin Sayısallaştırılması
(Digitization of Data by Numeric Mapping Techniques)

Bu makalede öncelikle Ensembl veri tabanından elde edilen ekson ve intron dizilimleri dokuz haritalama tekniklerine göre sayısal biçime dönüştürülmüştür. Bu teknikler 2.bölümde anlatılan, Tamsayı Haritalama Tekniği, Reel Haritalama Tekniği, Moleküler Kütle Haritalama Tekniği, DNA Walk Haritalama Tekniği, Çiftli Sayısal Haritalama Tekniği, Karmaşık1 Haritalama Tekniği, Karmaşık2 Haritalama Tekniği ve EIIP Haritalama Tekniği yöntemleridir. Tablo 1'de dokuz farklı sayısal haritalama tekniğinin örnek bir DNA diziliminde uygulanışı gösterilmektedir.

4.3. Verilerin Sınıflandırılması (Classification of Data)

4.3.1. Ayrık Fourier Dönüşümü Yöntemi ile Sınıflandırma
(Classification by Discrete Fourier Transform Method)

Sayısal sinyallere dönüşmüş DNA dizilimlerine Ayrık Fourier metodu uygulanmıştır. Eşitlik 1, 2, 3 uygulanarak ekson ve intron eğitim dizilerinden güç spektrumları bulunmuştur. Aynı şekilde T_3 değeri Eşitlik 5'e göre ekson ve intron eğitim dizilerinden elde edilmiştir. Daha sonra ekson 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 ekson sınıfı olarak nitelendirilir, küçük olması durumunda ise intron sınıfı olarak nitelendirilmektedir. Ayrık Fourier dönüşüm metodunda DNA diziliminin DFT spektrumun bulunmasında 4 farklı pencere kullanılmış ve rectangular, hamming, gaussian, blackman pencerelerinin sınıflandırma performansına etkisi gösterilmiştir. Tablo 2'de 9 haritalama tekniklerinin sınıflandırma performansına etkisi gösterilmiştir. Ekson ve intron sınıflandırma performansı Eşitlik 6 ve Eşitlik 7'e göre hesaplanmıştır. Doğruluk oranı için Eşitlik 8 kullanılmaktadır.

$$\text{Ekson Sınıflandırma}=\frac{\text{Doğru Bulunan Ekson Sayısı}}{\text{Toplam Ekson Sayısı}}\text{x}100\% \quad (6)$$

$$\text{Intron Sınıflandırma}=\frac{\text{Doğru Bulunan Intron Sayısı}}{\text{Toplam Intron Sayısı}}\text{x}100\% \quad (7)$$

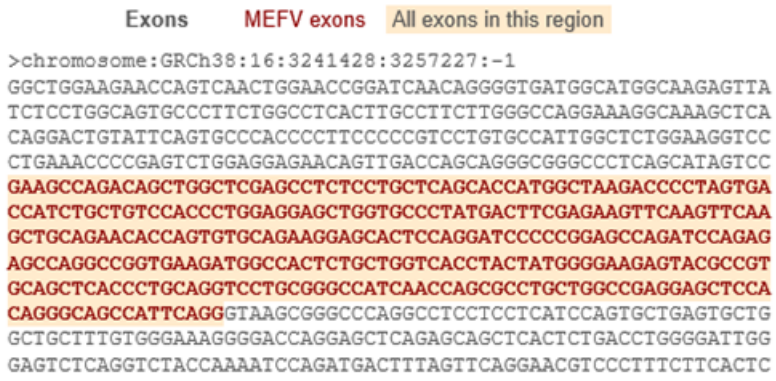
$$\text{Doğruluk}=\frac{\text{Doğru Bulunan Ekson Sayısı+Doğru Bulunan İntron Sayısı}}{\text{Ekson ve İntronların Toplam Sayısı}}\text{x}100\% \quad (8)$$

4.3.2. Destek Vektör Makineleri Yöntemi ile Sınıflandırma
(Classification by SVM Method)

Dokuz farklı haritalama tekniği ile sayısal sinyallere 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 belirlenmiştir.

4.3.3. K-En Yakın Komşu Algoritması ile Sınıflandırma
(Classification by K-NN Algorithm)

Haritalama teknikleri ile sayısal sinyallere dönüşmüş DNA dizilimlerine K-En Yakın Komşu algoritması(K-NN) uygulanmıştır. Bu çalışmada kullanılan K-NN algoritmasında k değeri 10 olarak alınmıştır.



Şekil 5. Ensembl veritabanındaki ekson görüntüsünden bir kesit (A sectional image from the exon in ensembl database)

Şekil 6. Ensembl veritabanındaki intron görüntüsünden bir kesit (A sectional image from the intron in ensembl database)

Tablo 1. DNA sayısal haritalama teknikleri (DNA digital mapping tecniques)

Method	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]
Integer Tekniği	purin(A,G) > purimidin(C,T) ise T=0, C=1, A=2, G=3 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 2 0 2 1 1 0 3 2]
Reel Tekniği	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 -1,5 1,5 0,5 0,5 -1,5 -0,5
Atomik Tekniği	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 78 66 58
Moleculer Tekniği	A=134, G=150, C=110, T=125	150 125 150 125 134 110 110 134 110 134 110 125 125 125 110 125 125 125 134 125 110 110 134 150125
DNA Walk Tekniği	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
Paired Numeric Tekniği	x(n)pürin ise (A ya da G) ui=1 dir, eğer x(n) purimidin ise (C ya da T) ui=-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 1 -1
Complex1 Tekniği	a=1+j, t=1-j, c=-1-j ve g=-1+j dir.	-1-j1-j-1-j1-j1+j-1+j-1+j1+j-1+j1+j-1+j1-j1-j1-j-1+j1-j1-j1-j1+j1-j-1+j-1+j1+j-1-j1-j
Complex2 Tekniği	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
EIIP Tekniği	A= 0,1260 G=0,0806 C=0,1340 T=01335	0,0806 0,1335 0,0806 0,1335 0,1260 0,1340 0,1340 0,1260 0,1340 0,1260 0,1340 0,1335 0,1335 0,1335 0,1340 0,1335 0,1335 0,1335 0,1260 0,1335 0,1340 0,1340 0,1260 0,0806

Tablo 2. Pencere fonksiyonlarına göre sınıflandırma performansları
(Performance classification by windowing functions)

Pencereleme Fonk.	Rectangular Pencere (%)				Hamming Pencere (%)			
Sayısal Haritalama Tekn.	Eşik Değer	Ekson	Intron	Doğruluk	Eşik Değer	Ekson	Intron	Doğruluk
İnteger Tekniği	1.3778e+007	87,10	88,25	87,68	4.6856e+005	92,75	99,66	96,21
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
Molekul 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	87,49
Paired Numerik Tekniği	2.9106e+005	87,80	88,02	87,91	7.8517e+004	96,29	94,87	95,58
Complex1 Tekniği	1.2601e+006	54,54	67,75	61,14	1.4822e+005	25,55	31,38	28,46
Complex2 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.	Gaussion Penceresi (%)				Blackman Penceresi (%)			
Sayısal Haritalama Tekn.	Eşik Değer	Ekson	Intron	Doğruluk	Eşik Değer	Ekson	Intron	Doğruluk
İnteger Tekniği	0.4583(1.0e+006)	86,12	86,38	86,25	4.6323e+005	85,25	87,12	86,18
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
Molekul 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	89,12	8.3740e+004	89,75	91,90	90,82
Paired Numerik Tekniği	8.0288e+004	92,92	91,60	92,26	3.2685e+005	89,00	86,58	87,79
Complex1 Tekniği	1.4493e+005	24,44	29,44	26,94	1.4116e+005	23,05	25,37	24,21
Complex2 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

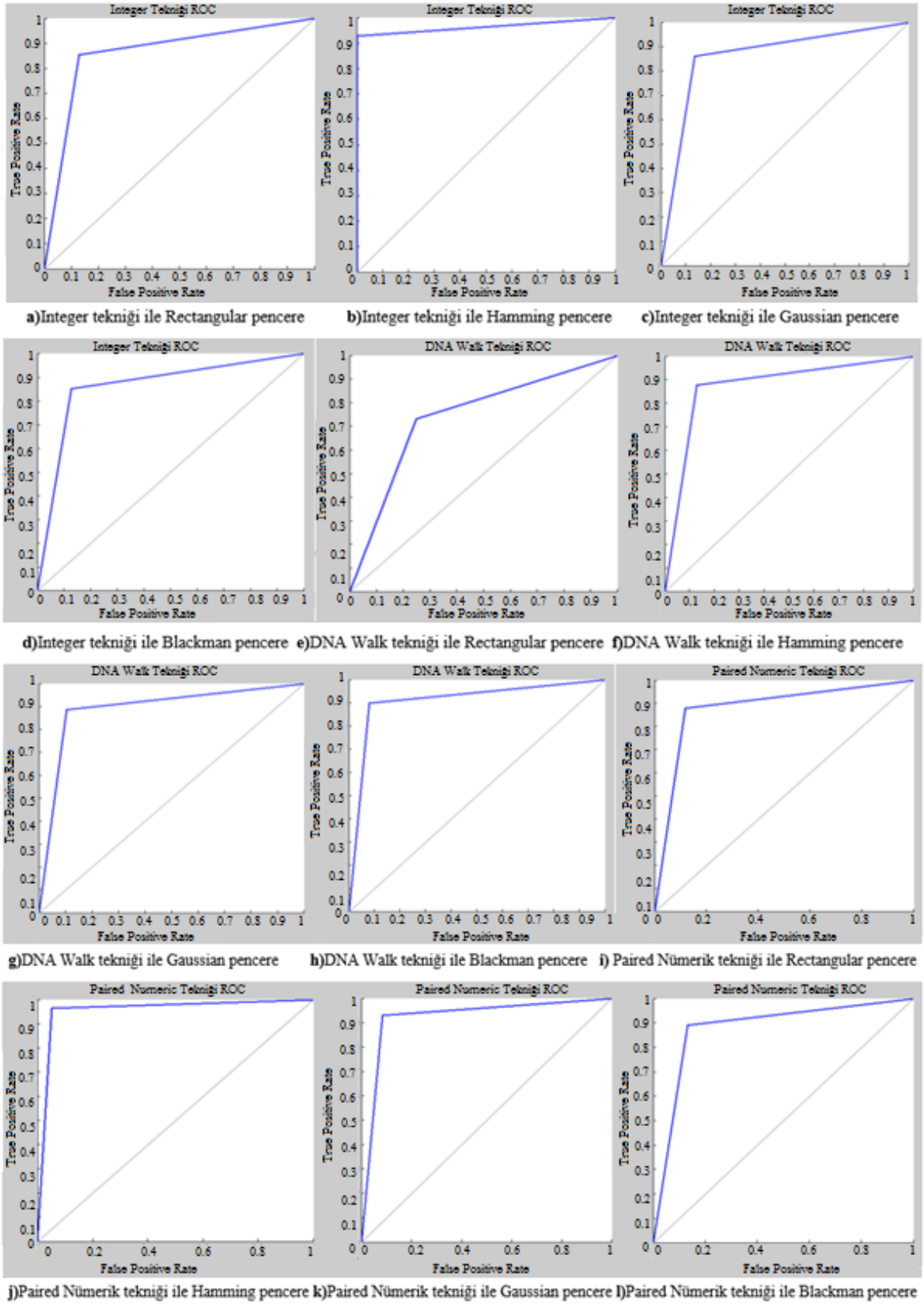
Tablo 3. Sayısal haritalama teknikleri için Fourier tabanlı yöntemin sınıflandırma performansının diğer makine öğrenme tabanlı yöntemlerle karşılaştırılması
(Comparison classification perfrmance of fourier based method with other machine-based methos for digital mapping techniques)

Sayısal Haritalama Teknikleri	Ayrık Fourier Tekniği (%)	Destek Vektör Makinesi (%)	K-En Yakın Komşu Algoritması (%)
Integer 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
Molekül Tekniği [17]	50,075	54,0816	59,88
DNA Walk Tekniği [18]	87,495	66,013	69,56
Paired Numerik Tekniği [19-20]	95,58	71,6326	75,68
Complex1 Tekniği [11]	28,465	52,33	56,13
Complex2 Tekniği [15]	27,985	50,66	54,17
EIIP Tekniği [21-22]	50,37	48,23	50,58

4.4. Sınıflandırma Başarımlarının Karşılaştırılması
(Comparion of Classification Performance)

Önerilen yaklaşım ve gerçekleştirilen uygulamalar ile Pencereleme Fonksiyonuna göre Sayısal Haritalama Tekniklerinin sınıflandırılma başarımlarının analizi yapılmış, deneysel sonuçları ise Tablo 2'de verilmiştir. Tablo 2'de görüldüğü üzere 4 farklı pencereleme fonksiyonuna göre %96,2 sınıflandırma başarımı ile Integer Tekniği ve %95,58 sınıflandırma başarımı ile Paired Numerik Tekniği en yüksek doğruluk oranlarına sahiptir. Diğer taraftan bu iki tekniğin doğruluk oranları, pencereleme fonksiyonlarına göre farklılık göstermektedir. Kullanılan Hamming penceresi fonksiyonunda bu iki teknik en yüksek sınıflandırma başarısı 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 Ayrık Fourier tabanlı yöntemle elde edilen sonuçlar, makine öğrenme tabanlı Destek Vektör Makineleri ve K-En Yakın Komşu algoritması yöntemleriyle elde edilen sonuçlar karşılaştırılmıştır. Tablo 2'de görüldüğü üzere en yüksek sınıflandırma başarısına sahip Hamming pencereleme fonksiyonu için %96,21 başarım oranı ile Integer tekniği, %87,49 başarım oranı ile DNA Walk Tekniği ve %95,58 başarım oranı ile Paired Numerik tekniğinin sınıflandırma başarımları diğer sayısal haritalama tekniklerine göre yüksek çıkmıştır.

Tablo 3'de görüldüğü üzere, DVM ve KNN gibi makine öğrenme tabanlı yöntemlerin ekson ve intron sınıflandırma performanslarının düşük olduğu, buna rağmen Ayrık Fourier Dönüşümü yönteminin ise diğer iki yönteme göre sınıflandırmada güçlü bir metot olduğu açıkça görülmektedir.



Şekil 7. Ekson ve intron sınıflandırmada 3 haritalama tekniğinin performans sonuç ROC eğrileri
(The performance results of ROC curve of 3 mapping techniques in exon and intron classification)

4.5. ROC Eğrileriyle Başarım Analizi
(Performance Analysis with ROC Curves)

Fourier tabanlı yaklaşımın uygulanması sonucuna göre en yüksek başarım oranlarına sahip Integer Tekniği, Paired Numerik Tekniği ve DNA Walk Tekniğinin ROC eğrileri sırasıyla Şekil 7'de gösterilmiştir. Şekil 7'deki ROC eğrilerine bakıldığında Integer tekniğinin Hamming penceresi fonksiyonunda, DNA Walk tekniğinin Blackman penceresi fonksiyonunda, Paired Numerik tekniğinin ise Hamming penceresi fonksiyonunda eğrinin altında kalan alan diğer pencereleme fonksiyonlarına göre daha büyüktür.

5. SONUÇLAR VE TARTIŞMALAR
(RESULTS AND DISCUSSIONS)

Sayısal haritalama teknikleri tarafından sayısallaştırılan DNA dizilimleri literatürde ekson ve intron sınıflandırma, DNA dizilimlerin karşılaştırılması, filo genetik ağaç oluşturulması ve kanser gibi genetik hastalıkların tespiti amacıyla kullanılmaktadır. Makalede kullanılan haritalama teknikleri ise bahsedilen bu amaçlardan biri olan ekson ve intron sınıflandırmada kullanılmıştır. Dolayısıyla kullanılan bu teknikler çalışmanın amacına ve bu amaç doğrultusunda kullanılan yönteme göre farklı performans gösterebilir. Makalede Paired Numeric, Integer ve DNA Walk tekniği DNA dizilimlerinden protein kod bölgelerini (ekson) belirlemek amacıyla Ayrık Fourier yönteminde diğer tekniklerden daha iyi sınıflandırma başarımı göstermiştir. Paired Numerik tekniği, DNA'nın yapısal özelliğini iyi yansıtması ve Ayrık Fourier tekniğinde karmaşıklığı azaltmasından dolayı protein kod bölgelerini bulmada diğer tekniklere göre yüksek performans göstermiştir. Integer 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 baz dizilimlerin kombinasyonlarının değişiminden dolayı (AG-GA veya AGA-GAG şeklinde) diğer sayısal haritalama tekniklerine göre daha yüksek performans sağlamıştır. Bu sayısal haritalama teknikleri DNA dizilimlerinde protein kod bölgelerini belirleme gibi uygulamalarının dışında hastalık teşhisi, DNA dizilimlerinin karşılaştırılmasında farklı performans değerleri gösterebilir.

6. SONUÇLAR (CONCLUSIONS)

Bir DNA diziliminde bulunan genlerin kodlanmayan bölümü intronu oluştururken, kodlanan bölümü ise eksonu oluşturur. Bir DNA diziliminin ekson ve intron olarak sınıflandırılması genetik çalışmalarında oldukça önem taşımaktadır. Bu makale çalışmasında, bir DNA diziliminin ekson ve intron olarak sınıflandırılmasında farklı sayısal haritalama tekniklerinin performanslarını karşılaştırılmıştır. Karşılaştırılma sonucunda insan türünün MEFV genine ait DNA dizilimleri, dokuz farklı haritalama tekniği ile sayısal dizilere dönüştürülmüştür. Dönüştürülen bu sayısal dizileri ekson ve intron olarak sınıflandırmak için Ayrık Fourier Dönüşümü Yöntemi kullanılmıştır. Ayrık Fourier Dönüşümü yönteminde 4 farklı pencereleme fonksiyonu kullanılmış ve bu pencereleme fonksiyonlarının da sınıflandırma başarımı üzerinde etkileri karşılaştırılmıştır. Uygulama sonucunda İnteger Tekniği ve Paired Numerik Tekniği dört pencereleme fonksiyonunda diğer sayısal tekniklere göre daha yüksek sınıflandırma performansı göstermiştir. DNA Walk Tekniği ise Rectangular pencereleme fonksiyonu dışında diğer 3 pencereleme fonksiyonu için yüksek başarım oranları göstermiştir. Hamming Pencereleme Fonksiyonunun sınıflandırma başarısı diğer pencereleme fonksiyonlarından daha yüksek olduğu gözlemlenmiştir. Ayrıca Fourier tabanlı yöntemle elde edilen sonuçlar makine öğrenme tabanlı Destek Vektör Makineleri ve K-En Yakın Komşu Algoritması yöntemleri ile karşılaştırılmıştır. Fourier tabanlı yöntemin diğer iki yönteme göre sınıflandırma başarımı oldukça yüksek çıkmıştır. Gelecek çalışmalarda ekson ve intronları belirlemede sınıflandırma performansını arttırabilmek için yeni bir sayısal haritalama tekniği geliştirilecektir ve geliştirilen bu tekniğin performansı varolan sayısal haritalama tekniklerinin sınıflandırma performansıyla karşılaştırılacaktır. Ayrıca DNA dizilimlerinden hastalık tespiti ile ilgili çalışmalarda geliştirilecek bu yeni sayısal haritalama tekniği kullanılacaktır.

KAYNAKLAR (REFERENCES)

1. Internet: http://schoolworkpelher.net/dna-mrna-introns-and-exons, Erişim Tarihi: 01.01.2015.
2. Kwan J.Y.Y., Kwan B.Y.M., Kwan H.K., Spectral Analysis of Numerical Exon and Intron Sequences, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine Workshops, Hong Kong, 876-877, 2010.
3. Marhon S.A., Kremer S.J., A dynamic representation-based, de novomethod for protein-coding region prediction and biological information detection, Elsevier, Digital Signal Processing 46, 10–18, 2015.
4. Zhang J., Yang C., DNA Sequence Recognition Based on the Markov Model, 6th International Conference on Biomedical Engineering and Informatics (BMEI 2013), 2013.
5. Mandal S.B., Saha S., Mandal A., Roy M., Prediction of Protein Coding Regions of a DNA Sequence through Spectral Analysis, IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision, 2012.
6. Xia J., Caragea D., Brown S.J, Prediction of Alternatively Spliced Exons Using Support Vector Machines, Int. J. Data Mining and Bioinformatics, 4 (4), 411-30, 2010.
7. Dror G., Sorek R., Shamir R., Accurate Identification of Alternatively Spliced Exons Using Support Vector Machine, Bioinformatics, 21 (7), 897-901, 2005.
8. Barman S., Saha S., Mandal A., Roy M., Prediction of protein coding regions of a DNA sequence through spectral analysis, Informatics, Electronics & Vision (ICIEV), 2012 International Conference, 18-19 May 2012.

9. Cristea P.D., Genetic Signal Representation and Analysis, SPIE International Conference on Biomedical Optics Symposium, 4623, 77–84, 2002.

10. Chakravarthy N., Spanias A., Lasemidis L.D., Tsakalis K., Autoregressive Modeling and Feature Analysis of DNA Sequences, EURASIP Journal of Genomic Signal Processing, 1,13-28, January 2004.

11. Cristea P.D., Genomic Signals of Reoriented ORFs**,** EURASIP J. Appl. Signal Process., 1, 132-137, 2004.

12. Berger J.A., Mitra S.K., Carli M., Neri A., New Approaches to Genome Sequence Analysis Based on Digital Signal Processing, IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), 1-4, October 2002.

13. Cristea P.D., Conversion of Nucleotides Sequences Into Genomic Signals, [J]. Cell. Mol. Med, 6, 279-303, April-June, 2002.

14. Dougherty E.R., Hmulevich I., Chen J., Wang Z.J., Genomic Signal Processing and Statistics, EURASIP Book Series in Signal Processing and Communications, Hindawi Pub. Corp, ISBN 977-5945-07-0, 2, 15-66, 2005.

15. Andersson J.D., Doolittle W.F., Nesbo C.L., Are There Bugs in Our Genome?, Science, 292, 1848-1850, 2001.

16. Todd Holden R., Subramaniam R., Sullivan E., Cheng C., Sneider G., Tremberger J.A., Flamholz, D. H., Leiberman, and Cheung, T. D., ATCG Nucleotide Fluctuation of Deinococcus Radiodurans Radiation Genes, Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE), 669417, 1-10, August 2007.

17. Buldyrev S.V., Dokholyan N.V., Goldberger A.L., Havlin S., Peng C.K., Stanley H.E., Viswanathan G.M., Analysis of DNA Sequences Using Methods of Statistical Physics, Physica A, Elsevier, 249, 430-438, 1998.

18. Berger J.A., Mitra S.K., Carli M., Neri A., Visualization and Analysis of DNA Sequences Using DNA Walks, Journal of the Franklin Institute, 341, 37-53, January-March 2004.

19. Buldyrev S.V., Goldberger A.L., Havlin S., Stanley H.E., Long-Range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis, Phy. Rev. E, 51 (5), 5084-5091, May 1995.

20. Akhtar M., Epps J., Ambikairajah E., Paired Spectral Content Measure for Gene and Exon Prediction in Eukaryotes, International Conference on Information and Emerging Technologies, ICIET 07, 1- 4, July 2007.

21. Nair A.S., Pillai S.S., A Coding Measure Scheme Employing Electron-Ion Interaction Pseudo Potential (EIIP), Journal of Bio-information, 1, 197–202, October, 2006.

22. Chakraborty S., Gupta V., DWT Based Cancer Identification Using EIIP, 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), 12-13 February 2016.

23. Yee Kwan J.Y., Ming Kwan B.Y., Keung Kwan H., Spectral Analysis of Numerical Exon ve Intron Sequences, 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops, 2010.

24. Shakya D.K., Saxena R., Sharma S.N., An Adaptive Window Length Strategy for Eukaryotic CDS Prediction, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10, 1241 – 1252, 2 July 2013.

25. Datta S., Asif A., A Fast DFT Based Gene Prediction Algorithm For Identification of Protein Coding Regions, ICASSP, 5, 653–656, 2005.

26. Internet: Başkent Üniversitesi, http://www.baskent.edu.tr/~aerdamar/LAB1.pdf Erişim Tarihi: 01.01.2015.

27. Internet: İstanbul Teknik Üniversitesi, http://web.itu.edu.tr/~baykut/lab/pdf/Deney_3.pdf, Erişim Tarihi: 01.01.2015.

28. Saberkari H., Shamsi M., Sedaaghi M., Golabi F., Prediction of protein coding regions in DNA sequences using signal processing methods, Industrial Electronics and Applications (ISIEA), 2012 IEEE Symposium on, 23-26 September 2012.

29. Ramachandran P., Lu W.S., Antoniou A., Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA, IEEE Transactions on Biomedical Engineering, 59, 1598-1609, June 2012.

30. Oppenheim A.V., Schafer R.W., Discrete Time Signal Processing, Prentice Hall, New Jersey, 1989.

31. Söderström T., Stoica P., System Identification, Prentice Hall, Cambridge, 1989.

32. Kayran A.H., Sayısal İşaret İşleme, İstanbul Teknik Üniversitesi, 1990.

33. Proakis J.G., Manolakis D.G., Digital Signal Processing, Prentice Hall, New Jersey, 1996.

34. Avcı K., Kaiser-Hamming Window and Its Performance Analysis For Nonrecursive Digital Filter Design, Journal of the Faculty of Engineering and Architecture of Gazi University, 29 (4), 823-833, 2014.

35. Kaya T., İnce M.C., Design of FIR Filter Using Modeled Window Function With Helping of Artificial Neural Networks, Journal of the Faculty of Engineering and Architecture of Gazi University, 27 (3), 599-606, 2012.

36. Karaarslan A., İskender İ., A Novel Method in Power Factor Correction Circuits Using Average Current Control Technique and Digital Signal Processor, Journal of the Faculty of Engineering and Architecture of Gazi University, 26 (1), 193-203, 2011.

37. Abo-Zahhad M., Ahmed S.M., Abd-Elrahman A.S., Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques, International Journal Information Technology and Computer Science, 8, 22-36, 2012.

38. Hota M.K., Srivastava V.K., DSP Technique for Gene and Exon Prediction Taking Complex Indicator Sequence, Proc. IEEE TENCON, 1-6, 2008.

39. Sahu S., Panda G. Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach, Genomic Proteomics&Bioinformatics, October 2010.

40. Hota M., Srivastava V., Identification of Protein Coding Regions Using Antinotch Filter, Digital Signal Processing, 22, 869-877, June, 2012.

41. Vaidyanathan P.P., Yoon B.J., The Role of Signal-Processing Concepts in Genomics and Proteomics, J. Franklin Inst. 341, 111–135, 2004.

42. Vaidyanathan P.P., Yoon B.J., Gene and Exon Prediction Using Allpass-Based Flters, Workshop on Genomic Signal Process. Stat., Raleigh, NC, 2002.

43. Mena-Chalco J., Carrer H., Zana Y., Cesar R.M., Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform, IEEE/ACM Trans. Comput.Biol. Bioinformatic., 5, 198–207, 2008.

44. Kotlar D., Levner Y., Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein-Coding Regions, Genome Res., 13, 1930–1937, 2003.

45. Ramachandran P., Lu W.S., Antoniou A., Location of Exons in DNA Sequences Using Digital Flters, Proceedings of IEEE, 2337–2340, 2009.

46. Akhtar M., Epps J., Ambikairajah E., Time and Frequency Domain Methods for Gene and Exon Prediction in Eukaryotes, Proc. IEEE ICASSP, 573–576, 2007.

47. Kwan H.K., Arniker S.B., Numerical Representation of DNA Sequences IEEE Inter, Conf. on Electro/Information Technology, EIT '09, Windsor, 307-310, 2009.

48. Cristea P.D., Representation and analysis of DNA sequences. in Genomic signal Processing and Statistics, EURASIP Book Series in Signal Processing and Communications, (Eds) Edward R. Dougherty et al Hindawi Pub., 2, 15-66, 2005.

49. Kwan J.Y.Y., Kwan B.Y.M., Kwan H.K., Novel Methodologies for Spectral Classification of Exon and Intron Sequences, EURASIP Journal on Advances in Signal Processing, 2012.

50. Das B., Türkoğlu İ., DNA Dizilimlerindeki Nükleotid Çiftlerinin Frekans Değerlerine Göre Farklı Sınıflandırma Yöntemleri ile Karşılaştırılması, Tıp Teknolojileri Ulusal Kongresi, 2014.

51. Law N.F., Cheng K., Siu W., On Relationship of Z-Curve and Fourier Approaches for DNA Coding Sequence Classification, Bioinformation, 242-246, 2006.

52. Akhtar M., Epps J., Ambikairajah E., On DNA Numerical Representations for Period-3 Based Exon Prediction, IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), 1-4, June 2007.

53. Saberkari H., Shamsi M., Sedaaghi M.H., Golabi H., Prediction of protein coding regions in DNA sequences using signal processing methods, IEEE Symposium on Industrial Electronics and Aplications (ISIEA), September 23-26, Bandung Indonesia, 2012.

54. Zhang L., Tian F., Wang S., A Modified Statistically Optimal Null Filter Method for Recognizing Protein-coding Regions, SciVerse ScienceDirect, Genomics Proteomics Bioinformatics 10, 166–173, 2012.

55. Ensembl Genbankası veritabanı, online erişim: http://www.ensembl.org

56. Yücesoy E., Nabiev V., Determination 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.

57. Sengur A., Multiclass Least-Squares Support Vector Machines for Analog Modulation Classification, Expert Systems with Applications, 36 (3), 6681-6685, 2009.

58. 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.

59. 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.

60. Kwan B., YM., Kwan J., YY., Kwan H.K., Spectral Classification of Short Numerical Exon and Intron Sequences, BMC Bioinformatics, DOI: 10.1186/1471-2105-12-S11-A13, 2011.