Journal of the Faculty of Engineering and Architecture of Gazi University 38:2 (2023) 707-719

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 ALL and CML malignancies being among the main types of leukaemia with graph neural networks and fuzzy logic algorithm

Fatma Akalın[1]*, Nejat Yumuşak[2]

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

Highlights:

- Impact of DNA mapping techniques on performance
- Evaluating the success of digital signal processing techniques
- Classification of ALL and CML malignancies

Keywords:

- DNA sequences
- Entropy-based/traditional mapping techniques
- Digital signal processing methods
- DGCNN approach
- Adaptive fuzzy logic algorithm

Article Info:

Research Article
Received: 12.11.2021
Accepted: 21.03.2022

DOI:

10.17341/gazimmfd.1022624

Correspondence:

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

Graphical/Tabular Abstract

It is realized to differentiate ALL and CML malignancies, which are the main types of leukaemia, using the BCR-ABL gene. The flow diagram of the alternative method, which is made in two different ways with the support of computerized diagnostic systems, is given in Figure A.

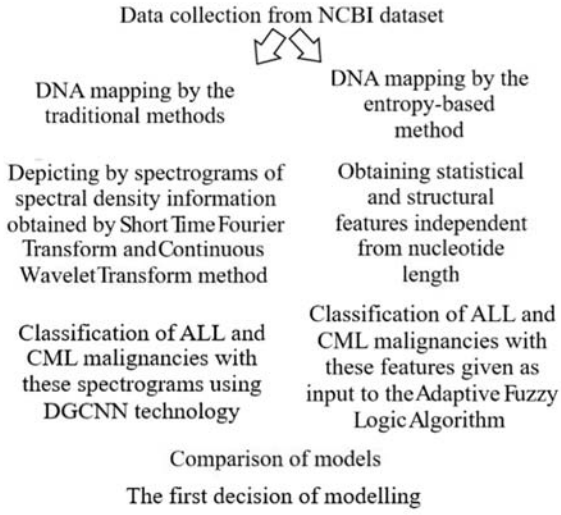


Figure A. Flow chart related to the classification of ALL and CML malignancies being among the main types of leukaemia

Purpose: To provide a classification of ALL and CML labelled spectrogram obtained using digital signal processing techniques on the DNA genome sequence

Theory and Methods: This study investigates ALL and CML malignancies with different nucleotide lengths obtained from the NCBI database. First, DNA sequences with a symbolic structure are digitized using conventional mapping techniques. Digitized DNA sequences are processed by Short-Time Fourier Transform and Continuous Wavelet Transform methods. The obtained spectral density information is depicted by means of spectrograms. Spectrograms with ALL and CML malignancies are classified with the DGCNN approach, which allows the hybrid use of CNN and GNN technologies. However, the expressions on the spectrogram do not provide a clear output as it is affected by the nucleotide length. Therefore, a study independent of the length of the sequences is carried out. In the first stage, DNA sequences are digitized with the Shannon Entropy-Based mapping technique, in which the codon distributions are taken into account. Then, the information extracted by the analyses performed on these numerical representations is then evaluated with the Adaptive Fuzzy Logic algorithm.

Results: This article provides the classification of ALL and CML malignancies from the NCBI database. The maximum success rate achieved in this study, which is carried out in two parts, was found to be 80%.

Conclusion: Multidisciplinary studies offered by the world of medicine and informatics carry out research to discover non-invasive and alternative methods in the realization of successful evaluations. In this context, in the current study, it is ensured that the diagnosis and treatment process is carried out effectively.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:2 (2023) 707-719

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

Lösemi hastalığının temel türlerinden ALL ve KML malignitelerinin graf sinir ağları ve bulanık mantık algoritması 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

- DNA haritalama tekniklerinin performans üzerinde etkisi
- Sayısal sinyal işleme tekniklerinin başarısının değerlendirilmesi
- ALL ve KML malignitelerinin sınıflandırılması

Makale Bilgileri

Araştırma Makalesi
Geliş: 12.11.2021
Kabul: 21.03.2022

DOI:
10.17341/gazimmfd.1022624

Anahtar Kelimeler:
DNA dizilimleri, entropi tabanlı / geleneksel haritalama teknikleri, sayısal sinyal işleme metotları, DGCNN yaklaşımı, uyarlanabilir bulanık mantık algoritması

ÖZ

Beyaz kan hücresi kanseri olan lösemi, yaşam kalitesini düşüren ve ilerleyen aşamalarda ölüme sebep olabilen maliyeti yüksek bir malignitedir. Farklı yaş gruplarında görülebilen bu hastalığın erken ve doğru teşhisinin sağlanması tedavi sürecini etkilemekte ve hastalığın ilerlemesini engellemektedir. Bu çalışmada löseminin temel türlerinden olan ALL ve KML malignitelerinin sınıflandırılması amaçlanmıştır. Genetik temelli maligniteler olan bu türlerin ayırt edilmesinde DNA'da bir mutasyon sonucunda beliren BCR-ABL geni analiz edilmiştir. Tıp dünyasında BCR-ABL geni üzerinden mevcut türlerin ayırt edilmesi hususunda PCR tekniği kullanılarak değerlendirmeler yapılabilmektedir. Ancak teşhis ve tedavi sürecindeki maliyetin ve zamanın indirgenmesi amacıyla disiplinlerarası çalışmalar da mevcuttur. İki aşamadan oluşan bu çalışmanın ilk aşamasında farklı nükleotit uzunluklarına sahip ALL ve KML DNA dizilimlerinin spektral yoğunluk bilgisi sinyal işleme teknikleri kullanılarak spektrogramlara yansıtılmıştır. Ardından CNN ve GNN teknolojilerinin hibrit yaklaşımı olan DGCNN teknolojisi ile ALL ve KML malignitelerine ait spektrogramlar sınıflandırılmıştır. Fakat nükleotitlerin farklı uzunluklarda olmasından dolayı spektrogramlar üzerinde net ifadeler elde edilememiştir. Çalışmanın ikinci aşamasında farklı uzunluklara sahip DNA dizilimleri, kodon dağılımlarının esas alındığı entropi temelli haritalama tekniği ile sayısallaştırılmıştır. Sayısallaştırılan bu dizilimler üzerinden çıkarılan istatiktiksel ve yapısal özellikler uyarlanabilir bulanık mantık algoritması ile sınıflandırılarak nükleotit uzunluğundan bağımsız bir çalışma gerçekleştirilmiştir. Böylece KML ve ALL malignitelerinin sınıflandırılmasında maksimum %80'lik bir başarı düzeyine ulaşılmıştır.

Classification of ALL and CML malignancies being among the main types of leukaemia with graph neural networks and fuzzy logic algorithm

HIGHLIGHTS

- Impact of DNA mapping techniques on performance
- Evaluating the success of digital signal processing techniques
- Classification of ALL and CML malignancies

Article Info

Research Article
Received: 12.11.2021
Accepted: 21.03.2022

DOI:
10.17341/gazimmfd.1022624

Keywords:
DNA sequences, entropy-based/traditional mapping techniques, digital signal processing methods, DGCNN approach, adaptive fuzzy logic algorithm

ABSTRACT

Leukaemia, a white blood cell cancer, is a costly malignancy that reduces the quality of life and can lead to death in the later stages. Early and accurate diagnosis of this disease, which can be seen in different age groups, affects the treatment process and prevents the progression of the disease. This study aims to classify ALL and CML malignancies, which are the main types of leukaemia. The BCR-ABL gene, which appears due to a mutation in the DNA, is analyzed to distinguish these types, which are genetically based malignancies. In the medical world, evaluations can be made using the PCR technique to distinguish the existing species over the BCR-ABL gene. There are also interdisciplinary studies to reduce the cost and time in the diagnosis and treatment process. In the first stage of this two-stage study, the spectral density information of ALL and CML DNA sequences with different nucleotide lengths is reflected on the spectrograms using signal processing techniques. Then, the spectrograms of ALL and CML malignancies are classified with DGCNN technology, which is a hybrid approach of CNN and GNN technologies. However, due to the different lengths of the nucleotides, clear expressions on the spectrograms could not be obtained. In the second stage of the study, DNA sequences with different lengths are digitized with the entropy-based mapping technique based on codon distributions. Statistical and structural features extracted from these digitized sequences are classified by the adaptive fuzzy logic algorithm, and a study independent of nucleotide length is performed. Thus, a maximum 80% success level is achieved in the classification of CML and ALL malignancies.

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

1. Giriş (Introduction)

Lösemi, kan hücrelerinin miktarındaki kontrolsüz artış ile ortaya çıkan ve vücudun kan üretim mekanizmasını etkileyen malignitedir. Farklı yaş gruplarında görülebilen bu malignite; akut miyeloid, akut lenfoblastik, kronik miyeloid ve kronik lenfositik olmak üzere 4 temel türe sahiptir. Bu türlerin ayrımı, hastalığın vücut içerisindeki ilerlemesinden ve hücre tipinde görülen farklılıklardan oluşmaktadır. Bu doğrultuda kanser hücrelerinin hızlı bir şekilde çoğaldığı lösemi türleri "akut" olarak adlandırılırken ilk belirtilerin yavaş oluştuğu ve yavaş yayılım gösteren lösemi türleri "kronik" olarak adlandırılmaktadır. "Miyeloid" ya da "lenfositik" şeklindeki tanımlamalar ise hücre tipinin farklılığını ifade etmektedir [1].

Lösemi hastalığı, kişinin yaşam kalitesini düşüren, iş gücü kaybına yol açan ve ilerleyen safhalarda ölümle sonuçlanma ihtimali olan bir kanser türüdür. Hastalığın tanısını koyabilmek için ilk aşamada gerçekleştirilen fiziksel muayene, hastalığın teşhisi hususunda yeterli bir netice sunmamaktadır. Tanı koymak için kan tahlillerinin ve biyopsilerin yapılması da gerekmektedir. Belirtilen bu işlemlerin yanı sıra hastalığın durumunun net olarak değerlendirilebilmesi için X ısınları, ultrasonlar ve CT taramalar gibi görüntülemelerin; akış sitometrisi, karaciğer fonksiyonu gibi testlerin varlığı da önem taşımaktadır. Öte yandan tanısı konulan lösemi hastalığının alt tiplerine karar verebilmek için mikroskop altına yerleştirilen hücrelerin incelenmesine imkân tanıyan periferik yayma tekniği, farklı amaçlar doğrultusunda gerçekleştirilen başka bir işlemdir [2]. Lösemi hastalığı kapsamında tüm bu süreçleri yaşayan hastalar ve doktorlar üzerindeki yükün azaltılması bilim insanlarının çalışmaları içerisinde konumunu korumaktadır.

Sunulan bu çalışmada, DNA dizilimleri üzerinde gerçekleştirilen analizler sonucunda lösemi hastalığının kronik miyeloid türünün ya da Ph-pozitif adı verilen spesifik akut lenfoblastik türünün ayrımı için çalışılmıştır. Bu iki lösemi türünün tanısında önemli yer tutan Philadelphia (Ph) kromozomu 1960 yılında keşfedilmiştir [3]. 22 numaralı kromozom üzerinde yer alan BCR geni ile 9 numaralı kromozom üzerinde yer alan ABL geninin 22 numaralı kromozom üzerinde gerçekleştirdiği bir mutasyon sonucunda oluşmuştur [4-5].

Ph kromozomu; KML tanısı almış hastaların %95'inde, yetişkin ALL tanısı almış hastaların %30'unda, çocuk ALL tanısı almış hastaların %5'inden daha az bir oranında ve nadiren AML hastalarında bulunan bir anomalidir [3].

Bu çerçevede KML ve ph-pozitif ALL türünün teşhisi için önemli bir biyobelirteç olan Ph kromozomunun tanımlanmasında kullanılan gerçek zamanlı PCR teknikleri, zaman içerisinde geleneksel sitogenetik tekniklerin yerini almıştır [6]. Ancak minimum maliyet ile sonuca ulaşmak, en kısa sürede en doğru yanıtı almak, uygulanılabilmesi kolay ve hastanın vücut bütünlüğünü bozmadan tedavi edilmesi amaçlanan non-invaziv metotların [7] ve alternatif PCR tekniklerinin keşfi üzerinde de halen çalışmalar devam etmektedir [8-11].

Öte yandan, genom teknolojisindeki son gelişmeler ile birlikte dizilim sayılarının miktarındaki artış, dizilimlerin analiz edilmesi için bilgisayar destekli sistemlerin gerekliliğini ön plana çıkaran diğer bir konudur. 2000'li yıllardan bu yana genomik veriler üzerinde hesaplamalı ve analitik olarak farklı birçok çalışma yapan bilişim dünyası, DNA parçaları içerisindeki gizli özelliklerin ve periyodiklik durumlarının elde edilmesi amacıyla genomik sinyal işleme alanını da kullanarak güçlü analizler sunmayı hedeflemiştir [12-13]. Aynı zamanda DNA dizilimleri üzerindeki mevcut düzensizliklerin tespiti ve genetik tabanlı hastalıklar üzerinde doğru tanının konulabilmesi için gerçekleştirilen bilgisayar destekli çalışmalar incelenmiş ve değerlendirmeler yapılmıştır.

Bu kapsamda, NCBI genom veri kümesinden elde edilen akciğer, meme, yumurtalık kanserine sahip DNA dizilimleri ile normal DNA dizilimlerine ilişkin analizin sağlandığı mevcut çalışmada sayısal haritalama tekniği ile sayısallaştırılan DNA yapısına, ayrık dalgacık yöntemi uygulanmıştır. Ardından dalgacık alanından elde edilen ortalama, medyan, standart sapma, çeyrekler arası aralık, çarpıklık ve basıklık gibi istatistiksel özellikler kullanılarak sınıflandırma gerçekleştirilmiştir [14]. Prostat, meme, kolon ve mide hücreleri ile sağlıklı hücrelerin genleri üzerinde gerçekleştirilen çalışmada Bayesian fusion tekniği ile ayrık fourier dönüşümünün kombinasyonu sonucunda yapılan analizlerde sağlıklı genlerin hasta genlerden ayrımında belirgin bir farklılık olduğu ifade edilmiştir [15]. Kanser hastalığının biyolojik deneyler kullanılmadan sınıflandırılabilmesi ve DNA parçaları içerisindeki gizli özelliklerin ve periyodiklik durumlarının ortaya çıkartılabilmesi amacıyla sayısal sinyal işleme yönteminin önerildiği çalışmada EIIP (Electron-Ion Interaction Pseudopotential) haritalama tekniği ile sayısallaştırılan DNA dizilimleri üzerinde ayrık dalgacık dönüşüm metodu kullanılarak protein kodlama bölgeleri üzerindeki mevcut anormalliklerin tahmin edilmesi amaçlanmıştır. Elde edilen sonuçların kanserli ve kanserli olmayan hücrelerin analizinde önemli bir rol oynadığı belirtilmiştir [13]. Prostat verileri kullanılarak kanser ile ilişkili olabilecek genetik dizilim varyantlarının verimli bir şekilde belirlenebilmesinin hedeflendiği çalışmada kanserli ve normal örnekler arasından genetik dizilim varyantlarının istatistiksel özellikleri karşılaştırılmıştır. Prostat kanserine sahip kişiler ile genetik dizilim varyantları arasında anlamlı bir istatiksel akış tanımlandığı belirtilmiştir [16]. DNA dizilimleri içerisinde silme(deletion), ters çevirme(inversion), ekleme(insertion), ve tekrar(duplication) gibi bazı yapısal farklılıkların gözlemlendiği çalışmada bu tür genomik anomalilerin tespit edilmesi için kullanılan mevcut DNA dizileme yöntemlerinin maliyetli olabilmesinden dolayı bilgisayar destekli bir yöntemin geliştirilmesi planlanmıştır. Bu doğrultuda çeşitli varyasyonların yerini tahmin etmek amacıyla sinir ağları tekniği kullanılarak genomik anomalilerin tespit edilmesini sağlayan bir yaklaşım önerilmiştir [17]. Farklı direnç özellikleri gösteren Mycobacterium tuberculosis'un genomik dizi analizi için dalgacık ayrışma katsayılarının enerjisinin incelendiği çalışmada, hesaplanan enerji miktarına dayanarak ilaca karşı gösterilen duyarlılıklar karşılaştırılmıştır. Elde edilen nihai çıktının geleneksel laboratuvar yöntemlerine göre daha kısa sürede gerçekleştiği ifade edilmiştir [18]. Hücrenin genetik yapısında gerçekleşen değişiklikler nedeni ile meydana gelen kanser hastalığının ele alındığı çalışmada ilk olarak hastalığın oluşmasında ve yayılmasında önemli bir faktör olan DNA dizilimleri üzerinde ayrık fourier dönüşümü ve anti-notch sayısal filtre modelinin hibrit kullanımı uygulanmıştır. Ardından DNA dizilimleri içerisinde yer alan protein kodlama bölgelerinin özelliklerine dayanarak kanserli olan ve kanserli olmayan örneklerin tanımlanması için istatistiksel teknikler kullanılmıştır. Destek vektör makineleri algoritması ile gerçekleştirilen sınıflandırmanın sunduğu çıktılar üzerinde yüksek doğruluk ve kesinlik gösterildiği ifade edilmiştir [12]. Kanser hastalığının ana nedenlerinden birisi olan genetik anormalliklerin incelendiği çalışmada, anormalliklerin analiz edilebilmesinde kullanılan dalgacık dönüşümü yöntemi ile değerlendirmeler yapılmıştır. DNA dizilimleri üzerinde meydana gelen büyük değişikliklerin tespitinde biyolojik deneylerin aksine daha uygun bir maliyet ve daha az bir zaman sunan sinyal işleme yöntemi ile kanser hastalığının teşhis edilmesinde tercih edilebilen bir çalışma olduğu ifade edilmiştir [19]. Kanser türlerini sınıflandırmak amacıyla dalgacık dönüşümü ile evrişimsel sinir ağı yöntemlerinin hibrit kullanımını gerçekleştiren çalışmada önerilen yaklaşımın state-of-the-art algoritmalarından daha iyi performans gösterdiği ve kanser

hastalarının prognostik riskini sınıflandırmada başarılı bir çözüm sunduğu belirtilmiştir [20]. Prostat kanseri hücreleri ile normal hücrelerin birbirinden ayırt edilebilmesinin amaçlandığı çalışmada EIIP haritalama tekniği kullanılarak sayısallaştırılan dizilimler, sinyal işleme tekniği ile birlikte kullanılan PCA modeli üzerinden test edilmiştir [21]. Kronik hepatit B virüsünün incelendiği çalışmada DNA dizilimine dayalı kolektif(ensemble) makine öğrenimi yöntemi ile karaciğer kanserinin tahmin edilmesi için bir değerlendirme yapılmıştır [22]. Prostat, meme ve kolon kanserlerine sahip genlerin kullanıldığı çalışmada entropi spekturumu kullanılarak aynı tip kanser genlerinin farklı tip kanser gruplarından ayırt edilmesi sağlanmıştır [23]. Yapılan çalışmalar değerlendirildiğinde, bilgisayar destekli genom dizileme teknolojisindeki gelişmeler ile birlikte hastalıkların erken teşhisine ve prognozuna yönelik yapılan araştırmalar ile biyolojik süreçlerdeki ilişkilerin ortaya çıkarılması sağlanarak verimli ve uygun maliyetli yaklaşımlar geliştirilmiştir.

Bu çalışmada, iki aşamadan meydana gelen bilgisayar destekli bir sistem inşa edilmiştir. İlk olarak NCBI veri kümesinden elde edilen ALL ve KML DNA genom dizilimleri üzerinde 8 farklı geleneksel haritalama tekniği kullanılarak sayısallaştırma işlemi gerçekleştirilmiştir. Ardından sayısallaştırılan DNA dizilimleri sırasıyla kısa zamanlı fourier dönüşümü ve sürekli dalgacık dönüşümü sinyal işleme teknikleri ile spektrogramlar olarak ifade edilmiştir. Sınıflandırma süreci, yeni bir yapay zekâ çerçevesi olarak önerilen ilişkiye duyarlı bir yapı olan GNN ve CNN metotlarının hibrit bir şekilde kullanılmasına imkân tanıyan DGCNN metodu ile gerçekleştirilmiştir. Ancak farklı uzunluklara sahip DNA dizilimleri spektrogramlar üzerinde net çıktılar üretememiştir. Dolayısıyla çalışmanın ikinci aşamasında, DNA'nın uzunluğundan bağımsız entropi tabanlı bir araştırma yapılmıştır. Kodon dağılımlarının sıklıkları üzerinden elde edilen istatistiksel ve yapısal değerlendirme sonuçları, bulanık mantık algoritması ile analiz edilerek sınıflandırma verimi arttırılmış ve %80 doğruluk oranına ulaşılmıştır.

2. Sayısal Haritalama Teknikleri (Numerical Mapping Techniques)

Organizmayı inşa eden ve canlılığın sürdürülmesinde önemli rol oynayan DNA; A, T, G ve C nükleotitlerinden meydana gelen sembolik bir dizilime sahiptir [24]. DNA'nın sembolik yapısı, dizilimlerin doğrudan işlenmesine engel olmaktadır. Bu amaçla literatürde önerilen farklı haritalama yaklaşımları ile dizilimlerin sayısallaştırılması planlanmıştır. Sabit haritalama tekniği içerisinde yer alan Reel Haritalama, Moleküler Kütle Haritalama, Tam Sayı Haritalama, Karmaşık Haritalama ile fiziko kimyasal özellik tabanlı haritalama tekniği içerisinde yer alan EIIP Haritalama, Atomik Haritalama, DNA-Yürüyüş Haritalama ve Eşleştirilmiş Haritalama çalışma kapsamında dizilimlerin sayısallaştırılmasında kullanılan tekniklerdir.

2.1. Sabit Haritalama Teknikleri (Fixed Mapping Techniques)

DNA dizilimlerini oluşturan bazların keyfi numerik değerler olarak ifade edilmesidir. Bu durum, 2 kısımda incelenmektedir. İlk kısımda; A, G, C ve T bazlarının dizilim içerisinde mevcut olup olmamasına göre bazlara 0 ve 1 sayıları atanırken ikinci kısımda; A, G, C ve T bazlarının dizilim içerisinde mevcut olup olmamasına göre bazlara reel veya karmaşık sayılar atanır. Reel Haritalama, Moleküler Kütle Haritalama, Tam Sayı Haritalama ve Karmaşık Haritalama sabit haritalama teknikleri içerisinde yer alan yöntemlerdir [25].

2.1.1. Reel haritalama tekniği (Real mapping technique)

DNA dizilimini oluşturan organik bazlara reel sayı temsillerinin atanmasıdır. Bu teknikte A, T, G ve C bazları sırasıyla -1.5, 1.5, -0.5 ve 0.5 değerleri ile ifade edilir. Örnek bir DNA dizilimi üzerinde reel haritalama tekniğine göre sayısallaştırılan dizilimin gösterimi Tablo 1'de verilmiştir [24-26].

2.1.2. Moleküler kütle haritalama tekniği (Molecular mass mapping technique)

DNA dizilimini oluşturan A, T, G ve C bazlarına sırasıyla 134, 125, 150 ve 110 değerlerinin atanmasıdır. Örnek bir DNA dizilimi üzerinde moleküler kütle haritalama tekniğine göre sayısallaştırılan dizilimin gösterimi Tablo 1'de verilmiştir [24-26].

2.1.3. Tamsayı haritalama tekniği (Integer mapping technique)

Pürin ve pirimidin bazlarına göre yapılan atama işlemidir. Bu kapsamda; A ve G pürin bazlarının toplam sayısının C ve T pirimidin bazlarının toplam sayısından büyük olması durumu incelenir. Bu koşulun sağlanması halinde; A, T, G ve C bazlarına sırasıyla 2, 0, 3 ve 1 değerleri atanır. Aksi bir durumda ise T bazının A bazından ve G bazının C bazından büyük olması araştırılır. Böyle bir koşulda A, T, G ve C nükleotitlerine sırasıyla 0, 2, 3 ve 1 değerleri atanır. Örnek bir DNA dizilimi üzerinde uygulanılan tamsayı haritalama tekniğine göre sayısallaştırılan dizilimin gösterimi Tablo 1'de verilmiştir [24-26].

2.1.4. Karmaşık haritalama tekniği (Complex mapping technique)

DNA dizilimini oluşturan nükleotitlerin tamamlayıcı özelliği kullanılarak 2 boyutlu bir düzlemde biyolojik özelliklere karşılık gelen matematiksel özelliklerin yansıtılmasını mümkün kılan bir haritalama tekniğidir. Bu kapsamda A, T, G ve C organik bazlarına sırasıyla 1+j, 1-j, -1+j ve -1-j değerleri atanır. Örnek bir DNA dizilimi üzerinde karmaşık haritalama tekniğine göre sayısallaştırılan dizilimin gösterimi Tablo 1'de verilmiştir [24-26].

2.2. Fiziko-Kimyasal Özellik Tabanlı Haritalama Teknikleri (Physico-Chemical Feature-Based Mapping Techniques)

DNA biyomoleküllerinin; biyofiziksel ve biyokimyasal özelliklerinin haritalanması sürecinde kullanılan teknikleri içermektedir. EIIP Haritalama, Atomik Sayı Haritalama, DNA-Yürüyüş Haritalama ve Eşleştirilmiş Sayısal Haritalama fiziko-kimyasal tabanlı haritalama teknikleri içerisinde değerlendirilmektedir [25].

2.2.1. EIIP haritalama tekniği (EIIP mapping technique)

DNA dizilimini oluşturan her bir nükleotidin EIIP temsilindeki yarı değerlik sayısı ile eşleştirilmesidir. Bu teknik ile DNA dizilimini oluşturan A, T, G ve C bazlarına sırası ile 0.1260, 0.1335, 0.0806 ve 0.1340 değerlerinin atanması sağlanır. Örnek bir DNA dizilimi üzerinde EIIP haritalama tekniğine göre sayısallaştırılan DNA diziliminin gösterimi Tablo 1'de verilmiştir [24-26].

2.2.2. Atomik sayı haritalama tekniği (Atomic number mapping technique)

DNA diziliminin bir dizi atom göstergesine dönüştürülmesi işleminde; A, T, G ve C organik bazlarına sırası ile 70, 66, 78 ve 58 değerleri atanır. Örnek bir DNA dizilimi üzerinde atomik sayı haritalama tekniğine göre sayısallaştırılan DNA dizilimi Tablo 1'de verilmiştir [24-26].

2.2.3. DNA-yürüyüş haritalama tekniği (DNA-walk mapping technique)

DNA-yürüyüş haritalama tekniğinde, varsayılan bir yürütecin her bir yürüyüşünün aşağı ve yukarı hareketi dikkate alınarak bir model oluşturulur ve seçilen baz çiftine göre DNA dizilimlerinin değerlendirilmesi sağlanır. Bu kapsamda seçilen baz çiftinin A ve C

Tablo 1. Örnek bir DNA diziliminin geleneksel haritalama teknikleri kullanılarak sayısallaştırılması [25]
(Digitizing of the sample DNA sequence using conventional mapping techniques)

	Haritalama Teknikleri (H.T.)	Haritalama tekniklerine göre örnek bir dizilim üzerinde gerçekleştirilen sayısal temsiller
Sabit Haritalama Teknikleri		… ATGCATGCAG …
	Reel H.T.	… -1.5, 1.5, -0.5, 0.5, -1.5, 1.5, -0.5, 0.5, -1.5, -0.5 …
		… ATGCATGCAG …
	Moleküler Kütle H.T.	… 134, 125, 150, 110, 134, 125, 150, 110, 134, 150 …
		… ATGCATGCAG …
	Tam Sayı H.T.	… 2, 0, 3, 1, 2, 0, 3, 1, 2, 3 …
		… ATGCATGCAG …
	Karmaşık H.T.	… 1+j, 1-j, -1+j, -1-j, 1+j, 1-j, -1+j, -1-j, 1+j, -1+j …
		… ATGCATGCAG …
	EIIP H.T.	… 0.1260, 0.1335, 0.0806, 0.1340, 0.1260, 0.1335 0.0806, 0.1340, 0.1260, 0.0806 …
Fiziko-Kimyasal Özellik Tabanlı Haritalama Teknikleri		… ATGCATGCAG …
	Atomik Sayı H.T.	… 70, 66, 78, 58, 70, 66, 78, 58, 70, 78 …
		… ATGCATGCAG …
	DNA-Yürüyüş H.T.	… -1, 0, 0, 1, -1, 0, 0, 1, -1, 0 …
	Eşleştirilmiş Sayısal H.T. (Hidrojen Bağı Enerji Kuralı)	… ATGCATGCAG … … -1, -1, 1, 1, -1, -1, 1, 1, -1, 1 …

olması durumunda sırasıyla -1 ve 1 değerleri atanırken T ve G olması durumunda sırasıyla 1 ve -1 değerleri atanır. Seçilen baz çifti haricindeki diğer bazlar için ise 0 değeri tayin edilir. Örnek bir DNA dizilimi üzerinde DNA-yürüyüş haritalama tekniğine göre sayısallaştırılan dizilim, Tablo 1'de verilmiştir [24-26].

2.2.4. *Eşleştirilmiş sayısal haritalama tekniği*
(Paired digital mapping technique)

Eşleştirilmiş sayısal haritalama, dizilimi oluşturuan A-T ve G-C bazlarına sırasıyla 1 ve -1 değerlerinin atanmasını sağlayan bir tekniktir. DNA diziliminin karmaşıklığının azaltılmasını mümkün kılan bu teknik kapsamında 7 farklı kural tanımlanmıştır [24-26].

2.2.4.1. *Pirimidin-pürin kuralı (Pyrimidine-purine rule)*

Sembolik DNA diziliminde pürin (A ya da G) bazlara 1, pirimidin (C ya da T) bazlara -1 değerinin atanmasıdır [25, 26].

2.2.4.2. *AA' kuralı (AA' rule)*

Sembolik DNA diziliminde A bazına 1 değeri atanırken diğer tüm bazlara -1 değerinin atanmasıdır [25, 26].

2.2.4.3. *TT' kuralı (TT' rule)*

Sembolik DNA diziliminde T bazına 1 değeri atanırken diğer tüm bazlara -1 değerinin atanmasıdır [25, 26].

2.2.4.4. *GG' kuralı (GG' rule)*

Sembolik DNA diziliminde G bazına 1 değeri atanırken diğer tüm bazlara -1 değerinin atanmasıdır [25, 26].

2.2.4.5. *CC' kuralı (CC' rule)*

Sembolik DNA diziliminde C bazına 1 değeri atanırken diğer tüm bazlara -1 değerinin atanmasıdır [25, 26].

2.2.4.6. *Hidrojen bağı enerji kuralı (Hydrogen bond energy rule)*

Sembolik DNA diziliminde, aralarında 3'lü hidrojen bağı bulunan G ve C organik bazlarına 1 değeri atanırken aralarında 2'li hidrojen bağı bulunan A ve T organik bazlarına -1 değerinin atanmasıdır [25, 26]. Örnek bir DNA dizilimi üzerinde eşleştirilmiş haritalama tekniğinin hidrojen bağı enerji kuralına göre sayısallaştırılan dizilimin gösterimi Tablo 1'de verilmiştir [25].

2.2.4.7. *Hibrit kuralı (Hybrid rule)*

Sembolik DNA diziliminde A veya C bazlarına 1 değeri atanırken T veya G bazlarına -1 değerinin atanmasıdır [25, 26].

Bu çalışmada sunulan geleneksel haritalama tekniklerinin avantaj ve dezavantajları Tablo 2'de verilmiştir.

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

Shannon tarafından tanıtılan entropi kavramı, karmaşıklığın bir ölçüsü olarak ifade edilmektedir. Meydana gelen olasılıklar üzerinden bilinmeyen olaylar kümesini işaret eden bu yapı kullanılarak x_i değerleri üzerinden belirsizlik hesaplanmaktadır. Bu yapının matematiksel ifadesi Eş. 1'de sunulmuştur.

$$S=-\sum_i p(x_i)\log(p(x_i)) \quad (1)$$

Ancak DNA dizilimlerinin sayısal gösterimi için yetersiz olduğu kabul edilen Shannon Entropisi, Ali Karcı tarafından geliştirilmiştir. Fraksiyonel shannon entropi temelli haritalama tekniği olarak önerilen bu yaklaşımın DNA dizilimindeki kodon dağılımları üzerinden entropisinin hesaplandığı matematiksel ifade Eş. 2'de sunulmaktadır [25, 27].

$$Sf=-\sum_i [(-p(x_i))^{\alpha_i} p(x_i)\log(p(x_i))] \quad (2)$$

Eş. 2'de verilen $p(x_i)$ her bir kodon için tekrarlanma sıklığını ifade etmektedir. α_i parametresi için [25, 27] çalışmalarında geliştirilen yeni bir tanımlama Eş. 3'te sunulmuştur.

Tablo 2. Geleneksel haritalama tekniklerinin avantaj ve dezavantajları [24-26]
(Advantages and disadvantages of traditional mapping techniques)

Haritalama teknikleri (H.T.)	Avantajları	Dezavantajları
Reel H.T.	A-T ve G-C organik bazları tamamlayıcı olma özelliğine sahiptir.	DNA diziliminde bulunmayan matematiksel özelliklerin tanıtılması gerekmektedir.
Moleküler Kütle H.T.	DNA dizilimlerinin çok boyutlu bir uzaya haritalanmasını sağlamaktadır.	Daha fazla araştırılması gerekmektedir.
Tam Sayı H.T.	Basit bir gösterime sahiptir. Verimli bir haritalama süreci sağlamaktadır.	DNA diziliminde bulunmayan matematiksel özelliklerin tanıtılması gerekmektedir.
Karmaşık H.T.	Nükleotitlerin eşlenik olma özelliğini yansıtarak daha doğru bir gen tahmini sağlamaktadır.	Zaman alanındaki baz yanılmasının (base bias) tanıtılması gerekmektedir
EIIP H.T.	DNA'nın fiziko-kimyasal özelliğini yansıtarak hesaplama yükünü azaltır ve gen ayrım yeteneğini geliştirir.	Bazı genomlar için kodlama bölgesinin tespiti hususunda başarısızlıkları mevcuttur.
Atomik Sayı H.T.	DNA'nın fiziko-kimyasal özelliğini yansıtır.	Çeşitli kodlama şemalarında farklı sonuçlar vermesinden dolayı araştırma kapsamında tekdüzeliğe neden olabilmektedir.
DNA-Yürüyüş H.T.	Uzun menzilli korelasyon bilgisini kullanarak nükleotit bileşimindeki değişiklikleri ortaya çıkarır.	1000 bazdan daha uzun olan dizilimler için uygun değildir.
Eşleştirilmiş Sayısal H.T.	Karmaşıklığı azaltılmış bir DNA yapısı sunmaktadır.	Karmaşıklığın azaltılması amacıyla sunduğu 7 kural içerisinden çalışma kapsamında başarılı olacak kuralın seçilmesi önemlidir.

$$\alpha_i = 1/\log(p(x_i)) \quad (3)$$

Bu kapsamda alfa değerinin sabit bir değer yerine genom dizisinin dikkate alınması suretiyle uyarlamalı bir değer olarak hesaplanması sağlanmıştır. Böylelikle kodon olasılıkları arasındaki ilişkinin gücü net bir şekilde yansıtılarak DNA diziliminin daha geniş sayısal temsili elde edilmiştir [25, 27].

3. DNA Analizinde Sinyal İşleme Yöntemleri
(Signal Processing Methods in DNA Analysis)

Sinyal işleme, genomik verilerin analiz edilmesi ve işlenmesi sonucunda elde edilen biyolojik bilgi vasıtası ile hastalıkların teşhis ve tedavisinde kullanılan, sistem tabanlı uygulamaların ortaya çıkmasında rol oynayan bir yöntemdir [14].

Bu yöntem kulanılarak gerçekleştirilen sinyal analizinde frekans ve genlik bileşenlerinin incelenmesi olağan bir durum olup olmadığının anlaşılabilmesi açısından önem taşımaktadır. Aynı zamanda biyolojik dizi analizinin sinyal işleme tekniklerindeki biyolojik deneylerin kullanılmasına göre daha çok avantaj barındırdığı [19]'da belirtilmektedir. Bu kapsamda DNA yapısının analizi için sayısallaştırılan dizilimler üzerinde gerçekleştirilen sinyal işleme teknikleri ile mevcut düzensizliklerin açığa çıkarılması hedeflenmiştir.

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

Sayısal sinyal işleme alanı içerisinde değerlendirilen ilk yöntem, Joseph Fourier tarafından tanıtılan fourier dönüşümü tekniğidir. Bu teknik sayesinde ilgili sinyalin frekans bileşenlerine analiz edilip bileşen yoğunluklarının ifade edilmesi fourier dönüşümü ile gerçekleştirilmektedir. Ancak bu yöntem ile sağlanılan analizlerde kaybedilen zaman bilgisinin yanı sıra sonsuz periyodiklik durumunun ve doğrusal olmayan sinyallerin işlenmesi güçlük içerdiğinden dolayı sinyallerin daha güçlü analizini sağlamak amacıyla kısa zamanlı fourier dönüşümü yöntemi geliştirilmiştir [19, 28, 29].

Kısa zamanlı fourier dönüşümü, sabit bir zaman penceresinin pencerenin sinyaldeki çarpımının ve fourier dönüşümünün kullanılması ile sinyalin frekans katsayılarını üreten bir tekniktir. Sinyallerin yerel analizinde daha güçlü olan kısa zamanlı fourier dönüşümünün matematiksel ifadesi Eş. 4'te sunulmaktadır [28].

$$KZFD(t,f) = \int_{-\infty}^{+\infty} x(t)w(t-K)e^{-2\pi i f t}dt = < g_{f,t}(t), x(t) > \quad (4)$$

Eşitlikte verilen f, frekansı; x(t), analiz edilen zaman serisini; w(t-K) zaman ekseninde K noktasına yerleştirilen pencere fonksiyonunu göstermektedir. Pencere fonksiyonu performansı da etkileyen önemli bir parametredir. Dolayısıyla bu çalışmada hamming, gausswin, chebwin, blacman ve bartlett fonksiyonları kullanılarak pencere fonksiyonlarının performans üzerindeki etkileri incelenmiştir. Gerçekleştirilen deneysel işlemler sonucunda eşleştirilmiş haritalama tekniğinin hidrojen bağı enerji kuralı, atomik, DNA-yürüyüş, EIIP, karmaşık, moleküler, reel ve tamsayı haritalama yöntemleri ile sayısallaştırılan DNA dizilimlerinin; hamming, gausswin, chebwin, blackman ve bartlett pencere fonksiyonları kullanılarak elde edilen spektrogramları üzerinde sınıflandırma süreci sağlanmıştır. Ulaşılan ortalama başarı oranları sırasıyla %55.62, %58.12, %53.75, %52.5 ve %58.75 olarak bulunmuştur. 8 farklı haritalama tekniği kullanılarak sağlanılan en yüksek ortalama başarı oranı bartlett pencere fonksiyonu olarak seçilmiştir. Bu fonksiyonun matematiksel ifadesi Eş. 5'te verilmiştir [30].

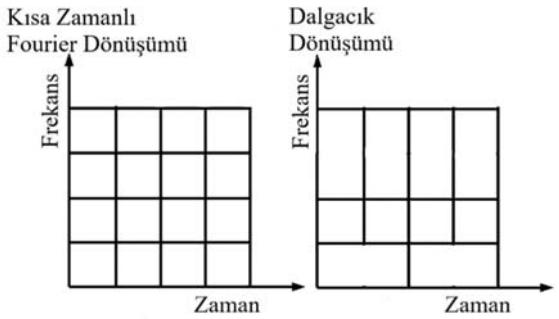
$$w(n) = \begin{cases} \frac{\frac{N-1}{2}+n}{\frac{N-1}{2}} & -\frac{N-1}{2} \le n \le 0 \\ 2 - \frac{\frac{N-1}{2}+n}{\frac{N-1}{2}} & 0 \le n \le \frac{N-1}{2} \end{cases} \quad (5)$$

Bartlett pencere fonksiyonu kullanılarak kısa zamanlı fourier dönüşümü ile ALL ve KML türlerine ait DNA dizilimleri spektrogram olarak ifade edilmiştir. Ancak geniş frekans aralıklarındaki sinyallerin analizi için uygun olmayan kısa zamanlı fourier dönüşümünün bu eksikliğini tolere edebilmek ve gerçekleştirilecek analizlerin gücünü

arttırabilmek amacıyla dalgacık dönüşümü yöntemi ile deneysel işlemler gerçekleştirilmiştir [31].

3.2. Dalgacık Dönüşümü (Wavelet Transform)

1909 yılında Alfred Haar tarafından geliştirilen dalgacık dönüşümü teorisi, fourier dönüşümünde ifade edilen problemlere çözüm olarak geliştirilmiştir. Bu yöntem ile sinyalin farklı parçalara bölünmesi işlemi, bir fonksiyonun iletilmesi ve ölçeklenmesi ile gerçekleşmektedir. Aynı zamanda fonksiyon, veri dizisi ile birlikte iletildiği için bilgi serisinin spekturumu her durum için hesaplanmaktadır. Farklı ölçeğe sahip fonksiyonlar için tekrarlanan bu sürecin sonunda bir dizi argüman-frekans bilgisi elde edilmektedir [32]. Kısa zamanlı fourier dönüşümü ve dalgacık dönüşümü yöntemlerinin karşılaştırılması Şekil 1'de sunulmuştur.



Şekil 1. Kısa zamanlı fourier dönüşümü ve dalgacık dönüşümü yöntemlerinin karşılaştırılması [33]
(Comparison of short-time fourier transform and wavelet transform methods)

Şekil 1 incelendiğinde kısa zamanlı fourier dönüşümünde zaman-frekans çözünürlüğü sabitken dalgacık dönüşümünde bu çözünürlüğün ayarlanabilme özelliğine sahip olduğu görülmektedir. Aynı zamanda dinamik sinyallerin varlığında güçlü bir zaman ve frekans çözünürlüğü sunan dalgacık dönüşümü yönteminin yüksek potansiyel barındırdığı görülmektedir. [13, 33].

3.2.1. Sürekli dalgacık dönüşümü (Continuous wavelet transform)

Sürekli dalgacık dönüşümü (SDD) yöntemi, kaydırılan dalgacık fonksiyonunun belirli bir ölçek ile çarpılmasından sonra zaman alanı boyunca toplanmasına imkân sağlamaktadır. Matematiksel ifadesi Eş. 6'da sunulmuştur [33].

$$SDD_{(s,\tau)}=\int_{-\infty}^{+\infty} g(t).\psi^{*}_{s,\tau}(t)dt \quad (6)$$

Denklemde verilen ¥s,τ(t), dalgacık fonksiyonunu; g(t), dönüşümü yapılacak fonksiyonu; τ, kaydırma parametresini; s, belirlenen ölçeğin parametresini ve *, kompleks eşleniği ifade etmektedir.

Dalgacık fonksiyonları, ana dalgacık fonksiyonu vasıtası ile ölçek ve kaydırma faktörlerinin kullanılması ile elde edilmektedir. Matematiksel ifadesi Eş. 7'de verilmiştir [33].

$$\psi^{*}_{s,\tau}(t)=\frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \quad (7)$$

Eş.7'de verilen $1/\sqrt{s}$ ifadesi farklı ölçeklere sahip normalizasyon faktörü olarak açıklanmaktadır. Bu denklemin Eş. 6'da yerine konulması ile Eş. 8'e ulaşılmaktadır [33].

$$SDD_{(s,\tau)}=\frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty} g(t).\psi^{*}_{s,\tau}\left(\frac{t-\tau}{s}\right)dt \quad (8)$$

Sunulan çalışmada sayısallaştırılmış DNA dizilimlerine, sürekli dalgacık dönüşümü yöntemi uygulanarak spektrogramlar elde edilmiştir.

4. Veri Kümesi (Dataset)

Bu çalışmada NCBI (Ulusal Biyoteknoloji Bilgi Merkezi) gen bankasından [34] tedarik edilen insan türüne ait ALL ve KML lösemi türlerine ilişkin BCR-ABL genleri üzerinde bir inceleme yapılmıştır. Kullanılan DNA dizilimlerinin ID numaralarına ve uzunluklarına ilişkin bilgi Tablo 3'te verilmiştir.

Tablo 3. NCBI veri kümesinden tedarik edilen nükleotit dizilimleri (Nucleotide sequences taken from the NCBI dataset)

ALL		KML	
Nükleotit numarası	DNA uzunluğu	Nükleotit numarası	DNA uzunluğu
FN820215.1	718	FN820207.1	377
FN820216.1	356	FN820208.1	205
FN820219.1	445	FN820209.1	516
FN820222.1	776	FN820210.1	767
FN820223.1	667	FN820211.1	698
FN820224.1	484	FN820212.1	545
FN820225.1	683	FN820213.1	195
FN820226.1	676	FN820214.1	723
FN820227.1	291	FN820217.1	561
FN820228.1	548	FN820218.1	272
FN820229.1	609	FN820220.1	262
FN820230.1	570	FN820221.1	879
FN820231.1	585	FN820237.1	708
FN820232.1	660	FN820238.1	835
FN820233.1	282	FN820239.1	621
FN820234.1	611	FN820243.1	761
FN820235.1	395	FN820248.1	708
FN820236.1	804	FN820250.1	218
FN820240.1	670	FN820252.1	564
FN820241.1	783	FN820255.1	357
FN820242.1	586	FN820259.1	212
FN820244.1	647	FN820260.1	422
FN820245.1	643	FN820261.1	489
FN820246.1	582	FN820262.1	373
FN820247.1	695	FN820265.1	994
FN820249.1	508	FN820266.1	707
FN820251.1	555	FN820267.1	809
FN820253.1	895		
FN820254.1	912		
FN820256.1	388		
FN820257.1	547		
FN820258.1	629		
FN820263.1	691		
FN820264.1	429		
FN820268.2	604		

Sunulan veri kümesinde, 282 ile 912 nükleotit uzunluğu arasında değişen ALL dizilimleri ve 195 ile 994 nükleotit uzunluğu arasında değişen KML dizilimleri üzerinden değerlendirmeler yapılmıştır.

5. Metodoloji (Methodology)

Sunulan çalışmada löseminin alt türlerinden olan ALL ve KML malignitelere ait DNA dizilimleri kullanılarak genetik temelli hastalıkların ayırt edilebilmesi hedeflenmiştir. Bu kapsamda sabit ve fiziko kimyasal özellik tabanlı haritalama teknikleri ile sayısallaştırılan DNA yapılarının özellikleri, sinyal işleme teknikleri vasıtasıyla spektrogram olarak ifade edilmiştir. ALL ve KML

malignileri için elde edilen spektrogramlar hem CNN hem de GNN teknolojisinin hibrit kullanımına izin veren DGCNN yaklaşımı ile sınıflandırılmıştır. Ancak farklı nükleotit sayısına sahip olan sayısallaştırılmış dizilimlerin spektrogram üzerinde sunduğu ifadeler, dizilimlerin farklı uzunluklarda olmasından dolayı net bir çıktı sunmamıştır. Dolayısıyla mevcut dizilimler için DNA'nın uzunluğundan bağımsız bir yaklaşım kullanılması planlanmıştır. İlk olarak kodon dağılımlarını esas alan shannon entropi tabanlı bir haritalama tekniği vasıtasıyla DNA dizilimleri sayısallaştırılmıştır. Bu tekniğin, sayısal temsillerin kodon olasılıkları arasındaki korelasyonu net bir şekilde yansıtarak DNA dizilimi için geniş bir aralık sunduğu [25] çalışmasında belirtilmektedir. Ardından sayısallaştırılan DNA dizilimleri üzerinde gerçekleştirilen analizler ile çıkarılan istatistiksel ve yapısal bilgiler, bulanık mantık algoritmasına girdi olarak verilmiştir. Elde edilen bulanık değerler üzerinde gerçekleştirilen sınıflandırma işlemi sonucunda ALL ve KML malignitelerinin tatmin edici bir oranda ayrımı sağlanmıştır.

5.1. Graf Sinir Ağı (Graph Neural Network)

Veri madenciliği, desen tanıma, optimizasyon ve analiz gibi birçok alanda öne çıkan yapay sinir ağları; tahmin, sınıflandırma ve veriler arasındaki ilişkilerin yorumlanmasında tercih edilen bir teknolojidir. Örneğin, son teknoloji evrişimsel sinir ağı kullanılarak sınıflandırma amacı ile düzenli tensörlere sahip olan veriler (görüntüler, videolar) üzerinde başarılı sonuçlar elde edilirken piksel değerleri rastgele atanmış düzenli tensor verileri üzerinde mevcut başarı oranının düştüğü ifade edilmektedir. Diğer taraftan kimya, biyoloji ve tıp alanları için moleküllerin modellenmesinde, ticari siteler için müşterilerin ve ürünlerin ilişkilendirilmesinde ya da sosyal ağların yapılandırılmasında kullanılan grafların düzensiz yapısı da net sonuçlar sunmamaktadır [35].

Dolayısıyla geleneksel sinir ağlarındaki birçok eksikliğin aşılabilmesi amacı ile graf sinir ağı çerçevesinde değerlendirilen uzaysal-zamansal graf sinir ağı yaklaşımı kullanılarak analizler gerçekleştirilmiştir.

5.1.1 Uzaysal-Zamansal graf sinir ağı (Spatial-temporal graph neural networks-STGNN-)

STGNN, zamansal ve uzaysal bağımlılığı aynı anda değerlendirebilen bir yaklaşımdır. Zamansal bağımlılığın modellenmesi için CNN yapısını, uzaysal bağımlılığın modellenmesi için evrişimsel graf yapısını kullanır [36]. Bu hibrit kullanım ile hem RAR hem de IIR bilgilerinin elde edilmesi sağlanarak güçlü bir yaklaşım oluşturulur [37]. Aynı zamanda artan veri miktarına karşın doğrusal bir zaman karmaşıklığı sunarak zamanın verimli kullanılmasını mümkün kılar [36].

Bu çalışmada sayısallaştırılan DNA dizilimleri üzerinde kısa zamanlı fourier dönüşümü ve dalgacık dönüşümü yöntemleri ile elde edilen spektrogramlar vasıtasıyla DGCNN yaklaşımının kullanımı gerçekleştirilmiştir [36].

DGCNN yaklaşımında ilk olarak ALL ve KML malignitelerine özgü spektrogramlara SIFT algoritması uygulanılmış ve spektrogramlardaki bireysel görüntü düzeyinde temsiller (D) açığa çıkartılmıştır. Ardından bu temsiller üzerinde k-ortalama (k-means) yöntemi ile N adet küme merkezi (X ϵ RNXD) belirlenmiştir. Böylelikle spektrogram içerisindeki ilişkileri tasvir eden küme merkezleri bağlamında her bir görüntü için komşuluk matrisi elde edilmiştir. Hem komşuluk matrisinin hem de görüntünün sahip olduğu küme özelliklerinin 4 katmanlı GNN yapısına verilmesi ile ilişkiye dayalı bir çıkarım sağlanmıştır. Çünkü spektrogramlar üzerindeki her bir pikselin çevresindeki pikseller ile bağlantılı olması, bilginin yayılmasını mümkün kılmaktadır. Mevcut bu durum mesaj geçiren sinir ağının (MPNN) varlığına işaret etmektedir. Graf evrişim işleminde gerçekleşen mesaj geçiş fonksiyonunun matematiksel ifadesi Eş. 9'da sunulmuştur [36].

$$h_v^{(k)}=U_k(h_v^{(k-1)},\sum_{u\,\epsilon N(v)} M_k(h_v^{(k-1)},h_u^{(k-1)},x_{vu}^e)) \quad (9)$$

Eş. 9'da, $h_v^{(0)}=x_v$ ve U_k () , M_k () öğrenilebilir parametreleri içeren fonksiyonları; k, iterasyon sayısını; v her bir pikseli ifade eden düğümü; u, komşu düğümünü temsil etmektedir [36].

Çalışmada tasarlanılan 4 katmanlı GNN yapısındaki düğüm özelliklerinin güncellenmesini gerçekleştirmek amacıyla Eş. 10-13'te gösterilen işlemler gerçekleştirilmektedir.

$$H^{(1)}=f_{relu}(\hat{A}XW^{(0)}) \quad (10)$$

$$H^{(2)}=f_{relu}(\hat{A}H^1W^{(1)}) \quad (11)$$

$$H^{(3)}=f_{relu}(\hat{A}H^2W^{(2)}) \quad (12)$$

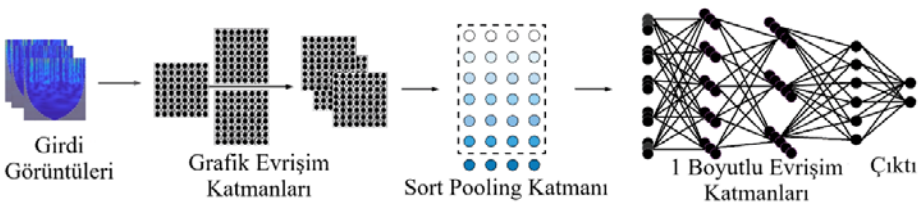
$$H^{(4)}=f_{relu}(\hat{A}H^3W^{(3)}) \quad (13)$$

Yukarıda verilen ifadelerde Â, komşuluk matrisinin normalize edilmiş temsilidir. Hem komşuluk matrisi hem de derece matrisi vasıtasıyla hesaplanmaktadır. Derece matrisinin matematiksel ifadesi Eş. 14'te verilmiştir [37]. Aynı zamanda f_{relu}, çıkışa relu fonksiyonunun uygulanıldığını ifade eder. H^l, l'inci katmanın özelliklerini belirtir.

$$\delta_{ij}\text{def} = \deg(v_i), \text{ if } i=j \quad (14)$$
$$0, \quad \text{otherwise}$$

GNN yapısı kullanılarak yerel altyapı özelliklerinin çıkarıldığı köşeler için sıralama yapıldıktan sonra elde edilen çıktılar önceden tanımlanan sırada köşe özelliklerinin sıralanmasını sağlayan sortpooling katmanına iletilir. Bu katmanda grafların yapısal rolüne uygun olarak sağlanan sıralama işlemi, 1 boyutlu geleneksel sinir ağı yapısına verilerek otomatik özellik çıkarım süreci gerçekleştirilir [35]. Kullanılan mimarinin temsili yapısı Şekil 2'de verilmiştir.

3 temel aşamadan oluşan yapay zekâ temelli DGCNN yaklaşımı ile gerçekleştirilen sınıflandırma sürecinde ALL ve KML spektrogramları üzerinde hem RAR hem IIR bilgilerinin elde edilip köşe düğümlerinin temsil edildiği bir sıralamanın gerçekleştirilmesi algoritmanın güçlü yanlarını oluşturmaktadır [35].



Şekil 2. CNN ve GNN teknolojilerinin hibrit kullanımını mümkün kılan DGCNN yaklaşımı [35]
(DGCNN approach that enables hybrid use of CNN and GNN technologies)

5.2. Uyarlanabilir Bulanık Mantık Algoritması (Adaptive Fuzzy Logic Algorithm)

Biyoenformatik, modern bilim alanları arasında hızlı büyüme eğrisine sahip olan disiplinlerarası çalışma alanıdır. Mikrodizi gen ekspresyon analizi (Microarray gene expression analysis), gen biyobelirteçlerinin tanımlanması (gene biomarkers identification), ve gen düzenleyici ağ çıkarımı (gene regulatory network inference) gibi farklı uygulama alanlarını barındırmaktadır. Bu uygulama alanları içerisinde DNA ve gen birimlerinin belirsiz yapılarına uygun bir şekilde bulanık değerler üreten bulanık mantık yaklaşımı güçlü bir çıkarım potansiyeli sağlamaktadır [38, 39]. Elde edilen potansiyel, farklı malignitelerin tespit edilmesi hususunda başarılı bir çıktı sunmaktadır [40].

Bu kısımda, kodon dağılımları üzerinde temellenen shannon entropi temelli haritalama tekniği kullanılarak ulaşılan sayısallaştırılmış dizilimlerin DNA uzunluğundan bağımsız olarak değerlendirilebilmesi için istatistiksel ve yapısal özellikler üzerinden bir analiz gerçekleştirilmiştir. Seçilen özelliklerin detayları aşağıda açıklanmaktadır.

5.2.1.1. Ortalama mutlak değer (Average absolute value)

Verinin genlik değeri için mutlak ortalamasının ölçüsüdür. Matematiksel ifadesi Eş. 15'te verilmiştir [25].

$$\frac{1}{N}\sum_{i=1}^{N}|i| \tag{15}$$

5.2.1.2. Standart sapma (Standard deviation)

Ortalamadan uzaklaşmanın ya da değişimin ölçüsü olarak tanımlanmaktadır. Matematiksel ifadesi Eş. 16'da verilmiştir [41].

$$S=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(i-i')^2} \tag{16}$$

5.2.1.3. Varyans (Variance)

Standart sapma sonucunda elde edilen matematiksel ifadenin karekök alınmamış haldir. Verilerin aritmetik ortalamadan sapma durumlarının karelerinin toplamını ifade eder. Matematiksel ifadesi Eş. 17'de verilmiştir.

$$V=\frac{1}{N-1}\sum_{i=1}^{N}|i-i'|^2 \tag{17}$$

5.2.1.4. Basit kare integral (Simple square integral)

Verinin enerjisini ifade etmek için kullanılmaktadır. Basit kare integral ölçütünün matematiksel ifadesi Eş. 18'de verilmiştir [25].

$$\alpha=\sum_{i=1}^{N}|X_i|^2 \tag{18}$$

5.2.1.5. Dalga boyu (Wavelength)

Genlik, frekans ve zaman ile ilişkili olan dalga boyu, zaman dilimi boyunca uzanan verinin kümülatif uzunluğunu ifade etmektedir. Matematiksel ifadesi Eş. 19'da verilmiştir [25].

$$\boldsymbol{\tau}=\sum_{i=1}^{N-1}|X_{i+1}-X_i| \tag{19}$$

5.2.1.6. Kovaryans (Covariance)

Rastgele atanmış değişkenler ile birlikte veriler üzerindeki değişimin incelendiği bir ölçüttür. Matematiksel ifadesi Eş. 20'de verilmiştir [42].

$$\mathrm{cov}(A,B)=\frac{1}{N-1}\sum_{i=1}^{N}|A_i-\mu_A|*|B_i-\mu_B| \tag{20}$$

5.2.1.7. Çarpıklık (Skewness)

Ortalama değere yakın dağılan verilerin, mevcut değerlerindeki asimetri ölçüsü olarak ifade edilir. Matematiksel ifadesi Eş. 21'de verilmiştir.

$$S=E(x-\mu)^3/\sigma^3 \tag{21}$$

Eş. 21'de verilen μ, verinin ortalaması; σ, verinin standart sapması olarak ifade edilmektedir [42].

5.2.1.8. Basıklık (Kurtosis)

Dağılımın aykırı bir değere ne kadar meyilli olduğunu ifade eden bir ölçüttür. Matematiksel ifadesi Eş. 22'de verilmiştir [41, 42].

$$k=E(x-\mu)^4/\sigma^4 \tag{22}$$

Eş. 22'de verilen μ, verinin ortalaması; σ, verinin standart sapması olarak tanımlanmaktadır.

5.2.1.9. Entropi (Entropy)

Veriler üzerindeki değerlerin rastgelelik/belirsizlik ölçüsüdür. Matematiksel ifadesi Eş. 23'te verilmiştir [41].

$$\mathrm{Ent}=-\sum_{i=1}^{M}p_i\log p_i \tag{23}$$

İstatistikte tercih edilen önemli bir ölçüttür.

5.2.1.10. GMDH ağları (GMDH networks)

İleri beslemeli, kendi kendini organize edebilen GMDH ağları, karışık sistemler için yüksek dereceli regresyon tipi modeller oluşturan bir yapıdır. Girdi ve çıktı arasındaki ilişki Kolmogorov–Gabor polinomu şeklinde tanımlandığından dolayı polinom sinir ağları (polynomial neural networks) olarakta adlandırılmaktadır. Kolmogorov–Gabor polinomunun matematiksel ifadesi Eş. 24'te verilmiştir [43, 44].

$$Y(x_1,\ldots,x_n)=a_0+\sum_{i=1}^{n}a_i x_i+\sum_{i=1}^{n}\sum_{j=i}^{n}a_{ij}x_i x_j+\sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{k=j}^{n}a_{ijk}x_i x_j x_k+ \tag{24}$$

Regresyon analizi ve karar düzenleme yöntemlerinin hibrit bir çalışması olarak önerilen GMDH yöntemi vasıtasıyla veriler üzerinden sayısal bir çıkarım yapılması sağlanmıştır.

5.2.2. ANFIS ağı (ANFIS network)

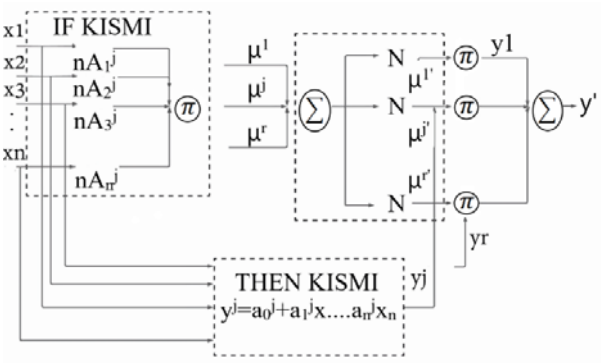
İstatististiksel ve yapısal özellikler ile elde edilen farklı deneysel ölçütler kullanılarak DNA dizilimleri içerisindeki bulanık çıktılar üzerinden çıkarım yapılabilmesi amacıyla ANFIS yöntemi kullanılmıştır [45].

ANFIS, yapay sinir ağlarının (ANN) bir türüdür ve Tagaki-Sugeno (TS) bulanık çıkarım sisteminin üzerinde temellenmektedir. ANFIS çerçevesine entegre edilen ANN ve bulanık mantık, IF THEN kuralına dayalı olarak doğrusal olmayan fonksiyonlara yaklaşık olarak yaklaşmaktadır. Dolayısıyla ANFIS modeli, gizli katmanlardaki nöronlar ile bulanık kuralların değiştirildiği bir ağ-tipine (network-type) sahiptir. Bu yöntemde yer alan bulanıklaştırıcı modül kullanılarak triangular, sigmoidal, gaussian gibi farklı bir üyelik fonksiyonu ile karakterize edilen bulanık setler için net girdi desenleri eşlenebilir. Aynı zamanda veri kümesinden öğrenme işlemi de TS

sistemi ile sağlanmaktadır. ANFIS verilen girdi-çıktı veri kümesi vasıtasıyla üyelik fonksiyonunu ve sonuç parametrelerini ayarlayabilme yeteneğine sahip bir yapıdır [45]. Sonuç kısmı için rastgele doğrusal bir fonksiyon içeren ANFIS yapısının matematiksel ifadesi Eş. 25'te verilmiştir [45].

$$R^j= \text{IF } x_1 \text{ is } A_1^j \text{ AND } x_2 \text{ is } A_2^j \text{ AND…AND } x_n \text{ is } A_n^j \text{ THEN } y^j=a_0^j+a_1^j x_1+\ldots+a_n^j x_n \quad (25)$$

Eş. 25'te R^j, IF ve THEN koşullarını dahil eden bulanık(fuzzy) kuraldır. N-boyutlu girdi vektörünün k'ıncı girdi değişkeni x_k'dır ve j'inci bulanık kuralında x_k ile ilişkilendirilen bulanık üyelik fonksiyonu, A_k^j'dir. Kesin girdi modellerini bulanık değerlere dönüştürmek ve bulanık kümeleri belirlemek için bu çalışmada gauss üyelik fonksiyonu seçilmiştir. Şekil 3'te temsili bir ANFIS yapısı sunulmaktadır [45].



Şekil 3. Temsili bir ANFIS yapısı [45]
(A representative ANFIS structure)

Sunulan yapının ilk katmanın da her bir bulanık küme için giriş değerlerinin üyelik derecesi elde edilmekte ve gauss üyelik fonksiyonu (GÜF) kullanılarak Eş. 26'da gösterilen bulanıklaştırma işlemi gerçekleştirilmektedir.

$$\eta A_k^j=\exp\left[-0.5(x_k-c_k^j/\sigma_k^j)^2\right] \quad (26)$$

Yukarıdaki ifade de c_k^j ve σ_k^j, k'ıncı girdi değişkenleri, j'inci GÜF'ün merkez ve genişliğini ifade eden temsilleridir [45].

ANFIS yapısının ikinci ve üçüncü katmanı sırasıyla ateşleme gücünü ve normalleştirme işlemini hesaplamak için modellenmiştir. Bu doğrultuda öncül(antecedent) kısımdaki VE(AND) işleminin her bir kural için sunduğu çıktı, Eş. 27'deki gibi hesaplanmaktadır [45].

$$\mu^j=\prod_{k=1}^{n} \eta A_k^j \quad (27)$$

Tüm bulanık kurallar göz önüne alındığında j. kuralın normalleştirilmiş sonucu Eş. 28'de verilmiştir [45].

$$\mu'^j=\mu^j/\sum_{j=1}^{R} \mu^j \quad (28)$$

Yukarıdaki ifade de μ'^j, son ağ çıkışındaki her bir kural için katkı derecesine Eş. 29'daki gibi karar veren ateşleme gücü olarak ifade edilmektedir [45].

$$y'=\sum_{j=1}^{R} \mu'^j y^j \quad (29)$$

Dördüncü katman, kuralların sonuçları için bir hesaplama sunmaktadır. Ek olarak son adımda gerçekleşen ANFIS'in durulaştırma adımı ile elde edilen net çıktı, kuralların sonuç bölümlerinin doğrusal kombinasyonu ile sağlanmaktadır [45].

6. Tartışma (Discussion)

Organizmayı inşa etmek ve canlılığını sürdürmek amacı ile devasa bilgi barındıran DNA, genetik temelli birçok hastalığın aydınlatılmasında önemli biyobelirteçtir [46]. Ancak DNA'nın sembolik bir yapıya sahip olması dizilimler üzerinde yapılacak matematiksel hesaplamalara engel olmaktadır. Dolayısıyla dizilimlerin sayısallaştırılması, yapılacak analizlerin başarısı açısından önemli bir konudur. Bu kapsamda sunulan çalışmada, lösemi hastalığının temel türleri içerisinde yer alan KML ve ph-pozitif ALL hastalıklarına sahip bireylerin DNA'sında gerçekleşen bir mutasyon sonucu oluşmuş philadelphia kromozomu ele alınmıştır. Bu kromozomun başarılı bir analizinin sağlanması açısından 8 farklı haritalama tekniği kullanılarak sayısallaştırma işlemi gerçekleştirilmiştir. Ardından bu sayısal temsiller içerisindeki gizli desenler, sinyal işleme yöntemleri olan kısa zamanlı fourier dönüşümü ve dalgacık dönüşümü teknikleri ile spektrogramlara yansıtılmıştır. ALL ve KML maligniteleri için elde edilen spektrogramlar hem CNN hem de GNN teknolojisinin hibrit

Tablo 4. Kısa zamanlı fourier dönüşümü kullanılarak elde edilen spektrogramların DGCNN yaklaşımı ile sınıflandırılmasının sonucunda elde edilen deneysel ölçütler
(Experimental criteria obtained as a result of classification with the DGCNN approach of spectrograms obtained using the short-time fourier transform)

Haritalama Teknikleri	Doğru Pozitif	Yanlış Pozitif	Yanlış Negatif	Doğru Negatif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Reel Haritalama Tekniği	9	2	6	3	0,55	0,60	0,50	0,54	0,56
Moleküler Kütle Haritalama Tekniği	7	4	4	5	0,55	0,57	0,50	0,72	0,63
Tamsayı Haritalama Tekniği	11	0	5	4	0,65	0,62	0,75	0,90	0,73
Karmaşık Haritalama Tekniği	6	5	6	3	0,55	0,57	0,50	0,72	0,63
EIIP Haritalama Tekniği	9	2	3	6	0,45	0,50	0,44	0,09	0,15
Atomik Sayı Haritalama Tekniği	10	1	5	4	0,75	0,75	0,75	0,81	0,77
DNA-Yürüyüş Haritalama Tekniği	8	3	7	2	0,55	0,57	0,50	0,72	0,63
Eşleştirilmiş Sayısal Haritalama Tekniği	9	2	5	4	0,65	0,64	0,66	0,81	0,71

kullanımına izin veren DGCNN teknolojisi ile sınıflandırılmıştır. Haritalama yaklaşımlarına ve sinyal işleme yöntemlerine göre elde edilen performans ölçütleri Tablo 4 ve Tablo 5'te verilmiştir.

Sayısallaştırılan DNA dizilimlerinin spektrogram üzerinde sunduğu ifadeler, dizilimlerin farklı uzunluklarda olmasından dolayı net bir çıktı sağlayamamıştır. Yetersiz görülen bu durum karşısında dizilimlerin uzunluğundan bağımsız bir yaklaşım kullanılması planlanmıştır. Bu nedenle kodon dağılımlarını esas alan shannon entropi tabanlı bir haritalama tekniği kullanılarak DNA dizilimlerinin sayısallaştırılması gerçekleştirilmiştir. Bu tekniğin sayısal temsillerinin kodon olasılıkları arasındaki korelasyonu net bir şekilde yansıtarak DNA dizilimi için geniş bir aralık sunduğu [25] çalışmasında belirtilmektedir. Ardından sayısallaştırılan DNA dizilimleri üzerinde gerçekleştirilen analizler ile çıkarılan istatistiksel ve yapısal bilgiler, bulanık mantık algoritması ile sınıflandırılmıştır. Elde edilen performans ölçütleri Tablo 6'da verilmiştir.

Sunulan değerlendirme ölçütleri incelendiğinde ALL ve KML türü için bulanık mantık algoritması tarafından yapılan doğru etiketleme sayılarının doğru ve yanlış tüm etiketleme sayılarına oranını ifade eden başarı oranı maksimum %80 olarak bulunmuştur. Başarı oranının yanı sıra farklı değerlendirme ölçütleri kullanılarak sonuçlar değerlendirilmiştir. Bu doğrultuda, ALL olarak etiketlenen görüntülerin bulanık mantık algoritması tarafından tahmin edildiği görüntülerin toplam sayısı olan doğru negatif ölçütü, ALL olarak etiketlenen görüntülerin bulanık mantık algoritması tarafından KML olarak tahmin edildiği görüntülerin toplam sayısı olan yanlış pozitif ölçütü, KML olarak etiketlenen görüntülerin bulanık mantık algoritması tarafından ALL olarak tahmin edildiği görüntülerin toplam sayısı olan yanlış negatif ölçütü, KML olarak etiketlenen görüntülerin bulanık mantık algoritması tarafından KML olarak tahmin edildiği görüntülerin toplam sayısı olan doğru pozitif ölçütü, bulanık mantık algoritmasının KML türü için yaptığı doğru etiketleme sayısının KML türüne ait tüm görüntülere oranı olan duyarlılık ölçütü, bulanık mantık algoritmasının ALL türü için yaptığı doğru etiketleme sayısının ALL türüne ait tüm görüntülere oranı olan özgüllük ölçütü, bulanık mantık algoritması tarafından KML olarak tahmin edilen görüntülerin gerçek durumda KML olarak etiketlenip etiketlenmediğini tanımlayan ilişkinin oranı olan kesinlik ölçütü ve kesinlik ile duyarlılık değerlerinin harmonik ortalaması olan F ölçütü değerlendirilmiştir.

Doğruluk oranı, duyarlılık, özgüllük, kesinlik ve F ölçütüne ilişkin kriterlerin matematiksel ifadesi sırasıyla Eş. 30-34'te verilmiştir

$$\text{Doğruluk Oranı}=(DP+DN)/DP+YP+DN+YN) \quad (30)$$

$$\text{Duyarlılık} = DP/(DP+YN) \quad (31)$$

$$\text{Özgüllük}= DN/(DN+YP) \quad (32)$$

$$\text{Kesinlik} =DP/(DP+YP) \quad (33)$$

$$\text{F ölçütü} = (2*\text{kesinlik}*\text{duyarlılık})/(\text{kesinlik}+\text{duyarlılık}) \quad (34)$$

Bu makalede, lösemi malignitesinin temel türlerinden olan ph-pozitif ALL ve KML hastalıklarının sınıflandırılması için yapılan mevcut çalışma, geleneksel sitogenetik tekniklerin yerini alan kantitatif gerçek zamanlı PCR analizlerine kıyasla gerçekleştirilen bilgisayar

Tablo 5. Sürekli dalgacık dönüşümü kullanılarak elde edilen spektrogramların DGCNN yaklaşımı ile sınıflandırılmasının sonucunda elde edilen deneysel ölçütler
(Experimental criteria obtained as a result of classification with the DGCNN approach of spectrograms obtained using the continuous wavelet transform)

Haritalama Teknikleri	Doğru Pozitif	Yanlış Pozitif	Yanlış Negatif	Doğru Negatif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Reel Haritalama Tekniği	6	5	4	5	0,60	0,60	0,60	0,81	0,68
Moleküler Kütle Haritalama Tekniği	8	3	6	3	0,60	0,63	0,55	0,63	0,63
Tamsayı Haritalama Tekniği	10	1	6	3	0,75	0,68	1	1	0,81
Karmaşık Haritalama Tekniği	8	3	6	3	0,55	0,50	0,37	0,54	0,51
EIIP Haritalama Tekniği	1	10	1	8	0,75	0,75	0,75	0,81	0,77
Atomik Sayı Haritalama Tekniği	9	2	3	6	0,70	0,66	0,80	0,90	0,76
DNA-Yürüyüş Haritalama Tekniği	8	3	6	3	0,50	0,53	0,40	0,72	0,61
Eşleştirilmiş Sayısal Haritalama Tekniği	9	2	5	4	0,65	0,64	0,66	0,81	0,71

Tablo 6. Uyarlanabilir bulanık mantık yaklaşımı kullanılarak gerçekleştirilen sınıflandırma sonucunda elde edilen deneysel ölçütler
(Experimental criteria obtained as a result of classification made using the adaptive fuzzy logic approach)

Haritalama Teknikleri	Doğru Pozitif	Yanlış Pozitif	Yanlış Negatif	Doğru Negatif	Başarı Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Bulanık Mantık Yöntemi*	8	4	2	6	0,70	0,80	0,60	0,66	0,72
Bulanık Mantık Yöntemi**	10	2	2	6	0,80	0,83	0,75	0,83	0,83

DGCNN Approach1*=GMDH ağları ile elde edilen sayısal çıkarımın özellikler içerisinde yer almaması sonucunda elde edilen deneysel ölçütler
DGCNN Approach2**=GMDH ağları ile elde edilen sayısal çıkarımın özellikler içerisinde yer alması sonucunda elde edilen deneysel ölçütler

destekli bir sistemdir. Geçmişten günümüze kadar multidisipliner çalışmaların çatısı altında löseminin alt türlerinin sınıflandırılması için mikrodizi teknolojisinin yardımıyla bilgisayar destekli sistemler üzerinde [47- 52] birçok çalışma yapılmıştır. Mikrodizi teknolojisi, tıp ve biyoloji alanlarında tercih edilen ve binlerce genin nispi ekspresyon seviyelerinin aynı anda izlenmesine olanak tanıyan bir analiz yöntemidir. Ancak bu teknoloji hastalık ile ilişkili olmayan genlerden dolayı ekstra bir işlem yüküne neden olmaktadır. Öte yandan ilişkili genler üzerinde yanlış bir eğilime neden olma ihtimali de mevcuttur [53, 54].

Bununla birlikte mevcut makalede gerçekleştirilen sınıflandırma işlemi direkt olarak DNA dizilimleri üzerinden yapılmıştır. Bildiğimiz kadarıyla DNA dizilimleri kullanılarak ALL ve KML malignitelerinin sınıflandırılması hususunda bilgisayar destekli yapılan ilk çalışmadır. Dolayısıyla bu yönüyle literatüre katkı sağlayacağı düşünülmektedir.

7. Sonuçlar (Conclusions)

Tıp ve bilişim alanları, hastanın vücut bütünlüğüne zarar vermeden uygulanılması kolay tedavi yöntemleri geliştirmek amacıyla non-invazif metotların ve alternatif yöntemlerin keşfi üzerinde disiplinlerarası çalışmalar gerçekleştirmektedir.

Bu kapsamda sunulan makalede, incelenen lösemi hastalığının temel türlerinin ayırt edilmesi için hem uygulama da hem de literatürde birçok araştırma gerçekleştirilmiştir. Çalışmalarda sağlanan yenilikler üzerinde dikkate alınan zaman ve maliyet değişkenleri, farklı parametreleri etkileyebilme özelliğine sahip olduğu için mevcut değişkenlerin iyileştirilmesi kritik bir konudur. Bu doğrultuda iki ayrı bölümde şekillenen bir çalışma gerçekleştirilmiştir. İlk olarak DGCNN yöntemi ile veriler üzerinde hem RAR (relation-aware representation) hem de IIR (individual image-level representation) bilgilerinin elde edildiği bir yapı kullanılmıştır ve artan veri miktarına rağmen doğrusal zaman karmaşıklığı sunan bir çıktı elde edilmiştir [36, 37]. İkinci olarak nükleotit uzunluğundan bağımsız bir yaklaşım önerilmiştir. Uyarlanabilir bulanık mantık yöntemi vasıtasıyla üretilen bulanık değerler, nükleotit uzunluğundan bağımsız bir çalışmanın yapılmasını sağlamıştır [45]. Böylece farklı uzunluklara sahip tüm DNA dizilimleri için istikrarlı sonuçlar üretilmiştir.

Gelecekte, nükleotit dizilimlerinin sayısallaştırılması safhasında yeni bir sayısallaştırma tekniği önerilerek daha yüksek başarı oranlarına ulaşılması hedeflenmektedir.

Kaynaklar (References)

1. Aydın G., Quercetinin KML kök hücreleri üzerine sitotoksik etkilerinin moleküler düzeyde incelenmesi, Yüksek Lisans Tezi, Erciyes Üniversitesi, Sağlık Bilimleri Enstitüsü, Kayseri, 2017.
2. Healthline. A guide to leukemia. https://www.healthline.com/health/leukemia#treatment. 2021.
3. Arslan S., KML ve ALL tanılı hastalarda BCR / ABL füzyon geni mutasyonlarının taranması, Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Eskişehir, 2014.
4. Kitamura H. et al., A new highly sensitive real-time quantitative-PCR method for detection of BCR-ABL1 to monitor minimal residual disease in chronic myeloid leukemia after discontinuation of imatinib, PLoS One, 14 (3), 1–13, 2019.
5. MedlinePlus. BCR-ABL genetic test. https://medlineplus.gov/lab-tests/bcr-abl-genetic-test/. 2021.
6. Paiva A.S. et al., Detection of the BCR-ABL Gene By the Real-Time PCR Method in Patients with Chronic Myeloid Leukemia in Rio Grande Do Norte, Brazil, Blood, 132 (1), 2018.
7. Uzoma I., Nna E., Detection and quantitation of bcr-abl1 fusion gene in saliva of chronic myeloid leukaemic patients in nigeria, Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, 2017.
8. Smitalova D., Dvorakova D., Racil Z., Romzova M., Digital PCR can provide improved BCR-ABL1 detection in chronic myeloid leukemia patients in deep molecular response and sensitivity of standard quantitative methods using EAC assays, Practical Laboratory Medicine, 25, 2021.
9. Yang R., Paparini A., Monis P., Ryan U., Comparison of next-generation droplet digital PCR (ddPCR) with quantitative PCR (qPCR) for enumeration of Cryptosporidium oocysts in faecal samples, International Journal for Parasitology., 44 (14), 1105–1113, 2014.
10. Maier J., Lange T., Cross M., Wildenberger K., Niederwieser D., and Franke G.N., Optimized Digital Droplet PCR for BCR-ABL, Journal of Molecular Diagnostics,21 (1), 27–37, 2019.
11. Jennings L.J., George D., Czech J., Yu M., and Joseph L., Detection and quantification of BCR-ABL1 fusion transcripts by droplet digital PCR, Journal of Molecular Diagnostics, 16 (2), 174–179, 2014.
12. Khodaei A., Feizi-Derakhshi M.R., and Mozaffari-Tazehkand B., A pattern recognition model to distinguish cancerous DNA sequences via signal processing methods, Soft Computing, 24 (21), 16315–16334, 2020.
13. Chakraborty S. and Gupta V., DWT based cancer identification using EIIP, Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016, 718–723, 2016.
14. Liu D.W. et al., Automated detection of cancerous genomic sequences using genomic signal processing and machine learning, Future Generation Computer Systems, 98, 233–237, 2019.
15. Das J. and Barman S., Bayesian Fusion in Cancer Gene Prediction, International Journal of Computer Application (0975 – 8887) International Conference on Computing, Communication and Sensor Network (CCSN 2014), 5–10, 2014.
16. Wang B., Mohl J., Leung M. Y., Computational Prediction of Functional Effects for Cancer Related Genetic Sequence Variants, Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, 2999–3001, 2020.
17. Sawyer E., Banuelos M., Marcia R. F., and Sindi S., A Neural Network Approach for Anomaly Detection in Genomic Signals, 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2020 - Proceedings, 968–971, 2020.
18. Saini S. and Dewan L., Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis, SpringerPlus, 5 (1), 1–15, 2016.
19. Gayathri T. T., Analysis of Genomic sequences for prediction of Cancerous cells using Wavelet technique, International Research Journal of Engineering and Technology(IRJET), 4 (4), 1071–1077, 2017.
20. Zhao Y. et al., Uncovering the prognostic gene signatures for the improvement of risk stratification in cancers by using deep learning algorithm coupled with wavelet transform, BMC Bioinformatics, 21 (1),1–24, 2020.
21. Ghosh A. and Barman S., Prediction of Prostate Cancer Cells based on Principal Component Analysis Technique, Procedia Technology, 10, 37–44, 2013.
22. Muflikhah L., Widodo N., Mahmudy W.F, Solimun, Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method, 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020, 37–41, 2020.
23. Das J., Barman S., DSP based entropy estimation for identification and classification of Homo sapiens cancer genes, Microsystem Technologies, 23 (9), 4145–4154, 2017.
24. 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.
25. Das B., DNA dizilimlerinden hastalık tanılanması için işaret işleme temelli yeni yaklaşımların geliştirilmesi, Doktora Tezi, Fırat Üniversitesi, Fen Bilimleri Üniversitesi,Elazığ, 2018.
26. 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.
27. Das B., Turkoglu I. A novel numerical mapping method based on entropy for digitizing DNA sequences, Neural Comput. Appl., 29 (8), 207–215, 2018.

28. Ahmadi H.R., Mahdavi N., Bayat M., A novel damage identification method based on short time Fourier transform and a new efficient index, Structures, 33, 3605–3614, 2021.

29. Fidan H., Dalgacık dönüşümü tekniği ile motor arıza tespiti, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Üniversitesi, Isparta 2006.

30. 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 2018, 1–4, 2018.

31. Xia Y., Johnson B.K., Jiang Y., Fischer N., Xia H., A new method based on artificial neural network, Wavelet Transform and Short Time Fourier Transform for Subsynchronous Resonance detection, International Journal of Electrical Power and Energy Systems, 103, 377–383, 2018.

32. Valizadeh M., Sohrabi M., Ameri Braki Z., Rashidi R., and Pezeshkpur M., Investigation of spectrophotometric simultaneous absorption of Salmeterol and Fluticasone in Seroflo spray by continuous wavelet transform and radial basis function neural network methods, Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy, 263, 2021.

33. Volkan Öner İ., Yeşilyurt K., Yilmaz E.Ç., Wavelet Analiz Tekniği Ve Uygulama Alanları, Ordu Üniv. Bil. Tek. Derg, 7 (1), 42–56, 2017.

34. National Center for Biotechnology Information. Nucleotide database. https://www.ncbi.nlm.nih.gov. 2021.

35. Zhang M., Cui Z., Neumann M., Chen Y., An end-to-end deep learning architecture for graph classification, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 4438–4445, 2018.

36. Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P.S., A Comprehensive Survey on Graph Neural Networks, IEEE Transactions on Neural Networks and Learning Systems., 32 (1), 4–24, 2019.

37. Wang S. H., Govindaraj V.V., Górriz J. M., Zhang X., Zhang Y. D., Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network, Information Fusion, 67, 208–229, 2021.

38. Boyaci A.Ç., Solmaz M.B., Kabak M., A model proposal for occupational health and safety risk assessment based on multi-criteria hesitant fuzzy linguistic term sets: An application in plastics industry, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (2), 1041–1053, 2021.

39. [39] Öztürk M., Paksoy T., An new interval type-2 hybrid fuzzy rule-based AHP system for supplier selection, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (3), 1519–1535, 2020.

40. Raza K., Fuzzy logic based approaches for gene regulatory network inference, Artificial Intelligence in Medicine, 97, 189–203, 2019.

41. Thakur N., Juneja M., Classification of glaucoma using hybrid features with machine learning approaches, Biomedical Signal Processing and Control, 62, 2020.

42. Matlab. İstatistik komutları. https://www.mathworks.com/help/. 2021.

43. Wikipedia. Group method of data handling. https://en.wikipedia.org/wiki/Group_method_of_data_handling. 2021.

44. Ivakhnenko A.G., Polynomial theory of complex systems. IEEE transactıons on systems, man, and cybernetıcs, SMC-1 (4), 364–378, 1971.

45. Mahdevari S.,Khodabakhshi M.B, A hybrid PSO-ANFIS model for predicting unstable zones in underground roadways, Tunnelling and Underground Space Technology, 117, 2021.

46. 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.

47. Wang X., Gotoh O., Cancer classification using single genes., International Conference on Genome Informatics, 23 (1), 179–188, 2009.

48. Ghorai S., Mukherjee A., Dutta P. K., Gene Expression Data Classification by VVRKFA, Procedia Technology, 4, 330–335, 2012.

49. Maulik U., Chakraborty D., Fuzzy preference based feature selection and semisupervised SVM for cancer classification, IEEE Transactions on Nanobioscience, 13 (2), 152–160, 2014.

50. Begum S., Sarkar R., Chakraborty D., Sen S., Maulik U., Application of active learning in DNA microarray data for cancerous gene identification, Expert Systems with Applications, 177, 2021.

51. Ocampo-Vega R., Sanchez-Ante G., De Luna M.A., Vega R., Falcón-Morales L.E., Sossa H., Improving pattern classification of DNA microarray data by using PCA and Logistic Regression, Intelligent Data Analysis., 20, 2016.

52. Chen Y., Zhao Y., A novel ensemble of classifiers for microarray data classification, Applied Soft Computing Journal, 8 (4), 1664–1669, 2008.

53. Wang X., Simon R., Microarray-based cancer prediction using single genes, BMC Bioinformatics, 12, 2011.

54. Khorshed T., Moustafa M.N., Rafea A., Learning Visualizing Genomic Signatures of Cancer Tumors using Deep Neural Networks, Proceedings of the International Joint Conference on Neural Networks, 2020.