Journal of the Faculty of Engineering and Architecture of Gazi University 37:3 (2022) 1355-1371

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

Classification of exon and intron regions obtained using digital signal processing techniques on the DNA genome sequencing with EfficientNetB7 architecture

Fatma Akalın[1]*, Nejat Yumuşak[2]

[1]Information Systems Engineering Department, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, 54187, Turkey
[2]Computer Engineering Department, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, 54187, Turkey

Highlights:

- Impact of windowing functions on performance
- Evaluating success of numerical mapping techniques
- Classification of exon and intron regions on DNA genome sequences by using EfficientNetB7 architecture

Keywords:

- DNA sequences
- Numerical mapping techniques
- Digital signal processing methods
- Transfer learning
- EfficientNetB7 architecture

Article Info:

Research Article
Received: 22.03.2021
Accepted: 02.10.2021

DOI:

10.17341/gazimmfd.900987

Correspondence:

Author:Fatma Akalın
e-mail: fatmaakalin@sakarya.edu.tr
phone: +90 264 295 6450

Graphical/Tabular Abstract

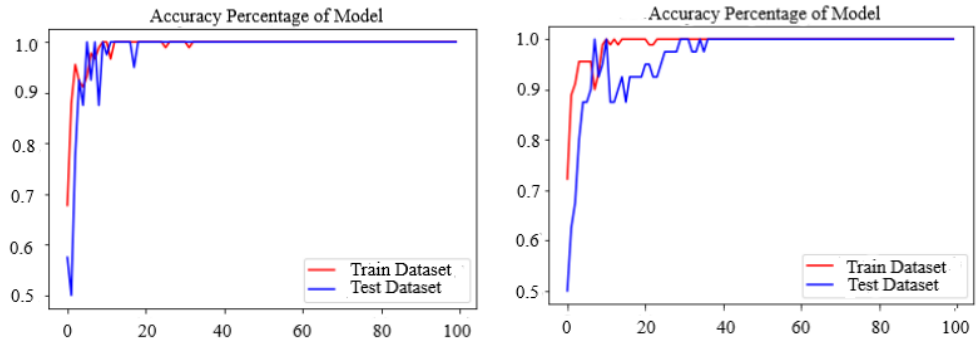


Figure A. The success rate obtained as a result of 100 epochs in the classification scope performed with the EfficientNet B7 architecture

Purpose: To provide a classification of exon and intron labelled images obtained using digital signal processing techniques on the DNA genome sequence.

Theory and Methods:
In this study, the gene sequences numbered EU447303.1, AM400881.1, AM886138.1, AM600680.1 with different nucleotide lengths obtained from the NCBI database was examined. Firstly, DNA sequences that consist of symbolic characters were digitized using numerical mapping techniques. Then, digitized DNA sequences were expressed in the frequency domain by the windowed Fourier transform method and regions tagged as exon and intron were classified by transfer learning architectures. EfficientNetB7 architecture that located of transfer learning architectures presented maximum success by hydrogen bond energy rule of paired mapping technique. Secondly, the DNA sequences digitized using the hydrogen bond energy rule of paired mapping technique were expressed with spectrograms labelled as exon and intron regions by the windowed short-term Fourier transform method. Subsequently, spectrograms classified by EfficientNetB7 architecture produced the most successful results with a %100 success rate on the BCR-ABL genes.

Results:
This article was conducted to differentiate exon and intron regions on the BCR-ABL gene obtained from the NCBI database. As a result of the classification carried out with EfficientNetB7 architecture, which is a transfer learning technique, the success rate was obtained as %100.

Conclusion:
A successful accuracy rate and a low loss value have been obtained in the scope of the studies performed on the BCR-ABL gene with numerical mapping techniques, digital signal processing methods and transfer learning techniques.

Journal of the Faculty of Engineering and Architecture of Gazi University 37:3 (2022) 1355-1371

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

DNA genom dizilimi üzerinde sayısal sinyal işleme teknikleri kullanılarak elde edilen ekson ve intron bölgelerinin EfficientNetB7 mimarisi ile sınıflandırılması

Fatma Akalın[1]*, Nejat Yumuşak[2]

[1]Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilişim Sistemleri Mühendisliği Bölümü, 54187 Serdivan, Sakarya, Türkiye
[2]Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, 54187 Serdivan, Sakarya, Türkiye

ÖNEÇIKANLAR

- Performans üzerinde pencereleme fonksiyonlarının etkisi
- Sayısal haritalama tekniklerinin başarısının değerlendirilmesi
- EfficientNetB7 mimarisi kullanılarak DNA genom dizilimleri üzerindeki ekson ve intron bölgelerinin sınıflandırılması

Makale Bilgileri

Araştırma Makalesi
Geliş: 22.03.2021
Kabul: 02.10.2021

DOI:
10.17341/gazimmfd.900987

Anahtar Kelimeler:
Sayısal haritalama teknikleri, sayısal sinyal işleme metotları, öğrenme aktarımı, efficientNetB7 mimarisi

ÖZ

Organizmayı inşa etmek ve canlılığını sürdürmek için yüksek miktarda bilgi barındıran DNA, önemli bir biyobelirteçtir. A,T,G ve C harflerinden oluşan sembolik bir dizilime sahip olan DNA genom parçası, protein üreten(ekson) ve protein üretmeyen(intron) kısımlardan meydana gelmektedir. Bu bölgelerin tanımlanması; kanserin gelişme durumunun incelenmesi, ilgili gen bölgelerinde mutasyonun gerçekleşip gerçekleşmediğinin izlenmesi ya da organizmanın büyüme ve gelişme durumlarının düzenlenmesi gibi farklı konuların aydınlatılmasında önemli bir role sahiptir. Bu kapsamda bilgisayar destekli sistemler ile ekson ve intron bölgelerinin doğru bir şekilde ayırt edilmesi hedeflenmiştir. Çalışmanın ilk aşamasında, farklı sayısal haritalama teknikleri ile sayısallaştırılan sembolik DNA dizilimleri üzerinde en başarılı sayısal haritalama tekniğine performans ölçütleri ile karar verilmiştir. Ardından ilk kısımda seçilen haritalama tekniği kullanılarak sayısallaştırılan DNA dizilimlerinin spektrogram olarak ifade edilmesi sağlanmıştır. Zamanla değişen bir sinyalin frekans spektrumunun görsel bir temsili olan spektrogramlar ekson ve intron bölgeleri olarak etiketlendikten sonra öğrenme aktarımı olan EfficientNetB7 mimarisi ile sınıflandırılmıştır ve sınıflandırma başarım oranı %100 olarak elde edilmiştir.

Classification of exon and intron regions obtained using digital signal processing techniques on the DNA genome sequencing with EfficientNetB7 architecture

HIGHLIGHTS

- Impact of windowing functions on performance
- Evaluating success of numerical mapping techniques
- Classification of exon and intron regions on DNA genome sequences by using EfficientNetB7 architecture

Article Info

Research Article
Received: 22.03.2021
Accepted: 02.10.2021

DOI:
10.17341/gazimmfd.900987

Keywords:
Numerical mapping techniques, digital signal processing methods, transfer learning, efficientNetB7 architecture

ABSTRACT

DNA is an important biomarker, containing enormous information for building the organism and maintaining its viability. DNA genome fragment with a symbolic sequence consisting of the letters A, T, G and C consists of protein-coding (exon) and non-coding (intron) parts. Identification of these regions plays an important role in different enlightening issues such as examining the development status of cancer, monitoring whether mutations occur in the relevant gene regions or regulating the growth and development of the organism. In this scope, it is aimed to distinguish the exon and intron regions correctly by computer-aided systems. In the first stage of the study, the most successful digital mapping technique on symbolic DNA sequences digitized with different numerical mapping techniques was decided by performance criteria. Then, the digitized DNA sequences using the mapping technique selected in the first part were expressed as spectrograms. Spectrograms, which are a visual representation of the frequency spectrum of a signal that changes over time, were labelled as exon and intron regions then were classified using the EfficientNetB7 model, a transfer learning architecture. At the end of the classification process, the success rate was obtained as %100.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *fatmaakalin@sakarya.edu.tr, nyumusak@sakarya.edu.tr / Tel: +90 264 295 6450

1. GİRİŞ (INTRODUCTION)

DNA (Deoxyrıbo Nucleic Acid) organizmayı inşa eden ve canlılığın sürdürülmesi için yüksek miktarda bilgi barındıran hücre parçasıdır [1]. Adenin, Timin, Guanin ve Sitozin nükleotitleri DNA genom dizisinin sembolik hale getirilmesinde kullanılan temel bileşenlerdir. Bu nükleotitler kullanılarak gen, protein, RNA ve DNA gibi çeşitli biyobelirteçler elde edilir [2].
Önemli bir biyobelirteç olan ve 4 temel nükleotitten meydana gelen DNA genom dizisi, genler ve genler arası boşluklardan oluşmaktadır. Protein sentezinden sorumlu olan gen bölgesi, ökaryot(cekirdekli) hücreler için protein kodlayan(ekson) ve protein kodlamayan(intron) bölgelerin ayrımına sahiptir [3].

Ökaryot hücrelerde protein kodlayan bölgeler, protein sentezi ile ilgili tüm bilgiyi barındıran kısımlardır. Bu bölgelerin doğru bir şekilde tanımlanabilmesi; büyümenin, gelişmenin nerede, nasıl ve ne zaman düzenleneceğini; hücrelerin çoğalma ve ölme durumlarını; kök hücrelerin nerede, nasıl bir değişiklik yaşayacağını; kanserin gelişme durumunun incelenmesini; ilgili genin mutasyon geçirip geçirmediğinin araştırılmasını; canlı organizmalarda protein oluşumundan sorumlu kodun belirlenebilmesini ve protein kodlanma süreçlerindeki değişikliklerin aydınlatılabilmesini sağlamada önemli bir rol oynamaktadır[4]. Buna karşın yaşam sürecini değerlendirmek için ayırt edici noktalardan biri olan DNA dizilimindeki ekson bölgelerinin tanımlanabilmesi, araştırmacılar tarafından zorlu bir problem olarak belirtilmiştir [1, 5, 6].

Genom teknolojisindeki son gelişmeler ile birlikte dizilim sayıları üzerindeki artış bu dizilimlerin doğru ve hızlı bir şekilde sonuçlanabilmesi için büyük bir problem olmuştur. Aynı zamanda ökaryotik hücrelerin gen bölgelerindeki karmaşık ekson ve intron yapılarının analizi, bilgisayar destekli sistemlerin ihtiyacını ortaya koymaktadır [7].

Bu kapsamda bilişim dünyasında 2000'li yıllardan bu yana genomik veriler üzerinde hesaplamalı ve analitik olmak üzere birçok farklı çalışma gerçekleştirilmiştir. DNA dizilimlerinin özelliklerinin analizi, genetik hastalıkların sınıflandırılması ve teşhisi gibi farklı açılardan değerlendirme süreçleri araştırmalarda mevcut olan konular içerisinde yer almıştır [8].

Diğer taraftan, bahsedilen çalışmaların yanı sıra yapılan incelemeler çerçevesinde genomik sinyal işleme (GSİ) farklı bir alan olarak ortaya çıkmıştır. Bu alan doğrultusunda yaşanan gelişmeler ile birlikte DNA parçaları içerisindeki gizli özelliklerin ve periyodiklik durumlarının ortaya çıkarılabilmesi amacıyla sayısal sinyal işleme (SSİ) yöntemlerinde gelişmeler yaşanmıştır [8, 9]. Gelişmeler ile birlikte DNA dizilimleri üzerinde sinyal işleme yöntemlerinden yararlanılarak çeşitli incelemeler yapılmıştır. Bu bağlamda ekson bölgelerindeki mevcut özelliklerin keşfedilmesinde ayrık zamanlı fourier dönüşümü (AZFD) yöntemini kullanan ilk kişilerden biri olan Tiwari ve ark. [10] genomik dizilerin periyodikliğinin ayrık fourier dönüşümünün spektral özelliği ile belirtilmesini sağlamıştır. Zaman içerisinde AZFD yöntemi için pencere yaklaşımları üzerinde temellenen değişiklikler de önerilmiştir. Böylelikle performans ve CPU zamanının iyileştirilmesi amaçlanmıştır. Öte yandan bilim dünyası farklı bir yöntem olan kısa zamanlı fourier dönüşümü (KZFD) ile ekson bölgelerinin tespitinde doğruluğun arttırılması amacıyla çeşitli çalışmalar da yapmıştır [8, 11]. Bu yöntem sayesinde zaman değerlerine karşılık gelen frekans bilgilerinin sağlanamamasından dolayı durağan olmayan serilerin analizinde kullanımı tatmin edici sonuçlar vermeyen AZFD yönteminin eksikliğinin giderilmesi sağlanmıştır [12]. Sunulan bazı araştırmalarda ise DNA dizilimleri üzerinde yapılan çalışmaların performansını arttırmak amacı ile filtre tasarımlarının gerçekleştirilmesi önerilmiştir [8].

Sayısal sinyal isleme süreçlerinin gerçekleştirildiği çalışmalardan esinlenerek oluşturulan bu makalede, biyolojik bilginin elde edilmesi ve ardından DNA zincirinin kodlanan ve kodlanmayan bölgelerindeki güç spektral yoğunluğunun tahmin edilmek suretiyle değerlendirilmesi amaçlanmıştır. Böylece DNA dizilimi üzerindeki protein kodlayan bölgelerin konumunun belirlenmesi ile genetik hastalıklar, kanser ve incelenen genin mutasyon geçirip geçirmediğinin bulunması gibi araştırılması istenen farklı durumların teşhis ve tedavisinde destek sağlanılması hedeflenmiştir [13-15]. Çünkü ekson bölgelerinin ayırt edilmesi için sinyal işleme yaklaşımına kıyasla geleneksel metotlarının kullanımı çok fazla işlem yükü oluşturmaktadır [5]. Aynı zamanda DNA örneklerinin kültür ortamı üzerinde değerlendirilmesi yaklaşık 3-4 hafta sürebildiği için bu alanda bilgisayar destekli bir çalışmanın gerekliliği kendini göstermiştir [5].

Nükleotit dizilimlerindeki bölgelerin tanımlanabilmesi için 2 farklı yaklaşım bulunmaktadır. İlk yaklaşımda aday eksonları tahmin etmek amacıyla desen eşleşmesi ve istatistiksel metotlar üzerinde temellenen programlar kullanılmıştır. Ancak bu programların çok sayıda yanlış ekson tahmini yaptığı ifade edilmiştir [16]. Diğer yaklaşımda ise nüklotit dizilimlerinin benzerliği üzerinde temellenen programlardan faydalanılmıştır. Bu program ile de gen tahmini üzerinde büyük başarılar sağlanılmasına karşın yetersiz temsil edilen transkriptleri ve doku aralığını tanımlamadaki güçlüğü belirtilmiştir [16].

Bu çalışmada NCBI (National Center for Biotechnology Information) genom veri kümesinden insana ait olan AM600680.1, AM886138.1, AM400881.1 ve EU447303.1 numaralarına sahip BCR-ABL genleri üzerinde ekson ve intron bölgelerinin sınıflandırılması hedeflenmiştir.

Bu doğrultuda çalışmanın birinci aşamasında, sembollerden oluşan DNA dizilimleri üzerinde işlem yapılabilmesi için ilk olarak farklı sayısal haritalama teknikleri ile DNA

dizilimlerinin sayısallaştırılması sağlanmıştır. Ardından pencerelenmiş fourier dönüşümü yöntemi ile ekson ve intron bölgelerinin frekans alanındaki karşılığı olan ekson ve intron etiketli spektrogramlar elde edilmiştir. Son olarak yapay zeka çatısı altında yer alan farklı öğrenme aktarım yöntemleri ile görüntülerin sınıflandırma sonuçları karşılaştırılmış ve başarıları değerlendirilmiştir. Bu bağlamda farklı nükleotit uzunluklarına sahip olan BCR-ABL genleri üzerinde, eşleştirilmiş sayısal haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılan DNA dizilimlerinin diğer sayısal haritalama tekniklerine göre daha güçlü olduğuna pencerelenmiş ayrık fourier dönüşümü tekniği ile karar verilmiştir. Dolayısıyla çalışmanın ikinci aşamasında eşleştirilmiş sayısal haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılan DNA dizilimleri pencerelenmiş kısa zamanlı fourier dönüşümü yöntemi kullanılarak ekson ve intron bölgeleri çerçevesinde zamanla değişen bir sinyalin frekans spektrumunun görsel temsili olan spektrogramlar şeklinde ifade edilmiştir [13]. Bu süreçte çeşitli öğrenme aktarımı mimarileri arasında gerçekleştirilen sınıflandırma sonucunda en başarılı mimari olarak değerlendirilen EfficientNetB7 ile ekson ve intron bölgeleri şeklinde etiketli spektrogramların sınıflandırılması sağlanmıştır.

Sunulan araştırmada önerilen yöntemin basit bir yaklaşım sunmasından dolayı gelecek çalışmalarda da tercih edilmesi beklenmektedir.

2. İLGİLİ ÇALIŞMALAR (RELATED WORKS)

DNA dizilimini meydana getiren gen bölgesi, ekson ve intron parçalarını barındıran biyolojik bir bileşendir. Bu bileşenlerin doğru bir şekilde tanımlanabilmesi canlı organizmada gerçekleşen ya da gerçekleşecek birçok olayın aydınlatılabilmesi açısından önem taşımaktadır.

Genomik veriler üzerinde sağlanan hesaplamalı ve analitik çalışmaların yanı sıra sinyal işleme yaklaşımlarının da zaman içerisinde ön plana çıktığı bilinmektedir [8]. Bu kapsamda son yıllarda ekson ve intron bölgelerinin doğru bir şekilde ayırt edilebilmesi amacı ile farklı çalışmalar yapılmıştır. [17] çalışmasında ökaryotik bir hücrede ekson ve intron bölgelerinin tanımlanması amacı ile yeni bir sayısal haritalama tekniği önerilmiştir. Önerilen dairesel haritalama (circular mapping) yaklaşımı ile her bir nükleotidin türüne ve kodonlardaki pozisyonuna bağlı olarak karmaşık sayı değeri ile eşleşen ilgili DNA diziliminin grafiksel bir temsili sunulmuştur. Bu çalışma ile ekson bölgelerinin tanınması hususunda tatmin edici sonuçlar elde edildiği ifade edilmiştir. [4] çalışmasında insan türünün MEFV genine ait DNA dizilimi değerlendirilmiştir. Sayısal haritalama teknikleri ile sayısallaştırılan nükleotit dizilimleri KNN(k-en yakın komşu) ve destek vektör makineleri algoritmalarının yanı sıra ayrık fourier dönüşümü metodu ile sınıflandırılmıştır. Ekson ve intron bölgelerinin tanınmasında yüksek bir başarı yüzdesinin elde edildiği belirtilmiştir. [3] çalışmasında ekson bölgelerinin tanımlanması için sayısallaştırılan DNA dizilimlerine kısa zamanlı ayrık fourier dönüşümü yöntemi uygulanmıştır. Konik tabanlı pencereleme yöntemi kapsamında voss, z eğrisi ve tedrahedron teknikleri ile başarılı bir çalışma olduğu ifade edilmiştir. [18] çalışmasında DNA dizilimini sayısallaştırmak için yeni bir haritalama tekniği önerilmiştir. Shannon denkleminin geliştirilmiş fraksiyonel türevi ile eşleşen ve aynı zamanda her bir kodona karşılık gelen bu metot ile tekil değer ayrıştırma, ayrık fourier dönüşümü, kısa zamanlı fourier dönüşümü yöntemleri kapsamında sınıflandırma işlemi gerçekleştirilmiştir. Elde edilen sonuçların mevcut haritalama tekniklerinden daha başarılı olduğu açıklanmıştır. [19] çalışmasında nükleotit dizilimleri üzerinden istatistiksel bilginin çıkarılması amacı ile dalgacık tabanlı zaman serisi yaklaşımı önerilmiştir. Bu kapsamda DNA diziliminden çeşitli bilgiler doğrultusunda elde edilen varyans bilgisi ile özellik vektörünün oluşturulması sağlanmıştır. Sonrasında ekson ve intron bölgelerinin sınıflandırılması amacı ile desen tanıma çerçevesi uygulanmıştır. Yapılan deneysel işlemler sonucunda uygun bir oranın elde edildiği belirtilmiştir. [15] çalışmasında farklı insan genlerinden elde edilen DNA dizilimlerinin güçlü ve zayıf hidrojen bağlarına sırasıyla 3 ve 2 numerik değerleri atanmıştır. Ardından sayısallaştırılan DNA dizilimlerine sayısal sinyal işleme süreci uygulanılarak elde edilen sonuçlar tartışılmıştır.

[20] çalışmasında protein kodlama bölgeleri için farklı bir teknik önerilmiştir. Bu teknik ile DNA dizilimindeki her bir nükleotitin dağılımı üzerinden varyans hesabı yapılarak ekson ve intron bölgelerinin ayrımının yapılabilmesi amacı ile yeni bir eşleme süreci sunulmuştur. Bu çalışmanın iyi bir performans gösterdiği belirtilmiştir. [21] çalışmasında ekson ve intron bölgelerini ayırt etmek amacı ile topolojik entropi hesaplaması, genomik sinyal işleme ve tekil değer ayrıştırması yöntemlerinin kullanıldığı bir yaklaşım önerilmiştir. Bu teknik sayesinde farklı türler arasında ayırt edici sonuçların elde edildiği ifade edilmiştir. [22] çalışmasında EIIP (Electron Ion Interaction Pseudo Potentials) yöntemi ile sayısallaştırılan DNA dizilimlerinden ekson bölgelerinin tahmini için ayrık fourier dönüşümüne dayalı spektral tahmin teknikleri uygulanılmıştır. Sunulan yöntemlerin ekson bölgelerini temsil eden pigler için saptanılabilirlik oranını arttırdığı açıklanmıştır. [5] çalışmasında, DNA'nın doğasında mevcut olan bulanık konfigürasyonlar nedeni ile uygulanılan gabor dalgacık dönüşümü yönteminin ölçeklendirme faktörünün bulanık kurala uyarlanması sağlanmıştır. Çalışmanın bulanık tabanlı öğrenme ile ekson bölgelerinin ayırt edilebilmesinde avantajlı bir konumda olduğu belirtilmiştir. [23] çalışmasında, 3 bazlı periyodiklik kavramının ekson bölgelerinin tespiti hususunda etkisi incelenmiştir. Ancak 3 bazlı periyodiklik özelliğinin kısa uzunluktaki ekson bölgelerinin belirginliğinde yeterli olmadığı ifade edilmiştir. [22] çalışmasında ekson bölgelerini temsil eden tepe noktaların saptanılabilirliğini iyileştirmek amacı ile AZFD yöntemine dayalı spektral tahmin teknikleri arasında karşılaştırmalı bir çalışma sunulmuştur. EIIP haritalama tekniği ile sayısallaştırılan DNA dizilimlerinin analizi için kullanışlı bir yazılım paketi geliştirildiği belirtilmiştir. [5]

çalışmasında DNA konfigürasyonundaki davranışların belirsiz olması nedeni ile bulanık yaklaşım üzerinde temellenen bir çalışma yapılmıştır. Ekson bölgelerinin hızlı ve kesin olarak saptanılabilir olmasının mevcut diğer tekniklere kıyasla daha verimli bir çözüm olarak sunulduğu yazarlar tarafından ifade edilmiştir.

[6] çalışmasında ekson bölgelerinin belirlenebilmesi amacıyla trigonometrik haritalama ile uyarlanabilir kaiser pencereleme yaklaşımının hibrit olarak kullanıldığı dirençli bir yaklaşım önerilmiştir. Mevcut algoritmalar kapsamında önerilen yaklaşım değerlendirildiğinde tahmin doğruluğunu iyi bir oranda iyileştirdiği buna karşın hesaplama karmaşıklığını optimize edemediğini açıklamıştır.

Bu çalışmada ise sayısallaştırılan DNA dizilerimlerine sinyal işleme teknikleri uygulanılarak ekson ve intron bölgelerinin sınıflandırılması sağlanmıştır. Sınıflandırma tıbbi görüntüleme alanında etkili sonuçlar veren ve güçlü bir sınıflandırma süreci sunan EfficientNetB7 öğrenme mimarisi ile gerçekleştirilmiştir. Zenginleştirilmiş veri kümesi üzerinde yapılan deneyler sonucunda yüksek başarı oranına ulaşılmıştır. Ancak derin öğrenme tekniğinin doğası gereği öğrenilemeyen bir durumun sınıflandırılması mümkün olmadığından dolayı yetersiz veri kümeleri üzerinde tatmin edici sonuçlar vermemiştir. Dolayısıyla veri kümesinin zenginleştirilmesi önemli bir kriter olarak çalışmada ifade edilmektedir.

3. ÇALIŞMADA KULLANILAN SAYISAL HARİTALAMA TEKNİKLERİ
(NUMERICAL MAPPING TECHNIQUES USED IN THE STUDY)

Genomik sinyal işleme; sayısal sinyal işleme, istatistik, matematik, desen tanıma gibi disiplinler arası metotları barındıran, biyoenformatik alanında saklı bilgileri ortaya çıkarmayı amaçlamayan güncel bir yaklaşımdır. Son zamanlarda bilim insanları; gen tespiti, dizi analizi, evrimsel analiz, genetik ağın modellenmesi, RNA tahmini gibi çeşitli çalışmalarda bu yaklaşımdan faydalanmıştır [2].

Ancak sembollerden oluşan DNA dizilimi üzerinde genomik sinyal işleme sürecini gerçekleştirmek için bu sembollerin sayısal dizilimlere dönüştürülmesi gerekmektedir. Dolayısıyla bu kısımda Reel Haritalama Tekniği, Moleküler Kütle Haritalama Tekniği, EIIP Haritalama Tekniği, Shannon Entropi Temelli haritalama tekniği ve Eşleştirilmiş Haritalama Tekniğinin Hidrojen Bağı Enerjisi kuralı kullanılarak DNA dizilimlerinin sayısallaştırılma süreçleri gerçekleştirilmiştir. Ancak belirtilen haritalama tekniklerinin seçilmesi hususunda dikkate alınan bazı ölçütler mevcuttur. Bu bağlamda A-T ve G-C organik bazlarının tamamlayıcı olma özelliğinden yararlanmak amacıyla Reel Haritalama Tekniği; DNA dizilimlerinin çok boyutlu bir uzayda haritalanması amacıyla moleküler kütle temsilinin kullanıldığı Müleküler Kütle Haritalama Tekniği; DNA'nın fizikokimyasal özelliğini yansıtarak hesaplama yükünü azaltan ve gen ayrım yeteneğini geliştiren EIIP haritalama tekniği[24]; DNA dizilimleri üzerinde kodon dağılımlarının entropisini hesaplayarak kodon olasılıkları arasındaki korelasyonu iyi bir şekilde açığa çıkaran ve DNA diziliminin nümerik temsili için daha geniş bir sayısal aralık sunan shannon entropi tabanlı sayısal haritalama tekniği [25]; dizilimde yer alan hidrojen bağlarından faydalanarak; G,C organik bazları açısından zengin ekson bölgeleri ile A,T organik bazları açısından zengin intron bölgelerinin tahmininde avantajlı bir durum oluşturan eşleştirilmiş haritalama tekniğinin hidrojen bağı enerjisi kuralı haritalama tekniklerinin seçiminde dikkate alınan önemli kriterlerdir [24]. Seçilen haritalama tekniklerinin ayrıntılı açıklaması aşağıda sunulmuştur.

3.1. Reel Haritalama Tekniği (Real Mapping Technique)

Sabit haritalama tekniği olan reel haritalama tekniği ile incelenen DNA dizilimleri üzerindeki organik bazlara A=-1,5, T=1,5, C=0,5 ve G=-0,5 atamaları yapılır. Tamamlayıcı özelliğine sahip olmasından dolayı tersine tamamlayıcı dizilim, ters dizilim ve tamamlayıcı dizilim şeklinde farklı dizi eşleşmeleri bu tekniğin çerçevesi içerisinde yer almaktadır [4].

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

Fiziko-kimyasal özellik tabanlı haritalama tekniği olan moleküler kütle haritalama tekniğinde verilen DNA diziliminin her bir organik bazına A=134, G=150, C=110,T=125 atamaları yapılarak sayısal bir dizilime dönüştürme işlemi gerçekleştirilmektedir [4].

3.3. EIIP Haritalama Tekniği (EIIP Mapping Technique)

EIIP temsilindeki yarı değerlik sayıları ile eşleştirilen fiziko-kimyasal özellik tabanlı haritalama tekniği olan EIIP haritalama tekniği ile verilen DNA dizilimindeki organik bazlara A=0,1260, G=0,0806, C=0,1340, T=0,1335 değerlerinin ataması yapılarak oluşan yeni dizilim kapsamında serbest elektron enerjisi dağılımlarının belirtilmesi sağlanmaktadır [4].

3.4. Shannon Entropi Temelli Haritalama Tekniği (Shannon Entropy Based Mapping Technique)

Shannon denkleminin kesirli bir türevi olan fraksiyonel Shannon entropi temelli bir yaklaşım [18] çalışmasında sunulmuştur. Geliştirilen bu teknik ile kodon dağılımlarının entropi değeri hesaplanmaktadır. Eş. 1'de önerilen yaklaşımın matematiksel ifadesine yer verilmiştir.

$$S_f = \sum_i \left[(-p(x_i))^{\propto i} \, \mathrm{p}(x_i) \log p(x_i)\right] \quad (1)$$

Denklem 1'de belirtilen $p(x_i)$ değeri her bir kodonun tekrarlanma sıklığıdır. $\propto_i$, genom dizisinden uyarlamalı olarak elde edilen bir değerdir ve matematiksel ifadesi Eş. 2'de gösterilmiştir.

$$1/(\log(p(x_i))) \quad (2)$$

Bu yaklaşım ile DNA diziliminin sayısal temsili için geniş bir aralığın elde edilmesi sağlanmıştır [18].

3.5. Eşleştirilmiş Sayısal Haritalama Tekniği (Paired Digital Mapping Technique)

Fiziko-kimyasal özellik tabanlı haritalama tekniği olan Eşleştirilmiş Sayısal Haritalama Tekniği ile verilen DNA dizilimindeki karmaşıklığın azaltılması hedeflenmiştir. Bu kapsamda 7 farklı kural tanımlanmıştır [4].

3.5.1. Pirimidin-pürin kuralı (Pyrimidine-purine rule)

Verilen nükleotit diziliminde pürin sayısının (A veya G) pirimidin sayısından (C veya T) yüksek olması durumunda A=1, G=1, C=-1, T=-1 değerleri atanırken tersi durumda A=-1, G=-1, C=1, T=1 değerleri atanmaktadır [4].

3.5.2. AA' kuralı (AA' rule)

Verilen nükleotit diziliminde A organik bazına 1 değerinin atanması durumunda diğer tüm organik bazların -1 değerini almasıdır [4].

3.5.3. TT' kuralı (TT' rule)

Verilen nükleotit diziliminde T organik bazına 1 değerinin atanması durumunda diğer tüm organik bazların -1 değerini almasıdır [4].

3.5.4. GG' kuralı (GG' rule)

Verilen nükleotit diziliminde G organik bazına 1 değerinin atanması durumunda diğer tüm organik bazların -1 değerini almasıdır [4].

3.5.5. CC' kuralı (CC' rule)

Verilen nükleotit diziliminde C organik bazına 1 değerinin atanması durumunda diğer tüm organik bazların -1 değerini almasıdır [4].

3.5.6. Hibrit kuralı (Hybrid rule)

Verilen nükleotit diziliminde A veya C organik bazlarına 1 değerinin atanması durumunda T veya G organik bazlarına -1 değerinin atanmasıdır [4].

3.5.7. Hidrojen bağı enerji kuralı (Hydrogen bond energy rule)

Aralarında 3'lü bağ bulunan G ve C organik bazlarına 1 değeri atanırken aralarında 2'li bağ bulunan A ve T organik bazlarına -1 değerinin atanmasıdır [4].

Bu çalışmada eşleştirilmiş sayısal haritalama tekniği içerisinde yer alan kurallar arasından hidrojen bağı enerji kuralı kullanılmıştır. Ekson bölgeleri G ve C organik bazları açısından zengin bir durumda iken intron bölgeleri A ve T organik bazları açısından zengin olduğu için kullanılan kural ekson ve intron tahmininde avantajlı bir durum oluşturmaktadır [24].

4. KULLANILAN SAYISAL SİNYAL İŞLEME YÖNTEMLERİ
(USED DIGITAL SIGNAL PROCESSING METHODS)

Genomik verilerin analiz edilmesinde büyük önem taşıyan sayısal sinyal işleme yöntemleri bu alanda yapılan araştırmalar için umut kaynağı olmuştur [1].

Bu kapsamda, insan türüne ait BCR-ABL geninin incelenmesi için fourier dönüşümü ve kısa zamanlı fourier dönüşümü şeklinde 2 farklı sayısal sinyal işleme yöntemi çalışmada kullanılarak DNA verilerinin analiz edilmesi planlanmıştır.

4.1. Pencerelenmiş Fourier Dönüşümü (Windowed Fourier Transform)

Fourier Dönüşümü yöntemi, bir zaman serisinin zaman ortamından frekans ortamına dönüştürülmesi olarak tanımlanır. Gerçekleştirilen dönüşüm işlemleri sonrasında zaman bilgisi kaybı yaşanırken zamana ait frekans bilgisi ile süreç yönetilmeye devam eder. Dolayısıyla durağan zaman serilerinin analizinde elde edilen tahminler durağan olmayan serilerin analizine göre daha tatmin edici sonuçlar vermektedir. Fourier dönüşümünün matematiksel ifadesi Eş. 3'te verilmektedir [12].

$$x(f) = \int_{-\infty}^{+\infty} x(t)\, e^{-2\pi i f t} dt \tag{3}$$

Verilen eşitlikte t, zamanı; f, frekansı; x(t), zaman serisini ve x(f), fourier dönüşümünü temsil etmektedir [12]. Ancak x(t)'nin periyodik olmayan bir sinyal olması durumunda sinyalin periyodik bir şekilde gösterilebilmesi amacı ile mevcut sinyalin yapay olarak uzatılması gerçekleştirilmektedir. Dolayısıyla sonlanma kısımlarında ek süreklilikler gerektiğinden dolayı bu probleme bir çözüm olarak pencerelenmiş fourier dönüşümleri sunulmuştur. Böylelikle zaman ve frekans bilgisinin kaybolmaması için kullanılan pencereler vasıtasıyla sinyallerin hem frekans hem de zaman bölgesi hakkında eş zamanlı bilgilerin elde edilmesi sağlanmıştır [26, 27]. Pencerelenmiş Fourier Dönüşümü, giriş sinyalinin bölgelere ayrılmasının ardından elde edilen her bölgenin frekans içerikleri ile analiz edilmesi kapsamında kullanılmıştır. Aynı zamanda kullanılan pencerelerin ortasındaki kısımların daha fazla vurgulanması sağlanırken sonlanma noktalarında ise daha az vurgulanma ile sönme gerçekleşmektedir. Sinyalin zamanda yerelleştirilmesi pencerenin etkisinden kaynaklanmaktadır [27].

Pencereleme fonksiyonları ile problemlerin bir kısmına çözüm üretilebildiği halde elde edilen bilginin seçilen pencerenin büyüklüğü ile ilişkili olması doğruluk değerini olumsuz etkileyebilmektedir [12]. Bu nedenle nihai durumda elde edilen bilginin farklı pencereleme fonksiyonları kullanılarak çeşitliliğin arttırılması sağlanmış ve kullanılan

fonksiyonlar içerisinden hibrit pencereleme fonksiyonları seçilerek sınıflandırmanın başarılı sonuçlar üretmesi hedeflenmiştir.

Kullanılan hibrit pencereleme fonksiyonlarının matematiksel ifadeleri Eş. 4-Eş. 12'de verilmiştir [28].

$$w1 = 0{,}5(cosh) + 0{,}5(bohman) \quad (4)$$

$$w2 = 0{,}5(cosh) + 0{,}5(cos3(x)) \quad (5)$$

$$w3 = 0{,}5(cosh) + 0{,}5(cos4(x)) \quad (6)$$

$$w4 = 0{,}5(cosh) + 0{,}5(hamming) \quad (7)$$

$$w5 = 0{,}5(\text{cosh}) + 0{,}5(\text{blackman}) \quad (8)$$

$$w6 = 0{,}5(cosh) + 0{,}5(optimized\ blackman) \quad (9)$$

$$w7 = 0{,}5(cosh) + 0{,}5(triangular) \quad (10)$$

$$w8 = 0{,}5(cosh) + 0{,}5(von - hann) \quad (11)$$

$$w9 = 0{,}5(cosh) + 0{,}5(welch) \quad (12)$$

Bu doğrultuda frekans alanında ifade edilen DNA dizilimlerinin aynı aralığa tekabül eden ekson ve intron bölgeleri için 9 hibrit pencereleme ile ayrık fourier dönüşümü çerçevesinde ortaya çıkan farklı frekans yoğunlukları dikkate alınmış ve genomik dizilerin periyodikliğinin pencere yaklaşımları üzerinde temellenen ayrık fourier dönüşümünün spektral özelliği ile betimlenmesi sağlanmıştır. Bu durum DNA dizilimlerinin analizi sırasında zaman alanından çıkarılamayan bazı gerekli bilgiler için frekans alanında verimli sinyal temsili sunmuştur. Böylelikle mevcut performans üzerinde bir artış elde edilmiştir [22, 8]**.**

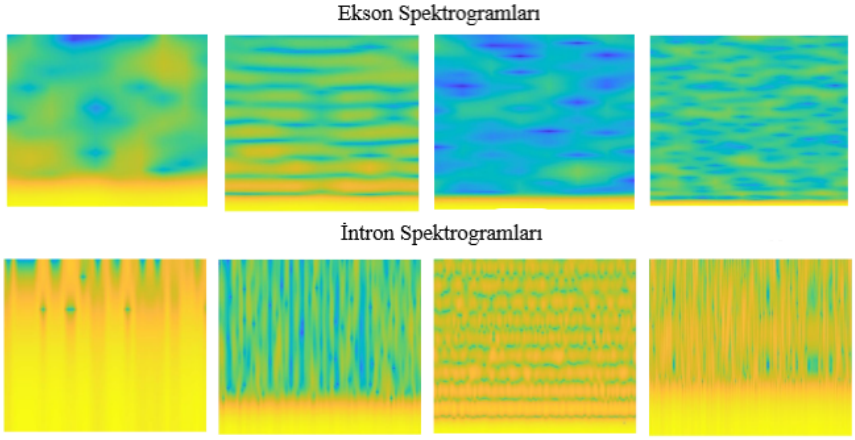
4.2. Pencerelenmiş Kısa Zamanlı Fourier Dönüşümü
(Windowed Short Time Fourier Transform)

4.1 kısmında anlatılan fourier dönüşümü, bir zaman serisine ait frekansların elde edildiği ancak zaman değerlerine karşılık gelen frekans bilgilerini sağlayamadığı için durağan olmayan serilerin analizinde kullanımı tatmin edici sonuçlar vermeyen bir tekniktir. Dolayısıyla mevcut probleme çözüm olarak kısa zamanla fourier dönüşümü geliştirilmiştir. Bu teknik ile analiz edilecek zaman serisi durağan olduğu kabul edilen küçük parçalara bölündükten sonra her bir parçanın pencere fonksiyonu ile çarpılması sağlanmaktadır. Ancak kısa zamanlı fourier dönüşümü ile frekansın sadece mevcut olduğu zaman aralığındaki bileşenleri elde edilebildiği için aynı anda hem iyi bir zaman hem de iyi bir frekans çözünürlüğü sunulamamaktadır [12]. Bu nedenle çalışmanın ikinci kısmında uygulanan kısa zamanlı fourier dönüşümünün gerçekleştirilebilmesi için öncelikle zaman eksenine sonlu bir pencereleme fonksiyonunun yerleştirilmesi ve pencerenin kaydırılması ile zaman serisinin fourier dönüşümünün alınması gerçekleştirilmiştir. Bu durum aynı zamanda frekans içeriğinin lokalize bir ölçüsünü de sağlamaktadır. Bahsedilen işlemin matematiksel fonksiyonu Eş. 13'te ifade edilmiştir [12].

$$KZFD(t,f) = \int_{-\infty}^{+\infty} x(t)w(t-K)\,e^{-2\pi ift}dt = < g_{f,t}(t), x(t) > \quad (13)$$

Eşitlikte verilen f frekansı, x(t) analiz edilen zaman serisini, w(t-K) zaman ekseninde K noktasına yerleştirilmiş pencere fonksiyonunu göstermektedir [12]. Pencerelenmiş kısa zamanlı fourier dönüşümünde performans, seçilen pencere fonksiyonuna bağlı olduğu için nihai durumda elde edilen bilginin barlet, blackman, blackman-harris, bohman, chebwin, gausswin, hamming, rectangular ve triang pencereleme fonksiyonlarının kullanılması ile çeşitliliğin arttırılması sağlanmış ve kullanılan pencereleme fonksiyonları ile elde edilen spektrogramlar sayesinde sınıflandırma süreci yönetilmiştir [12]. Sinyaller hakkında önemli bilgiler içeren spektrogramlar [13] Şekil 1'de verilmiştir.

Şekil 1'de verilen spektrogramlar, BCR-ABL geninin eşleştirilmiş sayısal haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılmasından sonra pencerelenmiş kısa zamanlı fourier dönüşümünün uygulanılması ile elde edilen ekson ve intron bölgelerini göstermektedir.



Şekil 1. Sayısallaştırılan DNA dizilimlerindeki ekson ve intron bölgelerinin spektrogram görüntüleri
(Spectrogram images of exon and intron regions in digitized DNA sequences)

5. GÖRÜNTÜ İYİLEŞTİRME SÜRECİ (IMAGE IMPROVEMENT PROCESS)

Çalışmanın ilk aşamasında, ikinci bölümde verilen sayısal haritalama teknikleri ile sayısallaştırılan DNA dizilimlerinin fourier dönüşümü ile frekans alanında gösterilmesi sağlanmıştır. Verilerin kenar, köşe gibi önemli noktalarının görüntü içerisinde nasıl bir konumda yer aldığının algılanması amaçlandığından dolayı kritik alan olan frekans sinyalinin belirginleştirilmesi hedeflenmiştir ve bu işlem için CLAHE (Contrast Limited Adaptive Histogram Equalization) yöntemi uygulanmıştır. Ardından 10 farklı öğrenme aktarımı mimarileri ile sınıflandırma süreci gerçekleştirilerek deneysel ölçütler üzerinden başarımları değerlendirilmiştir.

6. KULLANILAN SINIFLANDIRMA YÖNTEMİ (USED CLASSIFICATION METHOD)

Yapay zeka insanların sahip olduğu bilişsel yeteneklerin bilgisayarlara yüklenilmesi olarak ifade edilen geleneksel bilgisayar görme ve sinir ağları teknolojilerini içerisinde barındıran geniş bir yelpazedir.
Bu çalışmada ekson ve intron bölgelerini tanımlamak amacı ile sinir ağları çerçevesinde yer alan derin öğrenme metodu kullanılmıştır. Derin öğrenme, minimum 3 sinir ağı tarafından karmaşık desenlerin öğrenildiği bir yöntemdir. Aynı zamanda özelliklerin görüntüden otomatik olarak çıkarılmasını mümkün kılmasının yanı sıra yüksek boyutlu veri kümelerinde de kabul edilebilir sonuçlar üretmesi geleneksel makine öğrenme yöntemlerine kıyasla ön plana çıkan başarılı yanlarını oluşturmaktadır [29]. Ancak veri kümesinde az verinin olması durumunda yeterli bilgi derin öğrenme tekniği ile öğrenilemez [30]. Bununla birlikte gerekli donanıma sahip bilgisayarların mevcut olmaması durumunda istenilen mimarinin tasarlanılamaması, başarılı sonuçlar elde etmenin önünde bir engel oluşturmaktadır. Bu ve benzeri problemlerden dolayı derin öğrenme tekniğinin içerisinde yer alan öğrenme aktarımı teknolojisi kullanılmıştır [31]. Böylelikle, küçük veri kümelerinde bile yüksek başarıların elde edilebilmesinin önü açılmış ve güçlü donanıma sahip bilgisayarlarda tasarlanan mimarilerin kullanımı mümkün olmuştur [32].

6.1. Öğrenme Aktarımı (Transfer Learning)

Derin öğrenme yöntemi içerisinde yer alan öğrenme aktarımı farklı bir problem için tasarlanan mimarinin istenilen başka bir çalışmada kullanılması olarak tanımlanmaktadır. Bu teknoloji ile zamanın daha etkili bir şekilde değerlendirilebilmesi de mümkündür [32]. Çalışmada kullanılan 10 farklı öğrenme aktarımı mimarisi, derinlik sayıları ile birlikte Tablo 1'de sunulmuştur [33].

Bu çerçevede CLAHE yöntemi ile frekans alanında ifade edilen sinyallerin; kenar, köşe gibi önemli noktalarının görüntü kapsamında nasıl bir konumda yer aldığının belirginleştirilmesi sağlandıktan sonra boyutları 200x200 olarak yeniden şekillendirilmiştir. Ardından bu görüntüler Tablo 1'de yer alan öğrenme aktarımı mimarilerinin her birine girdi olarak verilmiştir. Mimarilerin sonuncu tam bağlantı katmanı amaca uygun yeni tam bağlantılı katman ile değiştirildikten sonra ekson ve intron görüntüleri kullanılarak gerçekleştirilecek sınıflandırmanın eğitim süreci için parametreler üzerinde ince ayar yapılmıştır. Her bir yenileme sırasında kayıp değeri iyileştirilerek maksimum başarı oranı ve minimum kayıp değeri elde edilmiştir [33]. Faydalanılan kayıp fonksiyonu, iki sınıf arasında uygulanılacak bir sınıflandırma süreci olduğundan dolayı ikili çapraz entropi yöntemi olarak seçilmiştir. Matematiksel ifadesi Eş. 14'te verilmektedir [34].

Tablo 1. Çalışmada kullanılan öğrenme aktarımı mimarileri (Transfer learning architectures used in the study)

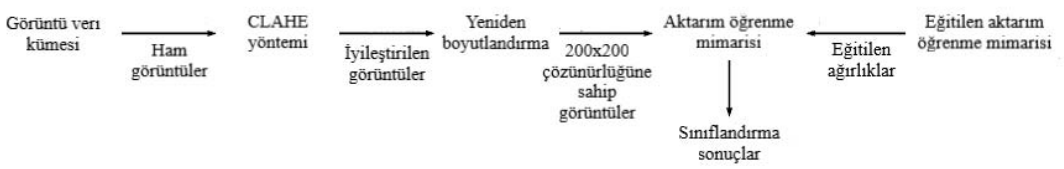
Öğrenme Aktarımı Mimarileri	Derinlik Sayısı
VGG16	16
VGG19	19
Xception	71
MobileNetV2	53
InceptionResnetV2	164
DenseNet169	169
DenseNet201	201
Resnet101	101
Resnet150	150
EfficientNetB7	813

$$İÇE(h,t)=-(h*\log(t)+(1-h)*\log(1-t)) \quad (14)$$

Yukarıda verilen denklemde h, hedef vektörünü ve t, tahmin edilen sınıf vektörünü belirtmektedir. Vektör elemanları ikili değerler olarak temsil edildiğinden dolayı 0-ekson ve 1-intron olarak tanımlanmıştır. Her bir öğrenme aktarımı mimarisi RMSprop optimize edicisi kullanılarak 0,00007 öğrenme oranı, 0,7 momentum ve 1e-07 epsilon değeri ile eğitilmiştir. Çalışmanın iş akışını gösteren blok şema Şekil 2'de sunulmaktadır [34].

6.2. Kullanılan Veri Kümesi (Used Data Set)

Bu makalede NCBI-Ulusal Biyoteknoloji Bilgi Merkezi gen bankası (https://www.ncbi.nlm.nih.gov) vasıtasıyla elde



Şekil 2 . Öğrenme aktarımı mimarileri için temsili veri akışı (Representative data flow for transfer learning architectures)[34].

edilen insan türüne ait BCR-ABL genleri üzerinde bir inceleme yapılmıştır. Çalışmada kullanılan AM600680.1, AM886138.1, AM400881.1 ve EU447303.1 numaralarına sahip genler farklı nükleotit dizilimlerinden oluşmaktadır. Aynı zamanda incelenen genlerin intron ve ekson bölgeleri gen bankasında verilen bilgiler ile belirlenilmiş ve Tablo 2'te sunulmuştur.

Tablo 2. Çalışmada kullanılan genlerin ekson ve intron bölgeleri (Exon and intron regions of the genes used in the study)

AM400881.1	
1-14	Ekson Bölgesi
15-71	İntron Bölgesi
72-527	İntron Bölgesi
AM600680.1	
1-29	Ekson Bölgesi
30-578	İntron Bölgesi
579-1114	İntron Bölgesi
1115-1180	Ekson Bölgesi
AM886138.1	
1-31	Ekson Bölgesi
32-280	İntron Bölgesi
281-790	İntron Bölgesi
791-853	Ekson Bölgesi
EU447303.1	
1-145	Ekson Bölgesi
274-488	Ekson Bölgesi

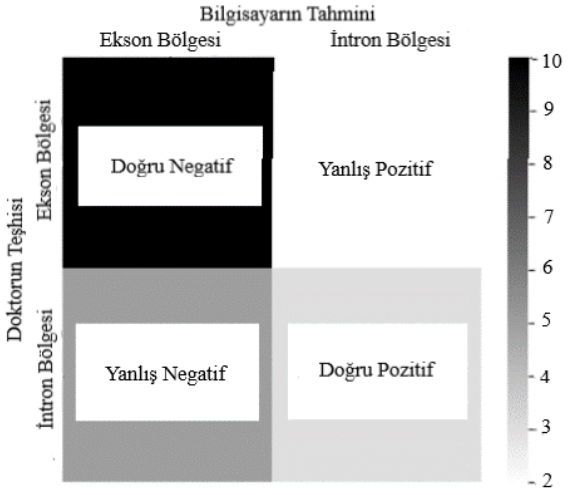
Bu kapsamda 13 ile 174 nükleotit uzunluğu arasında değişen ekson dizilimleri ve 56 ile 548 nükleotit uzunluğu arasında değişen intron dizilimleri üzerinde bir çalışma gerçekleştirilmiştir.

7. SINIFLANDIRMA BAŞARIMLARININ KARŞILAŞTIRILMASI (COMPARISON OF CLASSIFICATION ACHIEVEMENTS)

7.1. İlk Aşama (First Stage)

İki aşama şeklinde yürütülen bu çalışmanın ilk aşamasında; NCBI veri kümesinden insana ait AM600680.1, AM400881.1, AM886138.1 ve EU447303.1 numaralarına sahip 4 farklı BCR-ABL geni ele alınmıştır. Bölüm 2'de belirtilen sayısal haritalama teknikleri ile sayısallaştırılan farklı uzunluklardaki gen dizilimleri pencerelenmiş fourier dönüşümüne tabi tutularak frekans alanında ifade edilen bir tasviri sağlanmıştır. Ancak seçilen pencerenin büyüklüğü ile elde edilen bilginin ilişkili olması doğruluk değerini olumsuz etkileyebildiği için çalışmada farklı hibrit pencereleme fonksiyonlarının kullanılması ile çeşitliliğin arttırılması ve bu dezavantajın önüne geçilmesi hedeflenmiştir. Bununla birlikte elde edilen sinyal görüntülerinin belirsiz olmasından dolayı CLAHE yöntemi kullanılarak sinyallerin belirginleştirilmesi sağlanmıştır. Çünkü öğrenme aktarımı modelleri ile ilgili görüntü kapsamında kenar, köşe gibi önemli noktaların nasıl bir konumda yer aldığının algılanması amaçlandığından dolayı öğrenilmesi gereken ve kritik alan olan frekans sinyalinin belirginleştirilmesi önemli bir durum olarak değerlendirilmiştir.

Bu doğrultuda belirginleştirilen ekson ve intron etiketli veri kümesinin %65'i eğitim, %35'i test kümesi şeklinde ayrılmış ve 10 farklı öğrenme aktarımı mimarilerine verilerek her bir haritalama tekniği için aynı parametreler üzerinden sınıflandırma süreci gerçekleştirilmiştir. Sınıflandırma işleminden sonra elde edilen deneysel performans ölçütleri Şekil 3'te verilen karışıklık matrisi üzerinde açıklanmıştır.



Şekil 3. Deneysel performans ölçütlerinin değerlendirildiği örnek bir karışıklık matrisi
(A sample confusion matrix where experimental performance criteria are evaluated)

7.1.1. Doğru negatif (True negative)

Doktorlar tarafından ekson bölgesi olarak etiketlenen görüntüler içerisinden bilgisayarın ekson bölgesi olarak doğru tahmin ettiği görüntülerin toplam sayısını veren değerlendirme ölçütüdür.

7.1.2. Yanlış pozitif (False positive)

Doktorlar tarafından ekson bölgesi olarak etiketlenen görüntüler içerisinden bilgisayarın ekson bölgesi olarak yanlış tahmin ettiği görüntülerin toplam sayısını veren değerlendirme ölçütüdür.

7.1.3. Yanlış negatif (False negative)

Doktorlar tarafından intron bölgesi olarak etiketlenen görüntüler içerisinden bilgisayarın intron bölgesi olarak yanlış tahmin ettiği görüntülerin toplam sayısını veren değerlendirme ölçütüdür

7.1.4. Doğru pozitif (True positive)

Doktorlar tarafından intron bölgesi olarak etiketlenen görüntüler içerisinden bilgisayarın intron bölgesi olarak

doğru tahmin ettiği görüntülerin toplam sayısını veren değerlendirme ölçütüdür.

7.1.5. Doğruluk oranı (Accuracy rate)

Doğru tahmin edilen ekson ve intron bölgelerinin tüm görüntülere oranını ifade etmektedir.

$$(DP + DN)/(DP + YP + DN + YN) \quad (15)$$

Matematiksel hesabı Eş. 15'te verilen bu ölçütün yüksek oranlarda çıkması hedeflenmektedir.

7.1.6. Duyarlılık (Sensitivity/recall)

Doğru bir şekilde tahmin edilen intron bölgelerinin tüm intron görüntülerine oranını ifade etmektedir.

$$DP/(DP + YN) \quad (16)$$

Matematiksel hesabı Eş. 16'da verilen duyarlılık değerinin yüksek oranlarda çıkması hedeflenmektedir.

7.1.7. Özgüllük (Specificity)

Doğru bir şekilde tahmin edilen ekson bölgelerinin tüm ekson görüntülerine oranını ifade eden bir ölçüttür.

$$DN/(DN + YP) \quad (17)$$

Matematiksel hesabı Eş. 17'de verilen özgüllük değerinin yüksek oranlarda çıkması hedeflenmektedir.

7.1.8. Kesinlik (Precision)

Ekson bölgeleri protein üreten bölgeler olduğu için bu bölgelerin yanlış tahmin edilmesi yapılan çalışmanın güvenilirliği açısından kontrol edilmesi gereken bir husustur. Çünkü bilginin yer aldığı kısım burasıdır ve hatalı tahmin oranının yüksek olması sonucunda bir kayıp durumu meydana gelecektir. Bu bağlamda doğru bir şekilde tahmin edilen intron bölgelerinin fazla, yanlış tahmin edilen ekson bölgelerinin az çıkması beklenen bir durumdur.

$$DP/(DP + YP) \quad (18)$$

Matematiksel hesabı Eş. 18'de verilen kesinlik değerinin yüksek oranlarda çıkması hedeflenmektedir.

7.1.9. F ölçütü (F score)

Kesinlik ve duyarlılık değerlerinin harmonik ortalaması olarak ifade edilen F ölçütü uç değerlerin dikkate alınmasını sağlamaktadır.

F Skoru=2*kesinlik*duyarlılık/kesinlik+duyarlılık (19)

Matematiksel hesabı Eş. 19'da verilen F ölçütünün yüksek oranlarda çıkması hedeflenmektedir.

Tablo 3. Reel haritalama tekniği için elde edilen deneysel ölçütler (Experimental criteria obtained for the real mapping technique)

Öğrenme Aktarım Mimarileri	Doğru Negatif	Yanlış Pozitif	Yanlış Negatif	Doğru Pozitif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
VGG16	21	0	20	0	%51,2	0	%100	0	0
VGG19	21	0	20	0	%51,2	0	%100	0	0
Xception	3	18	1	19	%53,6	%95	%14,2	%51,3	%66,6
MobilenetV2	4	17	1	19	%56	%95	%19	%52,7	%67,7
InceptionResnetV2	15	6	3	17	%78,04	%85	%71,4	%73,9	%79
DenseNet169	0	21	0	20	%48,7	%100	0	%48,7	%65,5
DenseNet201	0	21	0	20	%48,7	%100	0	%48,7	%65,5
Resnet101	15	6	5	15	%73,1	%75	%71,4	%71,4	%73,1
Resnet150	16	5	4	16	%78,04	%80	%76,1	%76,1	%78
EfficientNetB7	18	3	6	14	%78,04	%70	%85,7	%82,3	%75,6

Tablo 4. Moleküler kütle haritalama tekniği için elde edilen deneysel ölçütler
(Experimental criteria obtained for the moleculer mass mapping technique)

Transfer Learning Mimarileri	Doğru Negatif	Yanlış Pozitif	Yanlış Negatif	Doğru Pozitif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
VGG16	21	0	20	0	%51,2	0	%100	0	0
VGG19	1	20	0	20	%512	%100	%0,047	%50	%66,6
Xception	8	13	4	16	%58,5	%80	%38	%55,1	%65,2
MobilenetV2	18	3	13	7	%60,9	%35	%85,7	%70	%46,6
InceptionResnetV2	1	20	2	18	%46,3	%90	%0,047	%47,3	%62
DenseNet169	4	17	2	18	%53	%90	%19	%51,4	%65,4
DenseNet201	6	15	2	18	%58,5	%90	%28,5	%54,5	%67,8
Resnet101	13	8	6	14	%65,8	%70	%61,9	%63,6	%66,6
Resnet150	6	15	6	14	%48,7	%70	%28,5	%48,2	%57
EfficientNetB7	11	10	5	15	%63,4	%75	%52,3	%60	%66

Tablo 5. EIIP haritalama tekniği için elde edilen deneysel ölçütler
(Experimental criteria obtained for the EIIP mapping technique)

Transfer Learning Mimarileri	Doğru Negatif	Yanlış Pozitif	Yanlış Negatif	Doğru Pozitif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
VGG16	14	7	4	16	%73,1	%80	%66,6	%69,5	%74,3
VGG19	12	9	6	14	%63,4	%63,6	%57,1	%60,8	%62,1
Xception	8	13	3	17	%60,9	%85	%38,09	%56,6	%67,9
MobilenetV2	21	0	18	2	%56,09	%10	%100	%100	%18,1
InceptionResnetV2	0	21	1	19	%46,3	%95	0	%47,5	%63,3
DenseNet169	17	4	6	14	%75,6	%70	%80,9	%77,7	%73,6
DenseNet201	11	10	4	16	%65,8	%80	%52,3	%61,5	%69,5
Resnet101	19	2	13	7	%63,4	%35	%90,4	%77,7	%48,2
Resnet150	12	9	6	14	%63,4	%70	%57,1	%60,8	%65,0
EfficientNetB7	14	7	5	15	%70,7	%75	%66,6	%68,1	%71,3

Tablo 6. Shannon entropi temelli haritalama tekniği için elde edilen deneysel ölçütler
(Experimental criteria obtained for the shannon mapping technique)Transfer Learning Mimarileri

	Doğru Negatif	Yanlış Pozitif	Yanlış Negatif	Doğru Pozitif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
VGG16	21	0	20	0	%51,2	0	%100	0	0
VGG19	21	0	20	0	%51,2	0	%100	0	0
Xception	12	9	1	19	%75,6	%95	%57,1	%67,8	%79,1
MobilenetV2	5	16	0	20	%60,9	%100	%23,8	%55,5	%70,9
InceptionResnetV2	16	5	4	16	%78,04	%80	%76,1	%76,1	%78
DenseNet169	13	8	5	15	%68,2	%75	%61,9	%65,2	%69,7
DenseNet201	2	19	1	19	%51,2	%95	%0,095	%50	%65,5
Resnet101	17	4	5	15	%78,04	%75	%80,9	%78,9	%76,9
Resnet150	14	7	8	12	%63,4	%60	%63,6	%63,1	%61,5
EfficientNetB7	15	6	5	15	%73,1	%75	%71,4	%71,4	%73,1

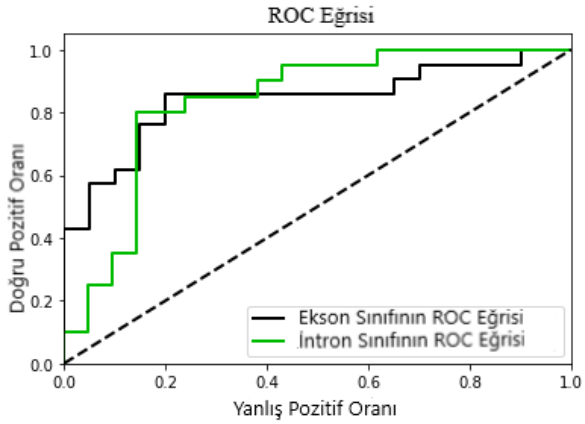
Tablo 7. Eşleştirilmiş haritalama tekniğinin hidrojen bağı enerji kuralı için elde edilen deneysel ölçütler
(Obtained experimental criteria for the hydrogen bond energy rule of the paired mapping technique)

Transfer Learning Mimarileri	Doğru Negatif	Yanlış Pozitif	Yanlış Negatif	Doğru Pozitif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
VGG16	21	0	20	0	%51,2	0	%100	0	0
VGG19	21	0	20	0	%51,2	0	%100	0	0
Xception	14	7	4	16	%73,1	%80	%66,6	%69,5	%74,3
MobilenetV2	19	2	12	8	%65,8	%40	%90,4	%80	%53,3
InceptionResnetV2	15	6	5	15	%73,1	%75	%71,4	%71,4	%73,1
DenseNet169	10	11	4	16	%63,4	%80	%47,6	%59,2	%68,0
DenseNet201	7	14	2	18	%60,9	%90	%33,3	%56,2	%69,1
Resnet101	13	8	3	17	%73,1	%85	%61,9	%68	%75,5
Resnet150	15	6	6	14	%70,7	%70	%71,4	%70	%70
EfficientNetB7	16	5	3	17	**%80,4**	%85	%76,1	%77,2	%80,9

Bu bilgiler ışığında Tablo 3-Tablo 7'de sırasıyla reel haritalama tekniği, moleküler kütle haritalama tekniği, EIIP haritalama tekniği, shannon entropi temelli haritalama tekniği ve eşleştirilmiş haritalama tekniğinin hidrojen bağı enerjisi kuralı kullanılarak sayısallaştırılan DNA dizilimlerinin frekans alanına aktarıldıktan sonra 10 farklı öğrenme aktarımı mimarisi kullanılarak gerçekleştirilen sınıflandırma süreci sonrasında elde edilen deneysel ölçütlerin sonuçları verilmiştir.

Tablo 3-Tablo 7 incelendiğinde en yüksek başarı oranının eşleştirilmiş haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılan DNA dizilimleri üzerinde uygulanılan EfficientNetB7 öğrenme aktarımı mimarisi ile sınıflandırılması sonucunda elde edildiği görülmektedir. Bununla birlikte sadece başarı yüzdesi ile değerlendirme yapılması modelin gücü hakkında sağlıklı sonuçlar vermediğinden diğer performans ölçütlerinin de incelenilmesi gereklidir. Dolayısıyla duyarlılık, özgüllük, kesinlik ve F skoru kriterleri de araştırılmıştır. Bu kriterlerin %100'lük bir oran sunması modelin başarısı açısından istenen bir durum olmasına karşın eşleştirilmiş haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılan DNA dizilimleri üzerinde EfficientNetB7 öğrenme aktarımı mimarisinin uygulanılması sonucunda elde edilen deneysel ölçütlerin sunduğu oranların da yöntemin durumu açısından

diğer oranlara kıyasla başarılı sonuçlar verdiği görülmektedir. Duyarlılık, özgüllük ve kesinlik kriterlerinin sunduğu yüzdelik sonuçların farklılığı ise tekniklerin ya da mimarilerin ekson veya intron bölgelerini tanımlama hususundaki gücünü ifade etmektedir. Bununla birlikte eşleştirilmiş haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılan DNA dizilimleri üzerinde çalıştırılan EfficientNetB7 öğrenme aktarımı mimarisi hem ekson hem de intron bölgelerinin saptanılabilirliği hususunda başarılı ve istikrarlı bir durum oluşturarak diğer mimarilere kıyasla bu çalışma kapsamında öne çıkmıştır. Aynı zamanda ekson ve intron bölgelerini ayırt etme performansı ROC eğrisi üzerinde Şekil 4'te gösterilmiştir.



Şekil 4. Ekson ve intron bölgelerini tanıma performansı (Recognition performance for exon and intron regions)

EfficientNetB7, tıbbi görüntüleme alanında etkili sonuçlar veren ve sınıflandırma süreçlerini güçlü bir şekilde yöneten bir mimari olduğundan bu çalışmada öne çıkmıştır [35]. Çalışmanın ilk aşamasında kullanılan tüm aşamalar Şekil 5'te gösterilmiştir. Tüm süreçlerin yer aldığı ilk aşamada en başarılı sayısal haritalama tekniği eşleştirilmiş sayısal haritalama tekniğinin hidrojen bağı enerji kuralı seçildiğinden çalışmanın ikinci aşamasında bu haritalama tekniği üzerinden çalışmalara devam edilmiştir.

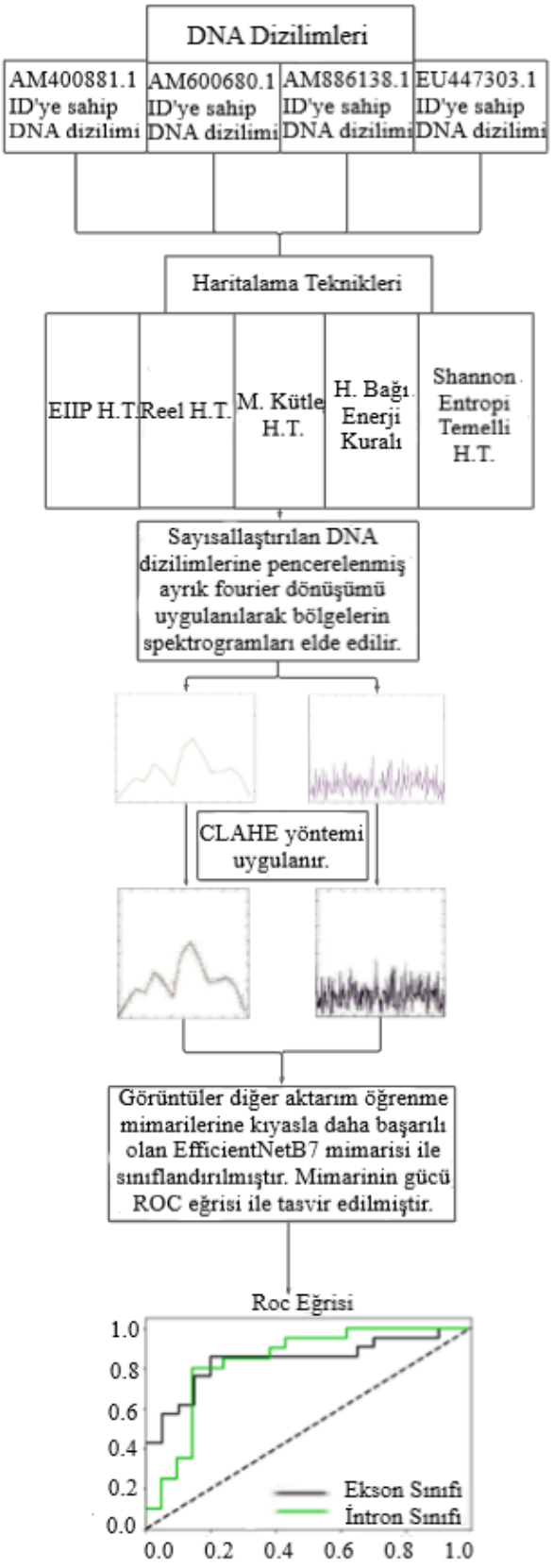
7.2. İkinci Aşama (Second Stage)

Çalışmanın ikinci aşamasında NCBI veri kümesinden insana ait olan AM400881.1, AM600680.1, AM886138.1, ve EU447303.1 numaralarına sahip 4 farklı BCR-ABL geni üzerinde inceleme yapılmıştır. Bu doğrultuda ilk aşamanın en başarılı tekniği olarak değerlendirilen eşleştirilmiş haritalama tekniğinin hidrojen bağı enerjisi kuralı ile sayısallaştırılan DNA dizilimleri üzerinde uygulanan kısa zamanlı fourier dönüşümü ile ekson ve intron bölgelerinin spektrogramları elde edilmiştir. Performansın seçilen pencere fonksiyonuna bağlı olmasından dolayı kısa zamanlı fourier dönüşümü yönteminde mevcut olan dezavantajlara kısmen de olsa çözüm bulmak amacı ile nihai durumda elde edilen bilginin barlet, blackman, blackman-harris, bohman, chebwin, gausswin, kaiser, hamming, rectangular ve triang olmak üzere 10 farklı pencereleme yöntemleri ile gösterilmesi sağlanılarak farklı ifadelerin elde edilmesi amaçlanmıştır.

Bu doğrultuda ilk aşamanın en başarılı öğrenme aktarımı olarak değerlendirilen EfficientNetB7 mimarisi ile ekson ve intron etiketli spektrogramların sınıflandırılması gerçekleştirilmiştir. Süreç AM400881.1, AM600680.1, AM886138.1, ve EU447303.1 numaralarına sahip DNA dizilimlerinin sırasıyla test veri kümesini oluşturması kalan dizilimlerin ise eğitim veri kümesini oluşturması şeklinde gerçekleştirilmiştir. Bu çerçevede 4 farklı sınıflandırma için uygulanılan eğitim ve test veri kümelerinin bilgileri Şekil 6'da gösterilmiştir.

Sınıflandırma süreçleri sonucunda elde edilen başarı oranlarının ve kayıp değerlerinin grafikleri Şekil 7 ve Şekil 8'de sunulmuştur.

Şekil 7 ve Şekil 8'de verilen grafikler incelendiğinde test veri kümesi olarak değerlendirilen, AM600680.1 ve AM886138.1 numaralı DNA dizilimlerinin 100 epok sonucunda ulaştıkları başarı oranları %100 olarak bulunmuştur. Değişen epok sayısı ile birlikte artan başarı yüzdeleri ve azalan kayıp değerleri, başarılı bir eğitim sürecinin gerçekleştiğini göstermektedir. Aynı zamanda farklı DNA dizilimleri üzerinde gerçekleşen test süreci sonunda istikrarlı bir duruma ulaşılması sağlandığından dolayı çeşitli veritabanlarından elde edilen gen dizilimleri üzerinde de başarılı ve istikrarlı sonuçlar elde edilmesi beklenmektedir.
Test veri kümesi olarak değerlendirilen AM400881.1 numaralı DNA diziliminde 100 epok sonucunda ulaşılan başarı oranı %66,67 seviyesinde kalmıştır. AM400881.1 numaralı DNA diziliminin ekson ve intron bölgeleri incelendiğinde tüm veri kümesi içerisinde yer alan en kısa ekson ve en kısa intron dizilimlerini içerdiği görülür. Diğer yandan test veri kümesi olarak değerlendirilen EU447303.1 numaralı DNA diziliminin 100 epok sonucunda ulaştığı başarı yüzdesi %55 seviyesinde kalmıştır. EU447303.1 numaralı DNA diziliminin ekson bölgeleri incelendiğinde ise eğitim kümesinde yer almayan en uzun dizilime sahip ekson bölgelerinden oluştuğu görülmektedir. Burada sırasıyla test kümesi olarak değerlendirilen AM400881.1 ve EU447303.1 numaralı DNA dizilimleri üzerinde gerçekleştirilen test sürecinde, test kümesinin daha önce görmediği çok farklı uzunluklara sahip dizilimler karşısında çıkarım yapılması istenmiştir. Bu durum epok sayısı değiştikçe test başarı yüzdesinin artarak istikrarlı bir duruma ulaşamaması ve test kayıp değerinin gittikçe azalan bir eğri konumunu oluşturamaması ile sonuçlanmıştır. Ek olarak eğitim kümesinin elde ettiği başarı yüzdesi ile test kümesinin elde ettiği başarı yüzdesi arasındaki farkın yüksek olması, ulaşılan %66,67 ve %55 başarı yüzdelerinin de aşırı öğrenme ile elde edilen bir oran olduğu yorumunu vermektedir. Bu bağlamda dizilimlerin sınıflandırılmasında kullanılan derin öğrenme yöntemi ile eğitilen modelin bilmediği bir davranış karşısında net bir saptama yapması mümkün olmadığı için AM400881.1 ve EU447303.1 numaralı DNA dizilimlerinin başarı oranına katkısı değerlendirmeye alınmamıştır.



Şekil 5. Çalışmanın ilk aşamasında uygulanan süreçler
(Processes applied in the first stage of the study)

Çalışmanın ikinci aşamasında kullanılan tüm aşamalar Şekil 9'da gösterilmiştir. NCBI veri kümesinden tedarik edilen BCR-ABL gen dizilimleri için bölgelerin doğru bir şekilde tanınabilirliğin farklı bir veri kümesi çerçevesinde mümkün olduğu ikinci aşama sonucunda elde edilen yüksek başarı oranı ile ispat edilmiştir.

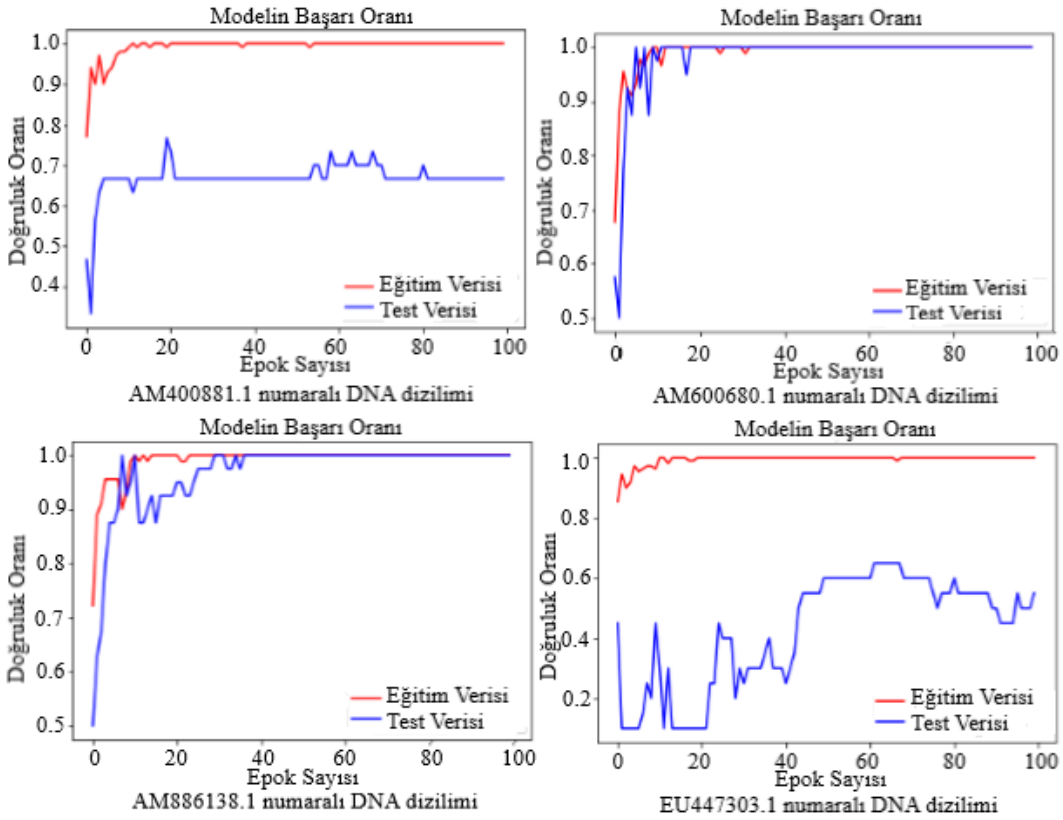
8. TARTIŞMA (DISCUSSION)

Çalışmanın amacı, gen dizilimleri kapsamında araştırılan ekson ve intron bölgelerinin tanınmasını sağlayarak hastalık tespitinde; genlerin mutasyon geçirip geçirmediklerinin keşfinde; büyümenin, gelişmenin; nerede, nasıl ve ne zaman düzenleneceğinin kontrolünde; hücrelerin çoğalma ve ölme durumlarının incelenmesinde; kök hücrelerin nerede, nasıl bir değişiklik yaşayacağının bulunup ortaya çıkarılmasında bilgisayar destekli güçlü bir sistemin oluşturulmasını sağlamaktır[4].

Bu doğrultuda yapılan çalışmalar incelenerek karşılaştırmalar gerçekleştirilmiştir. [36] çalışmasında ökaryot genler kapsamında kodlama bölgelerinin tespiti için walsh kodları üzerinde temellenen bir çalışma önerilmiştir. Bu çerçevede NCBI veri kümesinden elde edilen AF099922.1 numaralı gen dizilimi üzerinde yapılan test işlemi sonucunda %94'lük bir başarı oranı elde edilmiştir. [6] çalışmasında DNA konfigürasyonundaki bulanık davranışlar ve nükleotitlerin genetik kodu nedeniyle protein kodlayan bölgelerin saptanılabilmesi için gabor dalgacık dönüşümü yöntemi kullanılmıştır. NCBI veri kümesinden elde edilen AF099922.1 numaralı gen dizilimi üzerinde test edilen 5 farklı ekson bölgelerinin tahmininde ulaşılan başarı oranları sırasıyla %68,8, %79,5, %93,4, %90,2 ve %94,3 olarak ifade edilmiştir. [4] çalışmasında Ensembl veritabanından elde edilen ENSG00000103313 erişim numarasına ait gen dizilimi üzerinde ekson ve intron bölgelerinin sınıflandırılması sağlanmıştır. Yapılan sınıflandırma sonucunda elde edilen başarı oranı ayrık fourier dönüşümü ile %96,20 olarak bulunmuştur. [37] çalışmasında entropi tabanlı haritalama tekniği ile sayısallaştırılan DNA dizilimleri üzerinde VGG16, VGG19 ve ResNet öğrenme aktarımı mimarileri kulllanılarak özellik çıkarımı sağlanmıştır. En yakın komşu algoritması ve destek vektör makineleri algoritmaları ile yapılan 10 kat çapraz doğrulama sonucunda elde edilen en yüksek başarı oranı %97,8 olarak bulunmuştur. [24] USCS Assembly gen bankasından elde edilen insan genomunun kısa dizilimleri (GRCh37/hg19) üzerinde yapılan çalışmada intron ve ekson bölgelerinin sınıflandırılması açısından sayısallaştırma teknikleri arasında bir karşılaştırma yapılmıştır. Sunulan çalışma sonucunda eşleştirilmiş haritalama tekniği; ekson bölgelerini tanımlama hususunda %92,2 oranında, intron bölgelerini tanımlama hususunda %72,3 oranında başarı elde etmiştir. [20] çalışmasında kodlama bölgelerinin tahmini için sayısal sinyal işleme temelli yeni bir teknik önerilmiştir. Bu kapsamda uyarlamalı temsil yöntemi ile NCBI veri kümesinden elde edilen AF099922.1 numaralı DNA diziliminin sayısallaştırılma aşamasından sonra sinyal işleme ile kodlama bölgelerinin tanımlanması sağlanmıştır.

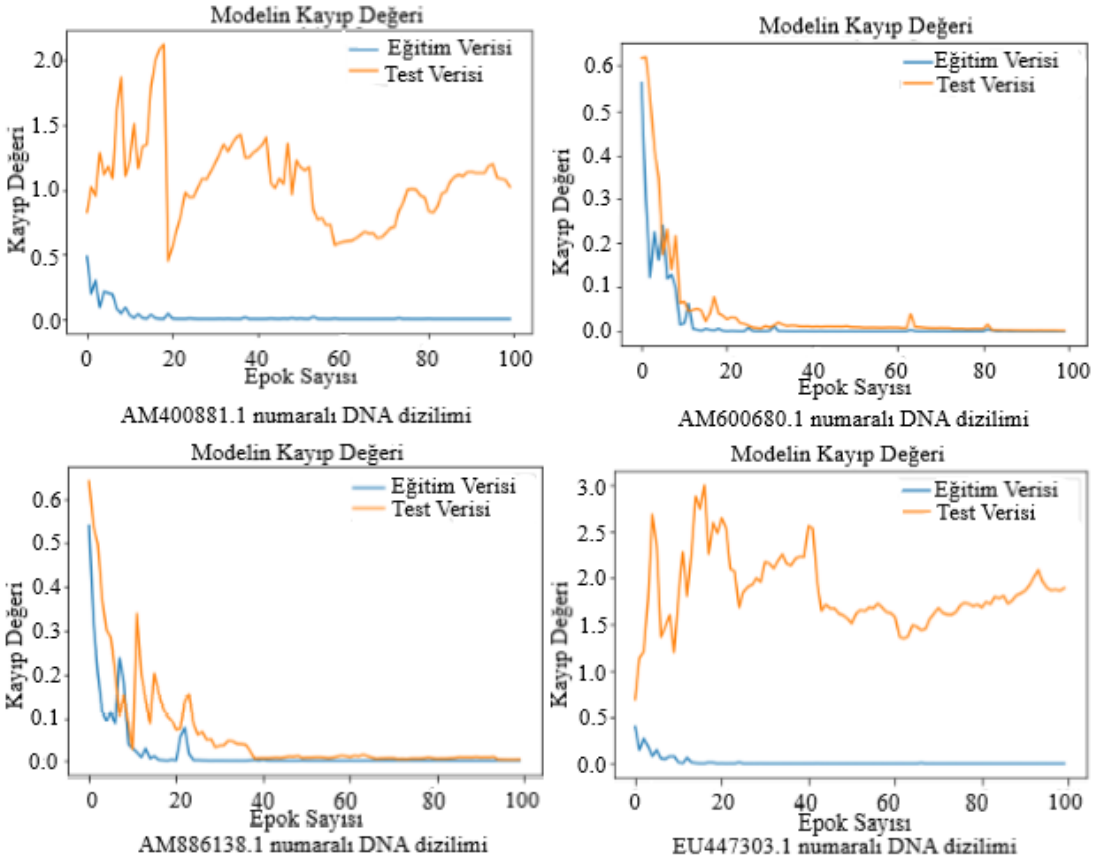
AM400881.1 numaralı DNA dizilimi	AM600680.1 numaralı DNA dizilimi	AM886138.1 numaralı DNA dizilimi	EU447303.1 numaralı DNA dizilimi
Test Veri Kümesi	Eğitim Veri Kümesi	Eğitim Veri Kümesi	Eğitim Veri Kümesi
AM400881.1 numaralı DNA dizilimi	AM600680.1 numaralı DNA dizilimi	AM886138.1 numaralı DNA dizilimi	EU447303.1 numaralı DNA dizilimi
Eğitim Veri Kümesi	Test Veri Kümesi	Eğitim Veri Kümesi	Eğitim Veri Kümesi
AM400881.1 numaralı DNA dizilimi	AM600680.1 numaralı DNA dizilimi	AM886138.1 numaralı DNA dizilimi	EU447303.1 numaralı DNA dizilimi
Eğitim Veri Kümesi	Eğitim Veri Kümesi	Test Veri Kümesi	Eğitim Veri Kümesi
AM400881.1 numaralı DNA dizilimi	AM600680.1 numaralı DNA dizilimi	AM886138.1 numaralı DNA dizilimi	EU447303.1 numaralı DNA dizilimi
Eğitim Veri Kümesi	Eğitim Veri Kümesi	Eğitim Veri Kümesi	Test Veri Kümesi

Şekil 6. Eğitim ve test veri kümeleri (Training and testing datasets)



Şekil 7. Dört farklı test kümesi üzerinde gerçekleştirilen sınıflandırma sonucunda elde edilen başarı oranları (The success rates obtained as a result of the classification performed on four different test sets)

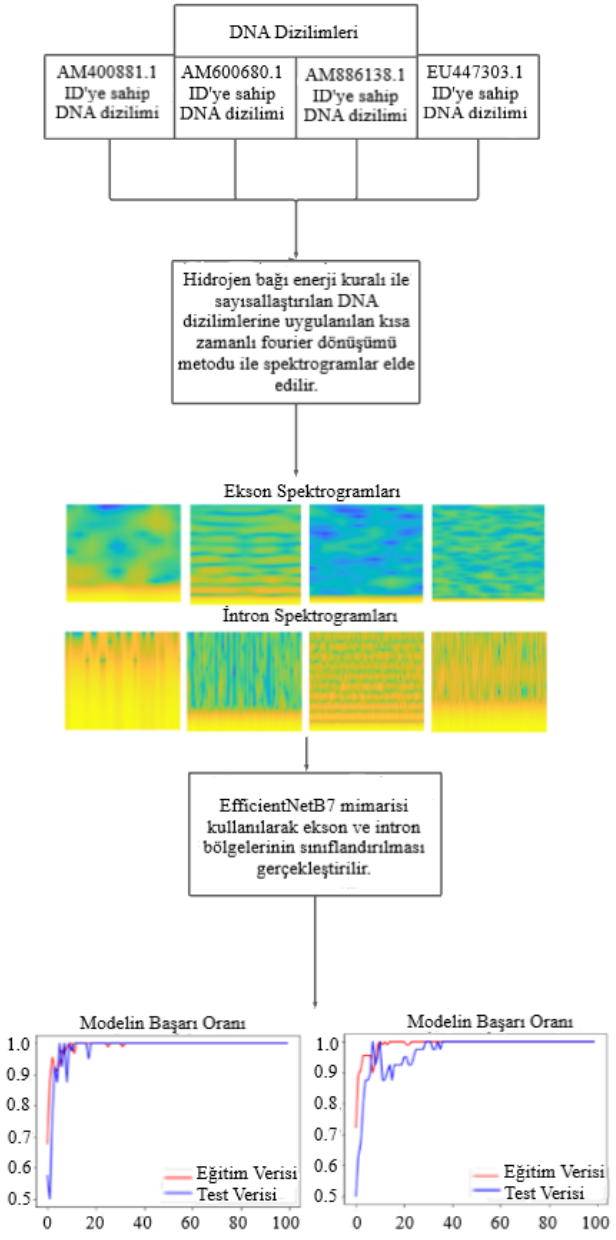
Asp67, HMR195 ve ALLSEQ gen dizilimleri üzerinde test edilen çalışmanın başarı oranları sırasıyla %79,%72 ve %75 olarak bulunmuştur. [19] çalışmasında ekson ve intron bölgelerinin belirlenebilmesi için DNA dizilimleri üzerinden



Şekil 8. Dört farklı test kümesi üzerinde gerçekleştirilen sınıflandırma sonucunda elde edilen kayıp değerleri
(The loss values obtained as a result of the classification performed on 4 different test sets)

istatistiksel bilginin çıkarılması hedeflenmiştir. Bu doğrultuda dalgacık tabanlı zaman serisi yaklaşımı önerilmiştir. Ardından destek vektör makineleri algoritması kullanılarak insan genomları üzerinde gerçekleştirilen sınıflandırma sonucunda elde edilen başarı oranları sırasıyla %83,8, %85,73, %86,5 ve %87,4 olarak belirtilirken önerilen yöntemin hiç görmediği bir veri kümesi üzerinde çalışması ile sağlanılan başarı oranları sırasıyla %85,95, %87,7, %88,5 ve %88,95 olarak bulunmuştur. [38] çalışmasında, ekson bölgelerinin doğru bir şekilde tanımlanabilmesi amacıyla zaman-frekans filtreleme yaklaşımı (S dönüşümü yöntemi) önerilmiştir. Bu doğrultuda NCBI veri kümesinden elde edilen AF099922.1 numaralı gen dizilimi üzerinde yapılan çalışma sonucunda ulaşılan ortalama performans %96 olarak bulunmuştur. Bu makalede sunulan çalışmada ise NCBI veri kümesinden elde edilen AM400881.1, AM600680.1, AM886138.1, ve EU447303.1 numaralarına sahip BCR-ABL genleri üzerinde analizler yapılmıştır. Analizlerin sonuçları incelendiğinde eğitim kümesinde yer alan farklı uzunluklara sahip ekson ve intron bölgelerinin eğitilmesinin ardından test kümesi üzerinde gerçekleştirilen tahmin süreci ile yüksek başarı oranı elde edilmiştir. Çünkü kullanılan derin öğrenme tekniği eğitim kümesinde öğrendiği verilerin test kümesindeki tahmini üzerinden çalışmaktadır. Dolayısıyla eğitim veri kümesi içerisinde spektrogram olarak ifade edilen ekson ve intron bölgelerinin farklı uzunluğa sahip dizilimlerinin, test veri kümesi üzerinde yapılacak tahmin için büyük önem taşıdığı bilinmektedir. Bu bağlamda sırasıyla test kümesi olarak seçilen AM600680.1 ve AM886138.1 numaralı DNA dizilimleri test edilirken eğitim veri kümesi içerisinde farklı uzunluklarda yer alan ekson ve intron dizilimlerinin bulunması sayesinde %100 başarı oranı elde edilmiştir.

Test kümesi olarak değerlendirilen AM400881.1 numaralı DNA diziliminin test edilmesi sonucunda %66,67 oranına ulaşılmıştır. Bu çerçevede DNA diziliminin ekson ve intron bölgeleri incelendiğinde tüm veri kümesi içerisinde yer alan en kısa ekson ve en kısa intron dizilimini bulundurduğu görülmektedir. Diğer yandan EU447303.1 numaralı DNA diziliminden oluşan test kümesi üzerinde gerçekleşen test sürecinde ise bu DNA diziliminin tüm DNA dizilimleri içerisinde yer alan en uzun ekson dizilimini barındırdığı görülmektedir. Eğitilen modelin bilmediği çok farklı uzunluğa sahip bir gen dizilimi üzerinde çıkarım yapması, önerilen derin öğrenme yönteminin doğası açısından sağlıklı sonuç vermeyeceği için AM400881.1 ve EU447303.1 numaralı DNA dizilimleri çalışmanın başarısının değerlendirilmesinde dikkate alınmamıştır. Dolayısıyla zenginleştirilen eğitim kümesi üzerinde önerilen modelin test edilmesi önemli bir kriter olarak belirlenmiştir.



Şekil 9. Çalışmanın ikinci aşamasında uygulanan süreçler
(Processes applied in the second stage of the study)

9. SONUÇLAR VE GELECEK ÇALIŞMALAR (CONCLUSIONS AND FUTURE WORKS)

Bu çalışmada sayısallaştırılan DNA dizilimleri üzerinde gerçekleştirilen sinyal işleme yaklaşımları ile dizilimlerde yer alan gizli bilgilerin açığa çıkarılması sağlanmıştır. Ardından ekson ve intron bölgelerinin tahmini için derin öğrenme çerçevesinde değerlendirilen öğrenme aktarımı yöntemi ile sınıflandırma süreci gerçekleştirilmiş ve değerlendirmeler yapılmıştır.

AM600680.1 ve AM886138.1 numaralı DNA dizilimleri için pencereleme yöntemleri ile zenginleştirilmiş veri kümesi üzerinde gerçekleştirilen çalışma sonucunda %100 başarı oranı elde edilmiştir. Ekson ve intron bölgelerinin teşhisinde incelenen çalışmalara kıyasla güçlü bir orana ulaşılması tıp dünyasında tercih edilebilecek bir çalışma olduğunu göstermektedir.

Gelecekte, DNA dizilimlerinden öğrenme aktarım mimarileri ile çıkarılan özellikler üzerinde bulanık mantık yöntemi uygulanılarak DNA'nın belirsiz yapısına uygun bir şekilde bulanıklaştırılan değerlerin farklı uzunluklara sahip olan ekson ve intron bölgelerinin tespitinde avantajlı bir durum oluşturacağı öngörülmektedir.

KAYNAKLAR (REFERENCES)

1. Barman S., Saha S., Mandal A., Roy M., Prediction of protein coding regions of a DNA sequence through spectral analysis, 2012 International Conference on Informatics, Electronics and Vision, 12–16, 2012.
2. Yu N., Li Z., Yu Z., Survey on encoding schemes for genomic data representation and feature learning-from signal processing to machine learning, Big Data Mining and Analytics, 1 (3), 191–210, 2018.
3. Hota M.K., Srivastava V.K., Performance analysis of different DNA to numerical mapping techniques for identification of protein coding regions using tapered window based short-time discrete Fourier transform, ICPCES 2010 - International Conference on Power, Control and Embedded Systems, 0–3, 2010.
4. Das B., Türkoglu I., Classification of DNA sequences using numerical mapping techniques and Fourier transformation, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (4), 921–932, 2016.
5. Das L., Das J.K. Nanda S., Detection of exon location in eukaryotic DNA using a fuzzy adaptive Gabor wavelet transform, Genomics, 112 (6), 4406–4416, 2020.
6. Das L., Nanda S., Das J. K., An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser window, Genomics, 111 (3), 284–296, 2019.
7. Meyer C., Scalzitti N, Jeannin-Girardon A., Collet P., Poch O., Thompson J.D., Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes, BMC Bioinformatics, 21 (1), 1–16, 2020.
8. Khodaei A., Feizi-Derakhshi M.R., Mozaffari-Tazehkand B., A pattern recognition model to distinguish cancerous DNA sequences via signal processing methods, Soft Computing, 24 (21), 16315–16334, 2020.
9. Chakraborty S., Gupta V., DWT based cancer identification using EIIP, Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, 718–723, 2016.
10. Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R., Prediction of probable genes by fourier analysis of genomic sequences, Bioinformatics, 13 (3), 263–270, 1997.
11. Raidi G. R., Gottlieb J., Lecture Notes in Computer Science: Preface, 3448, 2005.

12. Duran K., Yüksek Lisans Tezi, İTÜ, Fen Bilimleri Enstitüsü, İstanbul, 2013.

13. Anastassiou D., IEEE Signal Processing Magazine., 8–20, 2001.

14. Liu D., Jia R., Wang C., Arunkumar N., Narasimhan K., Udayakumar M., Elamaran V., Automated detection of cancerous genomic sequences using genomic signal processing and machine learning, Future Generation Computer Systems, 98, 233–237, 2019.

15. Roy M., Barman S., Spectral analysis of coding and non-coding regions of a DNA sequence by Parametric method, Proceeding of the 2010 Annual IEEE India Conference: Green Energy, Computing and Communication, 7–10, 2010.

16. Hsieh, S.J., Lin, C.Y., Chung, Y.S., Tang, C.Y., Comparative exon prediction based on heuristic coding region alignment, Proc. Int. Symp. Parallel Archit. Algorithms Networks,14–19, 2005.

17. Abo-Zahhai M., Ahmed S.M., Abd-Elrahman S.A., A new numerical mapping technique for recognition of exons and introns in DNA sequences, National Radio Science Conference NRSC, Proceedings, 573–580, 2013.

18. Das B., Turkoglu I., A novel numerical mapping method based on entropy for digitizing DNA sequences, Neural Computing and Applications, 29 (8), 207–215, 2018.

19. Gupta R., Mittal A., Singh K., Bajpai P., Prakash S., A Time Series Approach for Identification of Exons and Introns, 91–93, 2008.

20. Marhon S.A., Kremer S.C., Protein coding region prediction based on the adaptive representation method, Canadian Conference on Electrical and Computer Engineering, 000415–000418, 2011.

21. Li J., Zhang L., Li H., Ping Y., Xu Q., Wang R., Tan R., Wang Z., Liu B., Wang Y., Integrated entropy-based approach for analyzing exons and introns in DNA sequences, BMC Bioinformatics, 20, 11–13, 2019.

22. Dessouky A.M., Taha T.E., Dessouky M.M., Eltholth A.A., Hassan E., Abd El-Samie F., Non-parametric spectral estimation techniques for DNA sequence analysis and exon region prediction, Computer and Electrical Engineering, 73, 334–348, 2019.

23. Singh A.K., Srivastava V.K., The three base periodicity of protein coding sequences and its application in exon prediction, 2020 7th International. Conference Signal Processing and Integrated Networks, 64, 1089–1094, 2020.

24. Abo-Zahhad M., Ahmed S.M. Abd-Elrahman S.A., Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques, International Journal of Information Technology and Computer Science, 4 (8), 22–36, 2012.

25. Das B., Doktora Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ, 2018.

26. Toraman S., Türkoğlu İ., A new method for classifying colon cancer patients and healthy people from FTIR signals using wavelet transform and machine learning techniques, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (2), 933–942, 2020.

27. Aygün O., Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya, 2006.

28. Avci K., Coskun O., Spectral performance analysis of cosh window based new two parameter hybrid windows, 26th IEEE Signal Processing and Communications Applications Conference SIU, 1–4, 2018.

29. Hashimoto D.A., Ward T.M., Meireles O.R., The Role of Artificial Intelligence in Surgery, Advances in Surgery, 54, 89–101, 2020.

30. Narin A., İşler Y., Detection of new coronavirus disease from chest x-ray images using pre-trained convolutional neural networks, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (4), 2095–2107, 2021.

31. Gürkahraman K., Karakiş R., Brain tumors classification with deep learning using data augmentation, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (2), 997–1011, 2021.

32. Atila Ü., Uçar M., Akyol K., Uçar E., Plant leaf disease classification using EfficientNet deep learning model, Ecological Informatics, 61, 2021.

33. Elmas B., Identifying species of trees through bark images by convolutional neural networks with transfer learning method, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (3),1253–1269, 2021.

34. Sreng S., Maneerat N., Hamamoto K., Win K.Y., Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images, Applied Sciences (Switzerland)., 10 (14), 2020.

35. Muftuoglu, Z., Kizrak, M. A., Yıldırım, T., Differential Privacy Practice on Diagnosis of COVID-19 Radiology Imaging Using EfficientNet, International Conference on Innovations in Intelligent Systems and Application Proceedings, 2020.

36. Kumar M.R., Vaegae N.K, Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes, Biomedical Signal Processing and Control, 58, 2020.

37. Daş B., Toraman S., Türkoğlu I., A novel genome analysis method with the entropy-based numerical technique using pretrained convolutional neural networks, Turkish Journal of Electrical Engineering and Computer Sciences, 28 (4), 1932–1948, 2020.

38. Sahu S.S., Panda G., Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach, Genomics, Proteomics Bioinformatics, 9, 45–55, 2011.