T.C.
BATMAN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI

DOKTORA TEZİ

GEN DİZİLERİNİN TANIMLANMASI VE SINIFLANDIRILMASI AMACIYLA YAPAY ZEKÂ SİSTEMLERİNİN GELİŞTİRİLMESİ

Bahar ÇİFTÇİ

Ekim-2024
BATMAN

T.C.
BATMAN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI

DOKTORA TEZİ

GEN DİZİLERİNİN TANIMLANMASI VE SINIFLANDIRILMASI AMACIYLA YAPAY ZEKÂ SİSTEMLERİNİN GELİŞTİRİLMESİ

Bahar ÇİFTÇİ

Danışman
Doç. Dr. Ramazan TEKİN

Diğer Jüri Üyeleri

Doç. Dr. Yılmaz KAYA Doç. Dr. Melih KUNCAN Doç. Dr. M. Recep MİNAZ
Dr. Öğr. Üyesi Kaplan KAPLAN

Ekim-2024
BATMAN

TEZ KABUL VE ONAYI

Bahar ÇİFTÇİ tarafından hazırlanan "Gen Dizilerinin Tanımlanması ve Sınıflandırılması Amacıyla Yapay Zekâ Sistemlerinin Geliştirilmesi" adlı tez çalışması 18/10/2024 tarihinde aşağıdaki jüri tarafından oy birliği ile Batman Üniversitesi Lisansüstü Eğitim Enstitüsü Elektrik-Elektronik Mühendisliği Ana Bilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Doç. Dr. Melih KUNCAN	……..
Danışman Doç. Dr. Ramazan TEKİN	……..
Üye Doç. Dr. Yılmaz KAYA	……..
Üye Doç. Dr. M. Recep MİNAZ	……..
Üye Dr. Öğr. Üyesi Kaplan KAPLAN	……..

Yukarıdaki sonucu onaylarım.

Dr. Öğr. Üyesi Ömer Murat ÖTER
Lisansüstü Eğitim Enstitüsü Müdürü

ETİK BEYANI

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını beyan eder, aksinin ortaya çıkması durumunda her türlü yasal sorumluluğu kabullendiğimi bildiririm.

ETHICAL DECLARATION

I declare that all the information in this thesis has been obtained within the framework of ethical behavior and academic rules, and that the source of any statements and information that do not belong to me in this study prepared in accordance with the thesis writing rules has been fully cited, and I declare that I accept all kinds of legal responsibility in case of any contrary situation.

Bahar ÇİFTÇİ
18.10.2024

ÖZET

DOKTORA TEZİ

GEN DİZİLERİNİN TANIMLANMASI VE SINIFLANDIRILMASI AMACIYLA YAPAY ZEKÂ SİSTEMLERİNİN GELİŞTİRİLMESİ

Bahar ÇİFTÇİ

Batman Üniversitesi Lisansüstü Eğitim Enstitüsü

Elektrik-Elektronik Mühendisliği Ana Bilim Dalı

Danışman: Doç. Dr. Ramazan TEKİN

2024, 124 Sayfa

Dünya genelinde milyarlarca virüs türü bulunmakta ve en küçük parazit varlıklar olan virüsler ciddi bir tehdit oluşturmaktadır. Virüslerin geniş çeşitliliği ve hızlı evrimi göz önüne alındığında, bulaşma dinamiklerini daha iyi anlamak ve hedefe yönelik tedavilerin geliştirilmesini kolaylaştırmak amacıyla viral türlerin ve potansiyel konakçılarının hızlı ve doğru bir şekilde sınıflandırılmasına ihtiyaç duyulmaktadır. Bu kapsamda, çalışmada patojenik tek sarmallı RNA virüslerinden oluşan ve farklı viral türler ile konakçılar içeren PhyVirus veri seti incelenmiştir. Tez, üç ana bölümden oluşmakta olup her bölüm, genetik dizilerin sınıflandırılmasına farklı bir perspektiften yaklaşmaktadır. İlk bölümde, K-Mer kodlama yöntemi ile viral aileler ve konakçılar, Random Forest, Gradient Boosting, Extra Trees ve Tam Bağlantılı Derin Sinir Ağı (FCDNN) gibi Makine Öğrenmesi ve Derin Öğrenme algoritmaları kullanılarak sınıflandırılmıştır. FCDNN yöntemiyle virüs ailelerinin %99,60 başarı oranıyla tahmin edilmesi, çalışmanın önemli sonuçlarından biridir. Konak tahmininde ise en yüksek başarı %81,53 oranıyla ExtraTrees sınıflandırıcısı ile elde edilmiştir. Gen dizilerinde K-Mer kodlamaya dayanan farklı kelime uzunluklarının, viral ailelere ve konakçılara göre sınıflandırmaya etkisi değerlendirilmiş, sınıflandırma sonuçlarına ve literatür araştırmasına dayanarak konakçılar arasındaki akrabalık, genetik benzerlikler ve evrimsel ilişkiler incelenmiştir. İkinci bölümde, gen dizilerinin grafik ve görüntü tabanlı kodlama teknikleri (FCGR, DNAWalk, Gri Ölçekli Dönüşüm) kullanılarak sınıflandırılması gerçekleştirilmiştir. Bu teknikler, bir CNN modeli (InceptionV3) ile analiz edilmiş ve Gri Ölçekli Dönüşüm yöntemi ile %99,89 olarak doğruluk oranına ulaşılmıştır. DNAWalk uygulamasında gen dizisi yörünge görüntüleri %99,14 doğruluk oranıyla sınıflandırılmıştır. FCGR uygulamasında ise k'nın 3 ile 8 değerleri arasında gerçekleştirilen kodlamalarda en yüksek doğruluk %99,85 olarak elde edilmiştir. Bu tekniklerle yapılan kodlamalar, viral aileler ve konakların daha doğru sınıflandırılmasına olanak tanımıştır. Mevcut literatür incelendiğinde, gen dizilerinin farklı kodlama yöntemleriyle bir veri seti üzerinde uygulanıp bu yöntemlerin sınıflandırma performansına etkilerinin kapsamlı şekilde analiz edildiği başka bir çalışma bulunmamaktadır. Bu tez çalışmasının, bu alandaki önemli bir boşluğu doldurarak literatüre anlamlı bir katkı sunması amaçlanmaktadır.

Gen dizileri, çeşitli biyolojik ve teknik süreçlerden geçerek analiz için hazır hale getirilmektedir. Ancak bu süreçlerin herhangi bir aşamasında ortaya çıkabilecek hatalar, gen dizilerinde eksik verilere neden olabilmektedir. Literatürde sıkça tartışılan eksik veri tahmini, genellikle verilerin hizalanmış olmasını gerektiren mevcut yöntemlere dayanmaktadır. Tezin üçüncü bölümünde, eksik veri tahmin yöntemleri ele alınmış ve KNN-Imputation yöntemi için yeni bir yaklaşım geliştirilmiştir. PhyVirus veri setindeki gen dizilerinin farklı uzunlukları,

mevcut eksik veri tahmin yöntemlerinin doğrudan uygulanmasını engellemiştir. Bu sorun, geliştirilen KNN-Imputation yaklaşımıyla çözülerek çalışmaya özgün bir katkı sağlanmıştır.

Bu tez, genetik dizilerin kodlanması, sınıflandırılması ve eksik verilerin tahmini için yenilikçi yaklaşımlar geliştirmeyi ve bu yöntemlerin biyoinformatik araştırmalarda nasıl kullanılabileceğini ortaya koymayı amaçlamaktadır. Elde edilen sonuçlar, viral genom analizi ve sınıflandırma süreçlerine yeni metodolojik katkılar sunarak, bu alandaki bilimsel çalışmalara önemli bir referans niteliğinde olmayı hedeflemektedir.

Anahtar Kelimeler: RNA virüsleri, Viral sınıflandırma, Makine Öğrenmesi, Derin Öğrenme, K-Mer, FCGR, DNAWalk, KNN-Imputation

ABSTRACT

DOCTORAL THESIS

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE SYSTEMS FOR IDENTIFICATION AND CLASSIFICATION OF GENE SEQUENCES

Bahar ÇİFTÇİ

Batman University Graduate Education Institute

Department of Electrical and Electronics Engineering

Advisor: Assoc. Prof. Dr. Ramazan TEKİN

2024, 124 Pages

There are billions of virus species worldwide, and as the smallest parasitic entities, viruses pose a significant threat. Given the vast diversity and rapid evolution of viruses, there is a critical need for the rapid and accurate classification of viral species and their potential hosts to better understand transmission dynamics and facilitate the development of targeted treatments. In this context, the PhyVirus dataset, which consists of pathogenic Single-Stranded RNA viruses and contains different viral species and hosts, is analyzed in this study. The thesis consists of three main chapters and each chapter approaches the classification of genetic sequences from a different perspective. In the first section, viral families and hosts are classified using the K-Mer encoding method along with machine learning (ML) and deep learning (DL) algorithms, such as Random Forest, Gradient Boosting, Extra Trees, and Fully Connected Deep Neural Network (FCDNN). Prediction of virus families with FCDNN method with %99,60 success rate is one of the important results of the study. In host prediction, the highest success rate of %81,53 was obtained with the ExtraTrees classifier. The impact of different K-Mer word lengths on the classification of viral families and hosts was evaluated, and evolutionary relationships, genetic similarities, and host relatedness were examined based on classification results and literature review. In the second section, classification of genetic sequences was performed using graphical and image-based encoding techniques (FCGR, DNAWalk, and Grayscale Transformation). These techniques were analyzed with a CNN model (InceptionV3) and an accuracy rate of %99,89 was achieved with the Grayscale Transform method. In the DNAWalk coding method, the genetic sequence trajectory images were classified with an accuracy of %99,14. In the FCGR coding method, the highest accuracy of %99,85 was obtained with word lengths between 3 and 8. These methods allowed for more accurate classification of viral families and hosts. Upon reviewing the existing literature, no other study was found that comprehensively analyzes the effects of different encoding methods on classification performance using a single dataset. This thesis aims to fill a significant gap in the field and make a meaningful contribution to the literature.

Gene sequences are made ready for analysis through various biological and technical processes. However, errors that may occur at any stage of these processes may cause missing data in gene sequences. Missing data prediction, which is frequently discussed in the literature, is usually based on existing methods that require the data to be aligned. In the third part of the thesis, missing data prediction methods are discussed and a new approach for the KNN-Imputation method is developed. The different lengths of the gene sequences in the PhyVirus dataset prevented the direct application of existing missing data prediction methods. This issue was

resolved by the newly developed KNN-Imputation approach, which provided a unique contribution to the study.

This thesis aims to develop innovative approaches for encoding, classifying, and imputing missing data in genetic sequences and to demonstrate how these methods can be applied in bioinformatics research. The results obtained aim to be an important reference for scientific studies in this field by providing new methodological contributions to viral genome analysis and classification processes.

Keywords: RNA viruses, Viral classification, Machine Learning, Deep Learning, K-Mer, FCGR, DNAWalk, KNN-Imputation

ÖNSÖZ

Öncelikle doktora öğrenim sürecimin her aşamasında bana rehberlik eden, bilgi ve tecrübelerini paylaşarak akademik gelişimime büyük katkı sağlayan, hoşgörü ve sabrını benden hiç eksik etmeyen, kıymetli danışman hocam Sayın Doç. Dr. Ramazan TEKİN' e sonsuz teşekkür ve saygılarımı sunarım.

Tez izleme sürecinde yapıcı eleştirileri ve yönlendirmeleri ile çalışmama katkı sağlayan, aynı zamanda süreç boyunca motivasyonumu yüksek tutmamda önemli bir rol oynayan Tez İzleme Komitesi üyeleri Sayın Doç. Dr. Yılmaz KAYA ve Sayın Doç. Dr. Melih KUNCAN' a teşekkür ederim.

Hayatım boyunca her zaman yanımda olan, zor süreçlerimde bile yalnız hissettirmeyen ve bu hayatta sahip olduğum için çok şanslı hissettiğim sevgili aileme şükranlarımı sunarım. Bu günlere gelmemde onların dua ve desteklerinin büyük bir payı olduğu inancındayım.

Bu tezi, en büyük ilham kaynağım olan, çalışmalarımdan ötürü zamanından çaldığım, canımdan çok sevdiğim biricik kızım Asya'ya ithaf ediyorum...

Bahar ÇİFTÇİ
BATMAN-2024

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

SİMGELER VE KISALTMALAR

Kısaltmalar

A	: Adenin
ACT	: Actinopterygii
ARC	: Arachnida
ARN	: Arena
AUROC	: ROC Eğrisi Altındaki Alan
AVS	: Aves
BPCA	: Bayes Ana Bileşen Analizi
BLAST	: Basic Local Alignment Search Tool
C	: Sitozin
CGR	: Chaos Game Representation
CLC	: Calici
CNN	: Evrişimli Sinir Ağları
CNN-LSTM	: Evrişimli Sinir Ağları- Uzun Kısa Süreli Bellek
CRN	: Corona
ÇKA	: Çok Katmanlı Algılayıcı
DÖ	: Derin Öğrenme
DNA	: Deoksiribonükleik Asit
DP	: Doğru Pozitif
DSA	: Derin Sinir Ağları
DT	: Karar Ağacı
ET	: Extra-Trees
FCDNN	: Fully Connected Deep Neural Networks
FCGR	: Frequency Chaos Game Representation
FLO	: Filo
FLV	: Flavi
G	: Guanin
GAIN	: Generative Adversarial Imputation Network
GAN	: Generative Adversarial Nets
GB	: Gradient Boosting
GYSA	: Geri Yayılımlı Sinir Ağları
HMM	: Gizli Markov Modeli
HMS	: Homo sapiens
HNT	: Hanta
ICA	: Bağımsız Bileşen Analizi
IKNNImpute	: Yinelemeli KNNImpute
INS	: Insecta
IRD	: Grip Araştırma Veritabanı
KNN	: K-En Yakın Komşular
LASSO	: En Küçük Mutlak Shrinkage ile Seçim Operatörü
LBP	: Yerel İkili Desen
LDA	: Doğrusal Diskriminant Analizi
LR	: Lojistik Regresyon
MAR	: Rastgele Eksik Veri
MCAR	: Tamamen Rastgele Eksik Veri
MMM	: Mammalia
MÖ	: Makine Öğrenmesi
MNAR	: Rastgele Eksik Olmayan Veri

NB	: Naive Bayes
NCBI	: Ulusal Biyoteknoloji Bilgi Merkezi
NIAID	: Ulusal Alerji ve Enfeksiyon Hastalıkları Enstitüsü
OSA	: Olasılıksal Sinir Ağları
PCR	: Polimeraz Zincir Reaksiyonu
PCA	: Temel Bileşen Analizi
PHN	: Phenui
PRB	: Peribunya
PRM	: Paramyxo
RF	: Random Forest
RHB	: Rhabdo
RMSE	: Kök Ortalama Kare Hatası
RNN	: Tekrarlayan Sinir Ağları
RPT	: Reptilia
rRNA	: Ribozomal RNA
SLLSImpute	: Sequential Local Least Squares Imputation
SVC	: Doğrusal Destek Vektör Sınıflandırıcısı
SCDA	: Gürültü Giderici Otomatik Kodlayıcı
SKNNImpute	: Sıralı KNNimpute
SVM	: Destek Vektör Makinesi
T	: Timin
TGA	: Toga
tRNA	: Taşıyıcı RNA
U	: Urasil
UNC	: Unclassified
VC	: Virüs Konakları
VIPR	: Virüs Patojen Veri Tabanı ve Analiz Kaynağı
VF	: Virüs Aileleri
YN	: Yanlış Negatif
YP	: Yanlış Pozitif
YSA	: Yapay Sinir Ağları
YZ	: Yapay Zekâ

1. GİRİŞ

1.1. Doktora Çalışmasının Amacı ve Kapsamı

Biyoinformatik, günümüzde büyük ve karmaşık biyolojik verileri anlamlandırmak, bu bilgileri güvenli bir şekilde depolamak ve biyolojik araştırmalara yön vermek amacıyla giderek daha da önemli bir araştırma alanı haline gelmiştir. Genomik verilerin büyüyen hacmi ve artan karmaşıklığı, yeni analiz tekniklerine duyulan gereksinimi daha da artırmaktadır. Bu doğrultuda, bu tez çalışması, viral türlerin ve konakçılarının hızlı ve doğru bir şekilde sınıflandırılmasına odaklanarak, virüs bulaşma modelleri ve konakçı-patojen etkileşimleri hakkında derinlemesine bir bakış açısı sunmayı amaçlamaktadır. Bu sayede, salgın hastalıkların tahmini ve kontrol stratejilerinin geliştirilmesine önemli katkılar sağlanması hedeflenmektedir. Bu yaklaşımlar, virüs evrimi ve adaptasyonu konusunda literatüre katkı sağlarken, biyolojik araştırmalara da yeni bir perspektif kazandırmayı amaçlamaktadır.

Tez, üç ana bölümden oluşmaktadır ve her bir bölüm, genetik dizilerin sınıflandırılması ve analizine farklı bir perspektiften yaklaşarak, bu alanın derinlemesine anlaşılmasına katkıda bulunmayı hedeflemektedir. Tezin ilk bölümünde, filogenetik gen dizileri farklı k değerleri kullanılarak K-Mer ile kodlanmış ve bu diziler, çeşitli MÖ ve DÖ yöntemleriyle farklı eğitim-test oranları seçilerek sınıflandırılmıştır. Bu bölümde, K-Mer kodlama yönteminde kullanılan farklı k değerlerinin, sınıflandırma performansı üzerindeki etkisi analiz edilmiş ve bu değerlerin viral aileler ve konakçılar bazında sınıflandırma doğruluğunu nasıl etkilediği değerlendirilmiştir. Elde edilen sonuçlar, genetik dizilerin kodlanmasında K-Mer yönteminin potansiyel avantajlarını ve sınırlamalarını ortaya koymaktadır.

Tezin ikinci bölümünde, sayısal gösterim yöntemi olan K-Mer yöntemi dışında, grafik ve resim tabanlı farklı gen dizisi kodlama teknikleri kullanılarak filogenetik gen dizilerinin sınıflandırılması gerçekleştirilmiştir. Bu bölümde kullanılan teknikler arasında FCGR, DNAWalk ve Gri Ölçekli Dönüşüm teknikleri bulunmaktadır. Elde edilen veriler, InceptionV3 DÖ modeli kullanılarak sınıflandırılmıştır. Kodlama yöntemlerinin tercih edilme sebebi, farklı uzunluklardaki gen dizilerinden oluşan geniş veri setlerinin daha

etkili bir şekilde sınıflandırılmasını sağlamaktır. Bu bölüm, literatürdeki çeşitli gen dizisi kodlama tekniklerini ve bu tekniklerin sınıflandırma süreçlerine olan etkilerini, ayrıca sınıflandırma performansını nasıl geliştirdiklerini incelemektedir.

Tezin üçüncü bölümünde, filogenetik gen dizilerindeki eksik kısımların tahmin edilmesine odaklanılmıştır. Bu kapsamda, literatürdeki eksik veri tahmin yöntemleri incelenmiştir. Kullanılan veri seti, aynı viral aile içinde bile oldukça farklı gen dizisi uzunluklarına sahip olduğundan, literatürdeki mevcut yöntemlerin doğrudan uygulanması mümkün olmamıştır. Bu nedenle, KNN-Imputation yöntemi tercih edilmiş, ancak veri setindeki farklı uzunluktaki diziler nedeniyle bu yöntemin uygulanabilmesi için özel bir yaklaşım geliştirilmiştir. Bu yaklaşımda, aynı viral aileler içinde eksik veri içermeyen diziler arasında bir bölütleme işlemi önerilmiştir. Ardından, Etiket kodlama yöntemi kullanılarak kodlanan veriler, KNN-Imputation yöntemi ile atandıktan sonra yuvarlanmış ve sonrasında decode edilmişlerdir. Bölümde, eksik verilerin doğru bir şekilde tahmin edilmesi yoluyla veri setinin daha kapsamlı bir şekilde analiz edilmesi ve sınıflandırma doğruluğunun artırılması hedeflenmiştir. Elde edilen sonuçlar, eksik veri yönetimi ve tamamlanması için yeni yaklaşımlar sunmaktadır.

Bu kapsam doğrultusunda tez, genetik dizilerin farklı yöntemlerle kodlanması, sınıflandırılması ve eksik verilerin tahmini konusunda yenilikçi yaklaşımlar geliştirmeyi ve bu yaklaşımların biyoinformatik araştırmalarda nasıl uygulanabileceğini göstermeyi hedeflemektedir. Tezin sonuçları, viral genom analizi ve sınıflandırma süreçlerinde yeni metodolojik katkılar sağlamayı amaçlamakta ve bu alanda yapılan bilimsel çalışmalara önemli bir referans kaynağı olmayı hedeflemektedir.

1.2. Tezin Özgün Değeri

Biyoinformatik alanında kullanılan veri setleri genellikle karmaşık verilere sahip büyük veri setleridir. Bu çalışmada kullanılan PhyVirus veri seti ise, farklı viral aileler ve konakçı bilgileri ile aynı aile içinde dahi dizi uzunlukları açısından farklı uzunluklara sahip gen dizileri içeren bir veri seti olarak öne çıkmaktadır. Mevcut literatürde, PhyVirus veri seti hizalama tabanlı yöntemler kullanılarak incelenmiş ve bu yöntemlerle sadece RNA virüslerinin ortak genomik ve evrimsel özelliklerinin varlığı araştırılmıştır (Kustin & Stern, 2021). Ancak, bu çalışmada ilk defa YZ yöntemleri kullanılarak veri seti üzerinde sınıflandırma yapılmış ve literatürdeki diğer çalışmalarla karşılaştırıldığında

oldukça başarılı sonuçlar elde edilmiştir (Gunasekaran vd., 2021; Lopez-Rincon vd., 2020; Remita & Diallo, 2019).

Bu çalışmada 13 farklı viral aile ve 8 farklı konak türü temel alınarak, önceki çalışmalarda gerçekleştirilmeyen geniş kapsamlı bir sınıflandırma uygulanmıştır. Ayrıca, K-Mer kodlama yöntemi kullanılarak verilerde herhangi bir kayıp veya değişiklik olmaksızın analiz yapılabilmiş ve bu sayede genomik sınıflandırmada özellik seçiminin önemi ile bu yöntemin uyarlanabilirliğinin sınıflandırıcı performansı üzerindeki etkisi vurgulanmıştır.

Genetik dizilerin uzun ve karmaşık yapıları, ham halleriyle analiz edilmelerini zorlaştırmakta ve büyük veri setlerinin işlenmesi ve yorumlanmasında zorluklar oluşturmaktadır. Bu tez çalışmasının özgün katkılarından bir diğeri, PhyVirus veri setine dört farklı gen dizisi kodlama yöntemi uygulanarak, bu veri setindeki gizli bilgilerin, karmaşık kalıpların ve tekrarların keşfedilmesine olanak sağlamasıdır. Tez kapsamında gerçekleştirilen uygulama, kodlama yöntemlerinin çeşitli boyutlardaki genetik dizilerden oluşan veri setiyle etkileşimlerini incelemiş ve bu veri setindeki örüntülerin daha belirgin bir şekilde ortaya çıkmasını sağlamıştır. Böylece, genetik verilerin işlenmesi ve yorumlanmasında daha hassas ve etkili yöntemlerin geliştirilmesine imkân tanınmıştır. Mevcut literatür incelendiğinde, gen dizilerinin bu çalışmada olduğu gibi farklı kodlama yöntemleriyle bir veri seti üzerinde uygulanarak, bu yöntemlerin etkinliklerinin ve sınıflandırma performansına etkilerinin kapsamlı bir şekilde analiz edildiği başka bir çalışmaya rastlanmamıştır. Bu tez çalışması, bu alanda önemli bir boşluğu doldurarak, literatüre değerli bir katkı sunmuştur.

Gen dizileri, çeşitli biyolojik ve teknik süreçlerden geçerek analiz için hazır hale getirilmektedir. Ancak, bu süreçlerin herhangi bir aşamasında ortaya çıkabilecek hatalar, gen dizilerinde eksik verilere neden olabilmektedir. Eksik veri tahmini, literatürde sıkça tartışılan bir konu olmasına rağmen, mevcut yöntemler genellikle verilerin eşit uzunluklara sahip olmalarını gerektirmektedir. Bu çalışmanın özgün yanlarından biri, tez kapsamında kullanılan farklı uzunluklara sahip PhyVirus veri seti için uyarlanan bir yöntemle KNN-Imputation yönteminin eksik veri tahminine ilk kez uygulanmasıdır. Önerilen yöntemin, orantısız uzunluklara sahip PhyVirus gibi karmaşık veri setlerine uygulanabilir olması, literatüre önemli bir katkı sağlamaktadır.

1.3. Tezin Organizasyonu

Tez bölümlerinin organizasyonu aşağıdaki gibidir:

Bölüm 1'de doktora tezinin amaç ve kapsamı, özgün katkısı ortaya konulmuştur.

Bölüm 2'de canlılar için genetik bilgi; DNA'nın yapısı, RNA'nın yapısı, RNA çeşitleri, virüsler, Pozitif Anlamlı Tek Sarmallı RNA virüsleri, Negatif Anlamlı Tek Sarmallı RNA virüsleri, gen sekansı, sekans teknikleri, gen dizi alanları konularına yer verilmiştir.

Bölüm 3'te tez çalışmasının odaklandığı üç farklı alana dair literatür araştırmasına yer verilmiştir. Literatür araştırmasında gen dizilerinin MÖ ve DÖ ile sınıflandırılması, gen dizisi kodlama yöntemleri ve gen dizilerinde eksik veri atama yöntemleri alanlarındaki çalışmalara odaklanılmıştır.

Bölüm 4'te çalışmada kullanılan veri seti ile ilgili bilgilere yer verilmiştir.

Bölüm 5'de tez kapsamında DNA dizilerinin sınıflandırılmasında kullanılan çeşitli yöntemler olan Hizalama yöntemleri, MÖ yöntemleri, DÖ yöntemleri, Evrişimli Sinir Ağları (CNN), katman yapıları, model performans metrikleri hakkında detaylı bilgiler yer almaktadır.

Bölüm 6'da gen dizileri kodlama yöntemleri üç ana kısımda ele alınmıştır. Sayısal gösterim yöntemleri, Grafiksel gösterim yöntemleri, Resim gösterim yöntemleri şeklinde incelenmiştir.

Bölüm 7'de virüs ailelerine ve konaklarına dayalı sınıflandırma uygulaması gerçekleştirilmiştir. Önerilen yöntem aşamalar şeklinde açıklanarak virüs ailelerine ve konaklarına dayalı sonuçlar detaylı bir şekilde incelenerek değerlendirilmiştir.

Bölüm 8'de kodlama yöntemlerinin PhyVirus veri setine uygulamasının aşamaları detaylıca anlatılmış olup, uygulanan yöntemlerin başarıları gösterilip değerlendirilmiştir.

Bölüm 9'da gen dizilerinde eksik veriler, türleri, tahmin yöntemleri anlatılmıştır.

Bölüm 10'da PhyVirus veri setindeki eksik verilerin dağılımları, KNN-Imputation yönteminin PhyVirus veri setine uygulaması, sonuçları ve sonuçların değerlendirilmesi gerçekleştirilmiştir.

Bölüm 11'de sonuçlar genel olarak ele alınıp değerlendirilerek, tartışılmıştır.

2. GENETİK BİLGİ

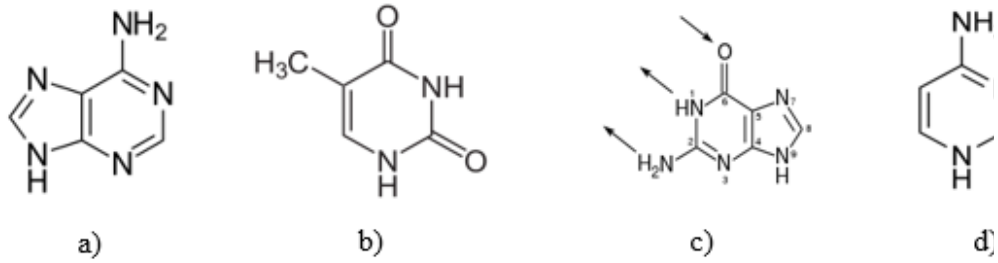
Genetik bilgi, bir organizmanın genetik materyalinde depolanarak, hücrelerin büyümesi, gelişmesi, çalışması ve üremesi için gerekli olan talimatları içeren biyolojik bilgilerden oluşmaktadır. Bu bilgi, nükleotidlerin belirli bir diziliminden meydana gelmekte ve proteinlerin sentezi için şifre sağlamaktadır. Genetik bilgi, canlıların genetik özelliklerini ve biyolojik işlevlerini belirleyen temel unsur olarak kabul edilmektedir. Canlı organizmaların biyolojik işlevlerini ve fiziksel özelliklerini belirlemekte ve genetik materyalin nesilden nesile aktarılmasını sağlamaktadır. Genetik bilginin doğru bir şekilde işlenmesi ve aktarılması, organizmaların sağlıklı gelişimi ve çevresel değişikliklere adaptasyonu için kritik bir öneme sahiptir.

Biyoinformatik, genetik bilgiye dayalı verilerin analiz edilmesi, yorumlanması ve depolanması için bilgisayar teknolojilerini ve matematiksel modelleri kullanan bir bilim dalı olarak tanımlanmaktadır. Genetik bilgi, biyoinformatik sayesinde biyolojik sistemlerin anlaşılmasında ve hastalıkların genetik temellerinin araştırılmasında kullanılmaktadır. Özellikle, genom dizileme teknolojilerinin gelişmesiyle birlikte biyoinformatik, büyük ölçekli genetik verilerin analizinde hayati bir araç haline gelmiştir.

Genetik bilginin hücrelerde doğru bir şekilde işlenmesi ve aktarılması, DNA ve RNA moleküllerinin yapısı ve işleviyle doğrudan ilişkilidir. Bu iki molekülün yapısı ve işlevleri, genetik bilginin nasıl kodlandığını, kopyalandığını ve hücre içinde nasıl kullanıldığını açıklayan temel mekanizmalar olarak değerlendirilmektedir.

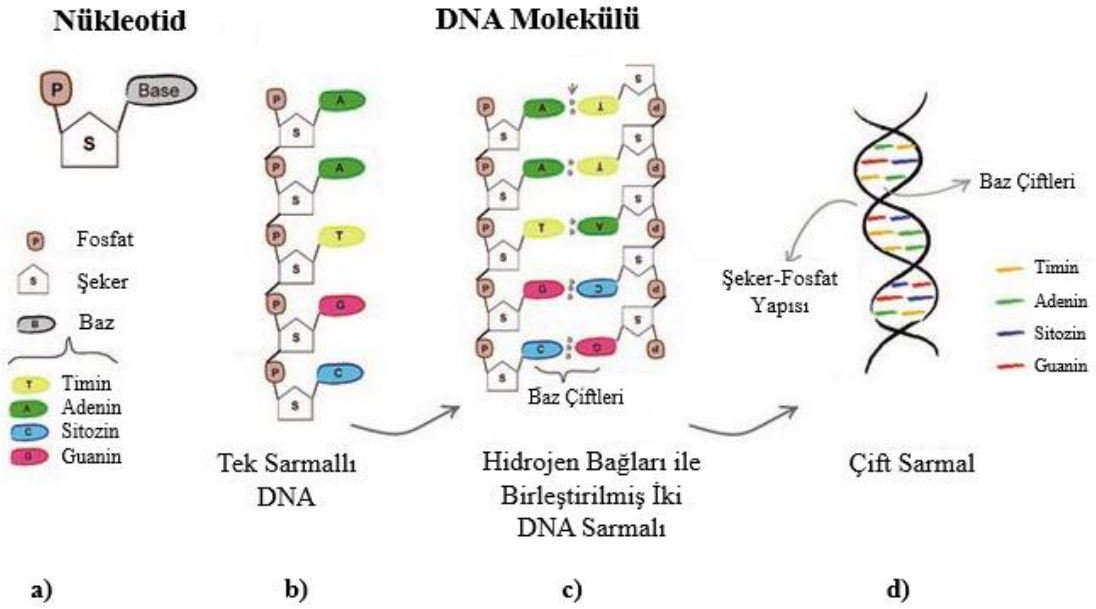
2.1. DNA

Deoksiribonükleik Asit (DNA), tüm canlılarda bulunan ve kalıtımı sağlayan bir moleküldür. Canlıların biyolojik özelliklerinin bir sonraki nesillere aktarılmasını mümkün kılan bu yapı, dört temel yapıtaşından, yani bazlardan meydana gelir. Bu bazlar, azot yapılı moleküller olup kimyasal olarak bazik özellik taşır. Adenin (A), Sitozin (C), Guanin (G) ve Timin (T); DNA'nın temel bazlarıdır.



Şekil 2.1. a) Adenin nükleotidinin baz yapısı, **b)** Timin nükleotidinin baz yapısı, **c)** Guanin nükleotidinin baz yapısı, **d)** Sitozin nükleotidinin baz yapısı (Berdis, 2022)

DNA, uzun baz dizilimlerinden oluşan iki iplikçikten meydana gelir ve bu dizilimlerin farklılıkları her canlıya özgüdür. DNA'nın iki iplikçikli sarmal yapısı, 1953 yılında James D. Watson ve Francis Crick tarafından tanımlanmıştır (Watson & Crick, 1953). Bu keşifleri onlara Nobel Kimya Ödülü kazandırmıştır. DNA'nın iki iplikçikli yapısı, birbiri etrafında dönen bir sarmal şeklindedir ve bu iplikçikler, bazların birbirine tamamlayıcı baz çiftleri oluşturmasıyla bir arada tutulur (Şekil 2.2). Adenin bazının tamamlayıcısı Timin, Timin bazının tamamlayıcısı Adenin; Guanin bazının tamamlayıcısı Sitozin, Sitozin bazının tamamlayıcısı ise Guanindir. Bu bazlar, hidrojen bağları aracılığıyla birbirine bağlanır.



Şekil 2.2. a) DNA nükleotidinin yapısı, **b)** DNA'nın tek iplikçik yapısı, **c)** DNA iplikçiklerinin hidrojen bağları ile bir araya gelmesi **d)** İki iplikçik yapısı (Cano Londoño vd., 2020)

DNA’nın azotlu bir baz, beş karbonlu bir şeker (deoksiriboz) ve bir fosfat grubundan oluşan temel yapıtaşı “nükleotid” olarak adlandırılır. Azotlu bazlar, genetik bilgiyi taşıyan ve DNA'nın çift sarmal yapısını koruyan ana unsurlardır. Beş karbonlu şeker ve fosfat grubu, nükleotidleri birbirine bağlayarak DNA’nın omurgasını meydana getirir. Bu bağlantılar zincir boyunca devam ederek uzun ve stabilize bir molekül oluşturur.

2.2. RNA

Ribonükleik Asit (RNA), ribo nükleotidlerden oluşan ve genellikle tek zincirli bir yapıya sahip olan bir moleküldür. RNA, beş karbonlu bir şeker (riboz) ve baz yapısından meydana gelir. RNA’nın yapısında yer alan bazlar Adenin, Guanin, Sitozin ve Urasil’dir. Urasil (U), yalnızca RNA'da bulunan ve DNA'da yer almayan bir bazdır. RNA, DNA'dan farklı olarak tek iplikçikli yapıda olup riboz şekeri içerir ve Timin yerine Urasil bazı bulundurur. DNA, genetik bilgiyi uzun süre saklama kapasitesine sahiptir ve bu nedenle kalıcı genetik materyal olarak işlev görür. RNA ise DNA kadar stabil değildir ve belirli ortamlarda hızla parçalanabilir. Bu nedenle, RNA genellikle kısa ömürlüdür ve geçici olarak işlev görür. RNA’nın üç ana türü vardır: Haberci RNA (mRNA), Ribozomal RNA (rRNA) ve Taşıyıcı RNA (tRNA).

mRNA; DNA'daki genetik bilginin proteinlere dönüşümünde önemli bir rol oynamaktadır. mRNA oluşumu için, DNA sarmalı açılır ve RNA polimeraz enzimi, DNA'nın bir kalıp ipliği üzerinden mRNA sentezler. Bu süreçte DNA’daki bilginin mRNA’ya aktarılması transkripsiyon olarak adlandırılmaktadır. Transkripsiyon tamamlandıktan sonra, mRNA’daki bilgi ribozomlarda proteinlere çevrilir ve bu aşama translasyon olarak bilinmektedir.

tRNA, mRNA’daki nükleotid dizisinin içeriğindeki bilgiyi aminoasitlerin tanıması ve ribozoma taşınması için kritik bir rol oynar. Her tRNA molekülü belirli bir aminoasidi bağlar ve ribozomlara getirir. Burada, mRNA’nın kodonlarına uygun antikodonlarla bağlanarak protein sentezine katkıda bulunur.

rRNA, ribozomların yapısında bulunan ve ribozomların proteinlerle birleşerek yapı oluşturmasını sağlayan temel bileşenlerden biridir. Ribozomlar, protein sentezinin gerçekleştiği hücresel organellerdir ve rRNA, bu sürecin temel bir parçasıdır. rRNA, ribozomların işlevini destekler ve protein sentezinin doğruluğunu sağlar.

2.3. Virüsler, Pozitif Anlamlı Tek Sarmallı RNA Virüsleri ve Negatif Anlamlı RNA Virüsleri

Virüsler, çoğalmak için bir konak hücreye ihtiyaç duyan en küçük parazit formudur ve dünya genelinde milyarlarca bulunmaktadır. DNA tarafından oluşturulan virüsler "DNA virüsleri" olarak adlandırılırken, RNA tarafından oluşturulanlar "RNA virüsleri" olarak adlandırılmaktadır. Baltimore sınıflandırma sistemi, büyük çeşitlilik gösteren bu virüsleri, genetik materyalleri ve replikasyon stratejileriyle ilgili ayırt edici özelliklere dayanarak sınıflandırmak için kullanmaktadır. David Baltimore tarafından 1971 yılında geliştirilen bu sınıflandırma şeması, virüsleri genomlarının doğasına ve replikasyonlarında yer alan mekanizmalara göre yedi sınıfa (1'den 7'ye) ayırmaktadır (Baltimore, 1971). Pozitif Anlamlı Tek Sarmallı RNA virüsleri (+ssRNA) Baltimore dördüncü gruba aittir ve Artı Sarmal veya Sens İplikçik olarak da bilinir. Negatif Anlamlı RNA virüsleri (-ssRNA), Baltimore beşinci gruba aittir ve Eksi İplikçik veya Antisens İplikçik olarak bilinir (D. Liu vd., 2019). +ssRNA virüsleri, doğrudan mRNA olarak görev yapan genoma sahiptirler. Bu, genetik bilginin viral RNA ile aynı yönde olduğu ve konakçı ribozomlar tarafından anında proteinlere çevrilmesine izin verdiği anlamına gelmektedir (D. Liu vd., 2019). Aksine, -ssRNA virüsleri viral mRNA'ya tamamlayıcı genoma sahiptirler. Bu durum, replikasyon sürecinde ek bir adım gerektirmektedir; çünkü translasyonun gerçekleşebilmesi için önce viral RNA'nın tamamlayıcı pozitif anlamlı bir mRNA'ya transkribe edilmesi gerekir (D. Liu vd., 2019). +ssRNA ve -ssRNA virüsleri arasındaki farklar Tablo 2.1’ de gösterilmiştir.

Tablo 2.1. +ssRNA ve -ssRNA virüsleri arasındaki farklar

	+ssRNA virüsleri	**-ssRNA virüsleri**
İçerik	Doğrudan mRNA üzerinde çalışan, genetik içerik olarak tek iplikli bir RNA içermektedir.	Tamamlayıcı mRNA dizisini oluşturan, genetik içerik olarak tek iplikli bir RNA içermektedir.
RNA Genomu	Pozitif anlamda RNA genomu mevcuttur.	Negatif anlamda RNA genomu mevcuttur.
Çoğalma Şekli	Çift sarmallı bir RNA ara ürünü aracılığıyla gerçekleşmektedir.	RNA'ya bağımlı RNA polimeraz yardımıyla gerçekleşmektedir.
Transkripsiyon İhtiyacı	Gerekli değildir.	Çeviri gerçekleşmeden önce, pozitif anlamda RNA'ya kopyalanmalıdır.
Viral mRNA'nın Rolü	Proteinlere kolayca çevrilebilmektedir.	mRNA'yı tamamlamaktadır.

2.4. Gen Sekansı ve Sekans Teknikleri

Gen sekansı, bir organizmanın genomundaki DNA veya RNA dizilimindeki nükleotidlerin sırasını belirleme işlemidir. DNA için bu nükleotidler A, C, G ve T iken, RNA için A, C, G ve U şeklindedir. Gen sekansı, biyolojik araştırmaların temelini oluşturarak, organizmaların genetik yapısını anlamamıza yardımcı olur. Yeni nesil sekanslama teknolojilerinin hızlı gelişimi, gen sekansı elde etme sürecini oldukça hızlandırmıştır. Bu teknolojiler, büyük ölçekli veri kümelerinin üretilmesine olanak tanır ve bu veriler, biyologlar tarafından analiz edilerek önemli biyolojik bilgiler çıkarılabilmektedir (H. Zhang vd., 2019). Gen sekanslama sonucu elde edilen büyük veri kümelerinin anlamlı hale getirilmesi için biyoinformatik bilim dalı kullanılmaktadır. Biyoinformatik yöntemleri sayesinde, genetik verilerdeki kalıplar ve ilişkiler tespit edilmekte ve bu bilgiler biyolojik süreçlerin daha iyi anlaşılmasına katkı sağlamaktadır.

Gen dizilerinin elde edilmesi ve bu dizilerden yeni analizler yapılması, bilimin gelişimine büyük katkılar sağlamıştır. 1960'lı yılların sonlarına kadar DNA'nın analizi oldukça zor bir görev olarak görülmekteydi. O dönemde, DNA'nın yapısını dolaylı olarak protein yapısı, RNA dizilimi veya genetik analiz yoluyla incelemek mümkündü. Ancak günümüzde, genomun belirli bir bölgesini detaylı şekilde incelemek ve kısa sürede tüm dizilimi öğrenmek mümkündür. Bir organizmanın DNA'sını elde etmek için izlenen süreç genellikle şu şekildedir: Öncelikle hücre zarı ve organellerin zarları parçalanır, ardından çeşitli enzimler veya çözeltiler kullanılarak DNA çökeltilir ve diğer makromoleküller DNA'dan ayrılır. Bu süreç, farklı hücre tiplerine göre değişiklik gösterebilir. Örneğin, bitki hücrelerinin zarını eritmek için kullanılan kimyasallar, hayvan hücreleri için kullanılanlardan farklı olabilir.

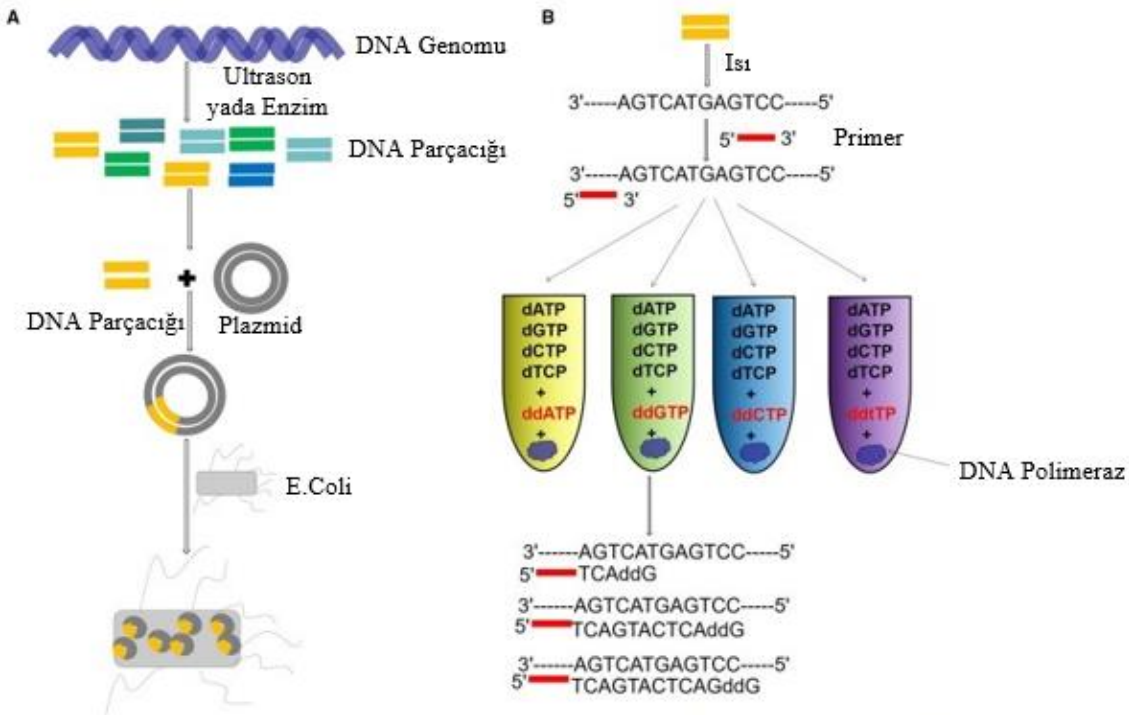
Günümüzde DNA dizi analizini gerçekleştirmek için üç ana yöntem kullanılmaktadır: Kimyasal Kırılma yöntemi (Maxam & Gilbert, 1977), Zincir Sonlanma yöntemi (Sanger vd., 1977) ve Yeni Nesil Dizileme yöntemleri (Margulies vd., 2005; Shendure vd., 2005; Schuster, 2007).

Kimyasal Kırılma yönteminde, DNA kimyasal maddeler aracılığıyla küçük parçalara ayrılır. Nükleotidler farklı kimyasallarla parçalanarak jel bir film üzerine yerleştirilir ve radyoaktif maddelerin değişen renklerine göre dizilim gerçekleştirilir. Bu yöntem çok spesifik ve hassas olsa da uzun dizilemeler için çok fazla zaman ve tehlikeli kimyasal madde gerektirmektedir.

Tablo 2.2. Kimyasal Kırılma Yönteminde kullanılan kimyasallar (Wood, 1983)

Özgül Baz	Baza özgül kimyasal	Baz ayırmada kullanılan kimyasal	Zincir kırmada kullanılan kimyasal
G	Dimetil sülfat	Piperidin	Piperidin
A+G	Asit	Asit	Piperidin
C+T	Hidrazin	Piperidin	Piperidin
C	Hidrazin+Baz	Piperidin	Piperidin
A>C	Baz	Piperidin	Piperidin

Zincir Sonlanma yönteminde, DNA parçası ısıtılarak iki iplikçikli sarmal yapı birbirinden ayrılır. DNA primeri eklenir ve ardından serbest nükleotidler ve DNA polimeraz enzimi ile tamamlayıcı bir DNA zinciri elde edilir. Nükleotidler, UV ışık altında görülebilen farklı renkte boyalarla etiketlenir. Zincir Sonlanma yöntemi devrim niteliğinde olsa da büyük ölçekli projeler için verimsiz kalabilmektedir.



Şekil 2.3. (A) Dizilenecek örnek DNA, (B) Zincir Sonlanma yöntemi (L. Zhang vd., 2021)

Yeni Nesil Dizileme yöntemleri ultra yüksek verim, ölçeklenebilirlik ve hız sunan bir sıralama teknolojisidir. İkinci Nesil ve Üçüncü Nesil Teknolojiler olarak ikiye ayrılır. İkinci Nesil Dizileme, çok sayıda küçük Zincir Sonlanma dizilimini aynı anda çalıştırarak daha hızlı, daha hassas ve düşük maliyetli sonuçlar üretir. Üçüncü Nesil Dizileme ise uzun DNA parçalarını dizileyebilme yeteneği ile öne çıkmaktadır, ancak genellikle biraz daha düşük dizi doğruluğuna sahiptir. İkinci Nesil Dizileme teknikleri, dizilemeden önce

Polimeraz Zincir Reaksiyonu (PCR) amplifikasyonuna ihtiyaç duymaktadır. PCR , belirli bir DNA parçasının milyarlarca kopyasını üretir. Bu teknik, DNA'nın istenilen bölgesinin çoğaltılmasını sağlayarak analiz için gerekli olan DNA miktarını artırmaktadır. İkinci Nesil Dizileme, Zincir Sonlanma yöntemine göre daha kısa sürelerde büyük miktarlarda kısa okumalar üretmektedir (Miyamoto vd., 2014). Üçüncü Nesil Dizileme ise, uzun DNA parçalarıyla ilgilenebilme yeteneği ile İkinci Nesil Dizilemeye göre avantaj sağlar, ancak daha düşük dizi doğruluğuna sahip olabilmektedir (Slatko vd., 2018).

Ayrıca Yeni Nesil Dizileme teknolojisinde, uzun okuma ve kısa okuma yaklaşımları bulunur. Kısa okuma yaklaşımları, daha düşük maliyetle yüksek doğrulukta veriler sunarken, uzun okuma yaklaşımları tam uzunlukta izoform dizilimi için daha uygundur. Ancak, uzun okuma yöntemleri, dizileme maliyetleri ve hata oranlarının yüksek olması gibi nedenlerle sınırlı bir kullanıma sahiptir.

Tablo 2.3. Kısa ve uzun okuma dizilemesi için avantajlar ve dezavantajlar tablosu (Slatko vd., 2018)

	Avantajlar	**Dezavantajlar**
Kısa okuma Dizileme	• Daha yüksek dizi doğruluğu • Ucuz • Parçalanmış DNA'yı dizileyebilme	• Yapısal varyantları çözememek, alelleri aşamamak veya oldukça homolog genomik bölgeleri ayırt edememek • Bazı tekrarlayan bölgelerin kapsamı sağlanamamakta
Uzun okuma Dizileme	• Tekrar dizileri nedeniyle kısa okuma dizisi ile karakterize edilmesi zor olan genetik bölgeleri dizileyebilme • Yapısal yeniden düzenlemeleri veya homolog bölgeleri çözebilme	• Okuma doğruluğu başına daha düşük • Nükleotid tahsisi yüksek hata oranları, ölçeklenebilirlik mevcudiyeti nedeniyle biyoinformatik zorluklar

Sonuç olarak, dizi analizi tekniklerinin tümü, genetik araştırmalarda ve biyomedikal bilimlerde önemli bir yere sahiptir. Teknolojinin gelişmesiyle birlikte, bu tekniklerin doğruluğu ve verimliliği artmakta, bilim insanlarına daha fazla olanak sunmaktadır.

2.5. Gen Dizi Alanları

Yeni Nesil Dizileme teknolojisi, biyoinformatikte devrim yaratarak laboratuvarların çeşitli uygulamalar gerçekleştirmesine ve biyolojik sistemleri detaylı bir

şekilde incelemesine olanak tanımıştır. Bu teknoloji ile üretilen veri hacminde muazzam bir artış olmuş ve daha önce cevaplanamayan birçok sorunun analiz edilip cevaplanması mümkün hale gelmiştir.

Yeni Nesil Dizileme teknolojisi, genomik analiz, bilgi erişimi, sağlık bilişimi ve anomali algılama gibi geniş bir uygulama yelpazesine sahiptir (Thermes, 2014). Verilerdeki bu artış ve sağladığı ayrıntılı bilgiler, özellikle DÖ gibi gelişmiş analitik tekniklerin gen dizisi verilerine uygulanmasının önünü açmıştır. Bu teknoloji ile tüm genomları hızla sıralamak, hedef bölgeleri derinlemesine sıralamak, yeni RNA varyantlarını keşfetmek veya gen ekspresyon analizi için mRNA'ları ölçmek için RNA dizilimini (RNA-Seq) kullanmak, genom çapında DNA metilasyonu ve DNA-protein etkileşimleri gibi epigenetik faktörleri analiz etmek mümkün hale gelmiştir. Yeni Nesil Dizileme teknolojisinin gen dizilerindeki uygulama alanları; Metagenomik, Farmakogenomik, Epigenetik, Tek Hücreli Transkriptomik şeklindedir (Schmidt & Hildebrandt, 2021).

Metagenomik, çevreden elde edilen gen dizi verilerinin analizini kapsar ve çevresel mikroorganizmaların ve biyolojik çeşitliliğin anlaşılmasında kritik öneme sahiptir. Metagenomik bağırsak mikrobunun analiz edilmesi, hastanelerdeki bakteri popülasyonları veya hastalarda viral evrim gibi ekosistemler hakkında önemli bilgiler verebilmektedir (Schmidt & Hildebrandt, 2021).

Farmakogenomik, gen dizi araştırmalarının bir diğer önemli alanıdır ve genetik varyantlar ile ilaçların etkileri arasındaki ilişkileri inceler. Hastalarda meydana gelen onkojenik değişiklikleri tanımlamada ve ilaç duyarlılığı varyasyonlarını belirlemede, ilaç yanıtını tahmin etmede, ilaç kombinasyonu etkinliği araştırmalarında yardımcı olur (Schmidt & Hildebrandt, 2021).

Tek Hücreli Transkriptomik, tek bir hücredeki RNA moleküllerinin tümünü analiz eder ve hücre içindeki gen ekspresyon profillerini inceleyerek hücresel heterojenliğin anlaşılmasına yardımcı olur. Özellikle kanser araştırmaları ve kök hücre biyolojisinde önemli uygulamalara sahiptir.

Epigenetik, genotip modifikasyonlarına dayanmayan fenotip değişikliklerini inceleyen bir alandır ve gen dizi verilerinden sinyallerin nicel tespitinde önemli bir rol oynamaktadır. Epigenetik araştırmalar, genetik materyalin modifikasyonlarını ve bu modifikasyonların fenotipik etkilerini inceleyerek hastalıkların gelişimini ve tedavi hedeflerini anlamada önemli bir rol oynar.

Yeni Nesil Dizileme teknolojileri, genom çapında DNA metilasyonu ve DNA-protein etkileşimleri gibi epigenetik faktörlerin analizinde kullanılmasının yanı sıra, insan mikrobiyomunu inceleme ve yeni patojenleri tanımlama imkânı da sunar. Bu teknolojiler, çevresel metagenomik çalışmalardan bulaşıcı hastalık gözetimine kadar geniş bir yelpazede araştırmacılara genetik içgörüler sağlar. Ayrıca, nadir varyantların, tümörlerin ve daha fazlasının incelenmesine olanak tanıyan dolaşımdaki DNA parçalarını saptayarak kanser tanı ve tedavisinde de önemli uygulamalara sahiptir (Smith, 2017).

Sağlık bilişimi ve klinik uygulamalarda, Yeni Nesil Dizileme yöntemleri önemli ilerlemeler sağlamıştır. Spesifik hastalıklarda altta yatan genomik ve genetik faktörlerin tanımlanması, kesin tanı koymada değerli bir araç olup, hasta yönetimi ve danışmanlığına rehberlik edebilmektedir (Smith, 2017). Genetik bilgi, risk altındaki aile üyelerinin belirlenmesinde de faydalı olabilir. Tüm genom dizilemesi ve bu dizilemenin tıbbi alanlardaki önemi, büyük popülasyon çalışmalarında ve spesifik hasta gruplarında nükleotid ve yapısal genomik varyantların analizinde temel bir rol üstlenmektedir. (Smith, 2017).

Yeni Nesil Dizileme teknolojilerinin sunduğu hız, hassasiyet ve düşük maliyet, genetik araştırmaların ve biyomedikal bilimlerin gelişiminde önemli bir etkiye sahiptir. Bu teknolojiler, genetik varyantların tespitinden çevresel metagenomik çalışmalara kadar geniş bir alanda araştırmacılara önemli bilgiler sunmaktadır. Yeni Nesil Dizileme teknolojilerinin sürekli gelişimi, genetik ve biyomedikal araştırmalarda daha fazla yeniliğe ve keşfe olanak tanımaktadır. Bu alanlardaki çalışmalar, genetik biliminin geleceğini şekillendirerek sağlık ve biyomedikal alanda daha doğru ve kişiye özel çözümler sunma imkanını beraberinde getirmektedir.

3. LİTERATÜR

Biyoinformatikte, büyük ölçekli ve karmaşık biyolojik verilerden bilginin nasıl elde edileceği, depolanacağı ve bu bilginin biyolojik araştırmalara nasıl rehberlik edeceği önemli bir çalışma alanıdır. Gen dizileri, özellikle viral diziler, biyolojik verilerin önemli bir kısmını oluşturur ve bu dizilerin doğru bir şekilde sınıflandırılması, virüslerin evrimsel ilişkilerinin anlaşılması ve yeni patojenlerin tespiti açısından büyük önem taşır. Bu bağlamda, gen dizilerinin viral ailelere göre sınıflandırılması, bu tür verilerin anlamlandırılması açısından birçok bilimsel ve klinik uygulamada kritik bir rol oynamaktadır. Bu sınıflandırma, hastalık patojenlerinin ve tedavilerinin belirlenmesi, salgın araştırmaları, biyogüvenlik, biyolojik savunma ve evrimsel araştırmalar gibi alanlara doğrudan katkıda bulunur. Bu doğrultuda, başlangıçta geliştirilen algoritmalar hizalamaya dayalı yaklaşımlara dayanmaktaydı; ancak, bu yöntemin dezavantajları nedeniyle son zamanlarda literatürde hizalamaya dayalı olmayan yaklaşımlar daha sık tercih edilmektedir. Özellikle gen dizilerinin analizi ve sınıflandırılması için hizalamaya bağlı olmayan yöntemler olan MÖ ve DÖ tekniklerinin kullanımında son yirmi yılda büyük ilerlemeler kaydedilmiştir.

A) Gen dizilerinin Makine Öğrenmesi ve Derin Öğrenme ile sınıflandırılması alanındaki literatür araştırması

Gen dizilerinin analizi ve sınıflandırılması için MÖ ve DÖ tekniklerinin kullanımı açısından önemli gelişmeler, yaklaşık 1990'ların sonlarına dayanmaktadır. Bu gelişmeler, biyolojik araştırmalara önemli bir yol gösterici olmuş ve genetik bilgiyi daha derinlemesine anlama ve biyolojik süreçlerin anlamını çıkarma konusunda büyük bir rol oynamıştır.

Gen dizilerinin ailelerine göre sınıflandırılmasıyla ilgili ilk çalışmalar MÖ teknikleri kullanılarak yürütülmüştür. Salzberg ve arkadaşları çalışmalarında; Haemophilus Influenzae ve Helicobacter Pylori ve farklı mikrobiyal genomlar üzerinde Gizli Markov Modeli (HMM) kullanarak sınıflandırma gerçekleştirmişlerdir (Salzberg vd., 1998). Farklı senaryolar ile gerçekleştirdikleri çalışmada en yüksek %97,80 doğruluk başarısı elde etmişlerdir. Statnikov ve arkadaşları çalışmalarında; kanser tiplerini kapsayan 11 veri setini sınıflandırmak için Çok Sınıflı Destek Vektör Makinesi (Çok

Sınıflı SVM), K-En Yakın Komşular (KNN), Geri Yayılımlı Sinir Ağları (GYSA) ve Olasılıksal Sinir Ağları (OSA) yöntemlerini kullanmışlardır (Statnikov vd., 2005). Elde edilen sonuçlarda; Çok Sınıflı SVM için %89,44, KNN için %77,16, GYSA için %67,73 ve OSA için %72,38 doğruluk başarısı elde etmişlerdir.

Belirli bir gen dizisinin tür veya alt tipine sınıflandırılması, virolojide önemli ve zorlu bir sorundur (Solis-Reyes vd., 2018). Bunun nedeni, aynı tür içindeki alt tiplerin genom dizilerinin bile, patogenezdeki farklılıklar, hastalık ilerleme oranları, ilaçlara ve aşıya duyarlılık gibi faktörler nedeniyle önemli ölçüde farklılık gösterebilmesidir. Örneğin, pandeminin erken dönemlerinden sonraki dönemlere kadar HIV-1 virüsünün alt tiplerinin genom dizileri arasında %15'e varan bir fark vardır. 2018 yılında Solis ve arkadaşları; HIV-1 virüsünün alt tiplerini içeren gen dizilerinin farklı K-Mer kelime uzunlukları ile çeşitli analizler gerçekleştirmişlerdir (Solis-Reyes vd., 2018). Toplam 15 sınıflandırma yöntemi ile yaptıkları çalışmalarda k=6 için en başarılı sonuçları elde etmişlerdir. Bu sonuçlar; Çok Sınıflı SVM' ler için %96,66, KNN için %93,97, Çok Katmanlı Algılayıcı (ÇKA) için %95,49 ve Lojistik Regresyon (LR) için %95,32 şeklindedir. Remita ve Diallo çalışmalarında; HCV virüsünün alt tiplerinden oluşan gen dizilerini sınıflandırmak için Multinomial Bayes, Markov, LR ve Doğrusal SVM yöntemleri kullanmışlardır (Remita & Diallo, 2019). Farklı K-Mer boyutları ile sonuçları değerlendirmişlerdir. Kullandıkları parametrelere bağlı olarak Doğrusal SVM için %100'e kadar doğruluk başarısı elde etmişlerdir. Diğer modeller de yüksek doğruluk başarısı elde etmişlerdir, ancak başarı oranları model türü ve K-Mer uzunluğuna bağlı olarak değişmektedir. Elde ettikleri sonuçlarda çoğu modelin %95'in üzerinde başarı gösterdiğini belirtmişlerdir. Cleydson ve arkadaşları; Geminiviridae virüs ailesinin alt tiplerini sınıflandırmak için SVM, Random Forest (RF) ve ÇKA MÖ yöntemlerini kullanmışlardır (Cleydson vd., 2017). En başarılı modelin. %96,6 ile RF olduğunu belirtmişlerdir.

Tabares ve arkadaşları çalışmalarında, gen ekspresyonu mikro dizi verilerinden oluşan tümör çeşitlerini sınıflandırmak için MÖ ve DÖ modellerini kullanarak, sonuçları karşılaştırmışlardır (Tabares-Soto vd., 2020). Kullandıkları modeller; Doğrusal Destek Vektör sınıflandırıcısı (SVC), LR, Doğrusal Diskriminant Analizi (LDA), Naive Bayes (NB), ÇKA ve Evrişimsel Sinir Ağlarıdır (CNN). Ayrıca RF gibi Karar Ağacı (DT) yöntemlerin de sonuçlarını analiz etmişlerdir. N-katlı çapraz doğrulama kullanarak elde ettikleri sonuçlarda MÖ modelleri arasında en yüksek başarı değerini LR için %90,60 ve CNN için %94,43 olarak tespit etmişlerdir.

DÖ teknikleri, 2010'ların ortalarından itibaren DNA dizilimlerinin sınıflandırılması için önemli hale gelmiştir. Ren ve arkadaşları çalışmalarında; DeepVirFinder adını verdikleri, geliştirdikleri bir CNN algoritmasını kullanarak viral gen dizisi tahmini yapmışlardır (Ren vd., 2019). Bu modeli, K-Mer kodlama yöntemi ile birleştirerek, Ulusal Biyoteknoloji Bilgi Merkezi'nden (NCBI) aldıkları prokaryotları (bakteriler ve arkeler) enfekte eden virüs genomlarını sınıflandırmak için kullanılmışlardır. Bu çalışmada %94'ün üzerinde doğruluk başarısı elde etmişlerdir. Benzer şekilde, Tampuu ve arkadaşları çalışmalarında ViraMiner adını verdikleri CNN tabanlı bir metodoloji geliştirerek çeşitli insan virom örneklerindeki potansiyel viral dizilimleri tanımlamayı amaçlamışlardır (Tampuu vd., 2019). ViraMiner mimarisi, Ren ve arkadaşlarının DeepVirFinder mimarisi üzerine inşa edilmiştir (Ren vd., 2019). Bu mimaride, ham gen dizilimleri Tek-Sıcak kodlama ile kodlanmaktadır. ViraMiner modeli ile test veri setinde 0.923'lik bir ROC eğrisi altındaki alan (AUROC) değeri ile yüksek bir performans elde etmişlerdir. Sukhorukov ve arkadaşları, şeftali, üzüm ve şeker pancarı için NCBI bitki virüsü verilerini sınıflandırmak amacıyla RF sınıflandırıcı ile birleştirdikleri CNN tabanlı bir modül sunmuşlardır (Sukhorukov vd., 2022). VirHunter olarak adlandırılan bu hibrit yöntemi, daha derin analiz için farklı K-Mer boyutları kullanarak geliştirmişlerdir. Modeli DeepVirFinder (Ren vd., 2019) ile karşılaştırıp sınıflandırma performansı açısından daha başarılı olduğunu belirtmişlerdir.

Zhang ve arkadaşları, Covid-19, AIDS, Influenza ve Hepatit C gibi viral hastalıklarla ilgili dört farklı veri kümesinde CNN, Derin Sinir Ağları (DSA) ve N-gram Olasılık Modelleri olmak üzere üç farklı algoritma kullanarak analizler yapmışlardır (X. Zhang vd., 2021). Ayrıca, Levenshtein mesafesine ve rastgele oluşturulan DNA alt dizilimlerine dayanan yeni bir özellik çıkarım yöntemi sunmuşlardır. Elde ettikleri sonuçlarda %86,82 ile %99,99 arasında değişen doğruluk oranları elde etmişlerdir. Özellikle, 3-gram özellik çıkarma yöntemiyle SVM algoritması kullanıldığında çoğu veri seti için en yüksek doğruluk başarısına ulaşmışlardır. Rincon ve arkadaşları (Lopez-Rincon vd., 2020) ise SARS-Covid-19 gen dizilimlerini CNN kullanarak sınıflandırmışlardır. 10 katmanlı çapraz ile %98,73 doğruluk oranı elde etmişlerdir.

Gunasekaran ve arkadaşları; DNA dizilerini sınıflandırmak için üç farklı DÖ mimarisi kullanmışlardır: CNN, CNN-LSTM (Evrişimli Sinir Ağı-Uzun Kısa Süreli Bellek) ve CNN-Çift Yönlü LSTM (Gunasekaran vd., 2021). DNA dizisinden özellik çıkarmak için mesafe ölçümüne dayalı yeni bir yaklaşım önermişlerdir. Bu modeli COVID-19, AIDS, İnfluenza ve Hepatit C virüslerini içeren bir veri setine

uygulamışlardır. Ek olarak, çalışmalarında Etiket kodlama ve K-Mer kodlama yöntemlerinin etkinliği karşılaştırmışlardır. k=6 değeri seçilerek K-Mer kodlama yöntemi ile gerçekleştirdikleri uygulamalarda; CNN ile %93,16, CNN-LSTM ile %93,02 ve CNN-Bidirectional LSTM sınıflandırma ile %93,13 doğruluk başarıları elde etmişlerdir. Dasari ve Bhukya, viral genom tahmini için CNN tabanlı EdeepVPP ve CNN-LSTM tabanlı EdeepVPP-Hibrit iki yaklaşım önermişlerdir (Dasari & Bhukya, 2022). On katlı çapraz doğrulama kullanarak insan metagenomik veri seti üzerinde analizler gerçekleştirmişlerdir. Analizler sonucunda AUC-ROC performans metriğine göre EdeepVPP modeli ile 98 başarı değerini ve EdeepVPP-Hibrit modeli için %99 başarı değerini elde ettiklerini bildirmişlerdir. Basu ve Campbell çalışmalarında; COVİD-19' un yirmi farklı varyantını sınıflandırmak için gen dizilerini K-Mer kodlama yöntemi ile kodladıktan sonra LSTM ile sınıflandırmışlardır (Basu & Campbell, 2023). Model başarısını %92,50 olarak tespit ettiklerini beyan etmişlerdir.

Poplin ve arkadaşları (Poplin vd., 2018) çalışmalarında, genotip verilerini kullanarak DeepVariant adını verdikleri CNN temelli bir yaklaşım önermişlerdir. Farklı memeli türlerine ait genom dizilerinden oluşan dört farklı veri seti üzerinde geliştirdikleri yaklaşımı değerlendirmişlerdir. Tüm veri setlerinde %90'ın üzerinde başarı oranı elde etmişlerdir.

Liang ve arkadaşları çalışmalarında (Liang vd., 2020); metagenomik verileri sınıflandırmak için DeepMicrobes ismini verdikleri DÖ yaklaşımını kullanmışlardır. DNA dizilerini dönüştürmek için Tek-Sıcak kodlama ve K-Mer yöntemlerini kullanmışlardır. K-Mer kodlama da k değerini 8 ile 12 aralığında seçmişlerdir. Sonuçlara göre özellikle k=12 için K-Mer kullanan modelin, diğer kelime uzunluklarından çok daha iyi performans gösterdiğini belirtmişlerdir. ResNet benzeri bir ağı, LSTM ağlarını ve Seq2species modellerini dahil edip gerçekleştirdikleri cins sınıflandırmasının değerlendirme sonucunu %96'nın üzerinde olduğunu bildirmişlerdir. Desai ve arkadaşları; insan bağırsak ve farklı ortamlardaki mikrobik ortamın anlaşılmasına yönelik çalışmaları için 16s rRNA gen dizilerini analiz edip sınıflandırmışlardır (Desai vd., 2020). Gen dizilerini sınıflandırmadan önce K-Mer yöntemini kullanarak kodlamışlardır. LSTM ve CNN hibrit modeller ile en yüksek %85 doğruluk başarısı elde etmişlerdir.

Bir virüsün konak organizmasını belirlemek, enfeksiyon kaynağını ve olası bulaşma yollarını anlamak için kritik öneme sahiptir. Bu bilgi, hastalığın nasıl tedavi edilmesi gerektiğini ve hangi önleyici tedbirlerin alınması gerektiği konusunda yönlendirmektedir. Yeni bir virüs tespit edildiğinde, hangi organizmalarda bulunduğunu

ve nasıl yayıldığını bilmek, salgınları kontrol etmek ve daha fazla yayılmayı önlemek için çok önemlidir. Ek olarak, virüslerin konakları, bu organizmaların evrimsel geçmişleri ve ekosistemlerdeki rolleri hakkında önemli bilgiler sağlamaktadır. Zhang ve arkadaşları çalışmalarında, NCBI viral genom veri tabanından 1.426 viral genom dizisini analiz edip ve konak tabanlı bir sınıflandırma uygulamışlardır (M. Zhang vd., 2017). Çalışmalarında, LR, SVM, Gaussian NB ve Bernoulli NB gibi MÖ yöntemleri, K-Mer kodlama yönteminde k'nın 1'den 8'e kadar olan değerlerini uygulayarak analiz etmişlerdir. k'nın farklı değerleri için farklı sonuçlar elde etmişlerdir. En başarılı model olarak tespit ettikleri RF için doğruluk başarısının %85 ile %100 arasında değiştiğini ifade etmişlerdir. Li ve Sun, gen dizilerinden Kuduz virüsü, Coronavirus ve Influenza A virüsünün konaklarını tahmin etmek için hizalama tabanlı yöntemlerin ve SVM yönteminin analiz sonuçlarını araştırmışlardır (Li & Sun, 2018). Hizalama tabanlı yöntemlerin başarılı sonuçlar gösterdiği, SVM'nin ise bazı durumlarda geride kaldığını ifade etmişlerdir. Influenza A için tüm yöntemlerin %60'ın üzerinde doğruluk oranı sağladığını, Coronavirus için %90'ın üzerine ulaştığını, Kuduz virüsü için ise ; hizalama tabanlı yöntemlerin daha yüksek başarı gösterdiğini belirtmişlerdir. Mock ve arkadaşları, çalışmalarında DSA ağlarını kullanarak üç farklı virüs türü (Influenza A virüsü, Kuduz Lyssa virüsü ve Rotavirüs A) için konak tahminini gerçekleştirmişlerdir (Mock vd., 2020). Geliştirdikleri DÖ modeli olan VIDHOP kullanarak gerçekleştirdikleri sınıflandırma uygulamasında oldukça yüksek doğruluk oranları elde etmişlerdir. Veri setleri üzerinde yapılan testlerde, modelin AUROC başarı değerleri %93 ile %98 arasında değişmiştir.

B) Gen dizisi kodlama yöntemleri alanındaki literatür çalışması

Gen dizisi kodlama yöntemleri, biyoinformatik ve genetik araştırmaların temel taşlarından biridir. Bu yöntemler, genetik verilerin daha verimli analiz edilmesine, sınıflandırılmasına ve görselleştirilmesine olanak tanımaktadır. Kodlama teknikleri, genetik bilginin çeşitli temsil biçimleriyle işlenmesini sağlamakta ve farklı analiz ihtiyaçlarına yanıt vermektedir. Gen dizisi kodlama yöntemlerinden sayısal gösterim yöntemleri; DNA dizilerinin bilgisayar destekli modellerle işlenebilir hale getirilmesi amacıyla kullanılmaktadır. Literatürde bu yöntemlerin, özellikle yüksek verimli genetik veri analizi için uygun olduğu belirtilmiştir. Grafiksel ve Resim gösterim yöntemleri ise giriş dizisinin çeşitli özelliklerini kodlayabilmekte, giriş dizisi uzunluğundan bağımsız olarak sabit boyutlu çıktı görüntüsü üretebilmekte ve her biyolojik diziden farklı bir imza üretebilmektedir. Sayısal gösterim yöntemiyle alakalı literatür özeti; bölüm 3.A 'da

anlatılmıştır. O yüzden bu bölümde grafiksel ve resim gösterim yöntemlerinin literatürdeki çalışmaları üzerinde durulacaktır.

Grafiksel gen dizisi kodlama yöntemlerinden olan Chaos game representation (CGR); genetik dizilerin belirli özelliklerini vurgulamak için kullanılan verilerin grafiksel dönüşümünü gerçekleştiren bir tekniktir. Bu yöntem ilk defa 1990'ların başında tanıtılmıştır (Jeffrey, 1990). Bu teknik devrim niteliğinde olmuştur. Çünkü kaos teorisi ve genetik kavramlarını bir araya getirilerek genetik dizilerin devasa, karmaşık yapıları, iki boyutlu bir uzayda temsil edilerek, benzersiz içerikler oluşturmaktadır. Ayrıca bu yöntem, diziler arasındaki benzerlikleri ve farklılıkları belirlemek için kullanılabilmekte, evrimsel ilişkileri veya işlevsel benzerlikleri tespit ederek bu alanda ilerlemelerin gerçekleştirilmesine olanak sağlamaktadır. CGR ile ilgili ilk uygulamalar; farklı organizmaların DNA dizileri arasındaki benzerlik ve farklılıklarını belirlemek için kullanıldığı genomların karşılaştırmalı analizi şeklindedir (Almeida vd., 2001; Deschavanne vd., 1999; Joseph & Sasikumar, 2006). Literatür incelendiğinde ilerleyen zamanlarda CGR yöntemi ile beraber başka yöntemlerin de entegre olarak kullanıldığı görülmüştür. Hoang ve arkadaşları çalışmalarında; İnsan Rinovirüsü, Grip ve HPV genomlarını içeren veri kümelerine CGR yöntemini uygulayarak DNA dizilerini dönüştürdükten sonra bu dizilerin koordinat bilgilerini kullanarak karmaşık sayılar şeklinde ifade etmişlerdir (Hoang et al., 2016). Daha sonra Dijital Sinyal İşleme alanında en çok kullanılan yöntemlerden biri olan Ayrık Fourier Dönüşümünü uygulayıp, her bir gen dizisi için ayrı ayrı güç spektrumu elde etmişlerdir. Fakat uygulamada farklı uzunluklara sahip gen dizilerinden oluşan bir veri seti ile çalıştıklarından dolayı, güç spektrumlarının uzunlukları da birbirinden farklı olmaktadır. Benzerlik karşılaştırmalarında bu durum bir sorun teşkil ettiğinden, farklı bir uzunluk eşitleme yöntemi önermişlerdir. Önerdikleri Eşit Ölçeklendirme Yöntemi için; veri setindeki en uzun dizi uzunluğunu belirledikten sonra, doğrusal enterpolasyon gerçekleştirerek tüm verilerin maksimum dizi uzunluğuna eşitlemesini gerçekleştirmişlerdir. Souza ve arkadaşları ise; SARS-CoV-2 Virüsüne ait Coronaviridae ailesinden altı viral türün örneklerini inceleyerek; Hoang ve arkadaşlarının gerçekleştirdiği çalışma (Hoang vd., 2016) ile aynı yolu izlemişlerdir (Souza vd., 2023). Fakat farklı güç spektrumlarındaki uzunluk sorunu için 64, 128 ve 256 boyutlarına ölçekledikleri farklı bir yöntem önermişlerdir. 64 boyutunda bir eşitleme için; dizi içerisindeki 64 adet en büyük değerli veri indislerini kaydedip geri kalan değerleri silmişlerdir. Diğer uzunluklar için de aynı

işlemi tekrarlamışlardır. Ardından bu indis değerlerine göre sırayla yerleştirerek bir boyut eşitleme gerçekleştirmişlerdir.

CGR yöntemi, genetik verilerin görselleştirilmesi ve analiz edilmesinde sunduğu benzersiz perspektiflerle birçok çalışmada önemli bir rol oynamıştır. Ancak, bu alandaki gelişmeler sadece CGR ile sınırlı kalmamış, daha kompleks gen dizisi kodlama ve analiz ihtiyaçlarına yanıt vermek üzere farklı yöntemlerin geliştirilmesine de yol açmıştır. Bu bağlamda, Frequency chaos game representation (FCGR) yöntemi, CGR' nin temel prensiplerini genişleterek genetik dizilerin daha detaylı ve yüksek doğrulukta analiz edilmesine imkân tanıyan bir diğer önemli yaklaşımdır. FCGR, genetik verilerin görsel olarak temsil edilmesinde daha yüksek çözünürlük ve daha fazla bilgi yoğunluğu sağlamasıyla dikkat çekmektedir. Literatürde, FCGR yöntemiyle elde edilen verilerin, özellikle DÖ algoritmalarıyla entegre edildiğinde, gen dizilerinin sınıflandırılmasında yüksek doğruluk oranlarına ulaştığı bildirilmektedir. Örneğin, Löchel ve arkadaşları; protein dizilerini görüntülere kodlamak için FCGR yöntemini kullanmışlardır (Löchel & Heider, 2021). Sınıflandırma için; DSA, SVM ve RF yöntemlerini kullanmışlardır. K-Mer kelime uzunlukları ile elde edilen FCGR görüntülerinin farklı k değerlerinde DSA yöntemi ile yüksek performans göstermiştir. Kullandıkları 17 farklı veri setinde, DSA için tüm sonuçların %90'ın üzerinde olduğunu bildirmişlerdir. Bazı veri setlerinde %99,5'a kadar yüksek doğruluk başarısı elde etmişlerdir.

Zhao ve arkadaşları; çocukların ağız örneklerinden alınan 16S rRNA gen dizilerinden çürük tespiti yapmak için FCGR uyguladıktan sonra CNN ve VGG16 ile sınıflandırmışlardır (C. Zhao vd., 2023). CNN ile %83'lük bir doğruluk ve VGG16 modeli ile de %87,5'lik bir doğruluk başarısı elde etmişlerdir. FCGR için k'nın 5, 6, 7 ve 8 değerlerini kullanmışlardır. k=8 de en yüksek doğruluk başarısını elde etmişlerdir fakat, artan k değerlerinde aşırı uyum gözlemlemişlerdir.

Cartes ve arkadaşları, 190.000'den fazla SARS-CoV-2 genom dizileri içeren büyük bir veri kümesi üzerinde çalışmışlardır (Cartes vd., 2022). FCGR yöntemi ile kodlama gerçekleştirdikten sonra, ResNet50 yöntemi ile sınıflandırmışlardır. k' nın 6, 7 ve 8 değerleri ile testlerini gerçekleştirmişlerdir. %96,22 genel doğruluk başarısı elde etmişlerdir.

Hammad ve arkadaşları; çeşitli insan Coronavirus gen dizilerini sınıflandırmak için hibrit bir DÖ yaklaşımı sunmuşlardır (Hammad vd., 2023). FCGR ile beraber derin özellik çıkarımı için AlexNet modelini kullanmışlardır. Daha sonra, ReliefF ve En Küçük Mutlak Shrinkage ile Seçim Operatörü (LASSO) algoritmalarını kullanarak özellik

seçimi gerçekleştirmişlerdir. En son olarak da DT ve KNN aracılığıyla sınıflandırma gerçekleştirmişlerdir. Önerdikleri hibrit DÖ yaklaşımı ile en yüksek doğruluk başarısını %99,71 olarak elde etmişlerdir.

Rizzo ve arkadaşları çalışmalarında; LeNet-5 mimarisinin değiştirilmiş bir versiyonunu kullanarak DNA dizisi sınıflandırmasına yönelik DÖ'ye dayalı bir yaklaşım sunmuşlardır (Rizzo vd., 2016). Bakterilere ait 3000 adet 16S ribosomal RNA gen dizisi içeren bir veri seti ile önerdikleri metodu uygulamışlardır. k'nın 5,6 ve 7 değerleri için FCGR görüntülerini oluşturmuşlardır. CNN ve SVM yöntemleri ile sınıflandırdıkları görüntülerdeki doğruluk başarıları sırasıyla; k=5 için %99,2, %99,1'dir.

FCGR'nin sağladığı bu detaylı analiz imkânı, genetik dizilerin farklı temsillerini ve grafiksel gösterim yöntemlerini geliştirme ihtiyacını da beraberinde getirmiştir. Bu noktada, gen dizilerinin başka bir görselleştirme yöntemi olan DNAWalk devreye girmektedir. DNAWalk yöntemi, genetik dizilerin grafiksel temsilinde farklı bir yaklaşım sunarak, dizilerin yapısal özelliklerini üç boyutlu bir uzayda haritalamaktadır.

İlk olarak Hamori ve Ruskin tarafından 1983 yılında tanıtılan bu teknik, nükleotid dizilimlerini birim vektörlerle temsil ederek gen dizilerinin uzaysal özelliklerini vurgulamaktadır (Hamori & Ruskin J., 1983). DNAWalk yönteminin bu çok boyutlu temsili, FCGR gibi yöntemlerle karşılaştırıldığında, genetik dizilerin daha karmaşık ve çok boyutlu ilişkilerini ortaya koyma potansiyeli sunmaktadır. Bu yöntem, özellikle DNA dizileri üzerindeki yerel ve genel nitelikleri keşfetmek için güçlü bir araç olarak kabul edilmiştir.

Gates'in gerçekleştirdiği bir çalışmada; dört nükleotidin her biri iki boyutlu bir uzayda (x,y) vektörlerle temsil edilmiştir (Gates, 1985). Bu iki boyutlu grafiksel DNA dizisi gösterimleri, DNA dizileri üzerinden doğrudan fark edilemeyen yerel ve genel nitelikler üzerine değerli bilgiler sunmaktadır. Gates'in grafik gösterimini geliştirmek için literatürde başka çalışmalar da gerçekleştirilmiştir (Leong & Morgenthaler, 1995; Nandy A., 1995). Fakat, bu gösterimlerde, kendileriyle örtüşme sonucu bozulma meydana gelmekte ve bu da bazı bilgilerin kaybolmasına yol açarak dejenerasyona sebep olmaktadır. Bu sebeple; DNA dizilerinin daha düşük dejenerasyonu için Guo ve arkadaşları çalışmalarında; dizileri ardışık bir vektör dizisi şeklinde göstermişlerdir (Guo vd., 2001).

Literatürde bazı araştırmacılar ise dinükleotidlerin daha fazla biyolojik bilgi bulundurması mantığından yola çıkarak, nükleotid çiftlerinin grafiksel gösterimine dayalı farklı yöntemler önerilmişlerdir (Qi & Fan, 2007; Wu vd., 2003; Zhang, 2009). Bu

yöntemlerde gen dizileri boyunca nükleotid çiftlerinin dağılımlarını temsil eden dinükleotid eğrileri çizilmektedir.

Literatürdeki grafiksel gösterimler ayrıca, DNA dizilerinin iki boyutlu olarak (Guo vd., 2001; Huang vd., 2008; X. Q. Liu vd., 2006), üç boyutlu (Cao vd., 2008; Qi vd., 2007), dört boyutlu (Chi & Ding, 2005), beş boyutlu (Liao vd., 2007), altı boyutlu (Liao & Wang, 2004) ve sekiz boyutlu (D. Zhang, 2019) gösterimleri ile de gerçekleştirilmiştir.

Literatürdeki bu çalışmaların tümü, herhangi bir sınıflandırma probleminde değil, dizi yapılarındaki benzerliklerin belirlenmesi için gerçekleştirilmiş uygulamalardır.

Kobori ve Mizuta çalışmalarında; DNA dizilerinin yörüngesinin, nükleotidler için önceden tanımlanmış vektörler ve ağırlık faktörleri ile iki boyutlu bir düzlemde çizildiği bir grafiksel gösterim sunmuşlardır. DNA dizileri arasındaki benzerlikleri tahmin etmek için nükleotidleri iki boyutlu vektörlerle değiştirip bunları ardı ardına bağlayarak DNA dizilerini ikili görüntüler olarak ifade etmişlerdir (Kobori & Mizuta, 2016). İkili görüntüler üzerinde 3x3 bitmap desenlerinin oluşma sıklıklarını hesaplayarak ve bitmap desenlerinin frekans histogramlarına dayanarak aralarındaki mesafe ölçümlerini gerçekleştirmişlerdir. Bu mesafe ölçümlerine göre 31 memeli türüne ait mitokondriyal genomları karşılaştırmışlardır.

Hossain ve arkadaşları çalışmalarında; bir DNA yörüngesinin belirli bir kısmını tekrar ziyaret etme sayısıyla ilişkili bilgileri dikkate alarak gri tonlamalı görüntüler oluşturmuşlardır (Hossain vd., 2021). Bu görüntülerde, dokunulan karelerin piksel değeri, yörüngenin kendisi tarafından yeniden ziyaret edilme sayısı olmuştur. Normal şartlar altında sadece siyah ve beyazdan oluşması gereken resimler bu yöntemde gri tonlamalı olarak gösterilmiştir. Ardından DNA benzerlik tespiti için Yerel İkili Desen (LBP) metodunu kullanıp çeşitli histogram bilgilerini elde etmişlerdir.

DNAWalk yöntemi, genetik dizilerin grafiksel temsilinde nükleotidlerin uzaysal dağılımını vurgulayan etkili bir yaklaşımdır. Ancak genetik verilerin temsilinde farklı yöntemler de geliştirilmiş olup, bu yöntemler genetik dizilerin başka açılardan incelenmesine olanak tanımaktadır. Bunlardan biri de DNA dizilerinin Gri Ölçekli görüntülere dönüştürülmesidir. Bu yaklaşım, nükleotid veya dinükleotid dizilerinin her birine belirli gri seviyeleri atayarak, genetik verilerin görsel bir temsilini sunmaktadır. Bu şekilde elde edilen Gri Ölçekli görüntüler, biyolojik dizilerin özelliklerini görsel olarak analiz etmeye ve sınıflandırma işlemlerine yeni bir yaklaşım sunmaktadır. Santamaria ve arkadaşları gerçekleştirdikleri çalışmalarında; farklı veri setlerindeki gen dizilerindeki

DNA nükleotid çeşitlerinin her birine bir gri değer ataması yaparak CNN modeli olan InceptionV3 ile sınıflandırmışlardır (Santamaría vd., 2019). Elde edilen başarı yüzdesi en yüksek %80,8'dir.

Delibas ve arkadaşları çalışmalarında (Delibas & Arslan, 2020); gri seviyeli görüntüler oluştururken dinükleotidlerin her birine sırasıyla 1-255 arasındaki gri seviye değeri arasında değerler atayarak bir görselleştirme tekniği önermişlerdir. Yazarlar, bu çalışmada bir sınıflandırma işlemi değil, filogenetik ağaç oluşturmak için DNA benzerlik tespiti gerçekleştirmişlerdir. Bu sebeple görüntülerin histogramlarını kullanmışlardır.

C) Gen dizilerinde eksik veri atama yöntemleri ile ilgili çalışmalar;

Veri setlerindeki eksik değerlerin giderilmesi için bazı araçlara ihtiyaç duyan araştırmacılarla birlikte, literatürde silmeden atamaya kadar çok sayıda teknik önerilmiştir (Yadav & Roychoudhury, 2018).

Eksik veri atama tekniklerinde, seçilen veri setinde mevcut bilgiler kullanılarak, karşılık gelen eksik değerleri atamak için uygun değerler ortalama, mod, medyan veya bazı klasik ve basit istatistiksel tahmin modelleri, karmaşık yöntemlerin yerine kullanılıp atamaları gerçekleştirilmektedir. Pigott, 2001 yılında gerçekleştirdiği çalışmasında; eksik verileri atamak için mod yöntemini kullanmıştır (Pigott, 2001). Ancak, Graham 2009 yılındaki çalışmasında, mod kullanılarak yapılan eksik veri atamanın, veri setindeki varyasyonu azaltabileceğini ve bu nedenle bazı durumlarda verinin doğal dağılımını bozabileceğini belirtmiştir (Graham, 2009). Benzer biyolojik fonksiyona sahip genlerin benzer biyolojik profile sahip olduğu olgusunun ardından, KNN-Impute algoritması önerilmiştir (Troyanskaya vd., 2001). Troyanskaya ve arkadaşları çalışmalarında, mayalardan oluşan gen dizi özelliklerine ait üç veri seti ile analizlerini gerçekleştirmişlerdir. Bu veri setleri zaman serileri, gürültülü zaman serileri ve zaman serisi olmayan diziler içermekte ve eksik verilerin oranı %1 ile %20 arasında değişen değerlerden oluşmaktadır. KNN-Imputation, SVDimpute ve satır ortalamaları yöntemleri veri setlerine uygulanmıştır. Yöntemlerin doğruluğu, Kök Ortalama Kare Hatası (RMSE) kullanılarak değerlendirilmiştir. KNN-Imputation yönteminin tahmin ettiği değerlerin orijinal değerlere ortalama sapması %6 ila %26 arasında değişmiştir.

Bras ve Menezes çalışmalarında Yinelemeli KNN-Impute yöntemini önermişlerdir (Brás & Menezes, 2007). Kullandıkları dört veri seti; zaman serileri, zaman serileri olmayan diziler ve karışık verilerden oluşmaktadır. Veri setlerindeki eksik verilerin oranı; <%5, %5–10, %10–20, %20–50 ve >% 50 şeklindedir. Ayrıca k'nın 5, 10, 15, 20 değerleri için analizler gerçekleştirmişlerdir. Zhang ve arkadaşları,

çalışmalarında Sıralı KNN-Impute yöntemini önermişlerdir (X. Zhang vd., 2008). Kullandıkları dört veri seti, gen ekspresyon verilerinden oluşmakta ve zaman serileri ile zaman serileri olmayan dizilerden oluşmaktadır. Veri setlerindeki eksik verilerin oranı ise %3,7, %3,3, %3,8 ve %3'tür. Ayrıca, eksik veriler için kullandıkları diğer yöntemler BPCAimpute, LLSimpute, KNN-Impute'dir. Keerin ve arkadaşları ise Küme tabanlı KNN-Impute (CKNN) yöntemini önermişlerdir (Keerin vd., 2012). Altı farklı, data matrix gen verileri ile çalışmışlardır. Veri setlerindeki eksik verilerin oranı %1, %2, %3, %4, %5, %10 ve %15'tir. Önerdikleri yöntemde Davies-Bouldin mesafe ölçütünü kullanmışlardır.

Oba ve arkadaşları, çalışmalarında %1, %2, %5, %10 ve %20 oranında eksik veri içeren beş farklı veri setinin eksik veri tahmini için Bayes Ana Bileşen Analizi (BPCA) yöntemini önermişlerdir (Oba vd., 2003). BPCA ve LLSImpute yöntemleri, KNN'ye oranla daha az verimlidir ve büyük veri kümelerinde uygulanmaları daha zor olabilmektedir. Zhang ve arkadaşları, çalışmalarında %3,7, %3,3, %3,8 ve %3 oranında eksik veri içeren dört farklı veri seti üzerinde çalışmışlardır. Bu veri setleri, gen ekspresyon verileridir ve zaman serileri ile zaman serileri olmayan dizilerden oluşmaktadır (X. Zhang vd., 2008). Eksik verileri atamak için BPCA, KNN-Impute, SKNNImpute, SLLSimpute ve LLSImpute yöntemlerini kullanmışlardır.

Stekhoven ve Bühlman; sürekli veriler, yalnızca kategorik veriler veriler ve karışık tip verilerden oluşan üç farklı veri seti üzerinde çalışmışlardır (Stekhoven & Bühlmann, 2012). Bu veri setlerindeki eksik veri oranları %10, %20 ve %30'dur. Çalışmalarında, karışık tiplerdeki veri setlerinde uygulanabilen bir yöntem olan MissForest'i önermişlerdir.

Saha ve arkadaşları, çalışmalarında gen ekspresyon verilerinden oluşan iki farklı veri seti üzerinde tahmin gerçekleştirmişlerdir. YSA tabanlı yeni bir eksik veri tahmin yöntemi (ANNImpute) önermişlerdir (Saha vd., 2017). Önerilen yöntem SVDImpute ve LLSImpute gibi algoritmalar ile karşılaştırılmıştır. Deneysel sonuçlarda, ANNImpute'un SVDImpute ve LLSImpute yöntemlerden daha iyi performans gösterdiğini belirtmişlerdir.

MÖ ve DÖ alanlarındaki son gelişmelerle birlikte, atama performansını iyileştiren yeni yöntemler geliştirilmiştir. DÖ tabanlı yöntemler, verinin karmaşık ve doğrusal olmayan yapısını modelleyebilme yetenekleri sayesinde eksik veri atamada giderek daha fazla tercih edilmektedir. Yoon ve arkadaşları (Yoon vd., 2018) çalışmalarında; GAN (Generative Adversarial Nets) (Goodfellow vd., 2014) yapısını uyarlayarak eksik verileri

tamamlamak adına bir yöntem önermişlerdir. Bu yönteme Generative Adversarial Imputation Network (GAIN) adını vermişlerdir. Bu yöntemi gen dizilerinden oluşan veri setlerinde uygulamamışlardır. Fakat; GAIN ve türevleri literatürde yaygınlaşarak genetik verilerde de kullanılmaya başlamıştır. Vinas ve arkadaşları çalışmalarında (Viñas vd., 2020) gen ekspresyon verilerindeki eksik verileri atamak için GAIN yöntemine dayalı bir GAIN-GTEx yaklaşımı önermişlerdir. Modeli analiz etmek için üç farklı kanserden elde edilmiş RNA-Seq verileri üzerinde eğitim gerçekleştirmişlerdir. Veri seti 15.201 RNA-Seq verisini içermektedir. Model, Medyan atama ve MisForest yöntemleri ile karşılaştırılmıştır. GAIN-GTEx'in atama performansı ve çalışma süresi açısından onlardan önemli ölçüde daha iyi performans gösterdiğini belirtmişlerdir. Zhao ve arkadaşları ise kanser hastalarına ait, %50 ve %60 oranında eksik veri içeren iki farklı veri seti üzerinde çalışmışlardır (S. Zhao, 2023). Veri setleri toplam 54.676 gen dizisinden oluşmaktadır. Eksik verilerin tahmini için GAIN ve ClueGAIN yöntemlerini kullanmışlardır. Hazra ve arkadaşları çalışmalarında, 8 kediye ait gen dizilerinden oluşan veri setindeki eksik gen dizilerini TGAN, CTGAN, TGAN-skip, AN-WGAN-GP, WGAN-GP, TGAN-skip-WGAN-GP yöntemleri ile tahmin etmişlerdir (Hazra vd., 2022).

Qiu ve arkadaşları çalışmalarında, DNA metilasyon verilerinden oluşan ve %15 oranında eksik veri içeren iki farklı veri seti üzerinde çalışmışlardır. Araştırmacılar tahmin için Varyasyonel Otomatik Kodlayıcı (VAE) yöntemini kullanmışlardır. Modeli değerlendirmek için KNN-Impute ve SVDImpute yöntemleri ile karşılaştırma yapmışlardır. Elde edilen sonuçlarda VAE test edilen tüm eksik senaryolarda KNN' den daha iyi RMSE' lere ulaşmakta ve çoğu senaryoda SVD'ye benzer veya daha iyi performanslara ulaşmıştır (Qiu vd., 2020).

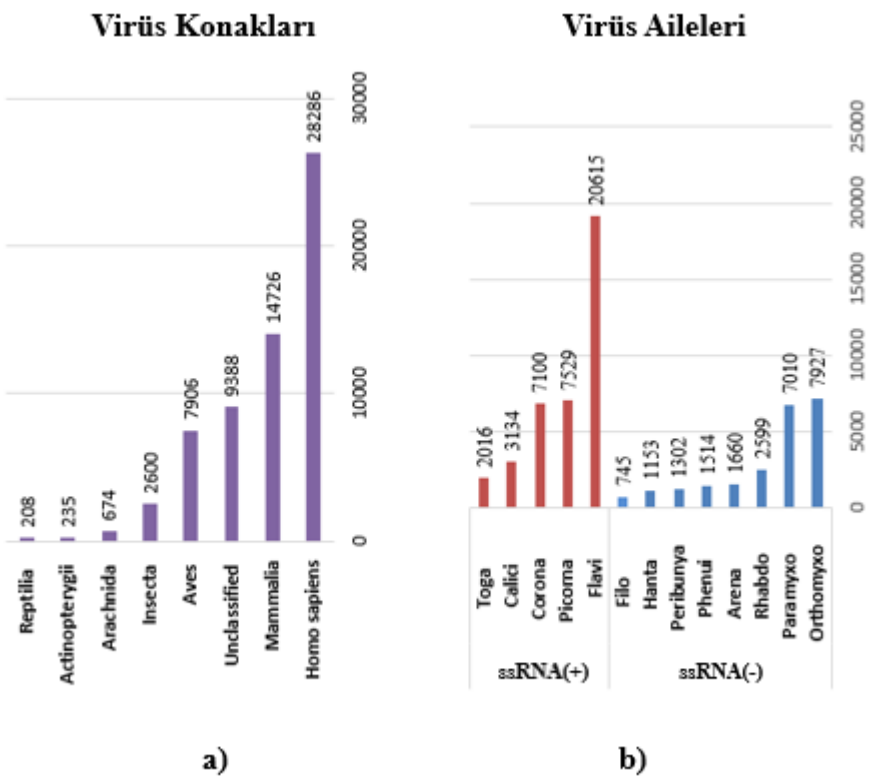
Gürültü Giderici Otomatik Kodlayıcı (SCDA), sağlam gizli özellikler oluşturmak için eksik değerlere sahip bozuk giriş verilerini kullanan bir tür Otomatik Kodlayıcıdır (Chen & Shi 2019). Bozuk giriş veya eksik veri değerlerini içerdiğinde, SCDA'lar, giriş verilerini yeniden yapılandırarak eksik veri noktalarının değerlerini tahmin edebilmektedirler (Vincent vd., 2008). Chen ve Shi (Chen & Shi 2019) çalışmalarında SCDA ile eksik verilerin tahminini gerçekleştirmişlerdir. SCDA modelini, genotip verilerinde çeşitli korelasyon veya bağlantı modellerini çıkarabilen bir evrişim katmanı kullanarak ve elde edilen bir ağırlık matrisi uygulayarak oluşturmuşlardır (Chen & Shi, 2019). Önerilen yöntemin analizi için bir maya ve insan genotip veri setini kullanmışlardır. SCDA yönteminin farklı kayıp veri senaryolarındaki performansını

değerlendirmek için, veri setindeki verilerin %5, %10 ve %20'sini rastgele sıfıra maskeleyerek üç set sentetik veri seti oluşturmuşlardır. Bu veri kümelerinin her biri için, verileri eğitim, doğrulama ve test için sırasıyla %65, %15 ve %20 veri içeren üç ayrı veri kümesine bölmüşlerdir. Ayrıca modeli karşılaştırmak adına satır ortalaması, KNN-Impute, SVDImpute gibi yöntemler uygulamışlardır. Sonuçlara göre SCDA modelinin diğer üç yönteme göre daha başarılı doğruluk sonucu verdiğini gözlemlemişlerdir.

4. MATERYAL: PHYVIRUS VERİ SETİ

Bu çalışmada kullanılan PhyVirus veri seti, Baltimore sınıfları dört ve beş (+ssRNA ve -ssRNA virüsleri) olan patojenik tek sarmallı RNA virüslerinden oluşmaktadır (PhyVirus | adi-stern, 2021). Kustin ve Stern çalışmalarında; RNA virüslerinde ek ortak genomik ve evrimsel özellikler olup olmadığını test etmek için geniş bir veri seti oluşturmuşlardır (Kustin & Stern, 2021). Bu veri seti için diziler Ulusal Alerji ve Enfeksiyon Hastalıkları Enstitüsü (NIAID) Virüs Patojen Veri Tabanı ve Analiz Kaynağından (VIPR) elde edilmiştir (Pickett vd., 2012). Ardından NIAID Grip Araştırma Veritabanından (IRD) grip dizileri ile genişletilmiştir (Y. Zhang vd., 2017). Konak bilgisi VIPR ve IRD'den alınmıştır. Virüs ailesi; +ssRNA ve -ssRNA olarak ayrılmaktadır ve replikasyon için konakçıya ihtiyaç duymaktadır. Yapısal olarak farkları bulunmaktadır. Veri setinde çift sarmallı virüslerin kafa karıştırıcı etkilerinden kaçınmak için tek sarmallı olan RNA virüslerini tercih etmişlerdir. PhyVirus veri setinde +ssRNA' ya ait beş, -ssRNA' ya ait sekiz virüs türü yer almaktadır. Bu virüslerin konakçıları yedi sınıfa ayrılmış olup, konağı belirlenmemiş olanlar ise "sınıflandırılmamış" şeklinde etiketlenmiştir. Veri setinde toplam 64.034 virüs dizisi bulunmaktadır. Veri setinde yer alan +ssRNA ailesine ait olan virüsler; Calici (3.134), Corona (7.100), Flavi (2.0615), Picorna (7.529), Toga (2.016) şeklindedir. -ssRNA ailesine ait olan virüsler; Arena (1.660), Filo (745), Hanta (1.153), Orthomyxo (7.927), Paramyxo (7.010), Peribunya (1.302), Phenui (1.514), Rhabdo (2.599) şeklindedir. Veri setindeki konak canlılar ise; insanlar/homosapiens (28.286), memeliler/mammalia (14.726), kuşlar/aves (7.906), böcekler/insecta (2.600), balıklar/actinopterygii (235), araknitler/arachnida (674), sürüngenler/reptilia (208) ve sınıflandırılmayanlar/unclassified (9.388) şeklindedir. Ayrıca, PhyVirus veri setindeki virüs gen dizileri oldukça geniş bir uzunluk aralığından oluşmaktadır. Gen dizileri en küçüğü 42, en büyüğü ise 13.176 nükleotid uzunluğundadır.

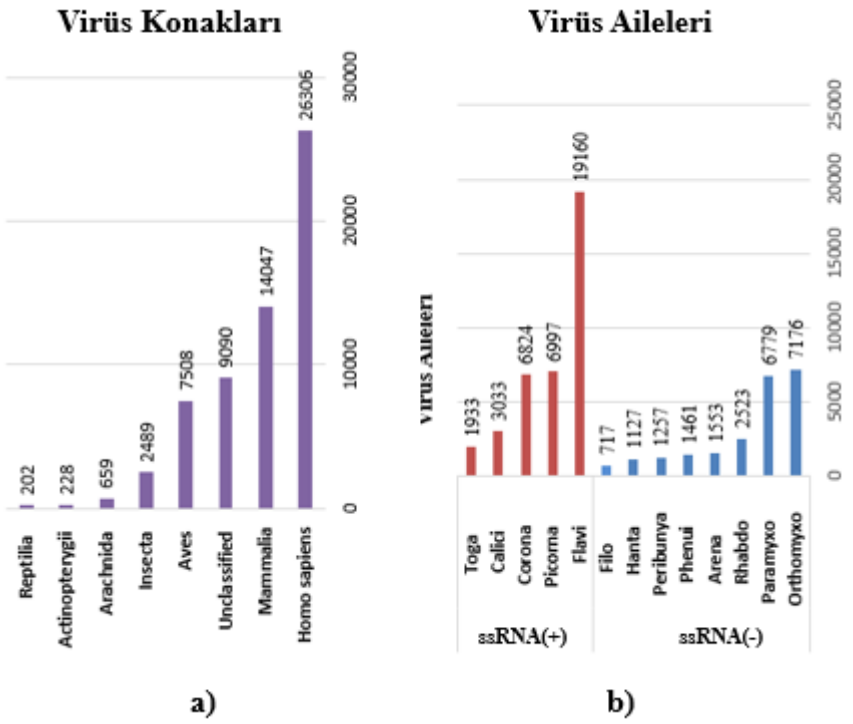
Şekil 4.1'de PhyVirus veri setinde, virüs konaklarının ve virüs ailelerinin sayısal dağılımı görülmektedir.



Şekil 4.1. PhyVirus veri setinde; **(a)** virüs konaklarının sayısal dağılımı, **(b)** virüs ailelerinin sayısal dağılımı

Gen dizi görüntüleri tarandıktan sonra, kayma çizikleri, lekelenme sorunları, çipteki kusurlar, hibridizasyon hatası, görüntü bozulması veya basitçe slayt üzerindeki toz gibi sorunlar sebebiyle kayıp değerler oluşabilmektedir (Oba vd., 2003). PhyVirus veri setinde de içerisinde hangi nükleotidin olduğu bilinemeyen bazı virüs dizileri yer almaktadır. Bu değerler, veri setinde "N" olarak işaretlenmiştir. Toplamda 64.034 virüs dizisinden 3.494 adet dizide kayıp veriye rastlanmıştır. Bu çalışmada 10. bölümde yer alan eksik veri atama uygulaması hariç diğer iki çalışma uygulamasında; kayıp değerlerin yer aldığı virüs dizileri çıkarılarak analizler gerçekleştirilmiştir. PhyVirus veri setinde, kayıp veriler içeren virüs gen dizileri çıkarıldıktan sonra veri setinde yer alan +ssRNA ailesine ait olan virüslerin sayısı; Calici (3.033), Corona (6.824), Flavi (19.160), Picorna (6.997), Toga (1.933) şeklindedir. -ssRNA ailesine ait olan virüslerin sayısı; Arena (1.553), Filo (717), Hanta (1.127), Orthomyxo (7.176), Paramyxo (6.779), Peribunya (1.257), Phenui (1.461), Rhabdo (2.523) şeklindedir. Veri setindeki konak sayıları ise; insanlar (26.306), memeliler (14.047), kuşlar (7.508), böcekler (2.489), balıklar (228), araknitler (659), sürüngenler (202) ve sınıflandırılmayanlar (9.090) şeklindedir. Şekil

4.2’de, PhyVirus veri setinde, kayıp veriler içeren virüs dizileri çıkarıldıktan sonra virüs konaklarının ve virüs ailelerinin sayısal dağılımı görülmektedir.



Şekil 4.2. PhyVirus veri setinde, kayıp veriler içeren virüs dizilimleri çıkarıldıktan sonra; **(a)** virüs konaklarının sayısal dağılımı, **(b)** virüs ailesinin sayısal dağılımı

5. DNA DİZİLERİNİN SINIFLANDIRILMASINDA KULLANILAN ÇEŞİTLİ YÖNTEMLER

DNA dizi sınıflandırması, biyolojik verilerin analizi ve yorumlanmasında kritik bir rol oynamaktadır (Ao vd., 2022). Bu süreç, genomik araştırmalarda yeni türlerin tanımlanması, evrimsel ilişkilerin belirlenmesi ve hastalıkların genetik temellerinin anlaşılması gibi birçok alanda çok önemli katkılar sağlamaktadır. Son yıllarda, DNA dizileme teknolojilerinde kaydedilen ilerlemeler sayesinde, genetik verilerin miktarında muazzam bir artış olmuştur. Yeni Nesil Dizileme yöntemleri, daha hızlı ve ekonomik bir şekilde büyük ölçekli genomik verilerin elde edilmesini sağlamaktadır. Birçok yeni virüs, bakteri veya genomik dizilimin bulunması ile DNA dizi veri tabanı hala gelişmektedir ve birçok çözülmemiş probleme katkıda bulunmaktadır (Soliman vd., 2022). Bu durum, biyolojik veri tabanlarının hızla genişlemesine yol açmış ve bu verilerin verimli bir şekilde sınıflandırılmasının gerekliliğini ortaya çıkarmıştır. Bunun yanı sıra, salgın hastalıkların kaynaklarının belirlenmesi ve kontrol altına alınması alanlarında önemli bir katkı sağlamaktadır. Örneğin, COVID-19 pandemisi sırasında, SARS-CoV-2 virüsünün genomunun hızlı bir şekilde dizilenmesi ve sınıflandırılması, virüsün yayılımını anlamada ve aşı geliştirme çalışmalarında kritik bir rol oynamıştır (Chang vd., 2020; R. Lu vd., 2020).

DNA dizilerinin hızlı ve doğru bir şekilde analiz edilmesi büyük önem taşımaktadır. Bu bağlamda, genetik dizilerin sınıflandırılmasında kullanılan yöntemlerin çeşitliliği ve etkisi de ön plana çıkmaktadır. Genetik dizilerin sınıflandırılmasında yaygın olarak kullanılan üç ana yöntem: Hizalama yöntemleri, MÖ yöntemleri ve DÖ yöntemleridir. Bu yöntemler, genetik verilerin giderek büyüyen veri tabanlarında daha verimli bir şekilde analiz edilmesine olanak tanır ve böylece biyolojik araştırmalara önemli katkılarda bulunur.

5.1. Hizalama Yöntemleri

DNA dizileri arasındaki benzerliklerin tespiti, biyolojik araştırmaların temelini oluşturmaktadır. Bu amaçla kullanılan Hizalama yöntemleri, dizilerin benzerlik gösteren

bölgelerini tespit etmek için dizilerin konumlarını karşılaştırmaya dayanmaktadır (Soliman et al., 2022). Bu yöntemler, dizilerin biyolojik olarak anlamlı bölgelerini tespit etmek ve karşılaştırmak için kullanılır. Genellikle Basic Local Alignment Search Tool (BLAST) (Altschul vd., 1990) ve ClustalW (Thompson vd., 1994) gibi araçlar kullanılarak gerçekleştirilen hizalama işlemleri, diziler arasındaki benzerlikleri ve farklılıkları belirlemekte etkili olmaktadır. Ayrıca çoklu dizi hizalaması, DNA, RNA ve protein dizileri arasındaki ortak işlevleri, yapıları veya ilişkileri belirlemede temel bir rol oynamaktadır.

Ancak, hizalama yöntemlerinin bazı dezavantajları bulunmaktadır. Bunlardan en önemlisi, yüksek hesaplama karmaşıklığıdır. Büyük veri setlerinde, özellikle de uzun DNA dizileriyle çalışıldığında, bu yöntemlerin hesaplama süreleri önemli ölçüde artmaktadır. Bu durum, hizalama yöntemlerinin performansını sınırlayarak daha verimli alternatif yöntemlerin araştırılmasını zorunlu kılmaktadır. Bu çerçevede, hizalamaya dayanmayan dizi sınıflandırma yöntemleri geliştirilmiştir. Bu yöntemler, hizalama işlemi olmaksızın diziler arasındaki benzerlikleri ve farklılıkları belirlemeye odaklanarak hesaplama sürecini basitleştirir ve böylece verimliliği artırır (Edgar, 2004).

5.2. Makine Öğrenmesi Yöntemleri

MÖ, bilgisayar sistemlerine verileri otomatik olarak öğrenme, analiz etme, desenleri tanımlama yeteneği kazandıran bir YZ alt dalıdır. Bu modeller, yapı olarak zamanla değişebilecek koşulları keşfedip öğrenir ve bunlara uyum sağlayarak modelin performansını artırmayı hedefler (Dixit & Prajapati, 2015). MÖ yöntemlerinin biyoinformatik verilerin analizinde kullanılması, bu alanda önemli iyileşmelere yol açmıştır. Genetik dizilerin analizi ve sınıflandırılmasında güçlü bir araç olarak öne çıkan MÖ, genetik dizileri hızlı ve hassas bir şekilde inceleyerek taksonomik çalışmalarda geniş bir uygulama alanı bulmuştur (Ao vd., 2022).

MÖ yöntemleri, büyük veri kümelerini işleme kapasiteleri ve desen tanıma yetenekleri sayesinde, hizalamaya dayalı yöntemlere göre genetik dizilerin analizinde daha hızlı ve daha doğru sonuçlar sunar. MÖ yöntemlerinde özellik seçimi, sınıflandırma başarısını en çok etkileyen faktörlerden biridir. Özellikle biyoinformatik gibi hassas alanlarda, büyük ve karmaşık veri kümelerindeki doğru özellik seçimi, MÖ uygulamalarının güvenilirliği açısından kritik bir faktördür. Bu, gelecekteki araştırmalar için de temel bir gereklilik olarak görülmektedir (Pfeifer vd., 2022). Ancak, DNA dizi

verilerinin belirgin özellikler taşımaması ve büyük veri miktarının işlenmesindeki zorluklar, MÖ yöntemleriyle yapılan analizleri karmaşık hale getirebilir. Bu nedenle, genetik dizilerin sınıflandırılması ve analizinde öznitelik çıkarımı süreci, veri setlerinin karmaşık ve gürültülü yapısından kaynaklanan zorlukların üstesinden gelmek için kritik bir adımdır (Dixit & Prajapati, 2015).

MÖ yöntemleri, farklı türdeki veriler ve problemler için çeşitli algoritmalar sunar. Her bir algoritma, belirli bir problem türü için kendi avantaj ve dezavantajlarına sahiptir. Bu nedenle, uygun algoritmanın seçimi, verinin türüne ve problemin bağlamına göre belirlenmelidir. MÖ algoritmaları, genellikle üç kategoriye ayrılmaktadır: Denetimli öğrenme, Denetimsiz öğrenme ve Yarı Denetimli öğrenme.

Denetimli öğrenme, etiketli veri kümeleriyle çalışarak modelin eğitilmesini sağlar. Bu algoritmalar, girdiler ve çıktılar arasındaki ilişkiyi öğrenerek yeni veriler için tahminler yapar. En çok kullanılan Denetimli öğrenme modelleri; Lineer Regresyon, LR, SVM, DT, RF, Extra-Trees (ET), Gradient Boosting (GB) ve KNN gibi yöntemlerdir.

Denetimsiz öğrenme, etiketlenmemiş verilerle çalışarak verilerdeki gizli desenleri ve yapıları ortaya çıkarmayı amaçlar. En çok kullanılan Denetimsiz öğrenme modelleri; K-Means Kümeleme, Hiyerarşik Kümeleme, Temel Bileşen Analizi (PCA), Bağımsız Bileşen Analizi (ICA) gibi yöntemlerdir.

Yarı Denetimli öğrenme hem etiketli hem de etiketlenmemiş verilerle çalışarak modelin eğitimini sağlar. Bu yöntem, etiketlenmiş verinin sınırlı olduğu durumlarda kullanışlıdır.

Sonuç olarak, MÖ algoritmaları arasından uygun yöntemin seçimi, verinin yapısına ve problem türüne bağlı olarak değişmektedir. Bu çalışmada, genetik dizilerin sınıflandırılmasında öne çıkan MÖ yöntemlerinden özellikle ağaç tabanlı algoritmalar tercih edilmiştir. Kullanılan yöntemler arasında RF, ET ve GB yer almaktadır.

5.2.1. Random Forest sınıflandırıcı

Ağaç tabanlı MÖ algoritmaları, düğümler ve dallarla görsel olarak temsil edilebilen, ağaç benzeri bir grafik yapısı kullanan yapıları nedeniyle "ağaç" adını almışlardır. Bu yaklaşımın en önemli modellerinden biri olan RF sınıflandırıcının öne çıkan özelliklerinden biri, aşırı uyum sorunlarını ele alma ve birden fazla modelin tahminlerini birleştirmeyi içeren topluluk öğrenimi (ensemble learning) ve torbalama (bootstrap aggregating) adı verilen teknikleri kullanarak performansı artırma yeteneğidir

(Breiman, 2001). Topluluk öğrenimi, birden fazla modelin tahminlerini birleştirerek genel performansı artırmayı amaçlar. Torbalama ise, farklı veri alt kümeleri üzerinde eğitilen birden fazla karar ağacını birleştirerek daha kararlı ve doğruluğu yüksek bir model oluşturur. Torbalama sürecinde, eğitim verileri rastgele ve tekrarlı olarak seçilir. Ardından, her bir alt veri kümesi için bir karar ağacı eğitilir. Bu süreç, modelin genelleme yeteneğini artırır ve aşırı uyum riskini azaltır.

RF algoritması, kendi içinde birbirinden bağımsız birçok alt karar ağacı üretir. Bu ağaçlar, eğitim verilerinin farklı alt kümeleri üzerinde eğitildikten sonra her biri bağımsız olarak tahminlerde bulunur. RF'deki her bir karar ağacında dallanma noktalarını belirlerken en sık kullanılan yöntemlerden biri Gini indeksi'dir. Gini indeksi, her bir düğümde sınıf dağılımını değerlendirerek o düğümdeki veri homojenliğini ölçer. Bir düğümdeki Gini indeksinin küçük oluşu, o düğümde sınıfların daha homojen olduğunu göstermektedir. Düğümdeki tüm örnekler tek bir sınıfa ait olduğunda Gini indeksi sıfıra ulaşır, bu da dallanmanın sona erdiği anlamına gelir (Watts vd., 2010). Eğer bir düğümdeki Gini indeksi büyükse; o düğümde tüm sınıfların eşit olarak karışmadığı anlamına gelmektedir. Gini İndeksi formülü Denklem 5.1'de görüldüğü gibidir. Bu formülde *t;* mevcut düğümü, *C*; toplam sınıf sayısını, p_i; *i*-inci sınıfın o düğümdeki toplam örnekler içindeki olasılığıdır.

$$Gini(t) = 1 - \sum_{i=1}^{C} {p_i}^2 \tag{5.1}$$

Bir düğüm iki alt düğüme (sol ve sağ düğüm) bölündüğünde, her bir alt düğüm için ayrı ayrı Gini indeksi hesaplanır ve toplam Gini indeksi, düğümlerin büyüklüklerine göre ağırlıklandırılır. Bu, bölünmenin kalitesini ölçer ve denklem 5.2'deki şekilde hesaplanır. Bu formül, bölünmenin ne kadar etkili olduğunu ölçen bir metrik sağlar. Amaç, her düğümde $Gini_S$ değerini minimize ederek veriyi daha homojen alt düğümlere ayırmaktır.

$$Gini_S = \frac{N_{sol}}{N} \times Gini_{Sol} + \frac{N_{sağ}}{N} \times Gini_{sağ} \tag{5.2}$$

Denklem 5.2'de N_{sol}; sol düğüme düşen verilerin sayısıdır. $N_{sağ}$ ise sağ düğüme düşen verilerin sayısıdır. **N**; orijinal düğümdeki toplam veri sayısıdır. Dolayısıyla, bölünme yapılmadan önce toplam veri sayısı; $N_{sağ}$ ve N_{sol} sayılarının toplamı kadardır. $Gini_{Sol}$; sol alt düğümdeki Gini indeksidir. $Gini_{Sağ}$; sağ alt düğümdeki Gini indeksidir.

RF, tüm ağaçların tahminlerini birleştirerek en yüksek skoru alan ağacın değerini sonuç olarak kabul eder (Svetnik vd., 2003). Bu şekilde, daha doğru tahminler elde edilmesi hedeflenmektedir. RF'nin bu özellikleri genetik diziler gibi genellikle yüksek boyutlu ve karmaşık bilgi içeren verilerin analizinde de son derece etkilidir. Genetik verilerin doğru bir şekilde sınıflandırılması, biyolojik ve tıbbi araştırmalar açısından büyük önem taşır. Bu bağlamda, RF algoritması, esnek yapısı ve yüksek doğruluk kapasitesi sayesinde gen dizilerinin karmaşık ilişkilerini etkili bir şekilde yakalayarak genetik verilerin analizinde yüksek performans sergileyebilmektedir.

5.2.2. Extra-Trees sınıflandırıcı

ET sınıflandırıcı ağaç tabanlı MÖ algoritmaları arasında yer alan, yapı itibariyle RF'ye benzeyen başka bir topluluk öğrenme yöntemidir (Geurts vd., 2006). ET, RF algoritmasıyla benzerlikler taşımasına rağmen, bazı temel farklılıklarla öne çıkar ve bu farklılıklar, ET'yi belirli veri kümelerinde daha dayanıklı ve verimli hale getirir. ET'nin basitliği, etkinliği ve hesaplama verimliliği, MÖ alanında önemli bir ilgi görmesine neden olmuştur.

ET ve RF algoritmaları arasındaki temel fark, rastgeleliği getirme şekillerindedir (Geurts vd., 2006). RF, her bir düğümde eğitim için rastgele bir özellik alt kümesi seçerken, ET buna ek olarak özellik bölmeleri için rastgele eşikler kullanır. Rastgele bölünme noktaları kullanıldığı için her ağacın yapısı farklı olur ve model genelinde daha fazla çeşitlilik sağlanır. Ağaçlar tamamlandığında, tüm ağaçların tahminleri birleştirilerek nihai tahmin yapılır. Bu süreç, her ağacın kendi sınıf tahminini gerçekleştirmesi ve ardından çoğunluk oyu ile nihai sınıfın belirlenmesiyle gerçekleşir. Denklem 5.3'te gösterildiği gibi M; toplam ağaç sayısını, $T_M(x)$; her bir ağacın tahminini ifade etmektedir.

$$y = mod(T_1(x), T_2(x), \dots, T_M(x)) \tag{5.3}$$

Ağaçlar, rastgele bölünmelerle oluşturulurken belirli durma koşullarıyla büyüme süreci sonlandırılabilir. Bu koşullar arasında, ağacın maksimum derinliğe ulaşması, bir düğümdeki tüm verilerin aynı sınıfa ait hale gelmesi veya düğümdeki örnek sayısının belirli bir minimum değerin altına düşmesi yer alır.

ET; her ne kadar rastgele bölünme noktaları kullansa da Gini indeksi gibi saflık ölçümleri hala uygulanabilmektedir. Bu ölçümler, alt düğümlerdeki verilerin homojenlik düzeyinin değerlendirilmesini sağlar. Ancak, bölünme noktaları rastgele seçildiği ve optimize edilmediği için karar verme aşamasında doğrudan kullanılamazlar.

ET'nin rastgele eşik seçimi, algoritmayı genetik dizi verilerinde bulunan gürültü ve aykırı değerlere karşı daha dayanıklı hale getirir. Aynı zamanda, ET basit ve etkili bir yöntem olup, daha az hesaplama kaynağı gerektirir. Bu özellik, özellikle büyük veri setleriyle çalışırken önemli bir avantaj sunar. ET, verilerdeki karmaşık ve doğrusal olmayan ilişkileri yakalama yeteneği sayesinde, genetik dizilerdeki etkileşimleri doğru bir şekilde analiz edebilir ve bu nedenle genetik verilerin doğru sınıflandırılması açısından güçlü bir araç olarak öne çıkar.

5.2.3. Gradient Boosting sınıflandırıcı

GB, topluluk öğrenme yöntemlerinin gücünden yararlanarak birden fazla modelin öngörülerini birleştirip genel performansı artıran etkili bir MÖ algoritmasıdır. GB, RF algoritmasının bağımsız ağaç yapısının aksine, ağaçları sıralı olarak oluşturma süreciyle öne çıkar ve her yeni ağacı teker teker tanıtarak modeli adım adım geliştirir (Friedman, 2001).

GB'nin temel özelliklerinden biri, sıralı ağaç oluşturma sürecidir. Bu süreçte her yeni ağaç, mevcut modelin yaptığı hataları düzeltmeye çalışır. Her iterasyonda model, hatalarını minimize edecek şekilde güncellenir. Bu yaklaşım, modelin doğruluğunu artırırken, aynı zamanda hataların kümülatif olarak azalmasını sağlar. Bu süreç, optimizasyon sürecinin gradyan iniş prensiplerine dayandığı bir sistemle çalışır (Bentéjac vd., 2021).

Denklem 5.4'te *D*, bir veri setini temsil etmektedir. Buna göre $x_i \in R^p$, i-inci veri noktasının *p*-boyutlu özellik vektörü ve y_i, sınıflandırmaya bağlı olarak hedef değerdir.

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \tag{5.4}$$

Denklem 5.5'de $L\big(y, F(x)\big)$ kayıp fonksiyonunu tanımlamaktadır. Burada modelin yaptığı tahmin ($F(x_i)$) ile gerçek değer (y_i) arasındaki fark gösterilmektedir.

$$L\big(y, F(x)\big) = \frac{1}{N}\sum_{i=1}^{N} l(y_i, F(x_i)) \quad (5.5)$$

Gradyan iniş yöntemi, kayıp fonksiyonunun eğimine dayanarak, her adımda modelin parametrelerini günceller ve bu sayede kayıp fonksiyonunun minimuma inmesini sağlar. Bu süreç, modelin performansını sürekli olarak iyileştirir ve daha doğru tahminler yapılmasına olanak tanır (Bentéjac vd., 2021). GB, karar ağaçlarından oluşan $F_M(x)$'i eğiterek hedef değerleri tahmin etmeye çalışır. Model, iteratif olarak denklem 5.6'de ifade edilen şekilde güncellenir. $T_m(x)$ karar ağacı modelini, γ_m; $T_m(x)$'in ağırlığını temsil etmektedir.

Bu yöntemde her bir adımda, önceki modele ek olarak bir yeni ağaç ($T_m(x)$) eklenir. Her iterasyonda m-inci modelin güncellemesi Denklem 5.6'da gösterilen şekilde gerçekleştirilmektedir. Burada γ_m; yeni eklenen ağacın katkısını ifade etmektedir ve denklem 5.7'de gösterilen şekilde hesaplanmaktadır.

$$F_m(x) = F_{m-1}(x) + \gamma_m T_m(x) \quad (5.6)$$

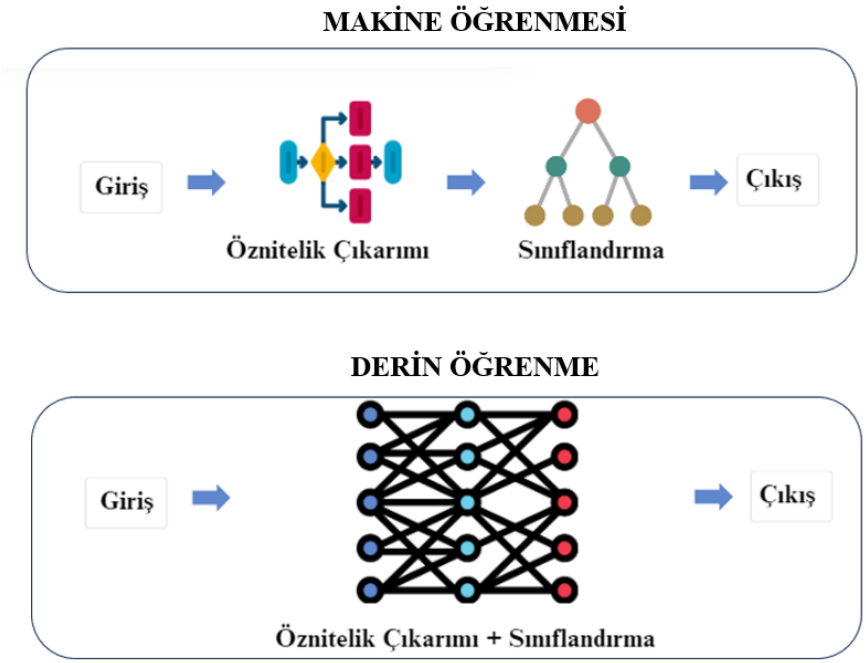
$$\gamma_m = arg_{\gamma}^{min} \sum_{i=1}^{N} l\,(y_i, F_{m-1}(x_i) + \gamma_m T_m(x_i) \quad (5.7)$$

5.3. Derin Öğrenme Yöntemleri

Son yıllarda, DÖ algoritmaları, büyük veri kümeleri üzerinden öğrenme kapasiteleri ve karmaşık tahminler yapabilme becerileriyle geleneksel MÖ algoritmalarına göre belirgin bir üstünlük sağlamaktadır (Hinton vd., 2012; Krizhevsky vd., 2012). Bu algoritmaların öne çıkan yönlerinden biri, verilerden karmaşık desenler ve özellikler öğrenebilme yetenekleridir. Katmanlı yapıları, düşük seviyeli özelliklerden başlayarak veriler üzerinde yüksek seviyeli soyutlamalar yapmalarını mümkün kılmakta ve bu çok katmanlı öğrenme süreçleri, modellerin daha hassas ve doğru tahminler yapmasını sağlamaktadır (Hinton vd., 2012). Özellikle biyoinformatik gibi karmaşık veri yapılarının analiz edildiği alanlarda, DÖ algoritmalarının kullanımı giderek daha yaygın hale gelmektedir.

DÖ algoritmaları, MÖ yöntemlerine kıyasla çeşitli avantajlara sahiptir. Öncelikle, bu algoritmalar katmanlar arasındaki öğrenme süreçleri sayesinde verilerin özelliklerinden soyutlamalara kadar kapsamlı bir öğrenme gerçekleştirebilmektedir.

Buna karşılık, MÖ algoritmaları genellikle manuel olarak belirlenmiş özniteliklere dayandığından, model performansı sınırlı kalabilmektedir.



Şekil 5.1. MÖ ve DÖ yapısı

DÖ algoritmalarının bir diğer önemli avantajı, veri ön işleme ve özellik çıkarımı gibi süreçlere olan ihtiyacı azaltmasıdır. MÖ yöntemlerinde, özellikle karmaşık ve yüksek boyutlu verilerle çalışırken özellik çıkarımı kritik bir adımdır; bu süreç hem zaman alıcıdır hem de uzman bilgi gerektirmektedir. DÖ algoritmaları ise özellik çıkarımını otomatik olarak gerçekleştirdiği için bu süreçlerde insan müdahalesine olan gereksinimi en aza indirmektedir (Şekil 5.1). Bu durum, genetik verilerin analizinde de büyük bir avantaj sağlamaktadır; çünkü genetik diziler genellikle yüksek boyutlu ve karmaşık veri yapılarından oluşur. Dolayısıyla, bu tür verilerin DÖ algoritmalarıyla analiz edilmesi, daha hızlı ve daha doğru sonuçların elde edilmesini mümkün kılmaktadır.

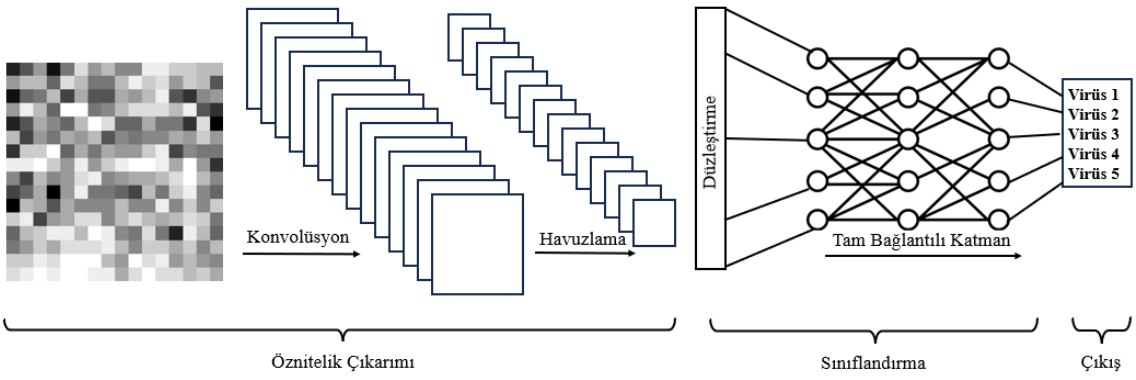
Biyoinformatik alanında, genetik dizilerin analizi ve sınıflandırılması, büyük ölçüde DÖ yöntemlerinin kullanımıyla ilerleme kaydetmiştir. Genetik dizilerdeki karmaşık yapılar ve ilişkiler, geleneksel MÖ yöntemleriyle tespit edilmesi zor olan bilgileri barındırmaktadır. DÖ algoritmaları ise bu tür karmaşık veri yapılarını otomatik olarak işleyip, önemli öznitelikleri çıkarabilmekte ve analiz sürecini hızlandırarak daha doğru sonuçlar elde edilmesini sağlamaktadır.

DÖ yöntemleri arasında Yapay Sinir Ağları (YSA), CNN'ler, Tekrarlayan Sinir Ağları (RNN), Autoencoder'ler, GAN ve Transformer Ağları gibi farklı mimariler bulunmaktadır. Bu mimariler, farklı veri türlerine ve problemlerine uygun çözümler sunmakta olup, özellikle biyoinformatik ve diğer bilimsel alanlarda giderek daha yaygın şekilde kullanılmaktadır.

5.3.1. Evrişimli sinir ağları

CNN, DÖ algoritmalarının en popüler ve etkili türlerinden biridir. CNN modelleri, bir görüntüyü girdi olarak alıp, bu görüntüdeki nesneleri veya belirgin özellikleri tanımlama yeteneğine sahiptir. Bu modeller, girdi görüntüsünü alıp, görüntüdeki çeşitli özelliklere ağırlıklar ve bias ataması yapar. Böylece, görüntüdeki özellikler arasındaki farklılıkları belirler ve analiz eder. Bu işleyişiyle, CNN, DÖ algoritmaları arasında güçlü bir araç olarak öne çıkar.

CNN mimarisi, Giriş katmanı, bir dizi Evrişim katmanı (Convulation layer), Aktivasyon katmanı, Havuzlama katmanı (Pooling Layer), Tam bağlantılı katman (Fully connected layer), Normalizasyon katmanı, Seyreltme (Dropout) katmanı ve Sınıflandırma katmanından oluşmaktadır. Bu katmanlar, veriler üzerinde belirli işlemler gerçekleştirerek, modelin öğrenmesini ve genelleme yeteneğini geliştirmesini sağlamaktadır. Şekil 5.2'de genel bir CNN mimarisinin yapısı gösterilmektedir.



Şekil 5.2. CNN mimarisi

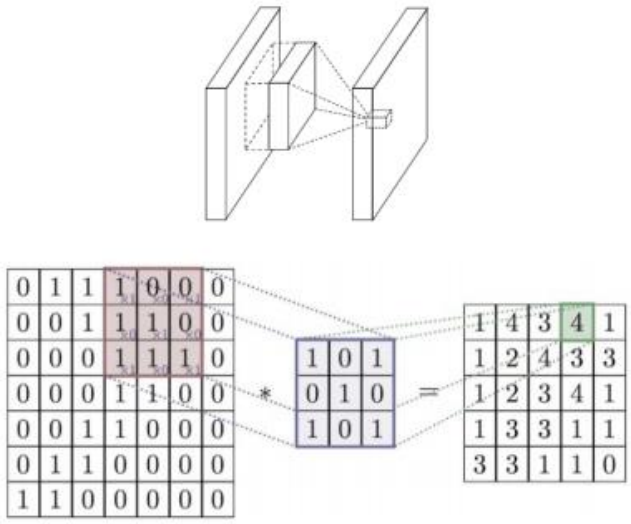
Piksel verilerinin ağırlıklarla çarpılması ve sonuçların toplanması işlemine evrişim denir. Bu, CNN'lerin temel işlemidir ve bu sebeple bu ağlara "Evrişimli sinir ağı" adı verilmiştir. Evrişim işlemi, görüntünün çeşitli bölgelerindeki piksellerin değerlerini

filtrelerle çarparak, bu bölgelerin önemini belirlemektedir. Bu işlemin tekrarlanması ve çeşitli filtrelerin kullanılması, görüntünün farklı özelliklerinin çıkarılmasına olanak tanımaktadır. Böylece, model hem düşük seviyeli (örneğin, kenarlar ve dokular gibi) hem de yüksek seviyeli (örneğin, nesneler ve yüzler gibi) özellikleri öğrenebilmektedir.

CNN'nin en büyük avantajlarından biri, parametrelerin sayısını önemli ölçüde azaltarak karmaşık görüntüleri işleyebilmesidir. Bu, modelin eğitim süresini kısaltmakta ve daha büyük veri kümeleri üzerinde daha etkili bir şekilde çalışmasını sağlamaktadır. Ayrıca, CNN'nin modüler yapısı, farklı görüntü işleme görevlerine kolayca uyarlanabilmesine olanak tanımaktadır. Bu özellikler, CNN'nin özellikle biyoinformatik alanında genetik verilerin görsel temsilinde ve bu verilerin sınıflandırılmasında kullanılmasını mümkün kılmaktadır.

Giriş katmanı: Giriş katmanı, verilerin sinir ağına tanıtıldığı ilk adımdır. Bu katman, görüntü verilerini ağın işlem yapabileceği bir formata dönüştürmektedir. Giriş katmanının boyutu, veri setinin boyutuna bağlıdır ve genellikle resmin piksel sayısını temsil etmektedir. Giriş verilerinin boyutu, CNN'in genel performansını ve sınıflandırma doğruluğunu doğrudan etkilemektedir. Bu nedenle, giriş katmanının veriyi doğru şekilde temsil etmesi ve ağın diğer katmanlarıyla uyumlu çalışması büyük önem taşımaktadır.

Evrişim katmanı: Evrişim katmanı, CNN'in temel yapı taşıdır ve giriş verisini işleme sürecinde kritik bir rol oynamaktadır. Bu katmanda, filtreler (kernel olarak da adlandırılmaktadır) kullanılarak giriş veri matrisinin (örneğin, bir görüntü) özellikleri çıkarılmaktadır. Her filtre, giriş verisi üzerinde kaydırılarak evrişim işlemi gerçekleştirilmekte ve bu işlem sırasında belirli bir girdi bölgesinin özellikleri çıkarılmaktadır (Şekil 5.3). Sonuçta elde edilen değerler, verinin yerel özelliklerini çıkararak daha yüksek seviyeli soyutlamalar yapılmasına olanak tanır ve modelin genel performansını artırır.



Şekil 5.3. CNN mimarisinde evrişim işlemi (Karaköse, 2019)

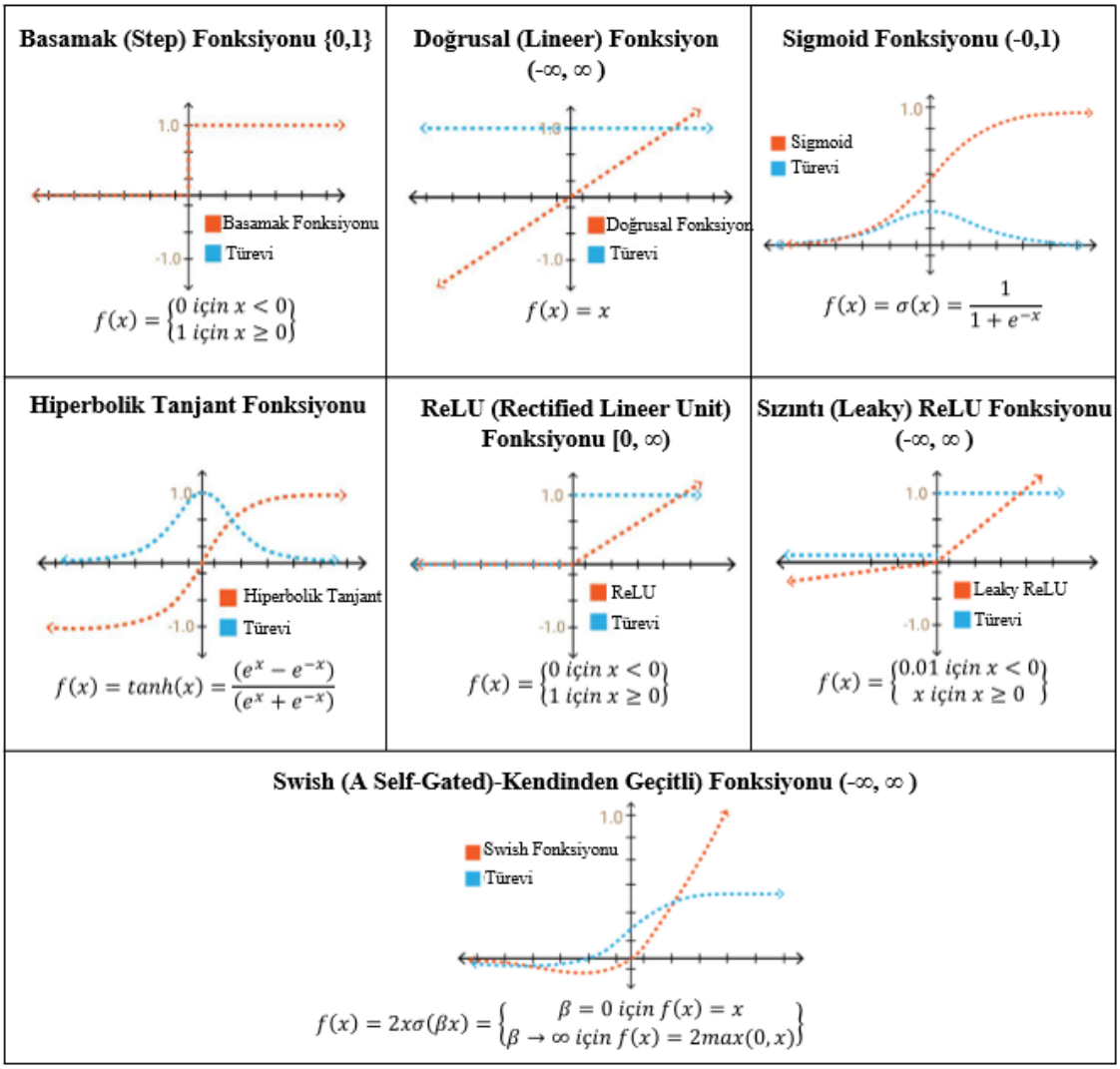
Evrişim işlemi genellikle denklem 5.7'de gösterilen şekilde ifade edilmektedir. Buna göre *I* giriş matrisini, *K* filtreyi, *i* ve *j* ise konumları belirtmektedir.

$$S(i,j) = (I * K)(i,j) = \sum_{m=0} \sum_{n=0} I(i+m, j+n) \cdot K(m,n) \quad (5.7)$$

Evrişim katmanı, verideki temel özelliklerin yanı sıra, derin katmanlarda daha karmaşık ve soyut özelliklerin öğrenilmesini sağlar. Filtre boyutları genellikle 3x3, 5x5 veya 7x7 gibi küçük matrisler şeklindedir ve her filtre, belirli bir özellik türünü algılamaya odaklanmaktadır. Bu işlem, verinin boyutunu küçültmek ve bilgi yoğunluğunu artırmak için kullanılır. Evrişim işlemi sırasında öğrenilen parametreler, modelin doğruluğunu ve genelleme kapasitesini artırarak, farklı veri türlerindeki yapısal ilişkileri yakalamaya yardımcı olur.

Aktivasyon Katmanı: Aktivasyon katmanı, CNN'deki her evrişim ve tam bağlantılı katmandan sonra kullanılan ve modelin doğrusal olmayan ilişkileri öğrenmesini sağlayan bir bileşen olarak görev yapmaktadır. Aktivasyon fonksiyonları, ağın her bir nöronundan çıkan değeri alır ve bu değerin bir sonraki katmana nasıl iletileceğini belirler.

Aktivasyon fonksiyonu olarak genellikle Basamak (Step), Doğrusal (Linear), Sigmoid, Hiperbolik Tanjant, Doğrultulmuş Doğrusal Birim (Rectified Linear Unit (ReLU)), Sızıntı (Leaky) ReLU ve Kendinden Geçitli (ASelf-Gated (Swish)) fonksiyonları kullanılmaktadır (Kaya vd., 2020). Sıklıkla kullanılan aktivasyon fonksiyonlarının grafikleri ve matematiksel denklemleri Şekil 5.4'de görülmektedir.



Şekil 5.4. Aktivasyon fonksiyonları ve matematiksel ifadeleri (Kaya vd., 2020)

Basamak (Step) fonksiyonu, temel bir doğrusal olmayan aktivasyon fonksiyonu olup, girdiye bağlı olarak çıktı değerini belirli bir eşik değerine göre 0 veya 1 olarak belirlemektedir.

Doğrusal (Linear) fonksiyonu, çıkışın girdinin bir doğrusal fonksiyonu olduğu basit yapıdaki bir aktivasyon fonksiyonudur.

ReLU (Rectified Linear Unit/Doğrultulmuş Doğrusal Birim) en yaygın kullanılan aktivasyon fonksiyonlarından biridir. ReLU fonksiyonu, her bir nöronun çıkışını pozitif olduğu sürece, olduğu gibi bırakmakta; negatif olduğu durumda ise sıfıra eşitlemektedir. Bu fonksiyon, negatif girdileri sıfıra eşitleyerek "vanishing gradient (kaybolan gradyan)" problemini büyük ölçüde azaltmaktadır. Bu sayede, derin ağların eğitimi daha hızlı ve verimli hale gelmektedir. Ancak ReLU'nun bazı dezavantajları da vardır; örneğin, tüm

negatif girdileri sıfıra eşitlediği için bazı nöronlar tamamen etkin olmayabilir. Bu sorun, "dead neuron (ölü nöron)" problemi olarak adlandırılmaktadır.

Sızıntı (Leaky) ReLU fonksiyonu, ReLU (Rectified Linear Unit) fonksiyonunun bir varyasyonudur. ReLU'nun ölü nöron problemine çözüm olarak geliştirilmiştir. ReLU, negatif girdileri sıfıra eşitlediği için bazı nöronlar tamamen aktif olmayabilir ve bu durum modelin performansını olumsuz etkileyebilir. Leaky ReLU, negatif girdiler için küçük bir eğim bırakacak şekilde tasarlanmıştır, böylece negatif değerler de sınırlı da olsa bilgi taşımaya devam edebilir.

Sigmoid aktivasyon fonksiyonu, çıkış değerlerini (0, 1) aralığına sıkıştıran doğrusal olmayan bir fonksiyondur. Genellikle, ikili sınıflandırma problemlerinde kullanılır. Sigmoid fonksiyonu, her bir nöronun çıkışını, 0 ile 1 arasında bir olasılık değeri olarak yorumlamayı mümkün kılmaktadır. Ancak, sigmoid fonksiyonunun önemli bir dezavantajı, kaybolan gradyan problemine neden olabilmesidir. Yüksek veya düşük değerlerde, türevleri çok küçük olabilmekte ve bu da derin ağların öğrenmesini zorlaştırmaktadır.

Hiperbolik Tanjant aktivasyon fonksiyonu, sigmoid fonksiyonuna benzemektedir, ancak çıkış değerleri (-1, 1) aralığındadır. Negatif ve pozitif girişler için daha geniş bir çıktı aralığı sunar ve bu nedenle sigmoid fonksiyonuna göre daha avantajlıdır.

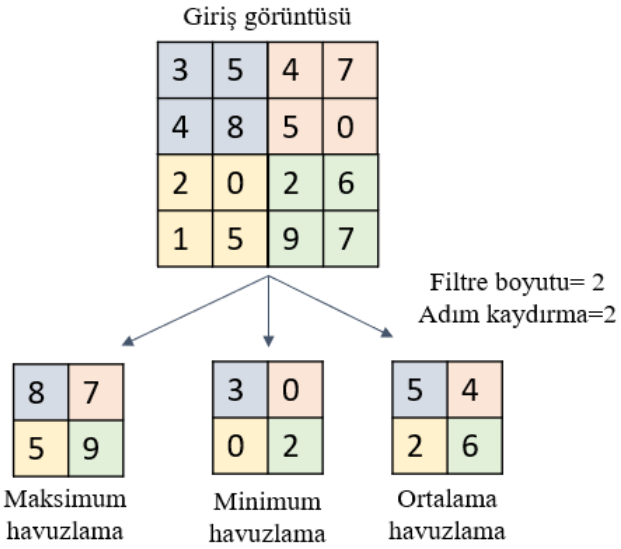
Swish Aktivasyon fonksiyonu; Google araştırmacıları tarafından önerilen ve sigmoid ile ReLU'nun özelliklerini birleştiren yeni bir aktivasyon fonksiyonudur. Negatif girdiler için yumuşak bir geçiş sağlamakta ve bu sayede modelin doğruluğunu artırmaktadır. Swish'in adaptif yapısı, ReLU'ya kıyasla bazı uygulamalarda daha iyi performans gösterebilmektedir. Swish, doğrusal olmayan aktivasyon fonksiyonları arasında, modern sinir ağı uygulamalarında giderek daha popüler hale gelmektedir.

Havuzlama Katmanı: Havuzlama işlemi, bir görüntünün daha küçük bir ölçekte örneklenmesini sağlayarak boyutunu azaltan bir yöntemdir. DÖ modellerinde havuzlama katmanı, genellikle veri boyutlarını küçültmek ve bu sayede hesaplama yükünü azaltmak için kullanılmaktadır (Kaya vd., 2020). Boyut küçültme işlemleri sırasında bazı bilgi kayıpları yaşanabilmekte ve bu da modelin performansında düşüşe yol açabilmektedir. Ancak havuzlama işlemi, modelin aşırı öğrenmesini (overfitting) önleyerek daha az hesaplama gerektiren daha verimli bir yapı oluşturmaktadır (Kaya vd., 2020).

Havuzlama katmanları genellikle 2x2 boyutunda filtreler kullanır. Bu katmanların temel rolü, özellik haritalarının çözünürlüğünü düşürerek veri hacmini küçültmektir. Aynı zamanda, öteleme ve dönme değişimlerine karşı dayanıklılık sağlayarak

sınıflandırma için gerekli bilgileri korumaktır. Havuzlama, hesaplama maliyetlerini büyük ölçüde azaltarak modellerin daha hızlı ve verimli çalışmasına katkıda bulunur. Eğitim sürecinde, havuzlama işlemi için genellikle geri yayılım algoritması kullanılmaktadır. Bu, modelin işlemleri daha hızlı ve verimli bir şekilde gerçekleştirmesine olanak tanır.

Havuzlama işleminin çeşitli türleri bulunmaktadır. En yaygın kullanılan havuzlama teknikleri Maksimum havuzlama, Ortalama havuzlama ve Minimum havuzlamadır. Maksimum havuzlama, her bir segmentin içindeki en yüksek piksel değerini alarak çıkış matrisi oluşturmaktadır. Şekil 5.5'te görüldüğü gibi, görüntüdeki en belirgin özellikleri korumaktadır. Minimum havuzlama ise, her segmentin en düşük piksel değerini alarak, detay kaybını en aza indirmeyi amaçlamaktadır. Ortalama havuzlama ise her segmentin ortalama piksel değerini alarak daha yumuşak bir genel görünüm sağlamaktadır. Bu işlemler, her segment için tekrarlanarak genel veri yapısı küçültülüp, bilgi yoğunluğu azaltılmaktadır.



Şekil 5.5. Havuzlama teknikleri

Tam Bağlantılı Katman: Bu katman, aynı zamanda yoğun katman olarak da bilinmekte ve genellikle DÖ modellerinin temel yapı taşlarından biri olarak kabul edilmektedir. Tam Bağlantılı Katman, çıktı katmanının boyutunu ve şeklini kontrol etmek için yaygın bir şekilde kullanılmaktadır. Tam bağlantılı katmanların en önemli özelliklerinden biri, her bir düğümün, önceki katmandaki tüm düğümlerle tam olarak bağlantılı olmasıdır. Bu özellik, bu katmanları esnek ve genel amaçlı bir hale

getirmektedir. Çünkü herhangi bir yapısal sınırlama getirmeyip modelin geniş bir veri yelpazesi üzerinde öğrenmeyi gerçekleştirmesine olanak tanımaktadır.

Tam bağlantılı katmanların temelinde nöronlar yer alır. Bu nöronlar, girdilerden gelen bilgiyi işlemek ve bir çıktı üretmek için kullanılmaktadır. Ağırlıklar, sinir ağının öğrenme sürecinde optimize edilen parametrelerdir. Ayrıca, her nöron, bias olarak bilinen ve düğümün çıkışını kaydırmak için kullanılabilen ek bir parametreye sahiptir. Bu parametreler, geri yayılım algoritması kullanarak öğrenilmekte ve ağın görünmeyen veriler üzerinde iyi bir tahmin doğruluğu sağlamasına yardımcı olmaktadır. Tam bağlantılı katmanlarda, her bir nöronun çıktısı, bir aktivasyon fonksiyonu aracılığıyla hesaplanmaktadır. Bu fonksiyonlar, sinir ağının farklı türde veri ilişkilerini öğrenebilmesini sağlar.

Tam bağlantılı katman yapısında önceki katmandan gelen bilgi, ağırlık matrisi ve bias vektörü kullanılarak girdilerden bir doğrusal yapı elde edilir. Daha sonra bu yapıya aktivasyon fonksiyonu uygulanarak doğrusal olamayan bir yapıya dönüşüm gerçekleştirilir. Denklem 5.8'de tam bağlantılı katmanın matematiksel ifadesi gösterilmektedir. Buna göre x giriş vektörü, W_{ij} ağırlık matrisi, b bias vektörü ve f aktivasyon fonksiyonudur.

$$z_i = f(\sum_{j=1}^{n} W_{ij} \cdot x_j + b_i) \tag{5.8}$$

Tam bağlantılı katmanlar, ağın esnekliğini arttırmakta fakat aynı zamanda modelin hesaplama maliyetini de yükseltmektedir. Çünkü her bir nöron, bir önceki katmandaki tüm nöronlarla bağlantılıdır ve bu da büyük veri kümelerinde hesaplama yükünü artırabilmektedir. Bu nedenle, DÖ modellerinde genellikle diğer katman türleriyle, örneğin evrişimsel katmanlar veya tekrarlayan katmanlarla birleştirilmektedirler. Bu kombinasyon hem modelin öğrenme kapasitesini arttırmakta hem de hesaplama maliyetini düşürmektedir.

Normalizasyon katmanı: Bu katman giriş verilerinin veya ara katman aktivasyonlarının dağılımını belirli bir ölçeğe getirerek modelin daha hızlı ve stabil bir şekilde öğrenmesini sağlar. Normalizasyon, sinir ağının her bir katmanındaki aktivasyonların belirli bir aralıkta kalmasını sağlamak için kullanılır ve bu sayede gradyanların çok büyük veya çok küçük olma sorununu önler. Modelin aşırı öğrenmesini engellemek ve genel performansını artırmak için kullanılmaktadır. Normalizasyon

katmanlarının en yaygın türlerinden biri “Batch Normalization (Toplu Normalizasyon)” dur. Bu teknik, her mini-batch içinde aktivasyonların ortalamasını sıfıra ve standart sapmasını bire ayarlayarak çalışmaktadır. Toplu Normalizasyon, derin ağların eğitim süresini önemli ölçüde azaltarak, modelin parametrelerini daha stabil hale getirmektedir. Toplu Normalizasyonun yanı sıra, “Layer Normalization (Katman Normalizasyonu)”, “Instance Normalization (Örnek Normalizasyonu)” ve “Group Normalization (Grup Normalizasyonu)” gibi diğer normalizasyon teknikleri de kullanılmaktadır. Bu teknikler, belirli uygulamalar ve ağ mimarilerine göre optimize edilerek seçilmektedirler.

Normalizasyon katmanı, modelin genel performansını artırmanın yanı sıra, kaybolan gradyan problemini azaltarak, derin ağların daha verimli ve etkili bir şekilde öğrenmesini sağlamaktadır. Ayrıca, modelin girdilerdeki küçük değişikliklere karşı daha az hassas olmasına yardımcı olarak, daha sağlam ve güvenilir bir performans elde edilmesine katkıda bulunur.

Seyreltme Katmanı: Seyreltme tekniği, sinir ağlarında aşırı uyumu önlemeye yönelik bir yöntemdir. Eğitim aşamasında rastgele seçilen bazı nöronları çıkararak bu nöronların geçici olarak ağın eğitiminin dışında bırakır. Bu teknik, sinir ağlarının daha genelleyici ve sağlam özellikler öğrenmesini sağlamakta, böylece modelin sadece eğitim verilerine değil, aynı zamanda daha geniş ve görülmemiş veri kümelerine de iyi bir şekilde genelleme yapmasını sağlamaktadır. Aşırı uyum, özellikle ağ çok büyükse, eğitim süresi çok uzunsa veya eğitim verisi sınırlıysa sıkça karşılaşılan bir sorundur. Seyreltme tekniği, bu problemi azaltarak modelin performansını iyileştirmeye yardımcı olmaktadır.

Seyreltme Katmanının temel çalışma prensibi, sinir ağını, belirli nöronların rastgele alt kümeleriyle eğitmeye zorlamaktır. Eğitim sırasında, her ileri geçişte, bir katmandaki bazı nöronlar rastgele seçilerek devre dışı bırakılır. Bu süreç, ağın farklı alt kümelerle eğitildiği yanılsamasını yaratır ve ağın çok sayıda varyasyonla karşı karşıya kalmasını sağlar. Bu çeşitlilik, modelin öğrenme sürecinde daha sağlam ve genelleyici özellikler kazanmasına olanak tanır. Özellikle büyük ağırlık değerlerine sahip bir sinir ağı, eğitim verilerine aşırı uyum sağladığını ve karmaşıklığı artırarak aşırı uyuma yol açtığını gösterebilir. Seyreltme tekniği, bu durumu hafifletmek için nöronları ve bağlantılarını rastgele olarak devre dışı bırakarak ağın daha basit ve etkili bir şekilde düzenlenmesini sağlar.

Seyreltme Katmanının bir diğer önemli özelliği, her epoch sırasında eğitimin daha hızlı gerçekleşmesini sağlamasıdır. Çünkü her bir iterasyon için sadece aktif nöronlar hesaplamalara katılarak hesaplama maliyetini düşürmektedir. Ancak, bu yaklaşımın bir

yan etkisi olarak yakınsama için gereken toplam iterasyon sayısı genellikle iki katına çıkmaktadır. Bu durum, ağın daha fazla eğitilmesini ve daha sağlam özellikler öğrenmesini sağlamaktadır. Seyreltme Katmanının bu yapısı, aslında farklı mimarilere sahip birçok sinir ağının paralel olarak eğitilmesine benzer bir etki yaratmaktadır. Eğitim sırasında, her katmanın güncellenmesi farklı bir yapılandırma ile gerçekleştirilerek modelin esnekliği arttırılır ve genelleme yeteneği geliştirilir.

Sinir ağlarında Seyreltme tekniğinin kullanımı, özellikle büyük ve karmaşık ağ yapılarında önemli avantajlar sunar. Bu katman, sinir ağlarının görme, konuşma tanıma, belge sınıflandırma ve hesaplamalı biyoloji gibi çeşitli denetimli öğrenme görevlerinde performansını artırarak daha etkili sonuçlar elde edilmesine katkıda bulunmaktadır (Szegedy vd., 2015).

Seyreltme, ağın herhangi bir gizli katmanında veya giriş katmanında uygulanabilir, ancak genellikle çıktı katmanında kullanılmaz. Bu, modelin son tahminini doğrudan etkilememek ve eğitim sırasında öğrenilen özelliklerin test sırasında tam olarak kullanılmasını sağlamak içindir.

Sonuç olarak, Seyreltme tekniği, sinir ağlarının aşırı uyum sağlamasını önlemek ve genelleme performansını artırmak için etkili ve yaygın bir düzenleme tekniğidir. Modelin daha sağlam ve genelleyici özellikler öğrenmesini sağlar, bu da gerçek dünyadaki veri setlerinde daha iyi performans gösteren modeller oluşturulmasına katkıda bulunur. Seyreltme tekniğinin sağladığı avantajlar, sinir ağlarının çeşitli uygulama alanlarında kullanılmasını teşvik eder ve modern DÖ sistemlerinin geliştirilmesinde kritik bir rol oynar.

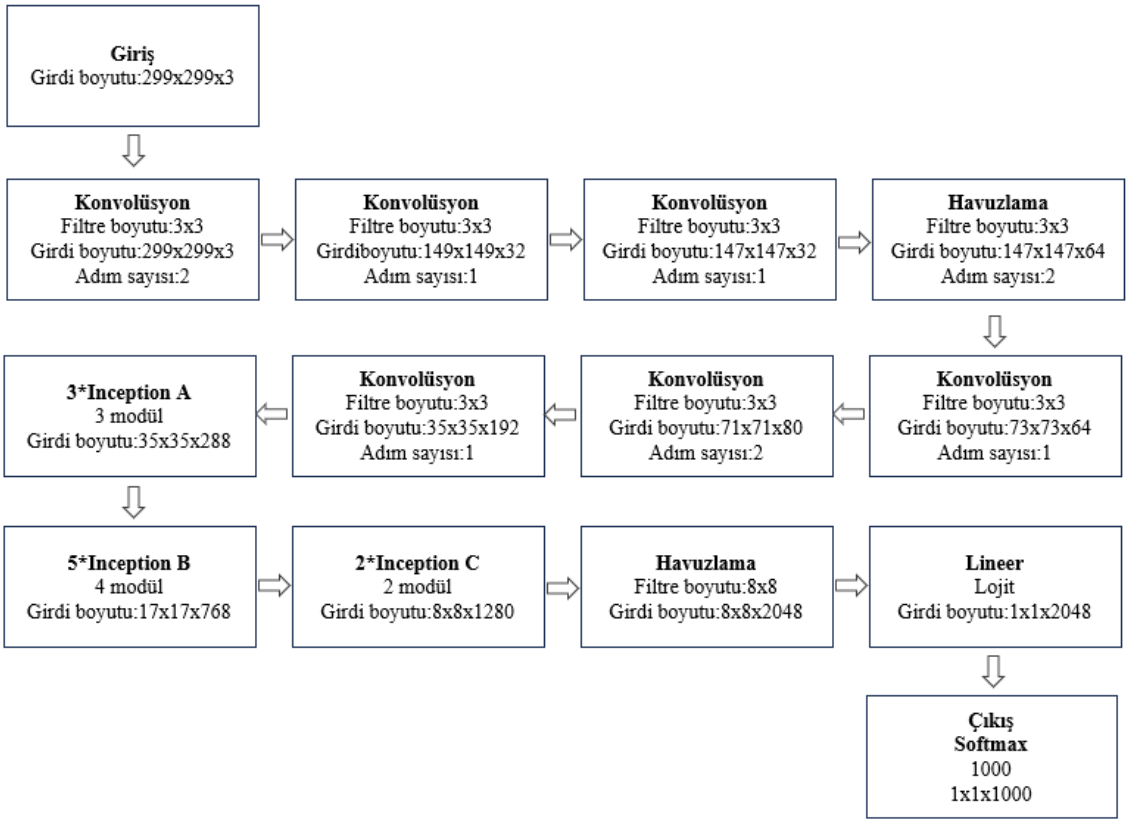
Sınıflandırma Katmanı: DÖ katmanlarından sınıflandırma katmanı, modelin son aşamasında yer alan ve girdileri belirli sınıflara ayırmak için kullanılan bir katmandır. Bu katmanın temel görevi, modelin önceki katmanlarında elde edilen özellikleri kullanarak girdileri önceden belirlenmiş sınıf sayısına göre sınıflandırmaktır. Sınıflandırma katmanı, modelin çıktısını oluşturan son katman olup, sınıflandırma problemlerinde nihai tahminlerin yapılmasını sağlar. Bu katmanda, sınıf sayısı kadar nöron bulunmaktadır ve her nöron, belirli bir sınıfı temsil etmektedir. Sınıflandırma katmanı, DÖ modellerinin performansı üzerinde doğrudan etkisi olan önemli bir unsurdur. Doğru bir sınıflandırma katmanı tasarımı, modelin eğitim ve test verileri üzerinde yüksek doğrulukta tahminler yapmasını sağlar.

5.3.2. InceptionV3

InceptionV3, çoğunlukla görüntü tanıma görevleri için kullanılan CNN modellerinden biridir. Google'daki araştırmacılar tarafından geliştirilen InceptionV3, Inception ailesinin bir üyesidir (Szegedy vd., 2015). CNN'lerin yapısını optimize ederek verimli hesaplama ile yüksek doğruluk sağlamak üzere tasarlanmıştır. InceptionV3'ün temel amacı, hesaplama yükünü azaltırken modelin önemli özellikleri yakalama yeteneğini korumak veya artırmaktır. Bunu başarmak için, model, büyük konvolüsyon işlemlerini daha küçük ve sıralı işlemlere ayırır. Böylelikle, daha büyük boyutlu konvolüsyonların hesaplama karmaşıklığı, küçük çaplı ve daha yönetilebilir işlemlere bölünerek azaltılmaktadır. Bu yöntem, aynı zamanda modelin daha hızlı ve daha verimli çalışmasına olanak tanır. Bu model, ImageNet veri seti üzerinde eğitilmiş olup, bir milyonun üzerinde görüntüden elde edilen geniş bir veri kümesinden öğrenim sağlamıştır. InceptionV3, toplamda 42 katman derinliğinde bir mimariye sahiptir ve 1.000 farklı kategoriye ayrılabilen sınıflandırma görevlerini başarıyla yerine getirebilmektedir.

InceptionV3, eğitim sürecini iyileştirmek ve aşırı uyumu önlemek için ara katmanlarına ek sınıflandırıcılar (Auxiliary Classifiers) yerleştirir. Bu sınıflandırıcılar, modelin ara katmanlarında ek düzenlemeler ve öğrenme gerçekleştirerek modelin genel performansını arttırmaktadırlar. Ancak, bu sınıflandırıcılardan elde edilen çıktılar, gerçek tahmin aşamasında kullanılmamaktadır; yalnızca eğitim sürecinde modelin genelleme yeteneğini geliştirmek amacıyla devreye girmektedirler.

InceptionV3, her biri aynı katmanda giriş verilerinin farklı yönlerini işlemek üzere tasarlanmış birden fazla Inception modülü içermektedir. Bu modüller, çeşitli boyutlardaki konvolüsyonel filtrelerden (örneğin 1x1, 3x3 ve 5x5 konvolüsyonlar) ve 3x3 maksimum havuzlama katmanlarından oluşmaktadır. Bu yapı, modelin farklı ölçeklerdeki özellikleri ve bilgileri yakalamasına olanak tanımaktadır. Küçük filtreler daha ince ayrıntıları yakalarken, daha büyük filtreler daha geniş kapsamlı bilgi özelliklerini öğrenmektedir. Bu modüler yapı sayesinde, model, çok çeşitli görüntü özelliklerini ve desenlerini etkili bir şekilde öğrenebilmektedir. Ağ, görüntü verilerini işlemek için 299x299 girdi boyutunu kabul etmektedir. InceptionV3 model mimarisi şekil 5.6'da görüldüğü gibidir.

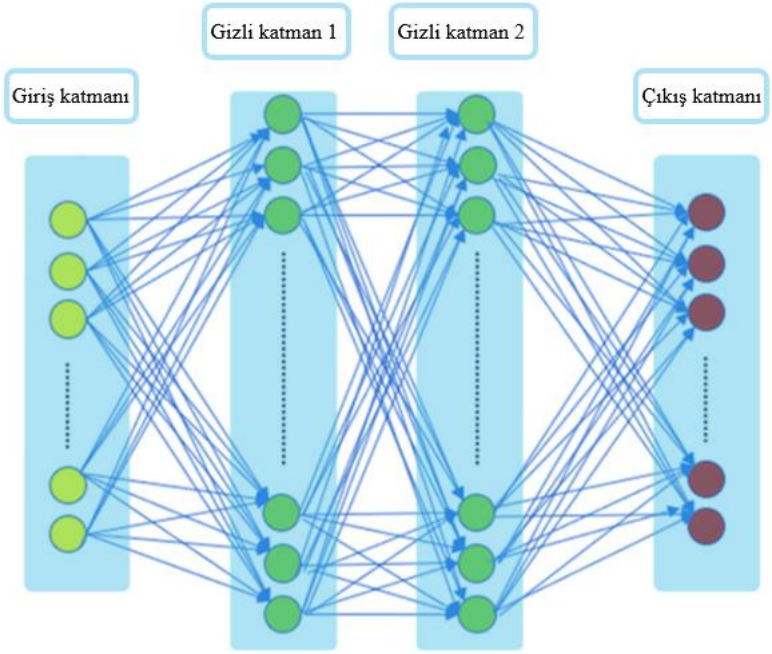


Şekil 5.6. InceptionV3 mimarisi

5.3.3. Fully connected deep neural network

Fully connected deep neural networks (Tam bağlantılı derin sinir ağları (FCDNN)), YSA'nın en temel ve yaygın kullanılan yapılarından biridir. Bu ağlarda, bir katmandaki her nöron, bir sonraki katmandaki tüm nöronlara bağlanır. Bu nedenle "tam bağlantılı" olarak adlandırılmaktadırlar. Bu özellik, ağın esnekliğini ve öğrenme kapasitesini artırırken, aynı zamanda veri setindeki karmaşık ilişkilerin öğrenilmesine olanak tanır.

FCDNN'ler, genellikle Giriş Katmanı, bir veya daha fazla Gizli Katman (Hidden Layers) ve Çıkış Katmanından (Output Layer) oluşur (Şekil 5.6). Giriş Katmanı; ağın ilk katmanıdır ve veri kümesindeki özelliklerin temsil edildiği yerdir. Bu katman, modelin aldığı ham veriyi işlemek için başlangıç noktasıdır. Gizli katmanlar; tam bağlantılı nöronlardan oluşur ve bu nöronlar bir önceki katmandaki tüm nöronlardan gelen sinyalleri alır. Gizli katmanlar, ağın karmaşık veri desenlerini öğrenmesini sağlar. Çıkış Katmanı ise ağın son katmanıdır ve modelin nihai tahminlerini veya çıktısını üretmektedir.



Şekil 5.7. FCDNN mimarisi

5.4. Model Performans Metrikleri

Performans ölçütleri bir sınıflandırma modelinin etkinliğinin kapsamlı bir değerlendirmesini sağlar (Jiao & Du, 2016). Bu çalışmada, performans kriterleri olan Doğruluk, Duyarlılık, Kesinlik ve F- Ölçütü değerleri; karmaşıklık matrisi ışığında, Tablo 5.1'de görüldüğü şekilde hesaplanan değerlerdir. Bu metriklerin birlikte kullanılması, modelin tahmin yeteneklerinin ve sağlamlığının kapsamlı bir şekilde değerlendirilmesini sağlamaktadır.

Tablo 5.1. Karmaşıklık matrisi

		Gerçek	
		Doğru	**Yanlış**
Tahmin Edilen	**Doğru**	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Yanlış	Yanlış Negatif (YN)	Doğru Negatif (DN)

DP, doğru olan verinin tahmin esnasında, doğru olarak sonuçlanması durumu, YP ise gerçekte yanlış olan verinin tahmin esnasında doğru sonuçlanması durumudur. YN; gerçekte doğru olan verinin tahmin esnasında yanlış olarak sonuçlanması durumu, DN ise gerçekte yanlış olan verinin tahmin esnasında yanlış olarak sonuçlanması durumudur.

Doğruluk; sınıflandırma modellerini değerlendirmek için temel bir ölçümdür ve doğru tahmin edilen örneklerin (hem gerçek pozitifler hem de gerçek negatifler) toplam örnekler içindeki oranını ifade etmektedir. Denklem 1'de görüldüğü gibi hesaplanır.

$$Doğruluk(\%) = 100 * \frac{DP+DN}{DP+YP+DN+YN} \tag{5.9}$$

Duyarlılık; toplam pozitif tahmin sayısına göre model tarafından yapılan doğru pozitif tahminlerin sayısını ölçen bir metriktir. Denklem 2'de görüldüğü gibi hesaplanır.

$$Duyarlılık = \frac{DP}{DP+YN} \tag{5.10}$$

Kesinlik; hassasiyet veya gerçek pozitif oran olarak da bilinir. Modelin doğru şekilde tanımladığı gerçek pozitif örneklerin oranını ölçer. Denklem 3'de görüldüğü gibi hesaplanır.

$$Kesinlik = \frac{DP}{DP+YP} \tag{5.11}$$

F- Ölçütü, Duyarlılık ve Kesinlik harmonik ortalamasıdır ve her iki kaygıyı dengeleyen tek bir metrik sağlar. Aşağıda Denklem 4'de görüldüğü şekilde hesaplanır:

$$F - Ölçütü = 2 * \frac{Duyarlılık*Kesinlik}{Duyarlılık+Kesinlik} \tag{5.12}$$

6. GEN DİZİLERİ KODLAMA YÖNTEMLERİ

DNA dizilerinin çeşitli yöntemler aracılığıyla kodlanması, biyoinformatik ve hesaplamalı biyoloji alanlarında büyük önem taşımaktadır. Kodlama yöntemleri, biyolojik verilerin, özellikle de genetik dizilerin analizinde matematiksel ve istatistiksel tekniklerin uygulanmasını kolaylaştırmaktadır. Bu sayede, genetik verilerin daha etkili bir şekilde işlenmesi ve yorumlanması mümkün hale gelmektedir.

Genetik dizilerin uzun ve karmaşık yapılarından dolayı ham halleriyle analiz edilmesi, büyük veri setlerinin işlenmesini ve yorumlanmasını zorlaştırabilmektedir. Büyük ölçüde zaman ve kaynak gerektiren işlemlerle karşılaşılır. Bu zorlukların üstesinden gelmek için DNA dizilerinin belirli formatlarda kodlanması, veri işleme sürecini hızlandırmakta ve hesaplama verimliliğini artırmaktadır. Kodlama yöntemleri, bu dizilerin daha küçük ve yönetilebilir bir biçimde temsil edilmesini sağlar. Böylece, dizi karşılaştırması, örüntü tanıma, sınıflandırma ve diğer analizler için daha verimli algoritmaların uygulanması mümkün hale gelir. Bu yöntemler, özellikle büyük veri setlerinin kullanıldığı genetik araştırmalar ve biyoinformatik projelerinde önemli avantajlar sunar. Buna paralel olarak, bu veri setlerinin verimli bir şekilde analiz edilmesi için MÖ ve DÖ algoritmaları giderek daha fazla tercih edilmektedir. Bu bağlamda, gen fonksiyonlarının tahmini, motiflerin tanımlanması ve organizmaların genetik yapılarına göre sınıflandırılması gibi görevlerde etkin bir şekilde kullanılmaktadırlar. Gen dizisi kodlama yöntemleri, verilerin MÖ ve DÖ algoritmaları ile yapılan sınıflandırma uygulamalarına uygun formatlarda sunulmasını sağlar. Bu uygulamalar, genetik, evrimsel biyoloji ve hastalık mekanizmalarının daha iyi anlaşılması açısından kritik bir rol oynar.

Kodlama yöntemleri, genetik dizilerin görselleştirilmesine de olanak tanır. Ham gen dizilerindeki saklı bilgiler, anlaşılamayan kalıplar ve tekrarlar, kodlama sayesinde açığa çıkarılabilmektedir. Bu görselleştirme teknikleri, biyologlar ve genetikçiler için değerli içgörüler sağlayarak genetik verilerin daha kolay yorumlanmasına yardımcı olmaktadır.

Ayrıca, genetik dizilerin kodlanması, verilerin sıkıştırılmasına ve depolama alanının optimize edilmesine katkıda bulunur. Büyük ölçekli genetik veri tabanları, bu sayede daha az yer kaplar ve veri yönetimi daha verimli hale gelir. Kodlama şemaları,

verilerin gerektiğinde orijinal ve yeni formatlarına dönüştürülmesini kolaylaştırır, bu da analiz süreçlerinde esneklik ve verimlilik sağlar.

DNA dizilerinin kodlanması, biyoinformatik ve hesaplamalı biyoloji alanlarında temel bir araç olarak kabul edilir. Bu teknikler, genetik verilerin daha kolay analiz edilmesini, yorumlanmasını ve görselleştirilmesini mümkün kılar (Gunasekaran vd., 2021). Genetik dizilerin kodlanması, verilerin daha etkili bir şekilde işlenmesine olanak tanıyarak bu alandaki araştırmaların hızını ve doğruluğunu artırır. Böylece, genetik verilerden elde edilen bilgiler, bilimsel araştırmalar ve klinik uygulamalar için değerli içgörüler sunar. Bu süreçler, genetik bilgilerin matematikçiler, bilgisayar bilimcileri, istatistikçiler ve biyoloji dışındaki alanlarda çalışan araştırmacılar tarafından da erişilebilir ve anlaşılabilir hale gelmesini sağlayarak disiplinler arası iş birliğini güçlendirir.

Literatürde birçok farklı türde gen dizisi kodlama yöntemi sunulmuştur. En genel haliyle gösterim şekillerine göre üç sınıfa ayrılır:

1. Sayısal gösterim yöntemleri
    - Etiket kodlama
    - Tek-sıcak kodlama
    - K-Mer kodlama
2. Grafiksel gösterim yöntemleri
    - Kaos oyun gösterimi (CGR (Chaos game representation))
    - Frekans kaos oyun gösterimi (FCGR (Frequency chaos game representation))
    - DNA yörünge görüntüleri (DNAWalk)
3. Görüntü gösterim yöntemleri
    - DNA Gri ölçekli ve Renkli görüntüler

6.1. Sayısal Gösterim Yöntemleri

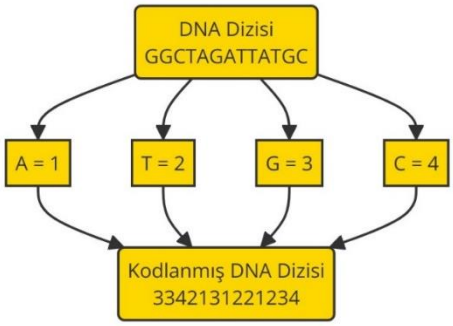
Kategorik gen dizilerinin sayısal gösterim yöntemleriyle temsil edilmesi, biyoinformatik ve genetik araştırmalar için büyük bir önem taşımaktadır. Bu teknikler, DNA ve RNA dizilerinin karmaşıklığını ve özelliklerini daha anlaşılır ve analiz edilebilir hale getirmektedir. Gen dizileri, doğaları gereği kategorik verilerdir ve doğal formda oldukça karmaşık ve uzun olabilmektedir. Sayısal gösterim yöntemleri, bu dizileri daha basit ve anlaşılır formatlara dönüştürerek analiz sürecini hızlandırmaktadır. Bu dönüşüm,

dizilerin çeşitli algoritmalar ve analiz araçları tarafından etkin bir şekilde işlenmesini sağlamaktadır.

MÖ ve DÖ modelleri, gen dizilerini işleyebilmek için sayısal gösterimlere ihtiyaç duymaktadır. Bu modeller, genetik verilerdeki belirli desenleri ve ilişkileri tespit ederek sınıflandırma ve tahmin yapma işlemlerini gerçekleştirmektedir. Sayısal gösterim sayesinde, biyoinformatik araçlar ve algoritmalar, büyük genomik veri setleri üzerinde daha doğru ve verimli sonuçlar elde edebilmektedir. Bu süreç, genetik araştırmaların hızlanmasına ve daha derinlemesine analizlerin yapılmasına olanak tanımaktadır. Literatürde en çok kullanılan yöntemler; Etiket kodlama (Label encoding), Tek-sıcak kodlama (One-hot encoding), K-Mer kodlama şeklindedir.

6.1.1. Etiket kodlama (Label encoding)

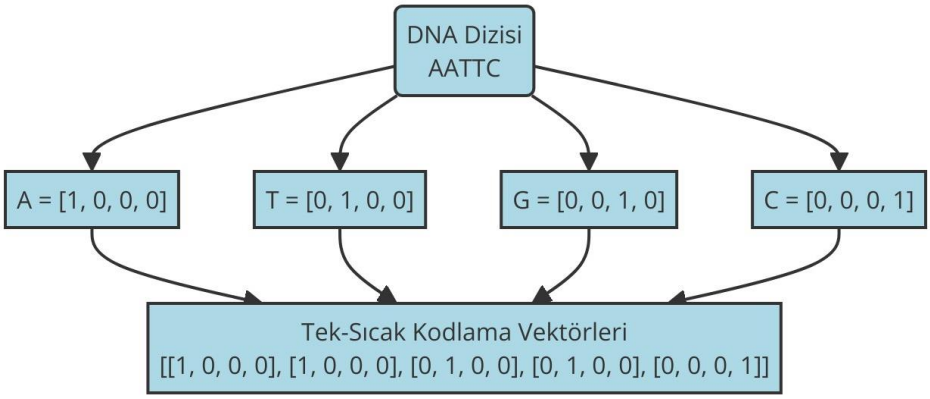
Etiket kodlama yöntemi, biyoinformatik ve genetik araştırmalarda DNA dizilimlerinin daha etkin ve anlaşılır bir şekilde analiz edilmesini sağlamak için kullanılan bir kodlama tekniğidir. Bu yöntem, özellikle DNA dizilimlerinde nükleotidlerin veya nükleotid dizilerinin belirli etiketlerle veya kodlarla temsil edilmesine dayanır. Tanımlanan nükleotid dizilerine karşılık gelen etiketler veya kodlar belirlenmektedir. Tanımlanan etiketler ve kodlar kullanılarak bir kodlama şeması oluşturulmaktadır. Bu şema, belirli nükleotid dizilerinin hangi etiketlerle veya kodlarla temsil edileceğini tanımlamaktadır. Kodlama şemasına uygun olarak, DNA dizilimleri etiketlerle veya kodlarla kodlanmaktadır. Yaygın olarak Etiket Kodlamasında Şekil 6.1'de görüldüğü üzere; A=1, C=2, G=3, T=4 olacak şekilde tüm veriler sayısallaştırılmaktadır. Bu şekilde, DNA dizilimleri daha basit ve anlaşılır bir biçimde temsil edilmektedir.



Şekil 6.1. Etiket kodlama yönteminin uygulanışı

6.1.2. Tek-sıcak kodlama (One-hot encoding)

Tek-sıcak kodlama, DNA dizilerini MÖ algoritmaları tarafından verimli bir şekilde işlenebilen sayısal bir formatta temsil etmek için biyoinformatikte yaygın olarak kullanılan bir tekniktir. Dört nükleotidi ikili bir matris formatına çevirir, bu da dizi analizi için çeşitli hesaplama yöntemlerinin uygulanmasını kolaylaştırmaktadır (Potdar vd., 2017). Tek-sıcak kodlamanın temel prensibi, her nükleotid için ikili bir vektör oluşturmaya dayanmaktadır; burada belirli bir nükleotidin varlığı, ilgili konumda '1' ve diğer tüm konumlarda '0' ile gösterilir. Örneğin gen dizisi; 4 × L matrisine dönüştürülüp; A = [1, 0, 0, 0], T/U = [0, 1, 0, 0], G = [0, 0, 1, 0] ve C = [0, 0, 0, 1] olarak gösterilebilir. Bu dönüşüm, DNA dizisinin, sayısal girdi verilerine ihtiyaç duyan çeşitli MÖ modellerinde kullanılabilmesini mümkün kılar. Nükleotid dizilerini sayısal bir formata dönüştürerek, dizi hizalaması, motif bulma ve gen ekspresyonu analizi gibi görevleri gerçekleştirmek için çeşitli hesaplamalı algoritmaların uygulanmasını sağlamaktadır. Ayrıca, bu kodlama yönteminin uygulanması kolaydır ve hesaplama açısından verimlidir, bu da onu büyük ölçekli genomik veri kümeleri için uygun hale getirmektedir.



Şekil 6.2. Tek-sıcak kodlama yönteminin uygulanışı

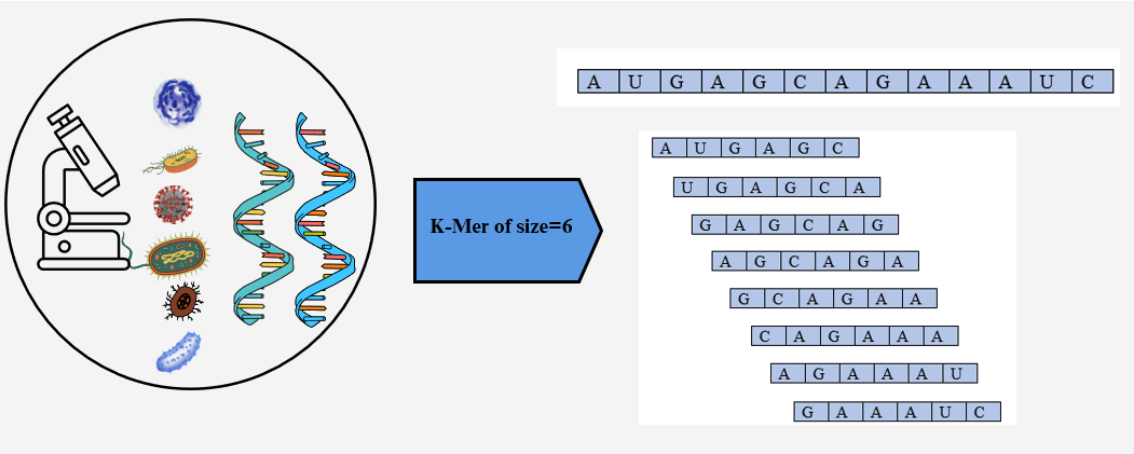
6.1.3. K-Mer kodlama

K-Mer kodlama yönteminde; belirlenen k değerine göre bir pencere şeklinde kelimeler oluşturulmaktadır ve DNA dizisi bazı ifadelere dönüştürülmektedir. Bu yöntem, metin sınıflandırma teknikleri kullanılarak analiz yapılmasını sağlamaktadır. k değerine göre sonuçlar değişiklik gösterebilmektedir. K-Mer kodlama yönteminde, belirlenen k değerine göre DNA dizisi bölümlere ayrılır ve bu bölümler "kelimeler"

olarak adlandırılır. Her bir kelime, belirli bir uzunlukta nükleotid dizisinden oluşmaktadır. Filogenetik uygulamalarda K-Mer yöntemi ilk olarak Blaisdell tarafından (Blaisdell, 1986, 1989) kullanılmıştır. Temel olarak, k değerine bağlı olarak oluşturulan kelimelerin frekansları tespit edilmektedir. Bu frekanslar, her bir kelimenin DNA dizisinde kaç kez tekrarlandığını göstermektedir. Dolayısıyla, k değerine göre frekans vektörü değişkenlik gösterebildiğinden, sonuçlarda da değişiklik olmaktadır (Compeau vd., 2011).

Genetik veri setlerinde yer alan gen dizileri, çok farklı nükleotid uzunluklarında olabilmektedir. Özellikle MÖ ve DÖ ile sınıflandırma yaklaşımlarında, dizi verilerini sadece kodlamak yeterli olmamaktadır. Tek-Sıcak Kodlama ve Etiket Kodlama gibi yöntemler, dizilerin hizalı olmasını gerektiren kodlama yöntemleridir. Yani, dizilerin nükleotid uzunlukları birbirinden farklı olduğunda, Tek-Sıcak Kodlama ve Etiket Kodlama gibi yöntemler uygulandığında ortaya çıkan kodlanmış yeni veri seti de değişken nükleotid uzunluğuna sahip olmaktadır. Bu nedenle, K-Mer kodlama yöntemi hizalama gerektirmemesi açısından avantaj sağlamaktadır. Şekil 6.3'de örnek bir dizi üzerinde k=6 için bir K-Mer kümesinin nasıl elde edildiği gösterilmektedir. DNA dizisi altı nükleotid uzunluğundaki kelimelere bölünmekte ve her bir kelimenin sayısı hesaplanmaktadır. Bu kelimelerden her birinin sayısını içeren sabit uzunluklu bir frekans vektörü üretilmektedir. Frekans vektörü, dizi uzunluğuna bakılmaksızın veri setindeki tüm gen dizileri için eşit olmaktadır. Bu şekilde, farklı nükleotid sayısına sahip gen dizilerinin kodlanması ve eşit boyutlara getirilmesi sağlanmaktadır. Ayrıca, k değerinin doğru seçimi kritiktir; çok küçük K-Mer'ler belirsizliklere yol açabileceğinden ve çok büyük K-Mer'ler yeterli dizi varyasyonunu yakalayamayacağından, uygun k değerinin belirlenmesi önem arz etmektedir (Bussi vd., 2021; Moeckel vd., 2024).

K-Mer yöntemi, biyoinformatik ve genetik araştırmalarda güçlü bir araç olarak kullanılmaktadır. DNA dizilimlerinin daha etkin ve anlaşılır bir şekilde analiz edilmesini sağlamaktadır. Özellikle büyük ölçekli genomik verilerin işlenmesinde ve analiz edilmesinde önemli avantajlar sunmakta ve yüksek verimlilik, doğruluk sağlamaktadır.



Şekil 6.3. K-Mer kodlama yönteminin uygulanışı

6.2. Grafiksel Gösterim Yöntemleri

Uzun dizilerdeki gizli kalıpları bulmak hem zor hem de değerli olabilmektedir. Grafiksel gösterim yöntemleri genetik verilerin daha anlaşılır, analiz edilebilir ve karşılaştırılabilir hale gelmesini sağlamaktadır. Bu yöntemler, dizileri grafiksel formatlarda sunarak, karmaşık biyolojik bilgilerin daha kolay yorumlanmasını sağlamaktadır. Genetik dizilerdeki tekrarlayan motifler ve düzenli desenler, biyolojik işlevler açısından büyük önem taşımaktadır. Grafiksel gösterimler, bu tür desenlerin ve motiflerin görsel olarak tanımlanmasını ve analiz edilmesini kolaylaştırmaktadır. Grafiksel gösterim, farklı genetik dizilerin karşılaştırmalı analizini kolaylaştırmaktadır. Farklı türler veya bireyler arasındaki genetik benzerlikler ve farklılıklar, görsel olarak daha kolay belirlenebilmekte ve analiz edilebilmektedir. Bu, evrimsel biyoloji ve filogenetik çalışmalar için özellikle önemlidir. Literatürde en çok kullanılan yöntemler; CGR, FCGR ve DNAWalk şeklindedir.

6.2.1. Kaos oyun gösterimi (Chaos game representation (CGR))

CGR, genetik dizilerin karmaşıklığını ve düzenliliğini analiz etmek için kullanılan etkili bir tekniktir. Bu yöntem ilk defa 1990'ların başında tanıtılmıştır (Jeffrey, 1990). Bu teknik devrim niteliğinde olmuştur. Çünkü kaos teorisi ve genetik kavramlarını bir araya getirilerek genetik dizilerin devasa, karmaşık yapıları, 2 boyutlu bir uzayda temsil edilerek, benzersiz bir içerik oluşmaktadır. Ayrıca bu yöntem, diziler arasındaki benzerlikleri ve farklılıkları belirlemek için kullanılabilecek, evrimsel ilişkileri veya

işlevsel benzerlikleri tespit ederek bu alanda ilerlemelerin gerçekleştirilmesine olanak sağlamıştır.

CGR, genetik dizilerin belirli özelliklerini vurgulayarak, DNA dizilerinin karmaşıklığını ve tekrarlayan motiflerini görselleştirmeye olanak tanımaktadır. CGR'nin temel prensibi, genetik dizileri bir nokta kümesi olarak temsil etmektir. Bu noktalar, belirli bir kurala göre oluşturulmaktadır. Örneğin, DNA dizilerindeki bazları temsil etmek için A, C, G ve T harfleri kullanılmaktadır. Her baz, bir koordinat düzlemindeki bir noktaya karşılık gelmektedir. CGR, bu noktaları oluşturmak için iteratif bir süreç kullanmaktadır.

Bu yöntemin temel adımları ve formülü aşağıda açıklanmıştır.

Başlangıç:

- Bir birim kare ile başlanılır.
- Karenin dört köşesi dört nükleotid (A, C, G, T) ile eşleştirilir. Örneğin:

A = (-1, 1)

T = (1, -1)

C = (-1, -1)

G = (1, 1)

İteratif Noktalama:

- Karenin ortasından başlanılır (başlangıç noktası, P_0); $P_0 = (0,0)$
- Dizideki her nükleotid için, yeni bir nokta, önceki nokta ile nükleotidi temsil eden köşe arasında belirli bir oranda yerleştirilir.

Formül:

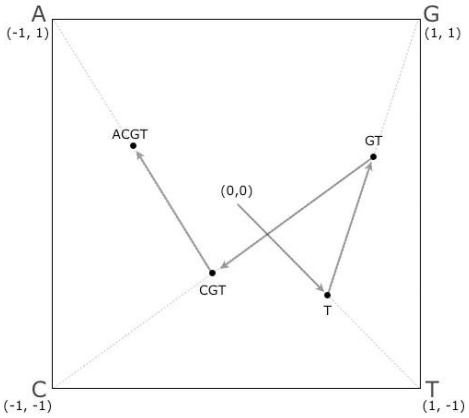
P_n mevcut nokta ve P_{n+1} sonraki nokta olmak üzere; P_{n+1} denklem 6.1'de görülmektedir.

$$P_{n+1} = (\frac{x_n + x_x}{2}, \frac{y_n + y_x}{2}) \quad (6.1)$$

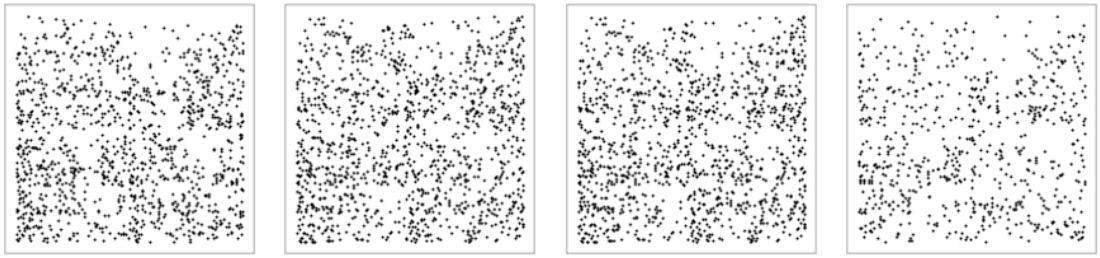
(x_n, y_n) P_n'nin koordinatları ve (x_x, y_x) X (A, C, G, T' den biri) nükleotidine karşılık gelen köşenin koordinatlarıdır.

Şekil 6.4'te örnek bir DNA dizisi "ACGT" için nokta koordinatları ve Şekil 6.5'te PhyVirus veri setindeki dört farklı Arena virüs gen dizilerine ait CGR noktalarının

görüntüleri gösterilmektedir. Şekil 6.4'te örnek bir DNA dizisi olan "ACGT" için; $P_0 = (0,0)$'dır. A, (-1, 1) noktasına atanmıştır. Bir sonraki nokta; $P_1 = (\frac{0+(-1)}{2}, \frac{0+1}{2})$ = (-0.5,0.5) olarak hesaplanır. C, (-1, -1) noktasına atanmıştır. $P_2 = (\frac{-0.5+(-1)}{2}, \frac{0.5+(-1)}{2})$ = (-0.75,-0.25) olarak hesaplanır. G, (1, 1) noktasına atanmıştır. $P_3 = (\frac{-0.75+1}{2}, \frac{0.25+1}{2})$ = (-0.125,-0.375) bulunur. T, (1, -1) noktasına atanmıştır. Bir sonraki nokta; $P_4 = (\frac{0.125+1}{2}, \frac{0.375+(-1)}{2})$ = (-0.5625,-0.3125) olarak hesaplanır.



Şekil 6.4. 'ACGT' sekansının CGR noktaları ile gösterimi

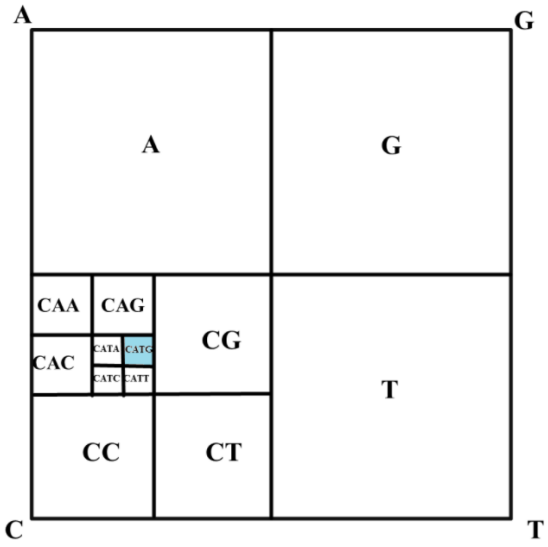


Şekil 6.5. PhyVirus veri setindeki dört farklı Arena virüs gen dizilerine ait CGR noktalarının gösterimi

6.2.2. Frekans kaos oyun gösterimi (FCGR (Frequency chaos game representation))

FCGR yöntemi, DNA dizilerini görselleştirmek ve analiz etmek amacıyla kullanılan etkili bir tekniktir. FCGR, DNA dizilerini belirli kurallara göre iki boyutlu bir matris şeklinde temsil ederek, dizilerin özelliklerini vurgulamaktadır.

Şekil 6.6' da 'CATG' oligomerinin konumu; her biri bir nükleotide karşılık gelen dört çeyreğe bölünmüş bir kare üzerinde gösterilmektedir. Karenin merkezinden başlanılmakta ve dizilimdeki her nükleotid için ilgili köşeye doğru yarıya kadar hareket edilerek yol işaretlenmektedir. Örneğin, C için merkezden sol alt köşeye doğru yarıya kadar ilerlenmekte ve bir nokta çizilmektedir. A için mevcut konumdan sol üst köşeye doğru yarıya kadar ilerlenmekte ve bir nokta çizilmektedir. T için geçerli konumdan sağ alt köşeye doğru yarıya kadar ilerlenmekte ve bir nokta çizilmektedir. G için mevcut konumdan sağ üst köşeye doğru yarıya kadar ilerlenmekte ve bir nokta çizilmektedir. Elde edilen nokta, 'CATG' oligomerinin konumudur.



Şekil 6.6. CATG oligomerinin konumu

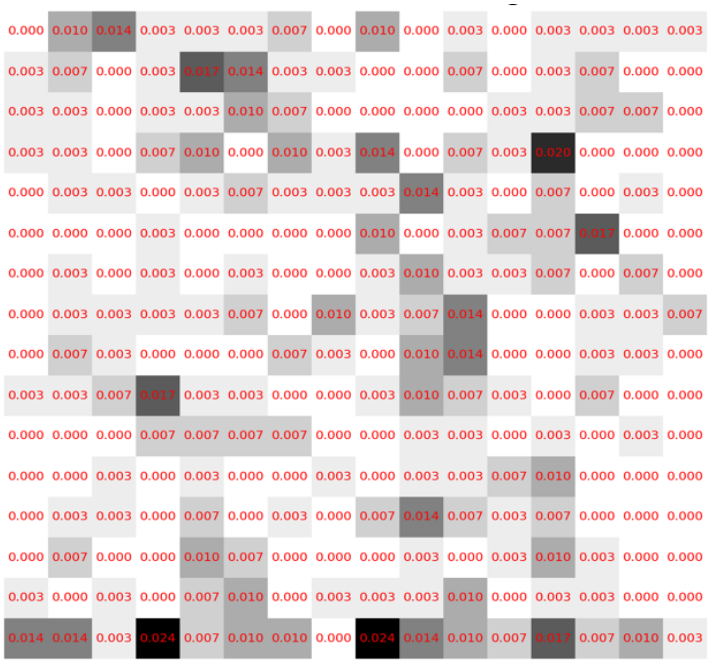
Dizideki her nükleotid için noktalar işaretlendikten sonra, karede benzersiz bir desen elde edilmektedir. FCGR yönteminde, bir DNA dizisi için K-Mer boyutu (k) belirlenmektedir. Benzersiz K-Mer'lerin sayısı (N) ise 6.2'de gösterilen formül ile hesaplanmaktadır:

$$N = 4^k \tag{6.2}$$

Bu K-Mer'leri temsil eden resmin boyutları, FCGR yönteminin iki boyutlu temsili nedeniyle kare matris olarak tanımlanmaktadır. Dolayısıyla, kare matrisin her bir kenarının uzunluğu (S), 6.3'de gösterilen formül ile hesaplanmaktadır:

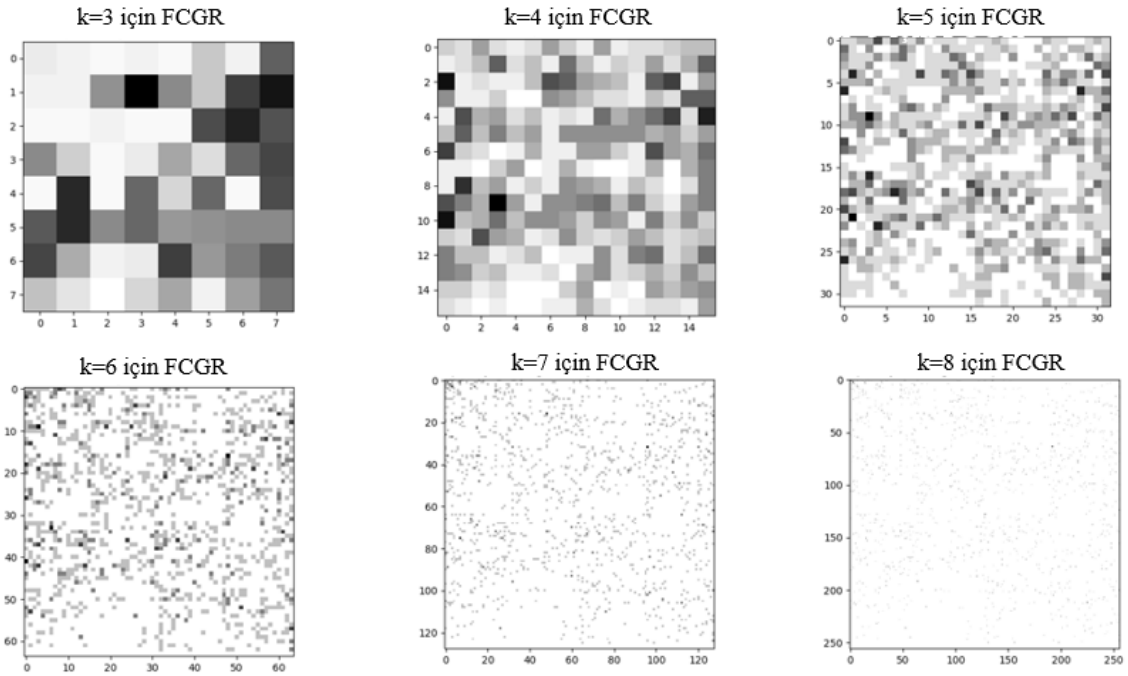
$$S = \sqrt{4^k} \tag{6.3}$$

Daha uzun DNA dizileri için; öncelikle DNA dizisinin tüm K-Mer'leri elde edilir. Ardından bu K-Mer'lerin frekansı hesaplanır. Her K-Mer'in frekansı iki boyutlu bir matrise yerleştirilir. FCGR matrisinin boyutları (S x S), K-Mer boyutuna bağlıdır. Şekil 6.7'de PhyVirus veri setindeki Arena virüsüne ait gen diziliminden elde edilen 4-Mer için FCGR görüntüsü yer almaktadır. k=4 için $S = \sqrt{4^4}$ formülünden 16x16 boyutunda bir matris elde edilmiştir. Elde edilen K-Mer'lerin görülme sıklıkların yani frekans değerlerine göre görüntüde koyu ve açık renkli olarak gösterilmiştir.



Şekil 6.7. PhyVirus veri setindeki Arena_L_1_981 virüsüne ait gen diziliminden elde edilen FCGR görüntüsündeki 4-Mer için frekans değerleri

Farklı K-Mer değerlerinde farklı boyutlarda matrislerde elde edilmekte ve gen dizilerine göre birbirinden tamamen farklı olan ayırt edici görüntüler oluşmaktadır. Şekil 6.8'de PhyVirus veri setinde Arena virüse ait aynı gen dizisinin k'nın 3, 4, 5, 6, 7 ve 8 değerleri için oluşan FCGR görüntülerine yer verilmiştir.



Şekil 6.8. Farklı k değerleri kullanılarak aynı Arena virüsü gen dizisinden elde edilen FCGR görüntüleri

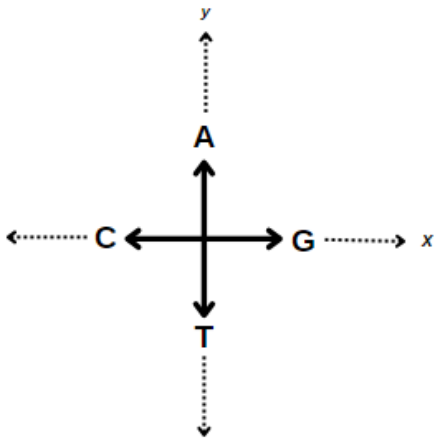
Uygulamada FCGR, daha ayrıntılı ve bilgilendirici görüntüler oluşturmak için daha uzun DNA dizileriyle birlikte kullanılmaktadır. Bu yöntem, büyük miktarda genetik bilgiyi kompakt ve anlaşılır bir şekilde temsil ederek, biyoinformatik araştırmalarda önemli avantajlar sağlamaktadır.

6.2.3. DNA yörünge görüntüleri (DNAWalk (DNA yürüyüşü))

DNAWalk yöntemi, DNA dizilerinin yapısal özelliklerini görselleştirmek ve analiz etmek için kullanılan güçlü bir tekniktir (Peng vd., 1992). Genellikle DNAWalk (DNA Yürüyüşü) olarak adlandırılan bu yöntem, bir DNA dizisinin iki, üç ve hatta daha fazla boyutlu bir uzaya haritalanarak grafiksel bir temsilini sağlamaktadır. DNAWalk yöntemi, araştırmacıların nükleotidlerin doğrusal düzenini uzamsal bir yörüngeye dönüştürerek genetik dizilerin karmaşık kalıplarını ve özelliklerini anlamalarına yardımcı olmaktadır (Arakawa vd., 2009).

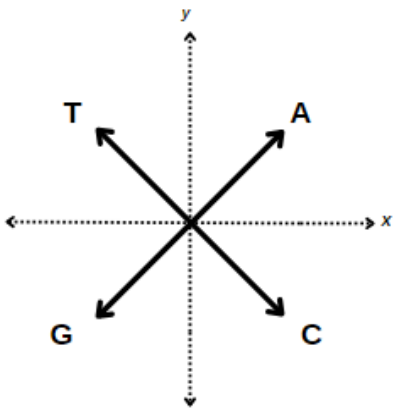
DNAWalk yönteminin temel ilkesi, bir DNA dizisindeki her nükleotide bir koordinat sisteminde belirli bir yön veya hareket atamaktır. Tipik olarak, dört nükleotid-A, C, G ve T farklı vektörlerle eşleştirilmektedir (Peng vd., 1992). Bir başlangıç noktasından başlayarak, genellikle iki boyutlu bir düzlemde, dizideki her nükleotid

sırayla okunur ve pozisyon atanan yöne göre güncellenir. Bu iteratif süreç, tüm DNA dizisini temsil eden bir yol veya yörünge ile sonuçlanır. İlk versiyonunda orjin (0,0) noktasından başlamak üzere, A (0,1), C (-1,0), G (1,0), T (0,-1) şeklinde bir yön belirlenmiştir.



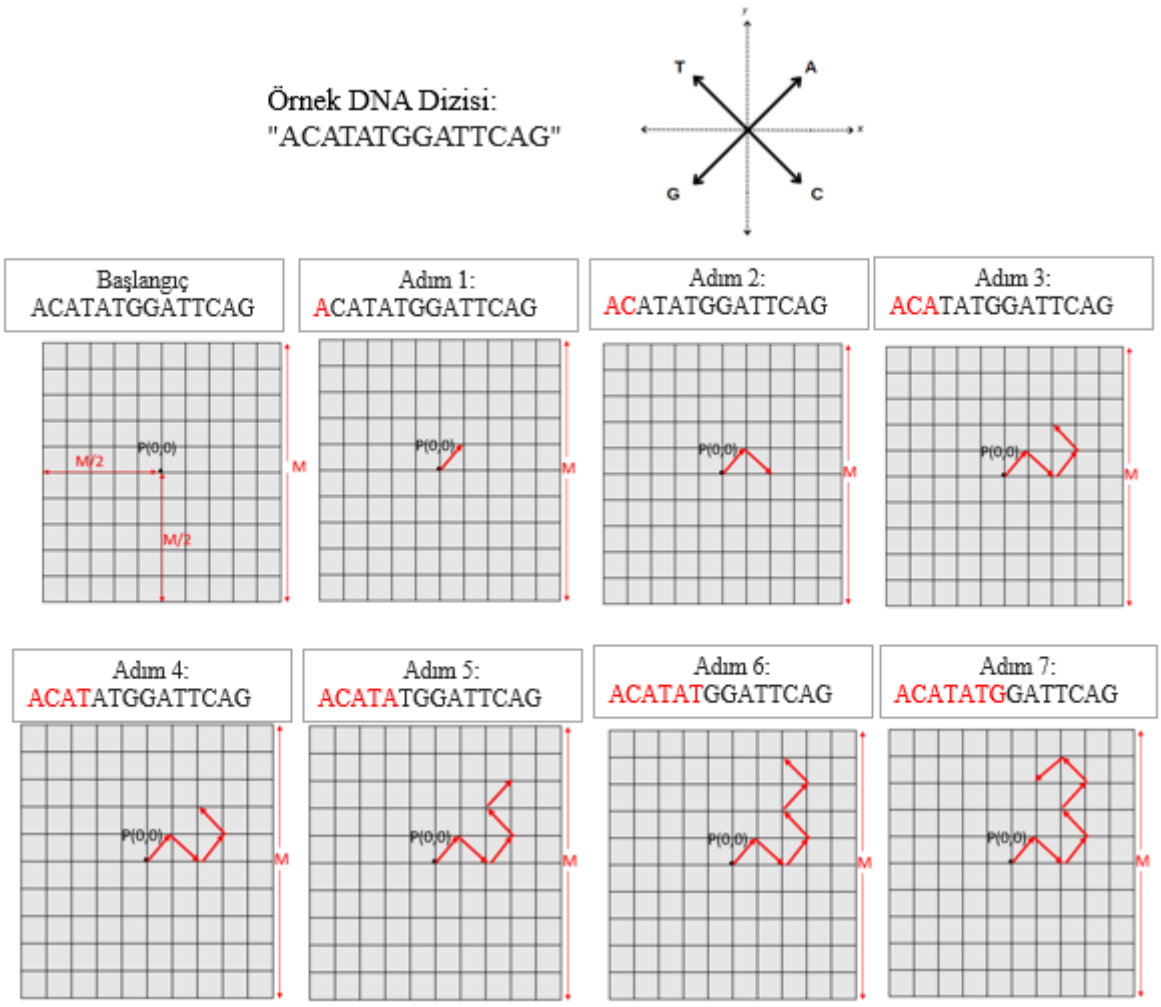
Şekil 6.9. Peng ve arkadaşlarının DNA Walk yönteminde kullandıkları vektörler (Peng vd., 1992)

Literatürde daha sonra birçok farklı gösterim olmuştur. En yaygın yaklaşımlardan biri, Kobori ve Mizuta tarafından gerçekleştirilen çalışmada kullanılmıştır (Kobori & Mizuta, 2016). Buna göre nükleotid vektör gösterimi A (1,1), C (1,-1), G (-1,-1), T (-1,1) şeklinde belirlenmiştir (Şekil 6.9).



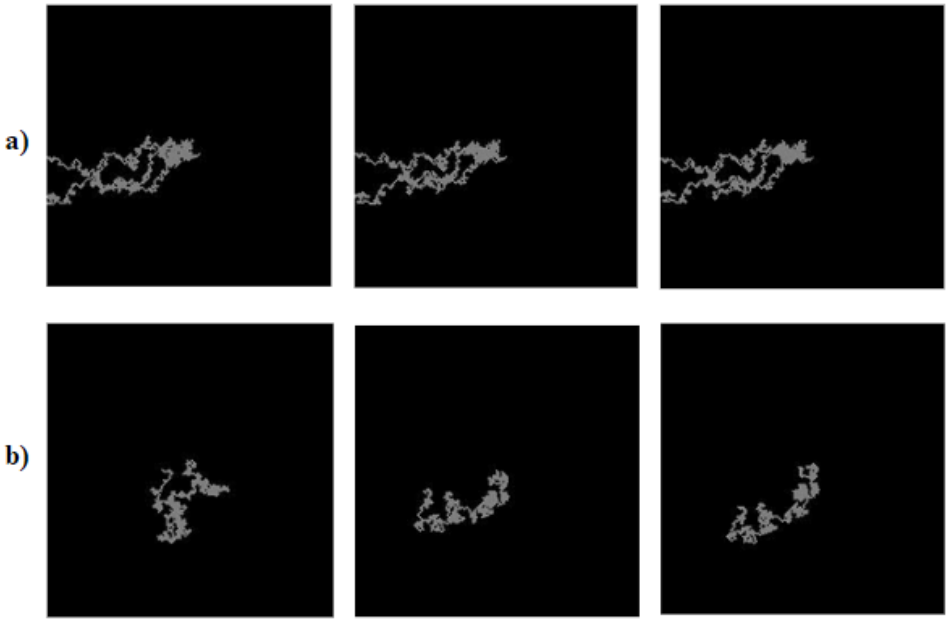
Şekil 6.10. Kobori ve Mizuta'nın DNAWalk yönteminde kullandıkları vektörler (Peng vd., 1992)

Şekil 6.11' de örnek bir gen dizisine (ACATATGGATTCAG) DNAWalk yönteminin adım adım uygulanması gösterilmektedir.



Şekil 6.11. Örnek bir gen dizisine (ACATATGGATTCAG) DNAWalk yönteminin adım adım uygulanması

Şekil 6.12'de PhyVirüs verisetinde yer alan Flavi virüs ailesine ve Arena virüs ailesine ait üç farklı virüs dizisinin Kobori ve Mizuta'nın kullandığı vektör gösterimiyle elde edilen görüntüler yer almaktadır.



Şekil 6.12. PhyVirus veri setinde yer alan **a)** Flavi virüs ailesine **b)** Arena virüs ailesine ait üç farklı virüs dizisinin Kobori ve Mizuta'nın kullandığı vektör gösterimiyle elde edilen görüntüler

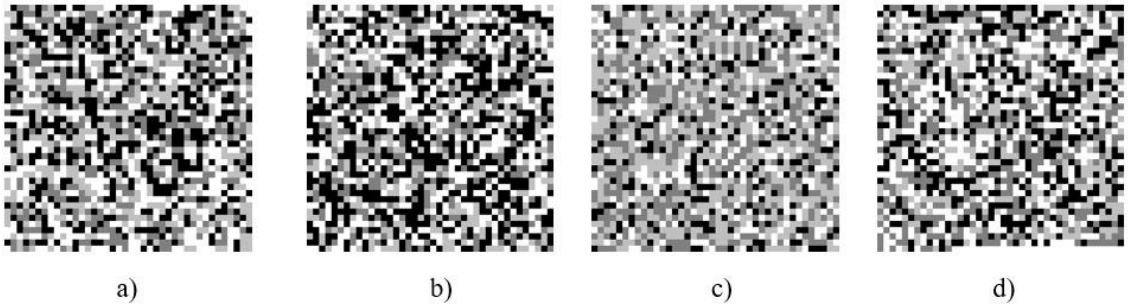
6.3. Görüntü Gösterim Yöntemleri

DNA dizilerinin görüntü formatına dönüştürülmesi, biyoinformatik araştırmalarında yeni ufuklar açmıştır. Bu görselleştirme yöntemleri, bilgisayarla görme gibi diğer alanlarda yaygın olarak kullanılan görüntü işleme tekniklerinin ve ML, DÖ modellerinin biyoinformatik verilerine uygulanmasına olanak tanımaktadır. Yüksek verimli dizileme teknolojilerinin sunduğu büyük verileri kullanarak, tüm genomları görüntüler şeklinde temsil etmek ve bu görüntüleri DÖ ve MÖ teknikleri ile analiz etmek mümkün hale gelmiştir. DNA gen dizilerini gri tonlamalı veya renkli görüntüler olarak görsel temsillere dönüştürmek; genetik materyal içerisindeki karmaşık yapıları ve desenleri anlamak için basit ve kolay bir yöntem olarak öne çıkmaktadır. Bu yaklaşım, sadece verilerin görsel olarak daha anlaşılır hale gelmesini sağlamakla kalmaz, aynı zamanda sınıflandırma, anormali tespiti ve evrimsel analiz gibi çeşitli görevlerde MÖ ve DÖ algoritmalarının kullanımını da kolaylaştırmaktadır.

6.3.1. DNA Gri ölçekli ve Renkli görüntüler

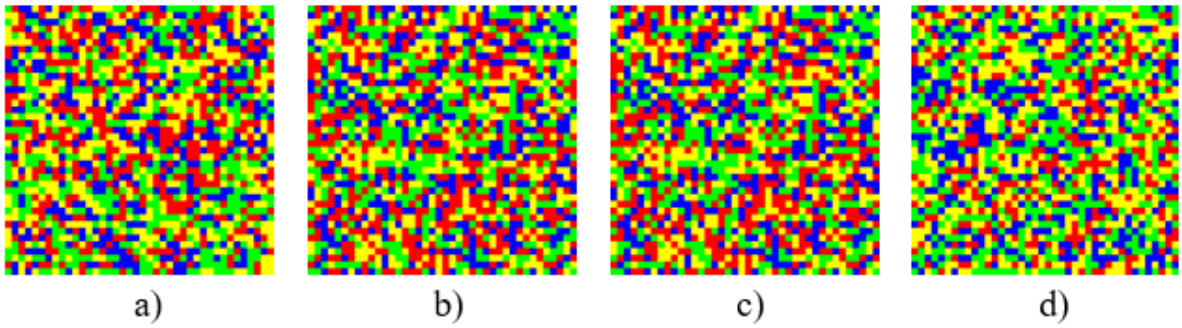
Gri ölçekli görüntü dönüştürme yönteminde, dizideki her nükleotid veya dinükleotid çiftine belirli bir gri değer atanmaktadır. Örneğin, her bir nükleotide 0 ile 255

arasında bir değer atanır. Burada 0 siyahı ve 255 beyazı temsil ederken, aralarında çeşitli gri tonları da yer almaktadır. Bu dönüştürme işlemi, her bir elemanın bir nükleotide veya bir çift nükleotide karşılık geldiği bir matrisle sonuçlanmakta ve matris daha sonra gri ölçekli bir görüntü olarak görselleştirilmektedir. Şekil 6.13'te PhyVirus veri setindeki farklı ailelere ait virüs dizileri; Gri ölçekli görüntüler şeklinde gösterilmektedir. Gri ölçekli görüntülerin avantajı, üretilmelerinin nispeten basit olması ve renkli görüntülere kıyasla daha az hesaplama gücü gerektirmesidir (Santamaría vd., 2019). Buna ek olarak, gri ölçekli görüntüler yüksek ayrıntı seviyesini korumakta ve bu da onları çeşitli analitik görevler için uygun hale getirmektedir.



Şekil 6.13. Phyvirus veri setinde **a)** Calici **b)** Corona **c)** Toga **d)** Rhapdo virüs ailelerine ait gen dizi örneklerinin Gri ölçekli görüntüler yöntemi ile gösterilmesi

Gri ölçekli görüntülerin ötesinde, DNA dizileri renkli görüntülere dönüştürülerek daha da zengin bir görsel temsil sağlanabilmektedir. Bu yaklaşımda, her nükleotid veya bir grup nükleotid belirli bir renkle eşleştirilmektedir. Örneğin, A için kırmızı, T için yeşil, C için mavi ve G için sarı renkleriyle temsil edilen bir model kullanılabilmektedir. Ortaya çıkan görüntü, dizinin bileşenlerini görsel olarak farklılaştıran renkli bir matristir. Şekil 6.14'te PhyVirus veri setindeki farklı ailelere ait virüs dizileri renkli görüntüler şeklinde gösterilmektedir. Renkli görüntüler, mutasyonları tanımlamada, anormallikleri tespit etmede ve evrimsel ilişkileri incelemede özellikle yararlı olabilecek nükleotidler arasında daha karmaşık desenleri ve korelasyonları yakalama avantajı sunmaktadır.



Şekil 6.14. PhyVirus veri setinde a) Calici b) Corona c) Toga d) Rhapdo virüs ailelerine ait gen dizi örneklerinin Renkli görüntüler yöntemi ile gösterilmesi

7. VİRÜS AİLELERİNE ve KONAKLARINA DAYALI SINIFLANDIRMA

Virüs ailelerine (VF) ve virüs konaklarına (VC) dayalı sınıflandırma uygulaması için, kapsamlı analizler gerçekleştirilerek, büyük filogenetik gen dizi veri kümelerinden biri olan PhyVirus veri seti ile testler gerçekleştirilmiştir. Önerilen metodun akış şeması şekil 7.1' de gösterilmiştir.

1. Aşama: RNA virüslerinin genetik materyallerinin çeşitli kimyasal ve fiziksel yöntemlerle elde edildikten sonra A, C, G ve T nükleotid dizileri şeklinde elde edilmektedir.

2. Aşama: Elde edilen nükleotid dizileri, virüs familya ve konak bilgileri korunarak biyoinformatik araçlar ile okunabilir ham FASTA formatında yer almaktadır. Gen dizi görüntüleri tarandıktan sonra, kayma çizikleri, lekelenme sorunları, çipteki kusurlar, hibridizasyon hatası, görüntü bozulması veya basitçe slayt üzerindeki toz gibi sorunlar sebebiyle kayıp değerler oluşabilmektedir (Oba vd., 2003). PhyVirus veri setinde de içerisinde hangi nükleotidin olduğu bilinemeyen bazı virüsler yer almaktadır. Bu değerler, veri setinde "N" olarak işaretlenmiştir. Toplamda 64.034 virüs dizisinden 3.494 adet dizide kayıp veriye rastlanmıştır.

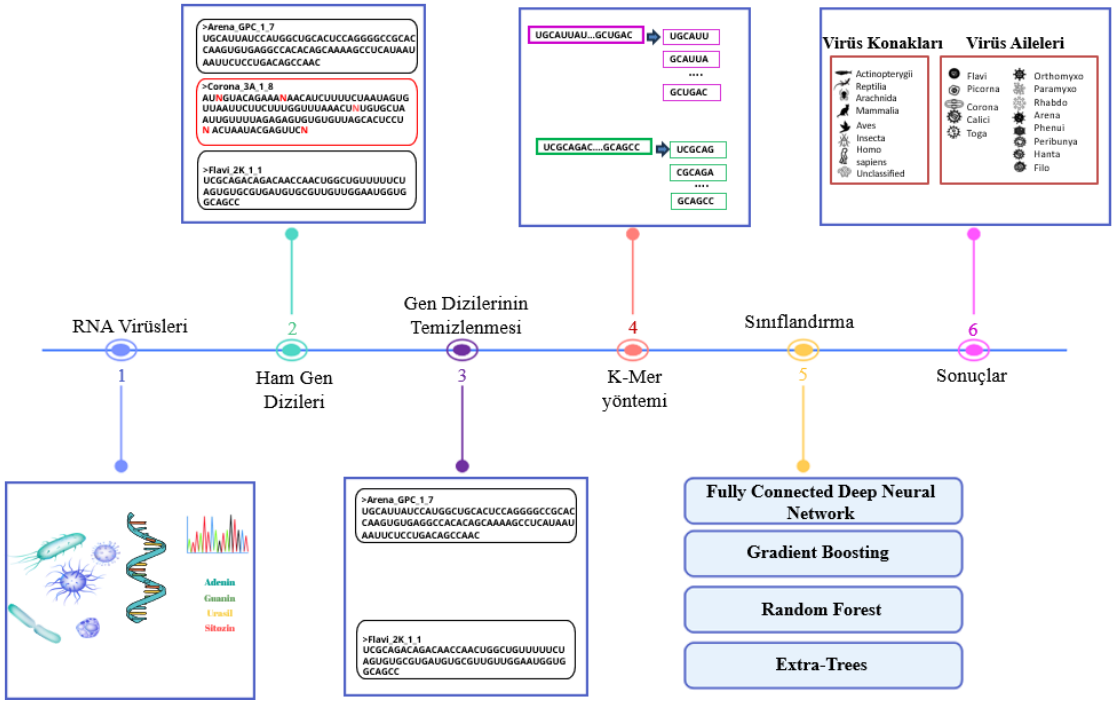
3. Aşama: 'N' şeklinde bilinmeyen nükleotid içeren virüs gen dizileri değerlendirme dışında bırakılmıştır. Ayrıca okumadan kaynaklı gen dizileri arasındaki boşluk veya farklı semboller veri seti içerisinde silinerek daha temiz bir veri seti elde edilmiştir.

4. Aşama: Farklı k değerleri ile K-Mer yöntemi uygulanarak virüs gen dizileri kodlanmıştır. PhyVirus veri setindeki gen dizileri oldukça geniş bir uzunluk aralığından oluşmaktadır. Gen dizileri en küçük 42, en büyük ise 13.176 nükleotid uzunluğundadır. Bu sebeple K-Mer Kodlama yöntemi tercih edilerek kelime uzunlukları k = 2,.., 12 aralığında seçilerek uygulamalar gerçekleştirilmiş olup her k değerine göre bir K-Mer kümesi elde edilmiştir.

5. Aşama: PhyVirus veri setinde gen dizilerine bağlı olarak 13 sınıflı virüs ailelerine (VF) ve 6 sınıflı virüs konaklarına (VC) göre sınıflandırma olmak üzere iki

temel uygulama gerçekleştirilmiştir. Her bir uygulamada RF, GB, ET gibi klasik MÖ yöntemleri ve DÖ yöntemi olan FCDNN sınıflandırma yöntemleri kullanılmıştır.

Eğitim ve test veri kümeleri için sırasıyla %80-20, %70-30, %60-40 ve %50-50 oranları kullanıldı. Train ve test veri kümelerindeki örnekler, orijinal veri kümesindeki sınıf oranları korunarak seçildi. Ayrıca, K-Mer kodlamasında kelime uzunluklarını belirleyen her k değeri için uygulamalar tekrarlandı. Bu çalışmada, VF ve VC' ye göre her iki temel uygulamada kullanılan sınıflandırıcıların parametreleri aynıdır. RF için, bölünme kalite ölçeği olarak "Gini" kullanıldı. Bunun yanında ağaç yapılı tüm sınıflandırıcılarda estimator sayısı 200, öğrenme oranı 0,05 ve maksimum derinlik 5 olarak alınmıştır.

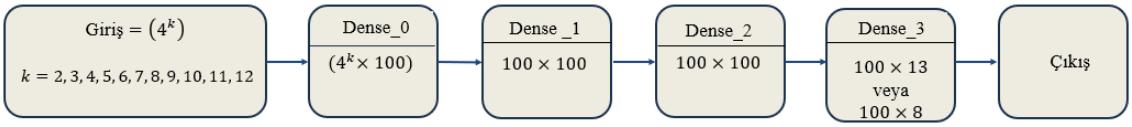


Şekil 7.1. Virüs aileleri ve Virüs konaklarına dayalı sınıflandırma uygulaması için önerilen metodun akış şeması

Bu çalışmada kullanılan FCDNN Sınıflandırıcı; bir giriş katmanı, iki gizli katman ve bir çıkış katmanından oluşmaktadır (Şekil 7.2). Çalışmada, K-Mer kodlama yöntemi farklı k değerlerine göre uygulandığı için, sınıflandırma aşamasında giriş katmanı için giriş değerleri 4^k şeklinde değişmektedir. Giriş katmanında 100 nöron, birinci gizli katmanda 100 nöron, ikinci gizli katmanda 100 nöron ve çıkış katmanında VF'ye göre sınıflandırma uygulamasında 13 adet nöron, VC'ye göre sınıflandırma uygulamasında 8

nöron bulunmaktadır. Çıkış katmanındaki her nöron, virüs sınıflarından birisini işaret etmektedir. Gizli katmanlarda ReLU aktivasyon fonksiyonu, çıkış katmanında ise verilen her bir eğitim örneği için bir Softmax aktivasyon fonksiyonu kullanılmıştır. Eğitim gerçekleştirilirken, geri yayılım algoritmasındaki gradyanların hesaplaması için Kategorik Çapraz Entropi Kayıp Fonksiyonu (Categorical Cross-Entropy) kullanılmıştır. Ayrıca optimizasyon tekniği olarak en etkili algoritmalardan biri olan Adam optimizasyon algoritması kullanılmıştır (Kingma & Lei Ba, 2015).

6. Aşama: Gen dizileri 13 sınıflı VF'ye ve 8 sınıflı VC'ye göre sınıflandırılmıştır. Farklı eğitim-test oranlarında, farklı k değerlerine bağlı ortalama doğruluklar ve standart sapmalar, sınıflandırıcı performansları gibi sonuçlar elde edilip, analizlerde bulunulmuştur. Sonuçlar değerlendirilip, çeşitli çıkarımlara ulaşılmıştır.



Şekil 7.2. Bu çalışmada kullanılan FCDNN modelinin mimarisi

7.1. Virüs Ailelerine Dayalı Sınıflandırma Sonuçları

Uygulamada kullanılan virüs aileleri (VF) ve kısaltılmış karşılıkları; ARN: Arena, CLC: Calici, CRN: Corona, FLO: Filo, FLV: Flavi, HNT: Hanta, ORT: Orthomyxo, PRM: Paramyxo, PRB: Peribunya, PHN: Phenui, PCR: Picorna, RHB: Rhabdo ve TGA: Toga şeklindedir.

Tablo 7.1'de VF sınıflandırma uygulamalarına ait her bir k değeri için ayrı ayrı Eğitim-Test oranlarına bağlı hesaplanan ortalama doğruluklar ve standart sapmalar sunulmuştur. Ayrıca Şekil 7.3'de k değeri için sınıflandırıcıların Train-Test oranlarına bağlı hesaplanan ortalama doğruluk değerlerinin grafikleri gösterilmektedir. Tablo 7.1 ve Şekil 7.3'de görüldüğü gibi FCDNN dışındaki diğer sınıflandırıcıların doğruluk değeri k değeri artıkça düştüğü, aksine FCDNN için k değeri arttıkça doğruluk değerinin yükseldiği görülmektedir.

k değeri arttıkça, ortaya çıkan özellik uzayı seyrekleşmektedir ve özellikler daha spesifik alt dizileri yansıtacak şekilde olmaktadır. Bu da ağaç tabanlı modeller olan GB, RF ve ET'nin performansının bir miktar düşmesine neden olmaktadır. Çünkü bu yöntemlerdeki ağaç yapıları; veri setinde istatistiksel olarak anlamlı olmayan özellikler

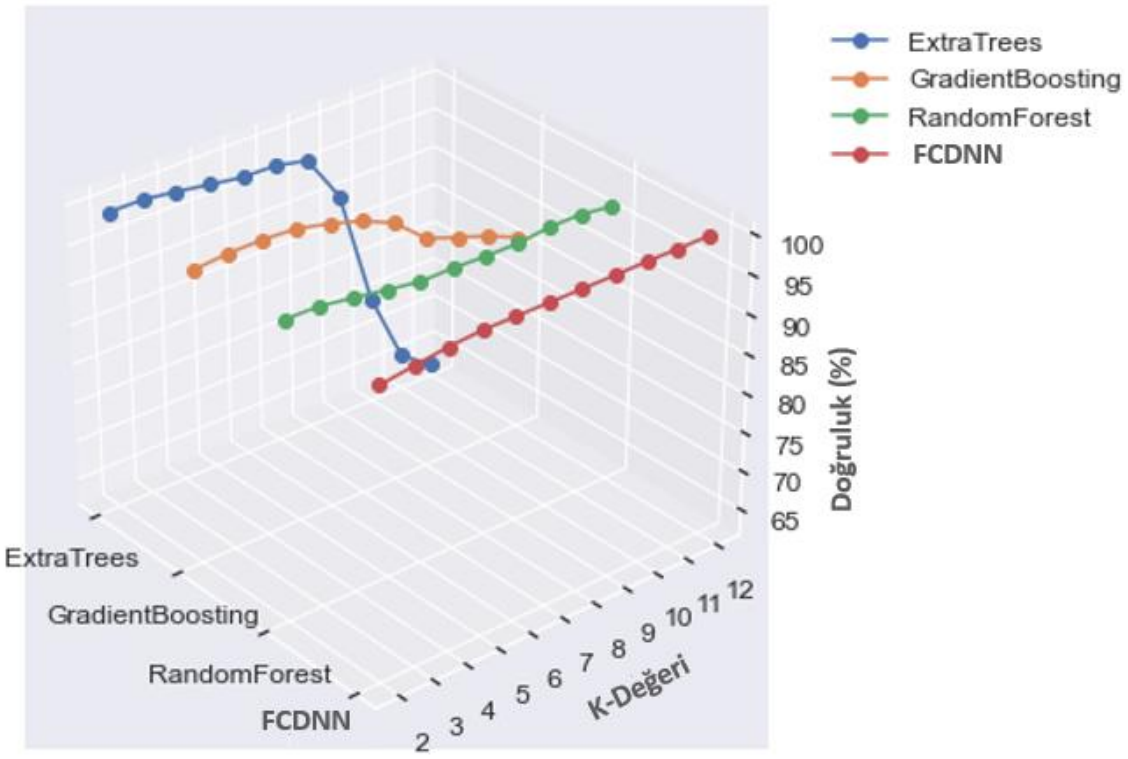
üzerinde bölünebilmektedir. Ancak, bir DÖ yöntemi olan FCDNN; verilerden özellikleri daha esnek ve sağlam bir şekilde soyutlayabilmekte, bu da potansiyel olarak FCDNN'nin neden daha yüksek k değerleriyle daha iyi performans gösterdiğini açıklamaktadır.

Bunun yanında Tablo 7.1'e göre, Train-Test oranlarına bağlı hesaplanan ortalama doğruluklar standart sapma değerlerinin minimum ±%0,05 (FCDNN sınıflandırıcı ve k=5 için) ve maksimum ±%0,56'dır (RF sınıflandırıcı ve k=12 için) olduğu görülmektedir.

Buna göre, Eğitim-Test oranlarının değişimlerinin sınıflandırıcı doğruluk değerleri üzerinde oldukça düşük (<%0,56) etkisi olduğunu söylenebilir. En yüksek ortalama doğruluk değeri %99,61 olarak FCDNN sınıflandırıcı ve k=5 ile elde edildiği görülmektedir.

Tablo 7.1. VF sınıflandırması için farklı k-Size (k-Boyutu) değerlerinde Eğitim-Test oranına dayalı ortalama doğruluklar ve standart sapmalar

		(%)Doğruluk Ort. ± Std. Sap.			
	k-Size	**RF**	**GB**	**ET**	**FCDNN**
VF Test Verileri	**k = 2**	%99,26 ±0,17	%98,75 ±0,12	%99,42 ±0,10	%98,22 ±0,31
	k = 3	%99,18 ±0,14	%99,12 ±0,17	%99,39 ±0,11	%98,79 ±0,31
	k = 4	%98,54 ±0,25	%99,16 ±0,07	%98,85 ±0,24	%99,33 ±0,27
	k = 5	%97,86 ±0,34	%98,91 ±0,20	%98,18 ±0,27	**%99,61 ±0,05**
	k = 6	%97,23 ±0,39	%97,94 ±0,16	%97,56 ±0,34	%99,60 ±0,09
	k = 7	%97,26 ±0,38	%96,73 ±0,22	%97,47 ±0,34	%99,59 ±0,08
	k = 8	%97,08 ±0,46	%94,87 ±0,30	%96,64 ±0,43	%99,55 ±0,06
	k = 9	%97,07 ±0,40	%91,36 ±0,37	%90,53 ±0,53	%99,57 ±0,09
	k = 10	%97,50 ±0,41	%89,88 ±0,14	%75,40 ±0,34	%99,60 ±0,06
	k = 11	%97,50 ±0,47	%88,48 ±0,18	%66,05 ±0,10	%99,57 ±0,04
	k = 12	%96,92 ±0,56	%86,67 ±0,12	%63,05 ±0,12	%99,56 ±0,07

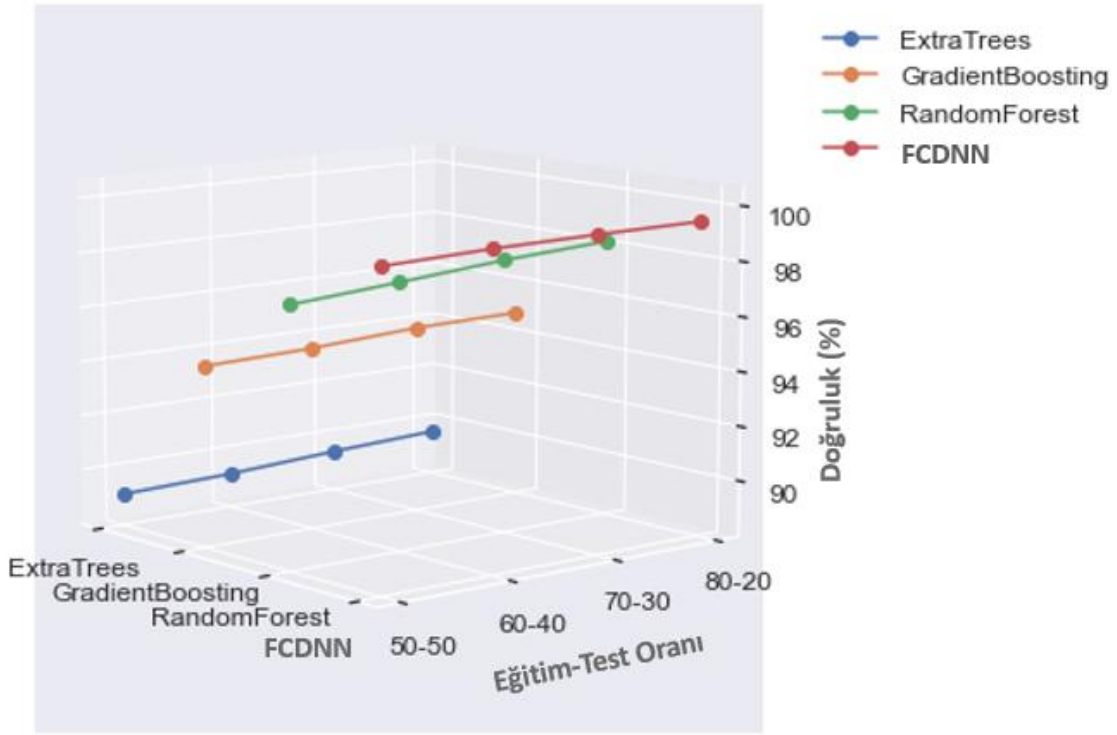


Şekil 7.3. VF sınıflandırmasında farklı K-Değerleri için sınıflandırıcı doğrulukları

Tablo 7.3'de VF sınıflandırma uygulamalarına ait her bir Eğitim-Test oranı için ayrı ayrı k-Size değerine bağlı hesaplanan ortalama doğruluklar ve standart sapmalar sunulmuştur. Ayrıca Şekil 7.4'te Eğitim-Test oranları için sınıflandırıcıların k değerine bağlı hesaplanan ortalama doğruluk değerlerinin grafikleri gösterilmektedir. Tablo 7.3 ve Şekil 7.4'te görüldüğü gibi tüm sınıflandırıcıların doğruluk değerinin Eğitim-Test oranları 80-20'den 50-50 doğru değiştikçe düştüğü görülmektedir. Bunun yanında Tablo 7.2'ye göre, k değerine bağlı hesaplanan ortalama doğruluklar standart sapma değerlerinin minimum ±%0,36 (FCDNN sınıflandırıcı ve TrnTst (Eğitim-Test)=70-30 için) ve maksimum ±%14,21 (RF sınıflandırıcı ve TrnTst=80-20 için) olduğu görülmektedir. Buna göre, k değerindeki değişimlerin Eğitim-Test oranlarındaki değişimlere kıyasla sınıflandırıcı doğruluk değerleri üzerinde daha yüksek (>%0,36) etkiye sahip olduğunu söylenebilir. En yüksek ortalama doğruluk değeri %99,42 olarak FCDNN sınıflandırıcı ve TrnTst=80-20 ile elde edildiği görülmektedir.

Tablo 7.3. Çeşitli Eğitim-Test oranlarına bağlı olarak farklı k-Size değerlerinde VF sınıflandırması için ortalama doğruluklar ve standart sapmalar

		(%)Doğruluk Ort. ± Std. Sap.			
	Eğitim-Test Oranı	**RF**	**GB**	**ET**	**FCDNN**
VF Test Verileri	**TrnTst_%80-20**	%98,15 ±0,73	%94,89 ±4,75	%89,57 ±14,21	**%99,42** ±0,45
	TrnTst_%70-30	%97,94 ±0,79	%94,84 ±4,75	%89,41 ±14,15	%99,41 ±**0,36**
	TrnTst_%60-40	%97,63 ±0,89	%94,62 ±4,76	%89,20 ±14,12	%99,37 ±0,46
	TrnTst_%50-50	%97,34 ±1,00	%94,52 ±4,75	%89,10 ±13,92	%99,25 ±0,58



Şekil 7.4. VF sınıflandırmasında farklı Eğitim-Test oranları için sınıflandırıcı doğrulukları

Virüs ailelerine göre sınıflandırma uygulamasında hem Tablo 7.2'ye ve hem de Tablo 7.3'e göre en yüksek doğruluk değerinin FCDNN sınıflandırıcı ile 80-20 Eğitim-Test oranlarını ve k=5 değeri için elde edildiği görülmektedir. Bu değerler esas alınarak gerçekleştirilen uygulamalarda sınıflandırıcıların başarı ölçütleri detaylı olarak Tablo 7.4'te sunulmuştur. Tablo 7.4' te görüldüğü gibi en yüksek doğruluk değeri FCDNN sınıflandırıcı ile %99,60 (kesinlik: %99,61, duyarlılık: %99,60, f-ölçütü: %99,60) olarak elde edilmiştir. En düşük doğruluk değeri ise RF sınıflandırıcı ile %98,26 (kesinlik: %98,31, duyarlılık: %98,26, f-ölçütü: %98,25) olarak elde edilmiştir.

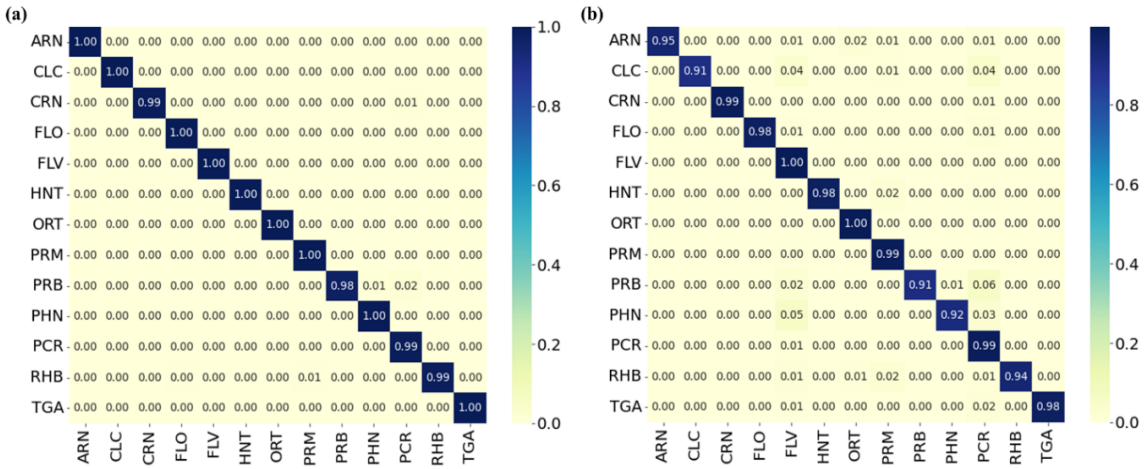
Tablo 7.4. VF'ye göre sınıflandırıcı performansları ($k = 5$ ve Eğitim-Test=%80-20)

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
RF	98,26	98,31	98,26	98,25
GB	99,13	99,15	99,13	99,13
ET	98,49	98,53	98,49	98,48
FCDNN	**99,60**	**99,61**	**99,60**	**99,60**

En yüksek ve en düşük doğruluk değerinin elde edildiği yöntemler olan FCDNN (Şekil 7.5 (a)) ve RF (Şekil 7.5 (b)) sınıflandırıcılara ait normalize edilmiş karışıklık matrisleri Şekil 7'da sunulmuştur. Karmaşıklık matrisleri, örnek sayılarının virüs ailelerine göre değişkenlik göstermesinden dolayı normalize edilerek sunulmuştur. Bu uygulamalara ait virüs ailelerinin her biri için detaylı doğru/yanlış sınıflandırma sayıları ise Tablo 7.5'te sunulmuştur. Tablo 7.5'te $k = 5$ ve TrnTst=80-20 için FCDNN ve RF sınıflandırıcıların Test veri setindeki tüm örneklerden (# Tüm Ö.) kaçını doğru (# Doğru Ö.), kaçını yanlış (# Yanlış Ö.) ve yanlış sınıflandırılan örnek sayılarının yüzdesi (Y. Yzde.) sunulmuştur. Tablo 7.5'e göre en yüksek doğruluk değerinin elde edildiği FCDNN sınıflandırıcı için en yüksek ($> \%1$) hatalı sınıflandırma RHB ve CRN virüs ailelerinde görülmektedir. Aksine FLO, HNT ve ORT virüs ailelerinde tüm örneklerin doğru sınıflandırıldığı görülmektedir. Diğer taftan en düşük doğruluk değerinin elde edildiği RF sınıflandırıcı için en düşük ($< \%1$) hatalı sınıflandırma sırasıyla ORT, FLV ve PRM virüs ailelerinde görülürken diğer virüs ailelerinde bu oranın nispeten daha yüksek olduğu görülmektedir.

Tablo 7.5. VF'ye göre en iyi sınıflandırıcıların doğru/yanlış sayıları ($k = 5$ ve Eğitim-Test=%80-20)

		FCDNN			RF		
VF	**# Tüm Ö.**	**# Doğru Ö.**	**# Yanlış Ö.**	**Y. Yzde.**	**# Doğru Ö.**	**# Yanlış Ö.**	**Y. Yzde.**
ARN	311	310	1	%0,32	295	16	%5,14
CLC	607	605	2	%0,33	551	56	%9,23
CRN	1365	1350	15	%1,10	1350	15	%1,10
FLO	143	143	0	%0,00	140	3	%2,10
FLV	3832	3827	5	%0,13	3827	5	%0,13
HNT	225	225	0	%0,00	220	5	%2,22
ORT	1435	1435	0	%0,00	1434	1	%0,07
PRM	1356	1355	1	%0,07	1344	12	%0,88
PRB	251	245	6	%2,39	229	22	%8,76
PHN	292	291	1	%0,34	268	24	%8,22
PCR	1399	1390	9	%0,64	1385	14	%1,00
RHB	505	498	7	%1,39	476	29	%5,74
TGA	387	386	1	%0,26	378	9	%2,33



Şekil 7.5. VF'ye göre normalize edilmiş Karışıklık Matrisleri (a) FCDNN sınıflandırıcı, (b) RF sınıflandırıcı

7.2. Virüs Konaklarına Dayalı Sınıflandırma Sonuçları

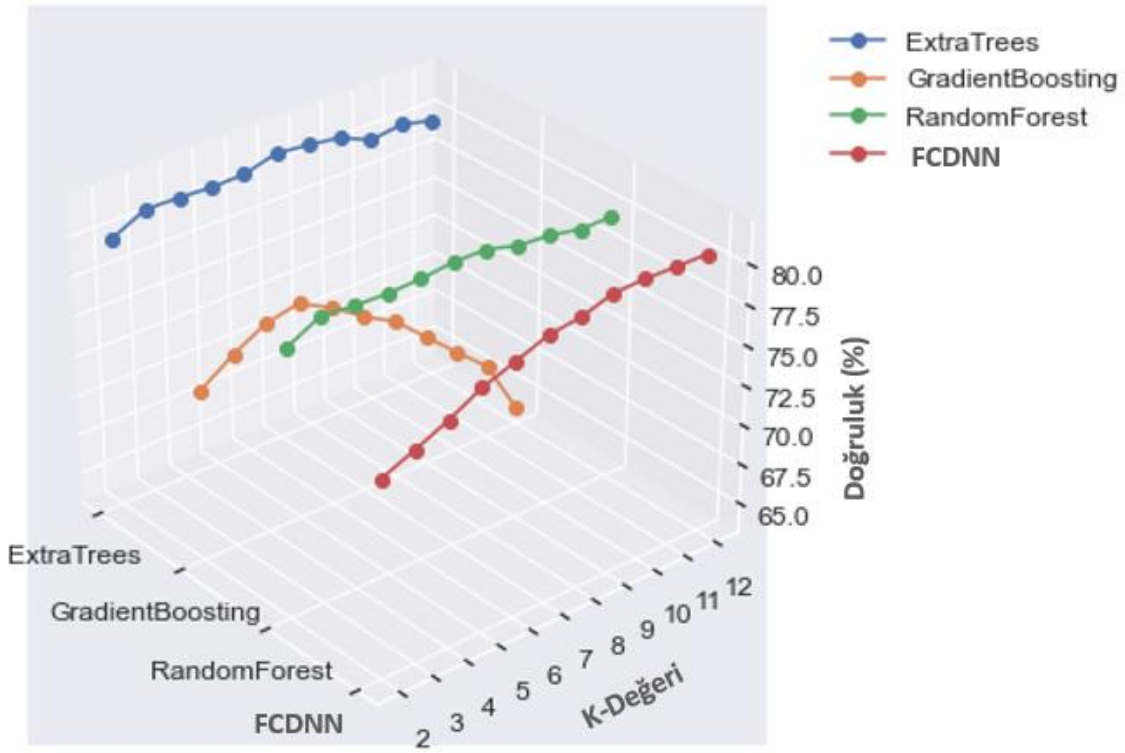
PhyVirus veri setinde eksik 3.494 satıra ek olarak bu uygulamada kayıt sayısı 1, 3 ve 7 olan sırasıyla Plantae, Nematoida ve Oomycota aileleri az veri sayısı nedeniyle değerlendirmeye dışında bırakılmıştır. Uygulamada kullanılan virüs konakları (VC) ve kısaltılmış karşılıkları; ACT: Actinopterygii, ARC: Arachnida, AVS: Aves, HMS: Homo sapiens, INS: Insecta, MMM: Mammalia, RPT: Reptilia ve UNC: Unclassified şeklindedir.

Tablo 7.6'da VC sınıflandırma uygulamalarına ait her bir k değeri için ayrı ayrı Eğitim-Test oranlarına bağlı hesaplanan ortalama doğruluklar ve standart sapmalar sunulmuştur. Ayrıca Şekil 7.6'da k değeri için sınıflandırıcıların Eğitim-Test oranlarına bağlı hesaplanan ortalama doğruluk değerlerinin grafikleri gösterilmektedir. Tablo 7.6 ve Şekil 7.6'da görüldüğü gibi ET ve RF sınıflandırıcıların doğruluk değeri k değeri artıkça neredeyse değişmediği (çok küçük bir değişim), aksine FCDNN sınıflandırıcının doğruluk değeri k değeri arttıkça yükseldiği görülmektedir. Benzer şekilde GB sınıflandırıcı doğruluk değeri k=5'e kadar yükselmiş daha sonra k=6 da düşmüştür. Bunun yanında Tablo 5'e göre, Eğitim-Test oranlarına bağlı hesaplanan ortalama doğruluklar standart sapma değerlerinin minimum ±%0,05 (RF sınıflandırıcı ve k=10 için) ve maksimum ±%1,14 (FCDNN sınıflandırıcı ve k=4 için) olduğu görülmektedir. Buna göre, VF uygulamaların aksine Eğitim-Test oranlarının değişimlerinin sınıflandırıcıların doğruluk değerleri üzerinde %1'e varan (>%0,36) değerlerde belirli bir

etkisi olduğunu söylenebilir. Tablo 5'e göre, Eğitim-Test oranlarının değişimlerinin doğruluk değeri üzerinde en düşük etkisi GB (std: min = ±%0,09, max = ±%0,36) ve en yüksek etkisi FCDNN (std: min = ±%0,15, max = ±%1,14) sınıflandırıcılarda olduğu görülmektedir. En yüksek ortalama doğruluk değeri %80,99 olarak ET sınıflandırıcı ve k=3 ile elde edildiği görülmektedir.

Tablo 7.6. VC sınıflandırma için farklı k-Size (k-Boyutu) değerinde Eğitim-Test oranlarına bağlı ortalama doğruluklar ve standart sapmalar

		(%)Doğruluk Ort. ± Std. Sap.			
	k-Size	**RF**	**GB**	**ET**	**FCDNN**
VC Test Verileri	**k = 2**	%79,83 ±0,42	%73,85 ±0,09	%79,99 ±0,50	%75,42 ±0,35
	k = 3	%80,86 ±0,51	%75,24 ±0,27	**%80,99** ±0,49	%76,27 ±0,73
	k = 4	%80,67 ±0,68	%76,35 ±0,32	%80,95 ±0,64	%77,17 ±1,14
	k = 5	%80,58 ±0,73	%76,77 ±0,13	%80,83 ±0,70	%78,26 ±0,67
	k = 6	%80,67 ±0,73	%75,69 ±0,21	%80,89 ±0,65	%78,96 ±1,03
	k = 7	%80,88 ±0,46	%74,31 ±0,36	%81,43 ±0,21	%79,72 ±0,34
	k = 8	%80,84 ±0,47	%73,18 ±0,31	%81,21 ±0,39	%79,96 ±0,77
	k = 9	%80,30 ±0,30	%71,34 ±0,51	%80,94 ±0,59	%80,58 ±0,15
	k = 10	%80,13 ±**0,05**	%69,53 ±0,25	%80,05 ±0,49	%80,63 ±0,56
	k = 11	%79,73 ±0,48	%67,77 ±0,33	%80,32 ±0,15	%80,58 ±0,54
	k = 12	%79,72 ±0,48	%64,15 ±0,35	%79,81 ±0,31	%80,54 ±0,61

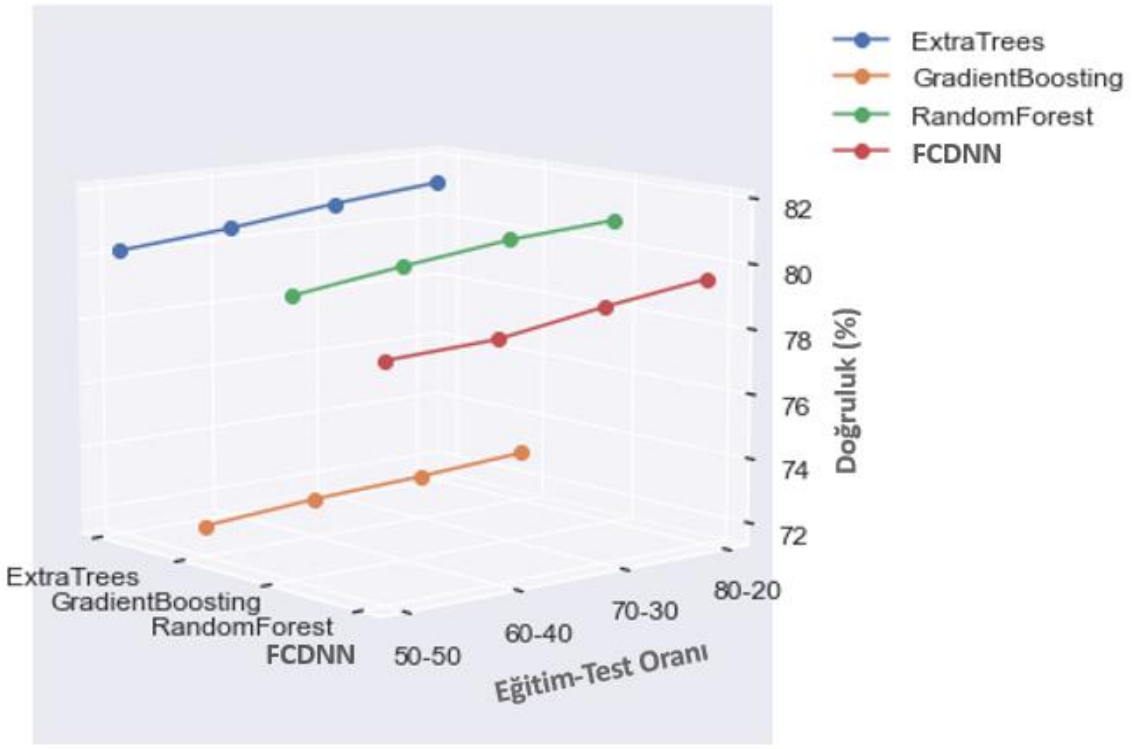


Şekil 7.6. VC sınıflandırması için farklı K-Size değerlerinde sınıflandırıcı doğrulukları

Tablo 7.7'de, VC sınıflandırma uygulamalarına ait her bir Eğitim-Test oranı için ayrı ayrı k-Size değerine bağlı hesaplanan ortalama doğruluklar ve standart sapmalar sunulmuştur. Ayrıca Şekil 7.7'de Eğitim-Test oranları için sınıflandırıcıların k değerine bağlı hesaplanan ortalama doğruluk değerlerinin grafikleri gösterilmektedir. Tablo 7.7 ve Şekil 7.7'de görüldüğü gibi tüm sınıflandırıcıların doğruluk değerinin Eğitim-Test oranları 80-20'den 50-50 doğru değiştikçe çok az da olsa düştüğü görülmektedir. Bunun yanında Tablo 7.7'ye göre, k değerine bağlı hesaplanan ortalama doğruluklar standart sapma değerlerinin minimum ±%0,42 (RF sınıflandırıcı ve TrnTst=50-50 için) ve maksimum ±%4,09 (GB sınıflandırıcı ve TrnTst=60-40 için) olduğu görülmektedir. Buna göre, k değerindeki değişimlerin Eğitim-Test oranlarındaki değişimlere kıyasla sınıflandırıcı doğruluk değerleri üzerinde daha yüksek etkiye sahip olduğunu söylenebilir. Tablo 7.7'ye göre, k değerlerindeki değişimlerin doğruluk değeri üzerinde en düşük etkisi RF (std: min = ±%0,42, max = ±%0,60) ve en yüksek etkisi GB (std: min = ±%3,89, max = ±%4,09) sınıflandırıcılarda olduğu görülmektedir. En yüksek ortalama doğruluk değeri %81,44 olarak ET sınıflandırıcı ve TrnTst=80-20 ile elde edildiği görülmektedir.

Tablo 7.7. VC sınıflandırması için farklı Eğitim-Test oranlarında k-Size değerlerine dayalı ortalama doğruluklar ve standart sapmalar

		(%)Doğruluk Ort. ± Std. Sap.			
	Eğitim-Test Oranı	**RF**	**GB**	**ET**	**FCDNN**
VC Test Verileri	**TrnTst_80-20**	%80,86 ±0,60	%72,86 ±3,89	**%81,20** ±0,58	%79,53 ±2,05
	TrnTst_70-30	%80,65 ±0,48	%72,64 ±3,98	%80,87 ±0,58	%79,11 ±1,66
	TrnTst_60-40	%80,24 ±0,48	%72,49 ±4,09	%80,46 ±0,59	%78,59 ±1,88
	TrnTst_50-50	%79,79 ±**0,42**	%72,24 ±3,97	%80,16 ±0,51	%78,44 ±2,11



Şekil 7.7. VC sınıflandırması için farklı Eğitim-Test oranlarında sınıflandırıcı doğrulukları

VC'ye göre sınıflandırma uygulamasında hem Tablo 7.7'ye ve hem de Tablo 7.6'ya göre en yüksek doğruluk değerinin ET sınıflandırıcı ile 80-20 Eğitim-Test oranlarını ve k=3 değeri için elde edildiği görülmektedir. Bu değerler esas alınarak gerçekleştirilen uygulamalarda sınıflandırıcıların başarı ölçütleri detaylı olarak Tablo 7.8'de sunulmuştur. Tablo 7.8'de görüldüğü gibi en yüksek doğruluk değeri ET sınıflandırıcı ile %81,53 (kesinlik: %80,60, duyarlılık: %81,53, % f-ölçütü: 80,86) olarak elde edilmiştir. En düşük sınıflandırma değeri ise GB sınıflandırıcı ile 75,52% (kesinlik: %73,45, duyarlılık: %75,52, f-ölçütü: %73,43) olarak elde edilmiştir.

Tablo 7.8. VC'ye göre sınıflandırma performansları (k=3 ve TrnTst=80-20)

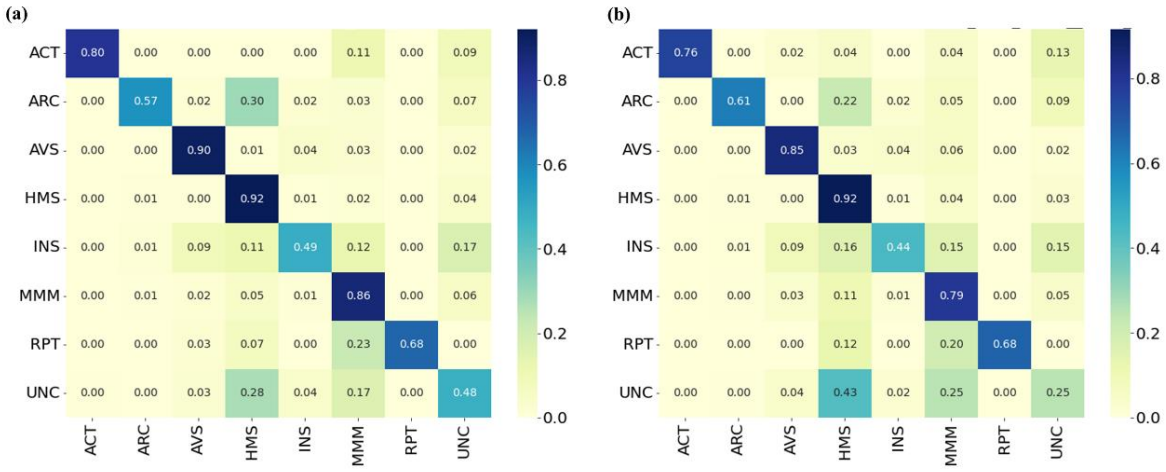
Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
RF	81,38	80,3	81,38	80,5
GB	75,52	73,45	75,52	73,43
ET	**81,53**	**80,6**	**81,53**	**80,86**
FCDNN	76,31	76,13	76,31	76,09

En yüksek ve en düşük doğruluk değerinin elde edildiği yöntemler olan ET (Şekil 7.8 (a)) ve GB (Şekil 7.8 (b)) sınıflandırıcılara ait normalize edilmiş karışıklık matrisleri Şekil 7.8' de sunulmuştur.

Karmaşıklık matrisleri, örnek sayılarının VF'ye göre değişkenlik göstermesinden dolayı normalize edilerek sunulmuştur. Bu uygulamalara ait virüs ailelerinin her biri için detaylı doğru/yanlış sınıflandırma sayıları ise Tablo 7.9'da sunulmuştur. Tablo 7.9'da k=3 ve TrnTst=80-20 için ET ve GB sınıflandırıcıların Test veri setindeki tüm örneklerden (# Tüm Ö.) kaçını doğru (# Doğru Ö.), kaçını yanlış (# Yanlış Ö.) ve yanlış sınıflandırılan örnek sayılarının yüzdesi (Y. Yzde.) sunulmuştur. Tablo 7.9'a göre en yüksek doğruluk değerinin elde edildiği ET sınıflandırıcı için en yüksek (>%50) hatalı sınıflandırma INS ve UNC virüs konaklarında görülmektedir. Diğer taraftan HMS ve AVT virüs konaklarında diğerlerine oranla daha düşük (<%10) hatalı sınıflandırma olduğu görülmektedir. En düşük doğruluk değerinin elde edildiği GB sınıflandırıcı incelendiğinde ise, en düşük (<%10) hatalı sınıflandırma yalnızca HMS virüs konağında görülürken diğer virüs konaklarında bu oranın çok daha yüksek olduğu görülmektedir.

Tablo 7.9. VC'ye göre en iyi sınıflandırıcı doğru/yanlış sayıları ($k = 3$ ve TrnTst=80-20)

		ET			GB		
VC	**# Tüm Ö.**	**# Doğru Ö.**	**# Yanlış Ö.**	**Y. Yzde.**	**# Doğru Ö.**	**# Yanlış Ö.**	**Y. Yzde.**
ACT	46	37	9	%19,57	35	11	%23,91
ARC	132	75	57	%43,18	81	51	%38,64
AVS	1502	1355	147	%9,79	1276	226	%15,05
HMS	5261	4843	418	%7,95	4819	442	%8,40
INS	498	242	256	%51,41	220	278	%55,82
MMM	2809	2419	390	%13,88	2231	578	%20,58
RPT	40	27	13	%32,50	27	13	%32,50
UNC	1818	872	946	%52,04	454	1364	%75,03



Şekil 7.8. VC'ye göre normalize edilmiş Karışıklık Matrisleri (a) ET sınıflandırıcı, (b) GB sınıflandırıcı

7.3. Virüs Ailelerine ve Konaklarına Dayalı Sınıflandırma Sonuçlarının Değerlendirilmesi

VF'ye göre yapılan uygulamalar için en yüksek ve en düşük doğruluk değerleri k=3 ve TrnTst=80-20 ile sırasıyla FCDNN (%99,60) ve RF (%98,26) sınıflandırıcılarla elde edilmiştir. Her bir virüs ailesi için ayrı ayrı doğru ve yanlış tahmin sayılarının verildiği Tablo 7.9 göz önüne alındığında, tüm virüs ailelerinin hatalı sınıflandırma oranlarının hem FCDNN (min=%0,00 ve max=%1,39) hem de RF (min=%0,07 ve max=%9,23) sınıflandırıcılar için %10'un altında olduğu görülmektedir. Gen dizilerinin okunması sırasında hata oranlarının yaklaşık olarak %0,1 ila %10 aralığında değiştiği (Poplin et al., 2018) göz önüne alındığında, virüs aileleri uygulamalarında kullanılan tüm sınıflandırıcıların performanslarının kabul edilebilir seviyelerde olduğu sonucunu ortaya koymaktadır. Bu durum Şekil 7.5'de verilen normalize edilmiş Karışıklık Matrislerinde de açık bir şekilde görülmektedir.

Diğer taraftan VC'ye göre yapılan uygulamalarda başarı VF'ye göre düşmüş olup, en yüksek ve en düşük doğruluk değerleri k=3 ve TrnTst=80-20 ile sırasıyla ET (%81,53) ve GB (%75,52) sınıflandırıcılarla elde edilmiştir. En yüksek hatalı sınıflandırma oranlarının (>%30) her iki sınıflandırıcı için sırasıyla UNC, INS, ARC ve RPT konaklarında olduğu görülmektedir (Tablo 7.9).

UNC olarak etiketlenmiş virüs konaklarında ET ve GB sınıflandırıcılar için sırasıyla %52,04 ve %75,03 oranlarında hatalı sınıflandırma olduğu görülmektedir. Bu etikete sahip virüs konaklarının tasnifi yapılamayan (Unclassified) örnekler olduğu ve içlerinde diğer virüs konaklarına ait örnekler içerebileceği göz önüne alındığında bu hata oranı beklenen bir durum olarak değerlendirilmektedir. Karışıklık matrislerine göre (Şekil 7.8 (a, b)) UNC örneklerinin öncelikle HMS (%28 ve %43) ve sonra MMM (%17 ve %25) konaklarıyla karıştırılmış olduğu görülmektedir. Benzer bir durumun INS etiketli konakları enfekte eden virüs genom dizlerinde görülmektedir. Bu etikete sahip örnekler Böcek (Insecta) türü konaklara ait olup, Şekil 7.8 (a, b) verilen karışıklık matrisleri göre her iki sınıflandırıcı için örneklerin sırasıyla %17 ve %15 oranlarında UNC olarak sınıflandırıldığı görülmektedir. UNC olarak sınıflandırılan hatalı örnekler için, tasnifi yapılmamış örnekler içinde INS ya da yakın genetik karakteristiğe sahip örneklerin olabileceği ve buna bağlı olarak modellerin yanılmış olabileceği değerlendirilmektedir.

Genel olarak birbirinden geniş ölçüde ayrılmış organizmaları enfekte eden virüslerin genetik yapılarının da birbirinden çok farklı olduğu görülmüştür. Diğer taraftan

virüsler ve konakları, evrimsel silahlanma yarışının bir parçası olarak birbirlerine gen aktardıkları anlaşılmaktadır (Aswad & Katzourakis, 2018). Ayrıca RNA virüslerinin genellikle DNA virüslerine kıyasla çok daha yüksek mutasyon oranlarına sahip oldukları (Sanjuán vd., 2010) da göz önüne alındığında, çalışmamızda hatalı tahmin edilen grupların nedenin buna dayandığı değerlendirilmektedir. Biraz daha açmak gerekirse, gen dizileri benzer olan konakları enfekte eden virüs türlerinin de birbirlerine benzer gen dizilerine sahip olabileceği ve bu nedenle modellerimizin hatalı tahminler yaptığı değerlendirilmektedir. Bu gerçekten yola çıkarak hatalı sınıflandırılan konak türlerine ilişkin değerlendirme aşağıda sunulmuştur.

INS etiketli konaklara ait virüslerin hatalı sınıflandırma oranları ET ve GB sınıflandırıcıları için sırasıyla %51,41 ve %55,82 olduğu görülmektedir. Şekil 7.8 (a, b) verilen karışıklık matrisleri göre her iki sınıflandırıcı için örneklerin MMM (%12 ve %15), ve HMS (%11 ve %16) konaklarıyla büyük oranda (>%10) karıştırılmış olduğu görülmektedir. Isawa ve ark. tarafından yapılan çalışmada (Isawa vd., 1998), böcek, memeli ve bitki Picorna virüslerinin ortak bir atayı paylaştıklarını güçlü bir şekilde ortaya koyulmuştur. Dolayısıyla INS etiketli konakçılara ait virüslerin MMM ve HMS olarak hatalı sınıflandırılmasının nedeninin virüslerin, uzak akraba konakçıları arasındaki genetik olarak benzerlikler olduğu tahmin edilmektedir (Wolf vd., 2018).

ARC etiketli virüs konaklarında hatalı sınıflandırma oranları ise her iki sınıflandırıcı için sırasıyla %43,18 ve %38,64 olduğu görülmektedir. Bu etikete sahip örnekler Araknit türü konaklara ait olup, Şekil 7.8 (a, b) verilen karışıklık matrisleri göre her iki sınıflandırıcı için örneklerin HMS (%30 ve %22) konaklarıyla büyük oranda (>%10) karıştırılmış olduğu görülmektedir. Yapılan çalışmalar, diğer eklembacaklı genomlarından farklı olarak Araknitlerin, insan genomuna çok benzeyen kısa eksonlar ve uzun intronlardan oluştuğunu göstermiştir (Sanggaard vd., 2014). Farklı türleri enfekte eden virüslerin tipik olarak kendi konakçılarının hücresel ve biyokimyasal ortamlarından faydalanmak için bağımsız olarak evrimleştikleri bilinmektedir (Wimmer & Goldbach, 1996). Dolayısıyla bu gerçekler ışığında, benzerlikleri saptanmış olan HMS ve ARC virüslerinin de konakçıları üzerinde evrimleşerek benzer genetik yapılara dönüşmüş olabileceği ve hatalı sınıflandırmanın buna bağlı oluştuğu öngörülmektedir.

Son olarak RPT etiketli virüs konaklarında hatalı sınıflandırma oranları her iki sınıflandırıcı için %32,50 olduğu görülmektedir. Bu etikete sahip örnekler Sürüngen türü konaklara ait olup, Şekil 7.8 (a, b) verilen karışıklık matrisleri göre her iki sınıflandırıcı için örneklerin MMM (%23 ve %20), ve HMS (%7 ve %12) konaklarıyla %7'nin

üzerinde bir oranda karıştırılmış olduğu ve hatalı sınıflandırıldığı görülmektedir. Bunun altında yatan sebebin, kardeş gruplar olan sürüngenler ve insanların benzer genetik bilgiler içermesi olarak görülmektedir (Janes vd., 2010).

8. PHYVIRUS VERİ SETİNDE KODLAMA UYGULAMALARI

Gen dizilerinin kodlanması, genetik verilerin daha verimli bir şekilde analiz edilmesi, depolanması ve yorumlanması için kritik bir süreçtir. Genetik bilginin doğru ve etkili bir şekilde temsil edilmesi, sınıflandırma açısından oldukça önemlidir. Bu çalışmada PhyVirus veri seti üç farklı yöntem ile kodlanarak bir CNN algoritması ile viral familyalar açısından sınıflandırılmışlardır. Önerilen yöntemin akış şeması şekil 8.1'de görülmektedir.

Aşama 1: Çalışmada kullanılan PhyVirus veri setindeki virüslere ait gen dizilerinin genetik materyalleri görülmektedir.

Aşama 2: PhyVirus veri setinde içerisinde eksik, ne olduğu bilinmeyen nükleotidler bulunmaktadır. Bunlar 'N' ile gösterilmektedir. Bu uygulamada, içerisinde 'N' bulunan virüs dizileri değerlendirme dışında bırakılmıştır.

Aşama 3: Bu aşamada virüs dizileri; FCGR, DNAWalk ve Gri Seviyeli Dönüşüm yöntemleri ile ayrı ayrı kodlanmıştır. PhyVirus veri setinde FCGR uygulaması için k'nın altı farklı değeri (k= 3, 4, 5, 6, 7, 8) ile kodlama gerçekleştirilmiş olup, toplamda altı farklı veri seti elde edilmiştir. k= 3 için 8x8'lik, k=4 için 16x16, k=5 için 32x32, k=6 için 64x64, k=7 için 128x128 ve k=8 için 256x256 boyutlarında matrisler elde edilmiştir.

DNAWalk uygulaması için Şekil 8.1'de görüldüğü şekilde nükleotidlerin her birine bir vektör tanımlaması (A= (1,1), T= (-1,1), C= (1,-1), G= (-1,-1)) yapıldıktan sonra DNAWalk uygulaması gerçekleştirilmiştir. Elde edilen 256x256 piksel boyutundaki görüntü boyutlarıyla yeni bir veri seti oluşturulmuştur.

Gri Seviyeli Dönüşüm uygulaması ile nükleotidlere birer gri seviye değeri (A=255, T=0, G=128, C=192) verilerek her bir virüs dizisi birbirinden farklı bir görüntüye dönüştürülmüş olup yeni bir veri seti elde edilmiştir.

Aşama 4: Her bir kodlama yönteminden elde edilen yeni veri setleri, sınıflandırılmak üzere bir CNN algoritması olan InceptionV3 ile sınıflandırılmıştır. Her bir InceptionV3 ile sınıflandırma uygulamasında kullanılan parametreler, veri setinin bölünmesi, performans değerlendirme metrikleri aynıdır.

Veri seti; %64 eğitim, %16 doğrulama ve %20 test olarak ayrılmıştır. Kullanılan InceptionV3 modeli daha önce ImageNet veri seti ile eğitilmiştir. ImageNet, 14

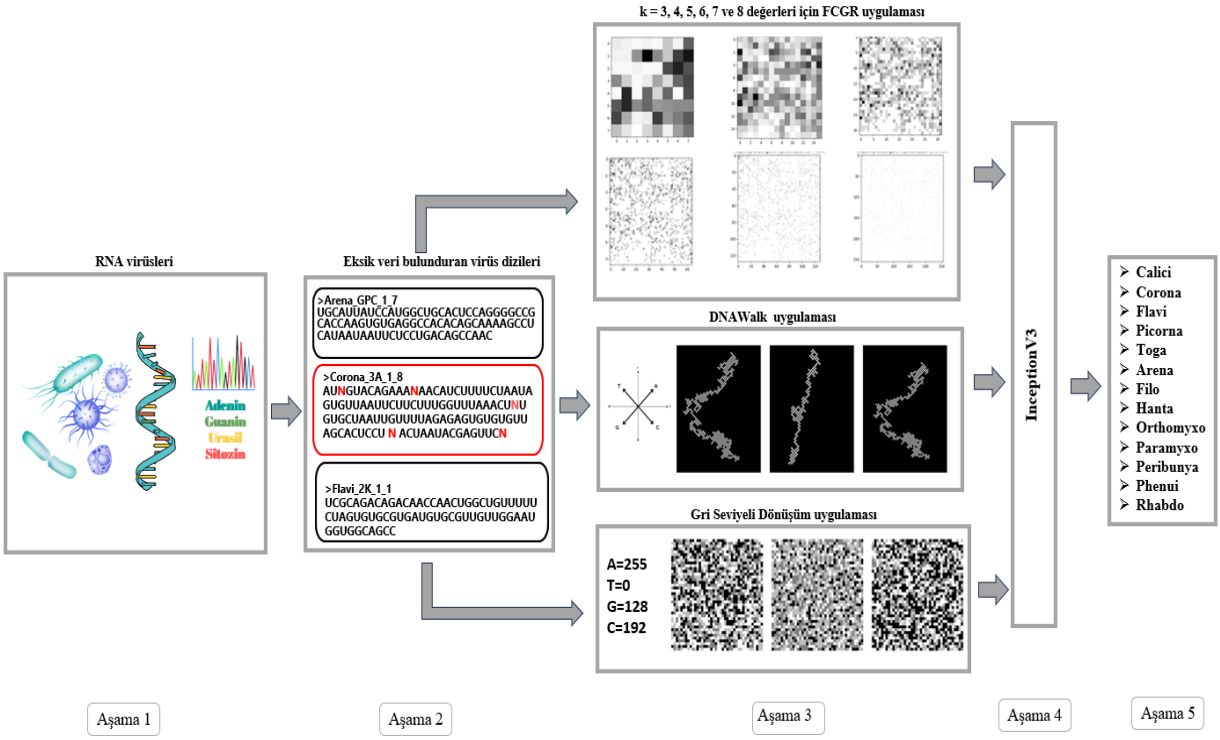
milyondan fazla görüntü içeren devasa bir veri setidir (Deng vd., 2010). Önceden eğitilmiş ağırlıklar kullanılması; modelin sıfırdan eğitim gerektirmeden diğer görevler için yeniden kullanılabilmesini sağlamaktadır. Çünkü milyonlarca görüntüden zengin bir özellik kümesi ile öğrenme gerçekleştirilmiştir (Zhuang vd., 2019). Bu durum, eğitim süresini kısaltmakta, daha yüksek doğruluk ve performans sağlamaktadır. Ayrıca, eğitim için gerekli hesaplama kaynaklarını azaltır ve daha iyi genelleme yeteneği sunmaktadır (Zhuang vd., 2019). Çalışmada sınıflandırma işlemi için belirlenen parametreler ve yöntemler detaylı bir şekilde optimize edilmiştir. İlk olarak, her bir epoch'da kullanılacak veri miktarını belirleyen batch size değeri 32 olarak seçilmiştir. Bu değer, modelin eğitim süresince yeterli bilgi almasını sağlarken, bellek kullanımını da dengede tutmaktadır. Optimizasyon işlemi için ise AdamW algoritması tercih edilmiştir. AdamW, weight decay'ı (ağırlık azalması) daha etkili bir şekilde düzenleyerek, modelin genel performansını artırmaktadır. Öğrenme oranı başlangıçta 0,001 olarak belirlenmiştir. Ancak, modelin daha stabil ve verimli bir öğrenme süreci geçirmesi için öğrenme oranının kademeli olarak azaltılması düşünülmüştür. Bu amaçla, adım_boyutu=3 ve gamma=0,8 olarak belirlenmiştir. Bu parametreler, her 3 epoch'ta bir Öğrenme Oranı'nın, belirli bir katsayı olan gamma ile çarpılarak azaltılmasını sağlamaktadır. Böylece, Öğrenme Oranı denklem 8.1'de gösterildiği gibi güncellenmektedir:

$$\text{Öğ}renme\ Oranı_{yeni} = \text{Öğ}renme\ Oranı_{eski} \times (gamma)^{floor(epoch/adım\_boyutu)} \quad (8.1)$$

Bu denklem, modelin başlangıçta hızlı bir şekilde öğrenmesini, ancak zamanla daha küçük adımlarla ilerleyerek sonuca daha yakın bir şekilde ulaşmasını sağlamaktadır. Gerçekleştirilen tüm testler 50 epoch boyunca sürdürülmüştür. Bu, modelin yeterli sayıda yineleme ile veriler üzerinde öğrenme sürecini tamamlamasına olanak tanımaktadır. Batch size, optimizer seçimi, öğrenme oranı ve epoch sayısı gibi parametreler dikkatli bir şekilde seçilmiş ve optimize edilmiştir. Bu parametrelerin her biri, modelin performansını maksimize etmek ve eğitim sürecini verimli hale getirmek için tasarlanmıştır. Çalışmada kullanılan kayıp fonksiyonu olarak (Cross-entropy) seçilmiştir. Label smoothing değeri 0,11 olarak seçilmiştir. Etiket düzeltme, sınıflandırma problemlerinde aşırı güvenli tahminlerin önüne geçmek için kullanılan bir tekniktir. Her eğitim ve doğrulama adımından sonra test veri kümesi kullanılarak yapılan testler ve ilgili model parametreleri belleğe kopyalanmıştır. Bu, modelin genelleme yeteneğini ve gerçek performansını izlemek için hayati öneme sahiptir. Ayrıca modelin yalnızca eğitim ve doğrulama

setlerinde değil, aynı zamanda yeni ve görülmemiş veriler üzerinde de ne kadar başarılı olduğunu belirlemeye yardımcı olmaktadır.

Aşama 5: Bu aşamada her bir veri setinin sınıflandırma sonuçlarına yer verilmiştir. 13 sınıflı virüs ailelerine dayalı sınıflandırma işlemi gerçekleştirilmiştir. FCGR ile gerçekleştirilen kodlama yönteminde farklı k değerleri ile elde edilmiş sınıflandırma sonuçları, k değerlerindeki değişime göre ve diğer kodlama yöntemleri ile kıyaslanarak değerlendirilmiştir. Benzer şekilde DNAWalk ve Gri Seviyeli Dönüşüm yöntemlerinin sonuçları da analiz edilmiştir.



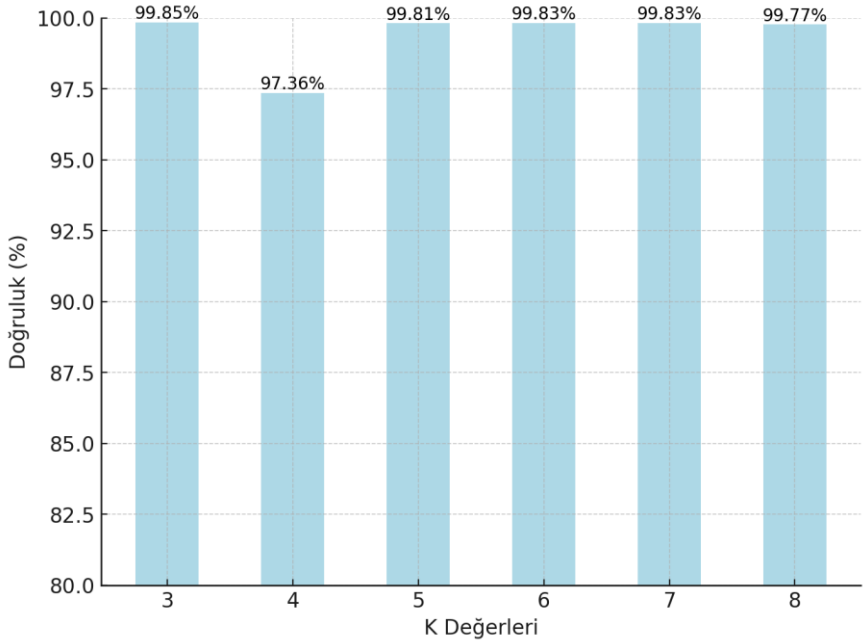
Şekil 8.1. Önerilen yöntemin akış şeması

8.1. Kodlama Yöntemleri Uygulama Sonuçları

PhyVirus veri setine FCGR, DNAWalk ve Gri Seviyeli Dönüşüm yöntemleri uygulanarak oluşturulan veri setleri, ayrı ayrı sınıflandırılmış ve elde edilen sonuçlar değerlendirilmiştir.

Uygulanan FCGR yönteminde K-Mer için k'nın altı farklı değeri ile altı farklı veri seti elde edilmiştir. Bu veri setleri ile InceptionV3 modeli kullanılarak altı farklı sınıflandırma uygulaması gerçekleştirilmiştir. Sınıflandırma esnasında tüm veri setleri

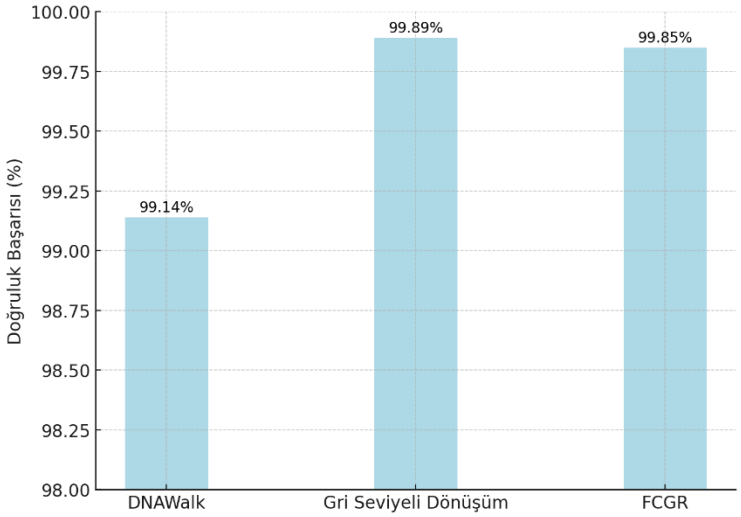
için kullanılan hiperparametreler aynıdır. Şekil 8.2'de k'nın farklı değerlerinde gerçekleştirilmiş sınıflandırma sonuçları görülmektedir.



Şekil 8.2. Farklı k değerlerinden elde edilen FCGR veri setlerinin InceptionV3 ile sınıflandırma sonuçları

Şekil 8.2' ye göre sınıflandırma sonucunda en yüksek başarı değeri (%99,85), k'nın 3 olduğu durumda gerçekleşmiştir. k=4'te diğer k değerlerine oranla keskin bir düşüş gerçekleşerek (%2,49), doğruluk başarısı %97,36'ya düşmüştür. k=5 değeri için doğruluk başarısı k=4'deki doğruluk başarısına oranla %2,45 yükselerek %99,81 değerine ulaşmıştır. k=6 değerinde, doğruluk başarısı %0,02 oranında yükselerek %99,83 olmuştur. k=7 değerinde doğruluk başarısı hiç değişiklik göstermemiş, k=6 değerindeki doğruluk başarısıyla aynı değeri göstermiştir. k=8 değerinde ise modelin doğruluk başarısı; k'nın 6 ve 7 değerlerine oranla %0,06'lık bir düşüş göstererek %99,77 olmuştur.

DNAWalk yöntemi kullanılarak kodlanmış PhyVirus veri setinin, InceptionV3 modeli ile gerçekleştirilen sınıflandırma başarısı %99,14'tür. Gri Seviyeli Dönüşüm yöntemi kullanılarak kodlanmış PhyVirus veri setinin sınıflandırma başarısı ise %99,14 olarak gözlemlenmiştir. Şekil 8.3'te veri setine uygulan üç farklı kodlama yönteminin doğruluk başarı grafiği yer almaktır. Şekil 8.3'deki grafik, FCGR yöntemi için en başarılı doğruluk değerini baz almıştır. Şekil 8.3'e göre her üç yöntemin de başarılı olduğu ve aralarında en başarılı kodlama yönteminin Gri Seviyeli Dönüşüm yöntemi olduğu görülmüştür.



Şekil 8.3. Farklı kodlama yöntemleri uygulanmış PhyVirus veri setinin InceptionV3 ile sınıflandırma sonuçları

8.2. Kodlama Yöntemleri Uygulama Sonuçlarının Değerlendirilmesi

Bu çalışma, farklı kodlama ve sınıflandırma yöntemlerinin PhyVirus veri seti üzerindeki performanslarını inceleyerek önemli sonuçlar elde etmiştir. PhyVirus veri seti, ailelerine göre K-Mer sayısal kodlama yöntemi uygulanarak en başarılı model olan FCDNN ile %99,60 doğruluk oranında sınıflandırılmıştır. Ancak, InceptionV3 modeli ile gerçekleştirilen sınıflandırmalarda kullanılan Gri Seviyeli Dönüşüm (%99,89) ve FCGR (%99,85) kodlama yöntemlerinin, FCDNN modeli ile gerçekleştirilen sınıflandırmalarda kullanılan K-Mer kodlama yöntemine göre daha yüksek başarı sağladığı görülmüştür. Bu bulgu, özellikle genetik verilerin görsel temsili ve sınıflandırmasında yüksek doğruluk oranları sağlayabilen kodlama yöntemlerinin önemini vurgulamaktadır.

FCGR yöntemiyle gerçekleştirilen sınıflandırmalar, kullanılan k değerlerinin sınıflandırma başarısı üzerinde belirgin bir etkiye sahip olduğunu göstermiştir. Şekil 8.2'de de görüldüğü gibi, k=3 değeriyle en yüksek başarı oranı olan %99,85 elde edilmiş, ancak k=4 değeriyle bu oran %97,36'ya düşmüştür. Bu durum, doğru k değerinin seçiminin model performansını önemli ölçüde etkileyebileceğini göstermektedir. k değeri arttıkça doğruluk oranında kademeli bir iyileşme görülmekle birlikte, bu artışlar istatistiksel olarak anlamlı bir farklılık yaratmamıştır.

InceptionV3 modelinin başarısının arkasındaki faktörlerden biri, önceden eğitilmiş ImageNet ağırlıklarını kullanmasıdır. Bu yaklaşım, modelin büyük bir veri seti

üzerinde önceden öğrenilmiş özellikleri yeniden kullanarak eğitim süresini kısaltmasına ve daha yüksek doğruluk ile genelleme yeteneğine sahip olmasına katkıda bulunmuştur. Modelin eğitim sürecinde kullanılan AdamW optimizasyon algoritması ve kademeli olarak azalan öğrenme oranı stratejisi, aşırı öğrenmeyi önlemiş ve modelin daha stabil ve verimli bir şekilde öğrenmesini sağlamıştır.

Sonuçlar, Gri Seviyeli Dönüşüm ve FCGR yöntemlerinin, PhyVirus veri setinin doğru bir şekilde sınıflandırılmasında etkili olduğunu ortaya koymaktadır. Gri Seviyeli Dönüşüm yöntemi en yüksek doğruluğu sağlamış, FCGR yöntemi ise yüksek performansı ile dikkat çekmiştir. Ancak, k değerinin doğru seçimi, sınıflandırma başarısı üzerinde kritik bir etkiye sahip olmaya devam etmektedir.

Literatürdeki çalışmalar, FCGR'nin farklı veri setleri ve sınıflandırma modelleri ile entegre edildiğinde yüksek doğruluk oranlarına ulaştığını göstermektedir (Cartes vd., 2022; Hammad vd., 2023). Çalışmamız da bu bulgularla paralellik göstermekte olup, InceptionV3 modeli ile FCGR yönteminin çeşitli k değerleri için sınıflandırma performansını değerlendirerek yüksek doğruluk oranları elde etmiştir. Çalışmamızda, Gri Seviyeli Dönüşüm yöntemi ile elde edilen sınıflandırma başarısının DNAWalk yöntemine göre daha yüksek olduğu görülmüştür, bu da literatürdeki diğer çalışmalarla uyumlu olarak Gri Seviyeli Dönüşüm yönteminin etkinliğini doğrulamaktadır.

9. GEN DİZİLERİNDE EKSİK VERİLER

Biyoinformatik, gen dizilerinin analizi ve sınıflandırılması gibi karmaşık süreçleri kapsayan bir alan olup, bu süreçlerin güvenilir ve tutarlı bir şekilde yürütülebilmesi için eksiksiz veri setlerine ihtiyaç duyulmaktadır. Ancak, DNA dizileme tekniklerindeki hızlı gelişmelere rağmen, bu veri setleri çoğunlukla eksik veri sorunlarıyla karşı karşıya kalmaktadır. Eksik veri, biyoinformatikte hem analizlerin doğruluğunu hem de sonuçların güvenilirliğini tehdit eden ciddi bir problemdir. Gen dizileri, çeşitli biyolojik ve teknik süreçlerden geçerek analiz için hazır hale getirilmektedir. Ancak, bu süreçlerin herhangi bir aşamasında ortaya çıkabilecek hatalar, gen dizilerinde eksik verilere yol açabilmektedir. Gen dizi görüntüleri tarandıktan sonra oluşabilecek kayma çizikleri, lekelenme sorunları, çip kusurları, hibridizasyon hataları, görüntü bozulması veya slayt üzerindeki toz gibi faktörler eksik veri probleminin başlıca nedenleridir (Oba vd., 2003). Bu sorunlar, elde edilen veri setlerinde bazı değerlerin kaybolmasına veya bozulmasına neden olabilmekte, bu da veri analizlerinin doğruluğunu ve güvenilirliğini ciddi şekilde etkileyebilmektedir

Biyoinformatikte eksik verilerin oluşmasına yol açan diğer etkenler arasında zorlu çalışma koşulları, kötü hava şartları, donanım arızaları, yazılım hataları ve insan kaynaklı hatalar yer almaktadır. Bu tür durumlar, gen örneklerinin toplanması, işlenmesi veya analiz edilmesi sırasında bazı değerlerin eksik kalmasına neden olabilir. Çeşitli araştırmalar, gen dizi veri kümelerinin farklı oranlarda, hatta %95'e varan eksik veri içerdiğini ortaya koymuştur (Dahl vd., 2016). Bu derece yüksek eksik veri oranları, biyoinformatik analizlerin güvenilirliğini ve geçerliliğini ciddi şekilde tehdit edebilmektedir.

Eksik veriler üç tür endişeye yol açmaktadır: birincisi performans kaybı, ikincisi verileri analiz etmedeki zorluklar ve üçüncüsü eksik ve mevcut değerler arasındaki tutarsızlıklar nedeniyle yanlı sonuçların üretilmesidir (Dubey & Rasool, 2021).

9.1. Eksik Veri Türleri

Eksik veriler, biyoinformatikte üç temel kategoride incelenmektedir: Tamamen Rastgele Eksik Veri (Missing Completely at Random- MCAR), Rastgele Eksik Veri (Missing at Random- MAR) ve Rastgele Eksik Olmayan Veri (Missing Not at Random- MNAR). Bu kategoriler, eksik verilerin oluşum nedenlerine ve bu nedenlerin veri setindeki diğer verilerle olan ilişkilerine göre tanımlanmaktadır.

Tamamen Rastgele Eksik Veri (MCAR): Bu tür eksik veriler, veri setinde herhangi bir özel nedene bağlı olmaksızın ortaya çıkmaktadır. Yani, eksik verilerin kayıp olma olasılığı, veri setindeki diğer herhangi bir veriyle ilişkisizdir. Bu durum, eksik verilerin veri setinin geneliyle ilgisiz olduğu anlamına gelmektedir. MCAR durumu, genellikle en nadir rastlanan eksik veri türüdür, ancak aynı zamanda en basit olandır çünkü bu tür eksik veriler, analizlerde minimal yanlılık oluşturmaktadır.

Rastgele Eksik Veri (MAR): Bu durumda, eksik veriler, veri setindeki gözlemlenen diğer verilerle ilişkili olabilmektedir, ancak eksik olan verinin kendisiyle doğrudan bağlantılı değildir. MAR, MCAR'dan daha az kısıtlayıcıdır çünkü eksik veriler, veri setindeki diğer mevcut verilerle ilişkilendirilebilmektedir. Örneğin, belirli bir genin ekspresyon verisi eksikse, bu eksiklik başka bir genin ekspresyon düzeyiyle ilişkili olabilir. MAR durumu, biyoinformatik analizlerde sıkça karşılaşılan bir eksik veri türüdür.

Rastgele Eksik Olmayan Veri (MNAR): MNAR, veri setindeki eksikliklerin hem eksik olan verinin kendisiyle hem de diğer verilerle ilişkili olduğu durumları ifade etmektedir. Bu tür eksik veriler, analizlerde en büyük yanlılığa neden olabilir çünkü eksiklik, analiz edilen verinin bir fonksiyonu olarak ortaya çıkmıştır. Örneğin, bir genin ekspresyon verisi belirli koşullarda sürekli olarak eksikse, bu eksiklik analiz sonuçlarını ciddi şekilde çarpıtabilmektedir.

Şekil 9.1. Eksik veri türleri (Newman, 2014)

9.2. Eksik Veri Tahmin Yöntemleri

Eksik verilerin tahmini ve bu eksikliklerin giderilmesi, biyoinformatikte önemli bir araştırma konusudur. Eksik veri tahmini, veri setindeki eksik değerlerin çeşitli istatistiksel analizler ve algoritmalar kullanılarak tamamlanması sürecidir. Geçmişten günümüze eksik verilerle başa çıkabilmek için birçok yaklaşım önerilmiştir. Bunlardan bir tanesi de eksik verilerin temizlenmesi işlemidir. Ancak, bu yaklaşım, analiz edilen veri setinin genel yapısını ciddi şekilde etkileyebilmekte ve elde edilen sonuçların güvenilirliğini azaltabilmektedir. Çalışmalar, eksik verilerin %10 ila %15'ten az olması durumunda, bu tür bir yaklaşımın kabul edilebilir olduğunu göstermiştir (Lin vd., 2017). Ancak, eksik veri oranı arttıkça, bu yöntemlerin uygulanabilirliği azalmaktadır.

Eksik verilerin tahmin edilmesi için geliştirilen yöntemler, basit istatistiksel tekniklerden karmaşık MÖ algoritmalarına kadar geniş bir yelpazeyi kapsamaktadır. Bu yöntemler arasında en yaygın olanları şunlardır:

- Ortalama, Mod, Medyan,
- Tahmini Ortalama Eşleştirme (Hot Deck Imputation) yöntemi,
- Knn-Imputation,
- MissForest,
- SVDImpute (Tekil Değer Ayrıştırımı) yöntemi,
- LLSImpute;
- Yapay Zekaya Dayalı Modeller

9.2.1. Ortalama, Mod, Medyan atama yöntemleri

Ortalama atama, eksik verilerin bulunduğu bir özniteliğin bilinen tüm değerlerinin aritmetik ortalamasının hesaplanarak eksik değerlerin bu ortalama ile doldurulması işlemi olarak tanımlanmaktadır. Bu yöntem, basitliği ve kolay uygulanabilirliği nedeniyle yaygın olarak kullanılmaktadır. Ancak, ortalama atama yöntemi, veri setinin varyansını azaltma eğiliminde bulunmakta, bu da analiz sonuçlarının yanlı olmasına yol açabilmektedir. Bu yöntem, MCAR durumunda nispeten etkili olabilmekteyken, MAR veya MNAR durumlarında yetersiz kalabilmektedir.

Mod atama yöntemi, bir özniteliğin en sık görülen değerinin (modunun) eksik değerler için kullanılması esasına dayanmaktadır. Bu yöntem, kategorik verilerde yaygın

olarak kullanılmakta ve eksik verilerin yerine en yaygın değerin atanmasıyla basit bir çözüm sunmaktadır. Ancak, mod atama yöntemi, eksik verilerin doğal varyasyonunu göz ardı edebilmekte ve veri setinin dağılımını bozabilmektedir. Ayrıca, mod değeri, veri setinde çok sık tekrar etmeyen bir değer olduğunda, bu yöntemin etkinliği azalmaktadır. Mod atama, özellikle MCAR durumunda etkili olabilmekteyken, MAR ve MNAR durumlarında yanıltıcı sonuçlar verebilmektedir.

Medyan atama yöntemi, bir özniteliğin tüm bilinen değerlerinin medyanının hesaplanarak eksik değerlerin bu medyan değeriyle doldurulması işlemi olarak tanımlanmaktadır. Medyan atama, özellikle simetrik olmayan veri setlerinde veya aykırı değerlerin bulunduğu durumlarda, ortalama atama yöntemine kıyasla daha güvenilir bir sonuç sunmaktadır. Çünkü medyan, veri setindeki aşırı uç değerlerden etkilenmemekte ve bu nedenle, veri setinin genel yapısını korumada daha etkili olmaktadır. Ancak, medyan atama yöntemi de eksik verilerin doğal varyasyonunu hesaba katmamakta ve bu nedenle bazı analizlerde yanlı sonuçlara yol açabilmektedir.

9.2.2. Tahmini Ortalama Eşleştirme yöntemi (Hot Deck Imputation)

Bu yöntem; eksik verilerin, aynı veri setindeki benzer gözlemlerden alınan değerlerle doldurulmasına dayanmaktadır (Kalton, 1982). Yani, eksik bir değere sahip olan bir gözlem, kendisine en çok benzeyen diğer gözlemlerden birinin değeri ile doldurulmaktadır. Bu işlem, genellikle aynı alt grup veya segment içinde benzer gözlemlerin bulunması prensibine dayanmakta olup, özellikle büyük veri setlerinde etkili sonuçlar sunabilmektedir.

Tahmini Ortalama Eşleştirme yöntemi, iki ana aşamadan oluşmaktadır. İlk aşamada, eksik değere sahip olan gözlem ile benzer özelliklere sahip diğer gözlemler arasından bir referans gözlem seçilmektedir. Bu referans gözlem, eksik değerin bulunduğu öznitelik için en yakın eşleşme olarak kabul edilmektedir. İkinci aşamada ise, referans gözlemde bulunan değer, eksik değerin yerini almak üzere kullanılır. Bu süreç, her bir eksik değer için tekrarlanarak veri seti tamamlanmaktadır. Bu yöntemin sınırlılıkları; veri setinin büyüklüğüne ve gözlemler arasındaki benzerliklerin doğru bir şekilde tanımlanmasına büyük ölçüde bağlıdır. Veri setinde yeterli sayıda benzer gözlem bulunmaması durumunda, yöntemin doğruluğu azalabilmektedir.

9.2.3. KNN-Imputation yöntemi

KNN algoritması, MÖ'nin temel taşlarından biri olup, sınıflandırma ve regresyon gibi farklı analiz yöntemlerinde yaygın olarak kullanılmaktadır. KNN algoritmasının bu esnek yapısı, eksik veri tahmini gibi veri bilimi problemlerine de uygulanmasını mümkün kılmaktadır. KNN-Imputation, eksik veri problemini çözmek amacıyla KNN algoritmasının adaptasyonunu ifade eder ve veri setindeki eksik değerlerin tahmin edilmesi için etkili bir yöntemdir (Troyanskaya vd., 2001). Eksik verileri doldurmak için veri setindeki mevcut gözlemlerin birbirine olan benzerliklerini kullanmaktadır. Eksik bir değere sahip olan bir gözlem, kendisine en yakın olan komşu gözlemler arasından seçilmekte ve bu komşu gözlemlerin ilgili özniteliklerdeki değerleri kullanılarak eksik veri tahmini yapılmaktadır. Veriler arasındaki yakınlık veya uzaklık mesafe değerleri yani gen benzerliğinin belirlenebilmesi Maksimum Sapma, Mahalanobis, Pearson Korelasyonu, Öklid Mesafesi ve Varyans Minimizasyonu gibi yöntemler ile belirlenmektedir.

KNN-Imputation yönteminin uygulanması, birkaç adımda gerçekleştirilmektedir. İlk olarak, eksik değere sahip olan gözlemin, veri setindeki diğer gözlemlerle arasındaki mesafeler hesaplanır. Ardından, belirlenen k sayısı kadar en yakın komşu gözlem seçilir. Bu komşuların ilgili öznitelikteki değerleri kullanılarak eksik değer tahmin edilir. Sürekli değişkenler için bu tahmin genellikle komşu gözlemlerin ortalaması alınarak yapılır, kategorik değişkenler için ise en sık tekrar eden değer (mod) kullanılır.

KNN-imputation algoritmasından sonra Sıralı KNN-Impute (SKNNimpute) (X. Zhang vd., 2008) ve Yinelemeli KNN-Impute (IKNNimpute) (Brás & Menezes, 2007), Küme tabanlı KNN-Impute (CKNN) (Keerin vd., 2012) varyantları ortaya çıkmıştır. CKNN yönteminin amacı önce kümeleme yöntemiyle verilerin korelasyonunu arttırmaktır. Kümeleme modelinde, mevcut tüm genleri kullanmak yerine; komşu araması yalnızca eksik genin en yakın olduğu kümeyle sınırlı olmaktadır, ayrıca zaman karmaşıklığı büyük ölçüde azalmaktadır (Keerin vd., 2012). Yöntemin en büyük avantajlarından biri, eksik verilerin tahmin edilmesinde esnek ve veri setine uyarlanabilir bir yaklaşım sunmasıdır. Veri setindeki mevcut bilgi kullanılarak eksik verilerin tahmini sağlanmakta, dolayısıyla yapay veri üretimi yerine veri setinin doğal yapısı korunmaktadır. KNN-Imputation, özellikle karmaşık veri setlerinde ve çok boyutlu veri analizi gerektiren durumlarda etkili bir çözüm sunar. Ayrıca, k sayısının doğru bir şekilde belirlenmesi durumunda, KNN-Imputation yönteminin tahmin performansı oldukça

yüksektir. Bu yöntem, veri analizi süreçlerinde eksik verilerin yönetilmesi için güçlü bir seçenek sunmakta olup, özellikle büyük ve karmaşık veri setlerinde, doğru uygulanması durumunda, yüksek doğrulukta tahminler elde edilmesini sağlamaktadır.

9.2.4. MissForest yöntemi

MissForest, RF tabanlı bir eksik veri atama yöntemi olup, MÖ temelli bir yaklaşıma dayanır (Stekhoven & Bühlmann, 2012). RF algoritmasının karışık verileri işleyebilme yeteneği ile ilk önce eğitim verileri eğitim gerçekleştirilip, ardından yinelemeli bir şekilde çalışarak eksik verilerin tahminini gerçekleştirebilmektedir. MissForest yönteminde ilk olarak, eksik veriler için başlangıç tahminleri yapılmaktadır. Genellikle bu tahminler, eksik olmayan değerlerin ortalaması ya da modundan elde edilmektedir. Verideki eksik olmayan değerler kullanılarak, her bir özellik için RF modeli eğitilmektedir. Daha sonra eksik değerler tahmin edilip ilgili hücrelere atanmaktadır. Tahmin edilen değerler güncellenmekte ve bu değerler üzerinden tekrardan model eğitilmektedir. Bu işlem, belirli bir hata eşiği ya da iterasyon sayısı sağlanana kadar devam etmektedir.

Bu algoritmanın da çeşitli avantaj ve dezavantajları bulunmaktadır. İteratif bir yöntem olduğu için, büyük veri setlerinde çalışması zaman alıcı olabilmektedir. Ayrıca, RF modeli, çok büyük veya çok karmaşık veri setlerinde yüksek hesaplama gücü gerektirebilmektedir. Fakat KNN-Impute algoritmasına kıyasla gürültülü verilere ve aykırı değerlere biraz daha dayanıklıdır.

9.2.5. SVDImpute yöntemi

SVDImpute yöntemi, veri kümesinin eksik hücrelerini atamak için verilen matrislerin tekil değer ayrıştırmasını uygulamaktadır. SVD, bir matrisin düşük dereceli bir yaklaşımını elde etmek için kullanılmakta ve bu sayede verideki temel yapıların anlaşılmasına olanak tanımaktadır. SVDImpute yönteminde; ilk adımda, eksik veriler için başlangıçta geçici bir tahmin yapılmaktadır. Bu tahmin, genellikle eksik olmayan verilerin ortalaması, medyanı veya en yakın komşuları kullanılarak yapılmaktadır. Bu adım, eksik veri hücrelerinin ilk kez doldurulmasını sağlamaktadır. Verinin tamamlanmış hali, SVD tekniği ile faktörize edilmektedir. SVD'den elde edilen matrisler, belirli sayıda

tekil değer kullanılarak yeniden birleştirilmektedir. Bu işlem, verinin düşük dereceli bir yaklaşımını oluşturmaktadır. Elde edilen düşük dereceli yaklaşım, eksik değerleri tahmin etmek için kullanılmaktadır. Eksik değerler tahmin edildikten sonra, süreç iteratif olarak tekrarlanmaktadır. Her iterasyon, eksik değer tahminlerini iyileştirmeyi ve daha doğru sonuçlar elde etmeyi amaçlamaktadır. İterasyonlar, belirli bir hata eşiği veya iterasyon sayısına ulaşılana kadar devam etmektedir.

SVDImpute tabanlı modeller özellikle büyük ve karmaşık veri setlerinde etkili bir şekilde çalışabilmektedirler. Fakat yüksek hesaplama maliyetine yol açabilmekte ve iteratif iyileştirme süreci zaman alıcı olabilmektedir. KNNimpute ile karşılaştırıldıklarında kayıp oranına ve verilerdeki gürültüye biraz daha duyarlıdırlar (Troyanskaya vd., 2001).

9.2.6. LLSImpute yöntemi

LLSImpute yöntemi; temel olarak Lineer Regresyon ve En Küçük Kareler Tahmini ilkelerine dayanmaktadır (Kim vd., 2005). Yöntem, eksik verileri tahmin etmek için verideki mevcut bilgilere dayalı olarak en iyi lineer ilişkiyi bulmayı amaçlamaktadır.

Bu yöntemde, benzer veriler ilk önce korelasyon katsayılarının mutlak değerleri büyük olan k-en yakın komşular veya k uyumlu veriler tarafından seçilmektedir. Tahmin En Küçük Kareler Regresyonu ile gerçekleştirilmektedir. Eksik değerlerin tahmin edilmesi süreci iteratif olarak tekrarlanabilmektedir. Her iterasyonda model yeniden eğitilmekte ve eksik veriler yeniden tahmin edilmektedir. Bu süreç, belirli bir hata eşiğine ulaşıldığında veya önceden belirlenmiş bir iterasyon sayısına ulaşıldığında tamamlanmaktadır.

9.2.7. Bayes Ana Bileşen Analizi (BPCA) yöntemi

BPCA yöntemi, PCA ve Bayes istatistiği temellerine dayanmaktadır (Oba vd., 2003). PCA, veri setindeki değişkenlerin boyutunu azaltarak, verideki en önemli bilgi bileşenlerini ortaya çıkarmaktadır. BPCA, bu bilgi bileşenlerini Bayesçi bir çerçeve içinde ele alarak, eksik verileri en iyi şekilde tahmin etmeyi amaçlamaktadır.

Bu yöntemde ilk adımda, veri seti Bayes modeli kullanılarak analiz edilmektedir. Bu model, verideki eksik değerleri tahmin etmek için kullanılan ana bileşenleri belirlemektedir. PCA'nın temel bileşenleri, Bayes yaklaşımıyla güncellenmekte ve eksik veriler üzerindeki belirsizlikler hesaba katılmaktadır. Her bir eksik değer tahmin edilirken, tahmin edilen değerin olasılık dağılımı da dikkate alınmaktadır. Bu sayede, tahmin edilen değerlerin doğruluğunun artması ve eksik veri atamasının daha güvenli hale gelmesi amaçlanmaktadır.

9.2.8. Derin Öğrenme tabanlı yöntemler

Eksik veri atama yöntemleri genellikle verinin doğrusal ilişkilerini temel alırken, son yıllarda gelişen DÖ teknikleri, eksik verilerin atamasında daha güçlü ve esnek çözümler sunmaktadır. DÖ tabanlı yöntemler, verinin karmaşık ve doğrusal olmayan yapısını modelleyebilme yetenekleri sayesinde eksik veri atamada giderek daha fazla tercih edilmektedir (Qiu vd., 2020; Viñas vd., 2020; Hazra vd., 2022). DÖ tabanlı eksik veri atama yöntemleri, genellikle YSA ve onların çeşitli türevlerine dayanmaktadır.

Autoencoder tabanlı yöntemler, verileri sıkıştırarak gizli temsiller oluşturmakta ve bu temsiller üzerinden verileri yeniden oluşturmaktadır. Eksik veriler, bu yeniden oluşturma sırasında tahmin edilmektedir (C. B. Lu & Mei, 2018).

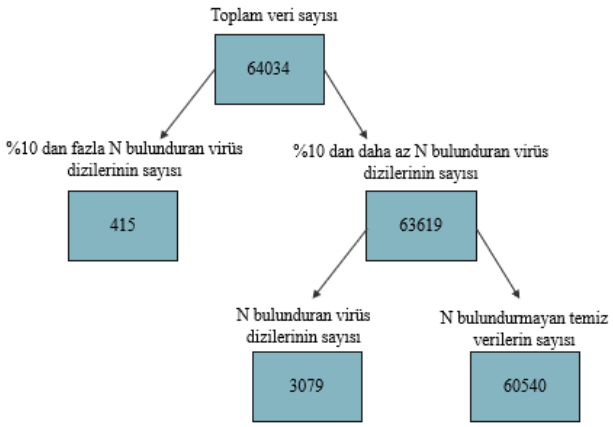
GAN'lar, eksik veri atama problemlerinde kullanılan bir başka DÖ yöntemidir. GAN'lar, iki yapay sinir ağı (üreteç (generator) ve ayırt edici (discriminator)) arasında bir rekabet süreciyle çalışmaktadır. Üreteç ağ, eksik verileri tahmin etmek için kullanılırken, ayırt edici ağ, tahmin edilen verilerin gerçek verilere ne kadar benzediğini değerlendirmektedir (Viñas vd., 2020; Hazra vd., 2022). Bu rekabetçi süreç, eksik verilerin daha doğru ve gerçekçi bir şekilde tahmin edilmesine olanak tanımaktadır.

Tekrarlayan Sinir Ağları (RNN), zaman serisi verileri gibi ardışık veri setlerinde eksik veri atama için kullanılan DÖ yöntemlerindendir. RNN'ler, verinin zaman içindeki bağımlılıklarını ve kalıplarını öğrenerek eksik verileri tahmin etmektedir.

10. PHYVIRUS VERİ SETİNDE EKSİK VERİLERİN TAHMİNİ

10.1. PhyVirus Veri Setindeki Eksik Verilerin Dağılımı

PhyVirus veri setinde toplamda 64.034 adet virüs dizisi yer almaktadır. En küçük dizi uzunluğu 42 en uzun dizi uzunluğu ise 13.176 karakter uzunluğundadır. Literatüre göre bir verinin %10'undan fazlası kayıp ise veri tahmin uygulamaları efektif çalışmamaktadır. Bu veri setinde Şekil 9.2'de görüldüğü üzere, %10 dan fazla N içeren virüs dizilerinin sayısı 415'tir. Dolayısıyla %10'dan daha az N bulunduran virüs gen dizilerinin sayısı ise 63.619'dur. Yani bu 63.619 adet gen dizisinde N bulundurmayan 60.540 temiz ve N bulunduran 3.079 adet eksik gözlem içeren gen dizisi bulunmaktadır. Dolayısıyla PhyVirus veri setinde %10'dan fazla N içeren eksik verilerin oranı %4,89'dur.



Şekil 10.1. Veri setindeki gen dizi verilerinin N bulundurma dağılımları

Veri seti içeriğinde en fazla %42 oranında N bulunduran gen dizisine rastlanmıştır. %10' dan fazla eksik değer bulunduran verilerin dağılımı genelde %10-%30 aralığındadır.

10.2. PhyVirus Veri Setindeki Farklı Boyutlu Dizilerde Bölütleme İşlemi

PhyVirus veri setinde gen dizilerinin boyutları çok geniş bir sayı aralığındadır. En küçük dizi uzunluğu 42 en uzun dizi uzunluğu ise 13.176 karakter uzunluğundadır. Benzer familyaların farklı familyalara oranla daha fazla dizi benzerliğine sahip olabileceği fikrinden yola çıkarak, aynı familya içerisinde atama işlemi gerçekleştirilmiştir. Aynı familyaya hatta konağa ait gen dizilerinde dahi, gen dizilerinin boyutu farklılık gözlenmiştir. Bu sebeple bölütleme işlemi gerçekleştirilmiştir. Bölütleme işlemi; aynı viral ailede, N içermeyen diziler arasında, eksik gen dizisine sahip virüs dizisi ile aynı uzunlukta veya daha uzun olan virüs dizilerinin tespiti gerçekleştirilerek yapılmaktadır. Ardından daha uzun olan kısımlar silinmektedir. Şekil 10.2'de, Picorna ailesine ait Picorna_Lab_2_1 virüs dizisi için atama işleminde, tüm veri seti içerisinde aynı ailede, N içermeyen diziler arasından, 54 dizi uzunluğuna eşit veya daha uzun olan virüs dizlerinin tespiti gerçekleştirilerek bölütleme gerçekleştirilmektedir.

phyVirus_sequence_i	Gen Dizilimi	viral_familya	dizi uzunlu
Picorna_VP1_4_163	AGACGCAAGCAACCGCTCGTCGACCTGCAAAACAACTGCTG	Picorna	42
Picorna_LAB_2_41	AACTTTGACCTGCTCAAGTTGGCTGGAGATGTGGAGTCCAACCCTGGG	Picorna	48
Picorna_LAB_2_44	AACTTTGGCCTGCTCAAGTTGGCTGGAGACGTGGAGTCCAACCCTGGG	Picorna	48
Picorna_LAB_2_50	AACTTTGACATGCTCAAGTTGGCTGGAGACGTAGAGTCCAACCCTGGG	Picorna	48
Picorna_LAB_2_1	GTACTGAACTTCGACCTGCTCAAGTTGTCAGGAGACNGTTGAGTCCAACCCTGGG	Picorna	54
Picorna_3B_VPG__3_39	CGGGCTTACAACCCAACTCTACCAGTGGCAAAACCTAAGGGTACATTCCCAGTAACTCAG	Picorna	60
Picorna_3B_VPG__3_40	AGAGCTTATAACCCAACACTCCCTGTTGCTAAAACAAAAGGAGCCTTCCCGGTCACACAG	Picorna	60
Picorna_2A_6_10	GGGCCTGGGGCAACAAACTCCTCACTTTTAGAGCAAGCAGGAGATGTTGAAGAAAATCCTGGA	Picorna	63
Picorna_2A_6_11	GGGCCAGGTGCTACAAATTTTTCCCTGTTGAAGCAAGCAGGAGACATTGAAGAAAATCCCGGG	Picorna	63

Şekil 10.2. Picorna_Lab_2_1 virüs dizisi için atama işleminde aynı veya daha uzun virüslerin tespit edilmesi

10.3. KNN-Imputation Yöntemi ile PhyVirus Veri Setinde Eksik Veri Tahmini ve Sınıflandırma

PhyVirus veri setindeki eksik verilerin KNN-Imputation yöntemi ile atanması için önerilen metodun akış şeması şekil 10.4' te gösterilmiştir.

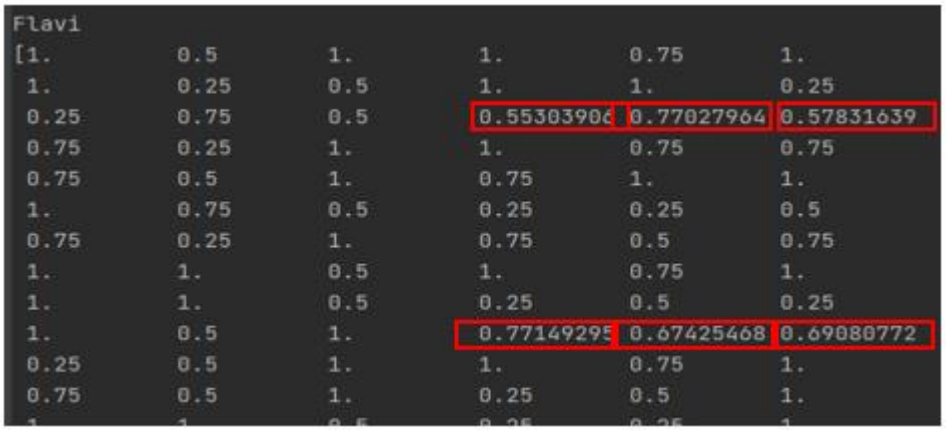
1. Adım: PhyVirus veri seti ham olarak FASTA formatındadır. Burada N olarak işaretlenmiş eksik veriler mevcuttur. Şekil 10.1'de görüldüğü üzere, %10'dan daha fazla N içeren gen dizileri analiz dışında bırakılmıştır. Geri kalan temiz ve %10'dan az N bulunduran veriler değerlendirilmiştir.

2. Adım: Eksik veri içeren gen dizisinin boyutu ile aynı veya daha uzun olan gen dizileri belirlenir. Daha uzun boyutlu olan gen dizilerinin boyutu, eksik veri içeren gen dizisinin boyutu ile bölünerek eşitlenir.

3. Adım: Yalnızca **s**ayısal verilerde atama gerçekleştirebilen KNN-Imputation yöntemi için; ayıklanmış ve bölütlenmiş olan gen dizilerinin Label Encoder yöntemi ile sayısallaştırılması bu adımda gerçekleştirilmektedir. ”A= 0,25, C=0,50, G=0,75, C=1,00, U=0,00” değerlerine göre kodlama gerçekleştirilmiştir. Eksik gen dizisinde yer alan N değerleri için Python NumPy içinde tanımlı “N=np.nan” değeri kullanılmıştır. Ayrıca varsayılan mesafe değeri Öklid ile hesaplanmaktadır.

4. Adım: Sayısal olarak kodlanmış olan ve eksik gen dizisi ile aynı boyutlara bölütlenmiş olan gen dizilerinden oluşan yeni veri kümesinin ((k komşu değeri= 5) ve (k komşu değeri= yeni veri kümesi uzunluğu) değerleri ile) KNN-Imputation yöntemi ile ataması gerçekleştirilmektedir.

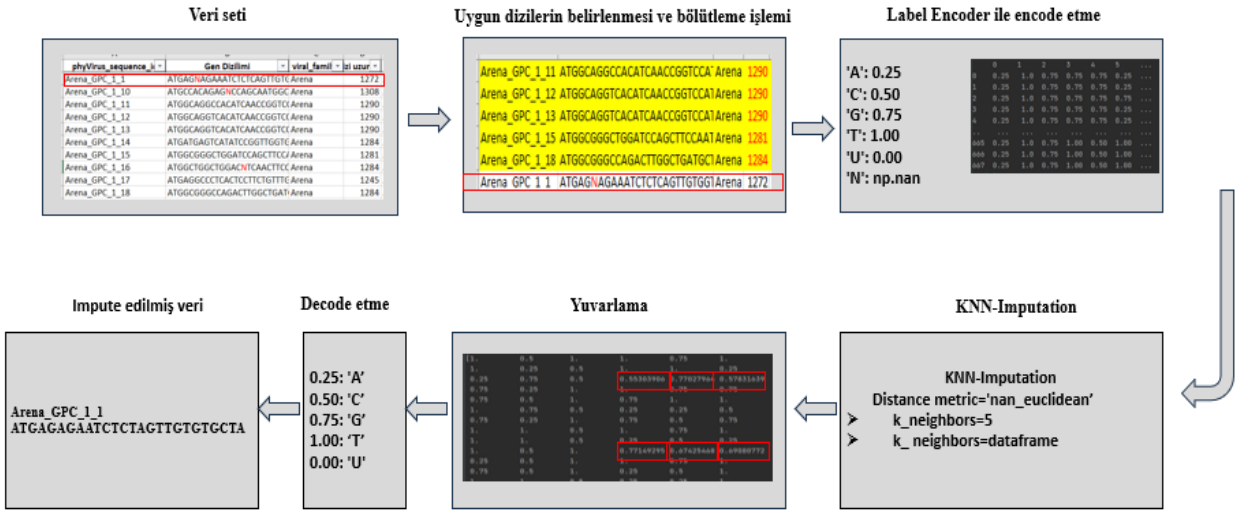
5. Adım: KNN-Imputation uygulandıktan sonra tahmin edilen N değerlerinden bazılarının, etiket kodlama yönteminde kullanılan 0,25, 0,50, 0,75 ,1,00, 0,00 değerlerinden farklı olduğu görülmüştür (Şekil 10.3). Bu sebeple, bu aşamada yuvarlama işlemi gerçekleştirilmiştir. Etiket değerlerine en yakın olacak şekilde değerlendirme gerçekleştirilmiştir.



Şekil 10.3. KNN-Imputation uygulandıktan sonra tahmin edilen N değerlerinden bazılarının küsuratlı görüntüleri

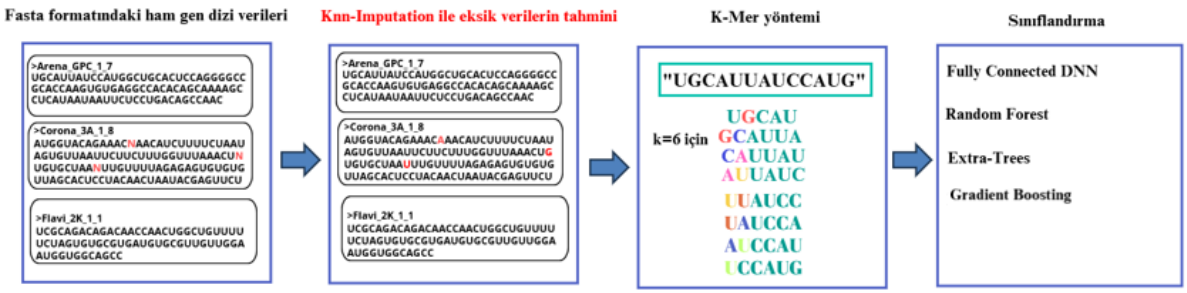
6. Adım: Imputation sonrası, elde edilen veriler decode edilmiştir.

7. Adım: Imputation sonuçlarının değerlendirmesi gerçekleştirilmiştir.



Şekil 10.4. Önerilen KNN-Imputation metodunun akış şeması

PhyVirus veri setindeki eksik veriler KNN-Imputation yöntemi ile atandıktan sonra elde edilen yeni veri seti ile sınıflandırma işlemi gerçekleştirilmiştir. Önerilen metodun akış şeması Şekil 10.5'te gösterilmiştir. Şekil 10.5'e göre, ilk aşamada Fasta formatındaki ham gen dizileri görülmektedir. İkinci aşamada, Ham gen dizileri Şekil 10.4'deki aşamalardan adım adım geçerek, KNN-Imputation yöntemi ile eksik olan gen dizileri tahmin edilmektedir. Üçüncü aşamada, elde edilmiş olan yeni ve temiz veri seti K-Mer yöntemi ile kodlanmaktadır. Dördüncü aşamada ise, kodlanmış olan gen dizileri FCDNN, RF, ET ve GB ile sınıflandırılmıştır.



Şekil 10.5. Eksik veri ataması yapılarak gerçekleştirilen sınıflandırma işleminin akış şeması

10.3. KNN-Imputation Uygulama Sonuçları

PhyVirus verisetinde, %10'dan daha az sayıda N içeren veriler için gerçekleştirilmiş olan KNN-Imputation uygulamasında varsayılan parametreler kullanıldığında k komşuluk değeri olarak '5' tercih edilmiştir. Ayrıca atama yapılacak

verinin listenin en altına yazılmasından dolayı daha iyi bir analiz için tüm veriler baz alınarak tahmin yürütülmesi amaçlanmış ve k komşuluk değeri tüm veri seti seçilmiştir. Fakat tahmin esnasında veri setinde eksik gözlemin bulunduğu dizinin uzunluğuna bağlı olarak yeni bir veri seti oluşturduğundan dolayı, KNN-Imputation uygulanan veri seti uzunluğu sürekli olarak değişmektedir. Bu yüzden ikinci analizde k değeri 'değişken'dir.

Imputation (Atama/Tahmin) gerçekleştirildikten sonra elde edilen yeni veri seti, dört sınıflandırma algoritması kullanılarak viral aileler baz alınarak sınıflandırılmıştır. Bu algoritmalar; FCDNN, RF, ET ve GB'dir. Veri setleri %80-%20 oranında eğitim ve test olarak ayrılmıştır. Tablo 10.1'de 'k=5' ve 'k=değişken' değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin FCDNN ile sınıflandırılma sonuçları yer almaktadır. Tablo 10.1'e göre; k=5 ile KNN-Imputation uygulanmış veri setinin FCDNN sınıflandırma başarısı, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısına oranla çok küçük bir miktar düştüğü (%0,8) gözlemlenmiştir. k= değişken olduğu durumda ise, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısı ile aynı başarıyı göstermiştir.

Tablo 10.1. k=5, k=değişken değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin FcDNN ile VF' ye dayalı sınıflandırılma sonuçları

	k=5	**k=değişken**	**Imputation yok**
Doğruluk	%98,80	%99,60	%99,60
Kesinlik	%98,81	%99,61	%99,61
Duyarlılık	%98,80	%99,60	%99,60
F-Ölçütü	%98,80	%99,60	%99,60

Tablo 10.2'de k=5 ve k=değişken değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin FCDNN ile sınıflandırılma sonuçları yer almaktadır. Tablo 10.2'ye göre; k=5 ile KNN-Imputation uygulanmış veri setinin ET sınıflandırma başarısı, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısına oranla çok küçük bir miktar düştüğü (%0,1) gözlemlenmiştir. k=değişken olduğu durumda ise, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısı ile aynı başarıyı göstermiştir.

Tablo 10.2. k=5, k=değişken ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin ET ile VF' ye dayalı sınıflandırılma sonuçları

	k=5	**k=değişken**	**Imputation yok**
Doğruluk	%98,39	%98,49	%98,49
Kesinlik	%99,40	%99,61	%99,61
Duyarlılık	%99,38	%99,60	%99,60
F-Ölçütü	%99,38	%99,60	%99,60

Tablo 10.3'de k=5 ve k=değişken değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin RF ile sınıflandırılma sonuçları yer almaktadır. Tablo 10.3'e göre; k=5 ile KNN-Imputation uygulanmış veri setinin RF sınıflandırma başarısı, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısına oranla az bir miktar düştüğü (%0,6) gözlemlenmiştir. k=değişken olduğu durumda ise, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısı ile aynı başarıyı göstermiştir.

Tablo 10.3. k=5, k=değişken değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin RF ile VF' ye dayalı sınıflandırılma sonuçları

	k=5	**k=değişken**	**Imputation yok**
Doğruluk	97,66	98,26	98,26
Kesinlik	98,66	98,31	98,31
Duyarlılık	97,65	98,26	98,26
F-Ölçütü	97,66	98,25	98,25

Tablo 10.4'te k=5 ve k=değişken değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin GB ile sınıflandırılma sonuçları yer almaktadır. Tablo 10.4'e göre; k=5 ile KNN-Imputation uygulanmış veri setinin GB sınıflandırma başarısı, imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısına oranla az bir miktar düştüğü (%0,23) gözlemlenmiştir. k=değişken olduğu durumda ise; imputation gerçekleştirilmemiş veri setindeki sınıflandırma başarısı ile aynı başarıyı göstermiştir.

Tablo 10.4. k=5, k=değişken değerleri ile KNN-Imputation uygulanmış veri setinin ve hiç imputation uygulanmamış veri setinin GB ile VF' ye dayalı sınıflandırılma sonuçları

	k=5	**k=değişken**	**Imputation yok**
Doğruluk	%98,90	%99,13	%99,13
Kesinlik	%98,93	%99,15	%99,15
Duyarlılık	%98,90	%99,13	%99,13
F-Ölçütü	%98,90	%99,13	%99,13

Tüm uygulama sonuçlarında k=5 ile gerçekleştirilen uygulamada sınıflandırma başarısı en çok %0,8 düşmüştür. k=değişken durumda, sonuçlarda herhangi bir değişiklik olmamıştır.

10.4. KNN-Imputation Uygulama Sonuçlarının Değerlendirilmesi

KNN-Impute, başarılı bir atama algoritması olmasına karşın, k değeri doğru ayarlandığında etkili sonuçlar verebilmektedir. Ancak, bu algoritma gürültüden kolayca etkilenmekte ve aykırı değerlerle karşılaşıldığında istenilen sonuçları veremeyebilir. Tüm verilerin depolanması ve her bir veri noktası arasındaki mesafelerin hesaplanması gerektiği için algoritmanın karmaşıklığı ve hesaplama maliyeti yüksektir.
PhyVirus veri setinde 64.034 adet veriden, 3.494'ünde eksik veri bulunmaktadır. Bu, %5,45 oranında bir eksik veriye karşılık gelmektedir. Gen dizilerinin virüs ailelerine göre sınıflandırma uygulamasında 3.494 adet eksik veri çıkarılarak toplam 60.540 adet veri ile çalışılmıştır. Fakat bu uygulamada; KNN-Imputation ile 3.494 eksik veriden 3.079'u tahmin edilmiştir. Oran olarak, eksik gen dizilerinin %11,87'si tahmin edilmemiştir. Tahmin edilen eksik verilerin tüm veri setine oranı ise %4,80'dir. k değeri değişken olduğunda, Tablo 10.1, 10.2, 10.3 ve 10.4'teki sonuçlara göre sınıflandırma yöntemlerinin başarısının değişmemesi, tahmin işleminin başarılı bir şekilde gerçekleştirildiğini göstermektedir. Çünkü verilere %4,80'lik bir katkının başarıyı büyük ölçüde etkilemesi beklenmemektedir. k=5 olduğunda ise sınıflandırma başarısının maksimum %0,8 oranında düşmesi, tahmin işleminin bazı dizilerde yanılgı içerdiğini göstermektedir. Bu durumun, k değerinin düşük seçilmesinden ve buna bağlı olarak daha az sayıda komşu ile karşılaştırma yapılarak tahmin işleminin yetersiz sayıda örnek ile gerçekleştirilmesinden kaynaklandığı düşünülmektedir.

KNN-Imputation yönteminin bazı sınırlamaları bulunmaktadır. Öncelikle, bu yöntem, veri setinin büyüklüğüne ve boyut sayısına bağlı olarak hesaplama açısından maliyetli olabilmektedir. Büyük veri setlerinde veya çok boyutlu veri alanlarında mesafe hesaplamaları zaman alıcı olabilmekte ve yüksek hesaplama gücü gerektirebilmektedir. Ayrıca, KNN algoritmasının temelinde yatan varsayım, eksik değere en yakın olan komşuların doğru tahminler sağlayacağıdır, ancak veri setindeki gürültü veya aykırı değerler bu tahminlerin doğruluğunu zayıflatabilmektedir.

KNN-Imputation yönteminin en büyük avantajlarından biri, eksik verilerin tahmin edilmesinde esnek ve veri setine uyarlanabilir bir yaklaşım sunmasıdır. Bu yöntem, veri setindeki mevcut bilgiyi kullanarak eksik verilerin tahmin edilmesini sağlamakta, dolayısıyla yapay veri üretimi yerine veri setinin doğal yapısını korumaya çalışmaktadır.

Literatürde imputation algoritmaları, eşit uzunlukta olan veriler arasında gerçekleştirilmektedir. Bu çalışmanın önemli yönlerinden birisi de önerilen yöntem ile neredeyse çok farklı uzunluk değerlerine sahip gen dizilerinin eksik verilerinin tahmin edilmesini sağlamasıdır.

11. TARTIŞMA

Literatür incelendiğinde, PhyVirus veri seti üzerindeki analizlerin daha önce hizalama tabanlı yöntemler kullanılarak gerçekleştirildiği ve yalnızca RNA virüslerinde ortak genomik ve evrimsel özelliklerin var olup olmadığı konularına odaklandığı görülmüştür (Kustin & Stern, 2021). Mutasyonel yanlılıkları vurgulamak için nükleotid yoğunluğu ve hizalamalar karşılaştırılmıştır.

Son zamanlarda yapılan çalışmalar, hizalamaya dayalı yöntemlerden hizalamadan bağımsız yaklaşımlara, özellikle de MÖ ve DÖ' ye doğru giderek daha fazla kaymaktadır. Bu tez çalışmasında, ilk defa PhyVirus veri seti; YZ yöntemleri kullanılarak analiz edilmiştir. Bölüm 7'de yer alan virüs ailelerine ve konaklarına dayalı sınıflandırma uygulamasında; 13 farklı viral aile ve 8 farklı konak türü temel alınarak, önceki çalışmalarda gerçekleştirilmeyen geniş kapsamlı bir sınıflandırma uygulanmıştır. Literatür incelendiğinde, çalışmaların genellikle aynı veya benzer uzunluktaki gen dizilerini içeren veri setlerini kullandığı görülmektedir. Bu çalışmayı literatürdeki diğer çalışmalardan ayıran özelliklerden biri de çok farklı uzunluklardaki dengesiz verilerle çalışılmış olmasıdır. Veri setindeki gen dizileri en küçük 42, en büyük ise 13.176 nükleotid uzunluğundan oluşmaktadır. Veri setinde bu şekilde değişken uzunluklarda genom dizilerinin bulunması ve gen dizilerinin kategorik değerlerden oluşması sınıflandırmayı zorlaştırmaktadır. Literatür araştırıldığında, analiz gerçekleştirebilmek için hizalama gerektiren yöntemler veri eşitleme gerektirmektedir. Veri kümesini eşitlemek için sonradan dizilerin eklenmesi aşırı işlem yüküne yol açmakta ve potansiyel olarak sınıflandırma performansını etkileyebilmektedir. Benzer şekilde, azaltmalar da veri kaybına yol açarak hatalı analitik sonuçlara neden olabilmektedir. Mock ve arkadaşlarının gerçekleştirdiği konak tabanlı bir sınıflandırma uygulamasında, üç konak türü, DÖ kullanılarak sınıflandırılmıştır (Mock vd., 2020). Veri setindeki farklı uzunluklardaki veriler için dolgu ve kırpma yöntemleri kullanılarak eşitlenmeleri sağlanmıştır. Bu tezde ise viral aile ve konak tabanlı sınıflandırma için K-Mer kodlama yapılmıştır. K-Mer yönteminin önemli bir özelliği; verileri herhangi bir kayıp veya değişiklik olmadan analiz edebilmesidir. Gerçekleştirilen sınıflandırma uygulamasında K-Mer kodlama yönteminin etkinliği, genomik sınıflandırmada özellik seçiminin önemi vurgulanmıştır. Ayrıca yöntemin uyarlanabilirliği ve farklı k değerinin sınıflandırıcı

performansı üzerindeki etkisi üzerinde durulmuştur. Literatürde DÖ, MÖ veya hibrit yöntemler kullanılarak farklı k değerleri üzerinde analizler gerçekleştirilmiştir. Genellikle 1 ila 15 arasında değişen k değerleri veya bu aralıktaki spesifik aralıklar analiz edilmiş ve sonuçlar değerlendirilmiştir (Basu & Campbell, 2023; Gunasekaran vd., 2021; Remita & Diallo, 2019; Solis-Reyes vd., 2018; Sukhorukov vd., 2022). Bu çalışmada da literatürle uyumlu olarak 2 ile 12 arasındaki k değerleri ile kodlamalar gerçekleştirilmiştir. Yapılan çalışmalar, başarılı k değerlerinin genellikle veri kümesinin uzunluğu ve kullanılan yöntemlerle ilişkili olduğunu göstermiştir (Basu & Campbell, 2023; Gunasekaran et al., 2021). Literatür incelendiğinde, önceki çalışmalarda 6 ila 9 arasındaki k değerlerinin genellikle daha başarılı olduğu görülmektedir (Gunasekaran vd., 2021; Remita & Diallo, 2019; Solis-Reyes vd., 2018). Bu çalışmada ise ağaç tabanlı MÖ yöntemleri olan GB, RF ve ET'nin, düşük k değerlerinde nispeten daha iyi performans sergilediği, buna karşın DÖ yönteminin k değeri arttıkça performansının iyileştiği tespit edilmiştir. VF'ye göre bulunan en yüksek doğruluk başarısı FCDNN ile %99,60 olarak tespit edilmiştir. Konak tahmininde ise en yüksek başarı %81,53 oranıyla ET sınıflandırıcısı ile elde edilmiştir. Literatürdeki diğer çalışmalara göre genel olarak daha yüksek başarılar elde edildiği görülmüştür (Gunasekaran vd., 2021; Lopez-Rincon vd., 2020; Remita & Diallo, 2019).

Bölüm 8'de gerçekleştirilen uygulamada PhyVirus veri seti; grafiksel gösterim yöntemi olan FCGR ve DNAWalk yöntemleri ile görüntü gösterim yöntemi olan Gri Ölçekli Dönüşüm yöntemi ile kodlanarak sınıflandırma uygulaması gerçekleştirilmiştir. Literatür incelendiğinde ise bu alandaki çalışmaların genelde yalnızca tek bir kodlama yöntemi üzerine yoğunlaşan çalışmalar olduğu görülmüştür (Cartes vd., 2022; Hammad vd., 2023; Löchel & Heider, 2021; Rizzo vd., 2016). Bu tez çalışmasında gerçekleştirilen FCGR uygulamasında k' nın 3 ile 8 değerleri arasındaki değerleri kullanılarak kodlamalar gerçekleştirilmiştir. Ardından bir CNN modeli olan InceptionV3 ile sınıflandırılmıştır. Zhao ve arkadaşları da çalışmalarında uyguladıkları FCGR yöntemi için k'nın 5, 6, 7 ve 8 değerlerini kullanmışlardır (C. Zhao vd., 2023). k=8 de en yüksek doğruluk başarısını (%87,5) elde etmişlerdir fakat, artan k değerlerinde aşırı uyum gözlemlemişlerdir. Carts ve arkadaşları da k'nın 6, 7 ve 8 değerleri ile testlerini gerçekleştirmişlerdir. Önerdikleri hibrit DÖ yaklaşımı ile %96,22 genel doğruluk başarısı elde etmişlerdir (Cartes vd., 2022). Rizzo ve arkadaşları çalışmalarında; k'nın 5, 6 ve 7 değerleri için FCGR görüntülerini oluşturmuşlardır. Gerçekleştirdikleri CNN uygulamasında k=5 için %99,2 doğruluk başarısı elde etmişlerdir. Bu çalışmada ise literatürden farklı olarak, belirlenen

aralıkta k'nın 3 değeri için en yüksek doğruluk başarısı %99,85 olarak elde edilmiştir. Ancak k'nın 4 değeriyle doğruluk başarısının %97,36'ya düştüğü görülmüştür. Bu durum, doğru k değeri seçiminin model performansını önemli ölçüde etkileyebileceğini göstermektedir. k değeri arttıkça doğruluk oranında kademeli bir iyileşme görülmekle birlikte, bu artışlar istatistiksel olarak anlamlı bir farklılık yaratmamıştır.

Literatürdeki bir diğer grafiksel gösterim yöntemi ise DNAWalk yöntemidir. Kobori ve Mizuta çalışmalarında; DNA dizileri arasındaki benzerlikleri tahmin etmek için nükleotidleri iki boyutlu vektörlerle değiştirip bunları ardı ardına bağlayarak DNA dizilerini ikili görüntüler olarak ifade etmişlerdir (Kobori & Mizuta, 2016). Hossain ve arkadaşları çalışmalarında; bir DNA yörüngesinin belirli bir kısmını tekrar ziyaret etme sayısıyla ilişkili bilgileri dikkate alarak gri tonlamalı görüntüler oluşturmuşlardır (Hossain vd., 2021). DNA dizileri arasındaki benzerlikleri ölçmek için DNA yörünge görüntülerine LBP yöntemini uygulamışlardır. Literatürdeki grafiksel gösterimler ayrıca, DNA dizilerinin iki boyutlu (Guo vd., 2001; X. Q. Liu vd., 2006; Huang vd., 2008), üç boyutlu (X. Q. Qi vd., 2007; Cao vd., 2008), dört boyutlu (Chi & Ding, 2005), beş boyutlu (Liao vd., 2007), altı boyutlu (Liao & Wang, 2004) ve sekiz boyutlu (D. Zhang, 2019) olarak gerçekleştirilmiştir. Literatürdeki bu alandaki çalışmaların tümü, herhangi bir sınıflandırma için değil, dizi yapılarındaki benzerliklerin belirlenmesi ve filogenetik ağaç oluşturabilmek için gerçekleştirilmiş uygulamalardır. Bu tez çalışmasının 8. bölümünde gerçekleştirilen DNAWalk uygulamasında, elde edilen gen dizisi yörünge görüntüleri, DÖ yöntemi olan InceptionV3 kullanılarak %99,14 doğrulukla sınıflandırılmıştır. Literatür incelendiğinde DNAWalk yöntemi uygulanarak elde edilmiş görüntülerden oluşan bir veri setinin, YZ yöntemleri ile analiz edildiği bir çalışmaya rastlanmamıştır. Bu çalışmanın bu alanda ilk olduğu ve önemli bir katkı sağladığı düşünülmektedir.

PhyVirus veri setine uygulanan bir diğer kodlama yöntemi ise görüntü tabanlı bir gösterim yöntemi olan DNA Gri Ölçekli dönüşüm yöntemidir. Diğer kodlama yöntemleri ile karşılaştırma yapabilmek amacıyla aynı sınıflandırma yöntemi (InceptionV3) kullanılmış ve bu yöntemler arasında en yüksek doğruluk oranı %99,89 olarak elde edilmiştir. Santamaria ve arkadaşlarının gerçekleştirdiği bir çalışmada, aynı sınıflandırma yöntemi ile %80,8 doğruluk oranına ulaşıldığı görülmektedir (Santamaría vd., 2019). Öte yandan, Delibaş ve arkadaşları, çalışmalarında nükleotidlerden ziyade dinükleotidlere gri değer atamış ancak bu görüntüleri sınıflandırma amaçlı kullanmamışlardır. Onlar, DNA benzerliğini tespit ederek filogenetik ağaç oluşturmak amacıyla görüntülerin histogramlarını kullanmışlardır (Delibas & Arslan, 2020).

Bu tez çalışmasında bölüm 7 ve 8'de gerçekleştirilen uygulamalarda gen dizileri kodlama-sınıflandırma aşamalarına geçmeden önce veri seti temizleme işlemine tabi tutulmuştur. Çeşitli sebeplerden ötürü gen dizilerinde eksik veriler oluştuğundan analizlerin hiçbirine eksik veri içeren virüslere ait gen dizileri dahil edilmemiştir. Bölüm 10'da gerçekleştirilen uygulamada ise tamamen bu eksik verilerin tahminine odaklanılmıştır. Bunun için veri setindeki mevcut gen dizileri kullanılarak, karşılık gelen eksik değerleri atamak için uygun değerler tahmin edilmiştir. Literatürde bu alanda önerilen yöntemlerin çoğu hizalama, aynı veya benzer uzunlukta veri gerektirmektedir. PhyVirus veri setindeki gen dizilerinin farklı uzunlukları, mevcut eksik veri tahmin yöntemlerinin doğrudan uygulanmasını engellemiştir. Bu sorun, geliştirilen KNN-Imputation yaklaşımıyla çözülerek çalışmaya özgün bir katkı sağlanmıştır. KNN-Imputation yöntemi; ilk defa Troyanskaya ve arkadaşları tarafından önerilmiştir (Troyanskaya vd., 2001). Gürültülü zaman serileri ve zaman serisi olmayan diziler üzerinde uygulanmıştır. Yöntemin biraz daha gelişmiş farklı varyasyonları da mevcuttur (Brás & Menezes, 2007; X. Zhang vd., 2008; Keerin vd., 2012). Fakat bu çalışmalar da eşit veya benzer uzunluktaki verilerden oluşan veri setlerine uygulanabilmiştir. Literatürdeki çalışmalarda KNN-Imputation için farklı k değerleri ile analizler gerçekleştirilmiştir. Örneğin Bras ve Menezes'in gerçekleştirdiği bir çalışmada k'nın 5, 10, 15, 20 değerleri için analizler gerçekleştirilmiştir. Bu çalışmada ise k'nın 5 değeri ve veri setindeki eksik gözlem içeren dizilerin uzunluğuna bağlı olarak değişen k uzunluk değerleri ile analizler gerçekleştirilmiştir. Literatürdeki çalışmalarda, yapay eksik gözlemler oluşturularak analizlerin bu verilere göre gerçekleştirildiği görülmektedir (Brás & Menezes, 2007; X. Zhang vd., 2008; Keerin vd., 2012;). Brás ve Menezes'in uygulamasında eksik veri oranları <%5, %5–10, %10–20, %20–50 ve >%50 olarak belirlenmiştir. Zhang ve arkadaşlarının çalışmasında ise bu oranlar %3,7, %3,3, %3,8 ve %3 olarak tanımlanmıştır (X. Zhang vd., 2008). Keerin ve arkadaşlarının analiz ettikleri veri setinde ise eksik veri oranları %1, %2, %3, %4, %5, %10 ve %15 olarak belirlenmiş ve bu oranlar üzerinden analizler yapılmıştır (Keerin vd., 2012). Bu çalışmada ise %10'dan daha az sayıda eksik veri içeren gen dizilerinin tahmini gerçekleştirilmiş olup, herhangi bir yapay eksik veri oluşturma işlemi gerçekleştirilmemiştir.

Eksik verilerin başarılı bir şekilde tahmin edilmesi, genetik analizlerin daha doğru ve kapsamlı bir şekilde yapılmasını mümkün hale getirmektedir. Bu durum, gen dizilerindeki benzerliklerin ve farklılıkların daha net bir biçimde ortaya konulmasını

sağlamaktadır. Literatürdeki çalışmalar, genetik olarak hiçbir türün veya soyun, bağlantılı olduğu diğer türlerden tamamen bağımsız olarak incelenemeyeceğini sıklıkla vurgulamaktadır (Janes vd., 2010). Virüsler, küçük genomları ve hataya açık replikasyon mekanizmaları nedeniyle genellikle yüksek adaptif yeteneklere sahiptirler (Simmonds vd., 2018). Belirli konakçılardaki uygunluklarını hızla optimize ederler ve enfekte ettikleri konakçılardan daha fazla değişikliğe uğrayabilmektedirler (Simmonds vd., 2018). Devam eden değişikliklere rağmen, virüslerin ve uyum sağlamayı amaçladıkları konakçı organizmaların yaklaşık üç milyar yıl önce ortak bir atadan mevcut ekosistemleri oluşturmak üzere ayrıştığına inanılmaktadır (Bamford vd., 2005). Bu durum, gen dizilerindeki benzerliklerin virüs konak ve aile sınıflandırmalarında kullanılabileceğini göstermektedir.

12. SONUÇLAR ve ÖNERİLER

Viral genomların değişken yapıları ve bu değişikliklerin farklı konaklara uyum sağlama süreçlerindeki rolü, biyoinformatik alanında yoğun ilgi gören bir araştırma konusudur. Gen dizileri üzerine yapılan detaylı analizler, viral ailelerin ve konaklarının doğru sınıflandırılmasının salgın hastalıkların öngörülmesi ve tedavi stratejilerinin geliştirilmesi açısından taşıdığı önemi ortaya koymaktadır. Teknolojik gelişmeler ve YZ yöntemlerinin bu genetik veriler üzerindeki uygulamaları, bilim dünyasında ve tıp alanında önemli ilerlemelere kapı açmıştır. Bu bağlamda, genetik verilerin doğru ve etkili bir şekilde sınıflandırılması, biyoinformatik araştırmalarının merkezinde yer alan kritik bir çalışma alanı haline gelmiştir. Virüslerin geniş çeşitliliği ve karmaşık doğasını yönetebilecek gelişmiş YZ araçlarına olan ihtiyaç, bu araştırmanın temel motivasyonunu oluşturmuştur. Bu tez çalışmasında, günümüzün önemli biyoinformatik problemlerinden biri olan viral ailelerin ve konak sınıflandırmalarının daha hızlı ve doğru bir şekilde gerçekleştirilmesine odaklanmıştır. Viral genomlar, yüksek adaptif kapasiteleri ve değişkenlikleri nedeniyle genetik analiz süreçlerinde önemli zorluklar barındırmaktadır. Bu zorlukların üstesinden gelmek hem bilimsel hem de halk sağlığı açısından büyük bir öneme sahiptir. Bu çalışmada MÖ ve DÖ yöntemlerinin viral genom analizinde nasıl kullanılabileceği kapsamlı bir şekilde incelenmiş, elde edilen bulgular doğrultusunda önemli sonuçlara ulaşılmıştır. Özellikle RNA virüslerinin sınıflandırılmasındaki zorluklar dikkate alındığında, bu modellerin sağladığı başarı potansiyeli dikkat çekicidir. Elde edilen bulgular, gen dizilerinin daha doğru ve hızlı bir şekilde analiz edilmesine olanak sağlayarak, biyoinformatik alanında yeni bir bakış açısı sunmaktadır. Bu çalışma, genetik verilerin analizinde MÖ ve DÖ yöntemlerinin ne kadar etkili olabileceğini gözler önüne sererken, gelecekteki araştırmalar için de yeni fırsatlar yaratmaktadır. Viral genomların doğru sınıflandırılması, halk sağlığı stratejilerinin geliştirilmesi ve bireyselleştirilmiş sağlık uygulamalarının şekillendirilmesinde önemli bir rol oynamaktadır. Bu doğrultuda geliştirilen yöntemler, yalnızca bilimsel çalışmalarda değil, gerçek dünyada da çeşitli uygulama alanlarına sahiptir. Özellikle halk sağlığı açısından kritik bir öneme sahip olan viral sınıflandırma süreçlerinde bu modeller, salgınların öngörülmesi, tedavi stratejilerinin geliştirilmesi ve hasta odaklı tedavi yaklaşımlarına katkı sağlama potansiyeli taşımaktadır.

Sunulan tez çalışması, biyoinformatik uzmanlarının genetik analiz süreçlerinde harcadığı zamanı azaltarak daha verimli, hızlı ve ekonomik sonuçlar elde edilebilecek bir karar destek sistemi sunmayı hedeflemektedir. Özellikle genetik profil analizlerinin bireyselleştirilmiş tıp uygulamalarına yönelik katkıları, gen tedavisi ve hasta yanıtlarının tahmin edilmesi açısından büyük bir önem taşımaktadır. Viral genomların doğru bir şekilde sınıflandırılması, sadece halk sağlığı müdahalelerinde değil, aynı zamanda farmakogenomik ve enfeksiyon hastalıkları yönetimi gibi alanlarda da kullanılabilir. Böylelikle, elde edilen veriler ışığında geliştirilen yöntemler, klinik uygulamalarda karar verme süreçlerini hızlandırabilir ve tedavi süreçlerini iyileştirebilir.

Her bilimsel araştırmada olduğu gibi, bu çalışmanın da belirli sınırlamaları bulunmaktadır. Kullanılan PhyVirus veri setinin dengesiz yapısı ve RNA virüslerinin yüksek mutasyon oranları, sınıflandırma süreçlerinde zorluklara neden olmuştur. Ayrıca, gen dizilerinin karmaşıklığı ve boyut farklılıkları, büyük verilerin işlenmesinde ekstra zorluklar doğurmuştur. MÖ ve DÖ modelleri, gen dizi kodlama yöntemleri ile oluşturulan farklı veri temsillerine rağmen, bazı durumlarda gen dizilerinin belirgin özelliklerini yeterince iyi yakalayamamış ve bu da bazı viral türler ve konakların sınıflandırma doğruluğunu olumsuz etkilemiştir. Bu sınırlılıklar, gelecek çalışmalar için önemli birer rehber niteliği taşımaktadır. Özellikle, veri ön işleme süreçlerinin geliştirilmesi ve daha dengeli veri setlerinin oluşturulması, sınıflandırma doğruluğunu artırmak için atılacak önemli adımlardan biridir.

Çalışmanın hedeflerinden biri, viral aileler ve konakların doğru sınıflandırılması yoluyla viral tehditlerin daha iyi anlaşılmasını sağlamaktır. Bu hedef büyük ölçüde gerçekleştirilmiş olup, elde edilen yüksek doğruluk oranları modellerin başarı düzeyini ortaya koymaktadır. Elde edilen bulgular, literatürde belirtilen diğer çalışmalarla tutarlıdır ve genetik veri analizinin gelecekteki araştırmalar için önemini koruyacağını işaret etmektedir. Ayrıca, genetik verilerin büyük boyutları ve karmaşıklığı göz önünde bulundurulduğunda, MÖ ve DÖ yöntemlerinin sınıflandırma süreçlerindeki etkinliğinin artırılabileceği öngörülmektedir. Gelecek çalışmalarda, bu modellerin performansını artırmak için Transformer tabanlı modeller gibi daha gelişmiş algoritmaların kullanılması önerilmektedir. Ayrıca, benzer virüs türlerini ayırt edebilmek için ek özelliklerin ve veri kodlama yöntemlerinin araştırılması da önemli bir adım olacaktır. Böylece, viral sınıflandırmaların doğruluğu artırılabilir ve genelleme yetenekleri daha da iyileştirilebilir.

Bu tez çalışması, gen dizisi kodlama yöntemlerinin çeşitli sınıflandırma modelleriyle birleşmesinin genetik veri analizi üzerindeki etkisini incelemiş ve eksik veri tahmini için yenilikçi yaklaşımlar geliştirerek bu yöntemlerin biyoinformatikteki uygulamalarını ortaya koymayı amaçlamıştır. Elde edilen bulgular, yalnızca bu alandaki teorik katkıları değil, aynı zamanda genetik araştırmalarda uygulamalı çözümler sunma potansiyelini de göstermiştir. Genetik verilerin YZ yöntemleri ile analizi, gelecekte hem biyoinformatik hem de tıbbi araştırmalar açısından önemini koruyacak olup, bu tez çalışması, bu alandaki ilerlemelere önemli bir referans kaynağı olmuştur.

KAYNAKLAR

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., & Fletcher, M. (2001). Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, *17*(5), 429-437. https://doi.org/10.1093/BIOINFORMATICS/17.5.429

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ao, C., Jiao, S., Wang, Y., Yu, L., & Zou, Q. (2022). Biological Sequence Classification: A Review on Data and General Methods. *Research*, *2022*. https://doi.org/10.34133/RESEARCH.0011/SUPPL_FILE/RESEARCH.0011.F1.PDF

Arakawa, K., Tamaki, S., Kono, N., Kido, N., Ikegami, K., Ogawa, R., & Tomita, M. (2009). Genome Projector: zoomable genome map with multiple views. *BMC bioinformatics*, *10*, 31. https://doi.org/10.1186/1471-2105-10-31

Aswad, A., & Katzourakis, A. (2018). Cell-Derived Viral Genes Evolve under Stronger Purifying Selection in Rhadinoviruses. *Journal of Virology*, *92*(19). https://doi.org/10.1128/JVI.00359-18

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, *35*(3), 235-241. https://doi.org/10.1128/BR.35.3.235-241.1971

Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2005). What does structure tell us about virus evolution? *Current Opinion in Structural Biology*, *15*(6), 655-663. https://doi.org/10.1016/J.SBI.2005.10.012

Basu, S., & Campbell, R. H. (2023). Classifying COVID-19 Variants Based on Genetic Sequences Using Deep Learning Models. *Springer Series in Reliability Engineering*, 347-360. https://doi.org/10.1007/978-3-031-02063-6_19/COVER

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937-1967. https://doi.org/10.1007/S10462-020-09896-5/TABLES/12

Berdis, A. (2022). Nucleobase-modified nucleosides and nucleotides: Applications in biochemistry, synthetic biology, and drug discovery. *Frontiers in Chemistry*, *10*, 1051525. https://doi.org/10.3389/FCHEM.2022.1051525/BIBTEX

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(14), 5155-5159. https://doi.org/10.1073/PNAS.83.14.5155

Blaisdell, B. E. (1989). Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *Journal of Molecular Evolution*, *29*(6), 526-537. https://doi.org/10.1007/BF02602924

Brás, L. P., & Menezes, J. C. (2007). Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular Engineering*, *24*(2), 273-282. https://doi.org/10.1016/J.BIOENG.2007.04.003

Breiman, L. (2001). *Random Forests*. *45*, 5-32.

Bussi, Y., Kapon, R., & Reich, Z. (2021). Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS ONE*, *16*(10). https://doi.org/10.1371/JOURNAL.PONE.0258693

Cano Londoño, N. A., Velasco, J. O., García, F. C., & Franco, I. B. (2020). *SDG 6 Clean Water and Sanitation*. 85-104. https://doi.org/10.1007/978-981-32-9927-6_7

Cao, Z., Liao, B., & Li, R. (2008). A group of 3D graphical representation of DNA sequences based on dual nucleotides. *International Journal of Quantum Chemistry*, *108*(9), 1485-1490. https://doi.org/10.1002/QUA.21698

Cartes, J. A., Anand, S., Ciccolella, S., Bonizzoni, P., & Vedova, G. Della. (2022). Accurate and fast clade assignment via deep learning and frequency chaos game representation. *GigaScience*, *12*, 1-11. https://doi.org/10.1093/GIGASCIENCE/GIAC119

Chang, T. J., Yang, D. M., Wang, M. L., Liang, K. H., Tsai, P. H., Chiou, S. H., Lin, T. H., & Wang, C. T. (2020). Genomic analysis and comparative multiple sequences of SARS-CoV2. *Journal of the Chinese Medical Association*, *83*(6), 537-543. https://doi.org/10.1097/JCMA.0000000000000335

Chen, J., & Shi, X. (2019). Sparse Convolutional Denoising Autoencoders for Genotype Imputation. *Genes*, *10*(9). https://doi.org/10.3390/GENES10090652

Chi, R., & Ding, K. (2005). Novel 4D numerical representation of DNA sequences. *Chemical Physics Letters*, *407*(1-3), 63-67. https://doi.org/10.1016/J.CPLETT.2005.03.056

Cleydson, J., Silva, F., Carvalho, T. F. M., Fontes, E. P. B., & Cerqueira, F. R. (2017). *Fangorn Forest (F2): a machine learning approach to classify genes and genera in the family Geminiviridae*. https://doi.org/10.1186/s12859-017-1839-x

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). Why are de Bruijn graphs useful for genome assembly? *Nature biotechnology*, *29*(11), 987. https://doi.org/10.1038/NBT.2023

Dahl, A., Iotchkova, V., Baud, A., Johansson, S., Gyllensten, U., Soranzo, N., Mott, R., Kranis, A., & Marchini, J. (2016). A multiple phenotype imputation method for genetic studies. *Nature genetics*, *48*(4), 466. https://doi.org/10.1038/NG.3513

Dasari, C. M., & Bhukya, R. (2022). Explainable deep neural networks for novel viral genome prediction. *Applied Intelligence*, *52*(3), 3002-3017. https://doi.org/10.1007/S10489-021-02572-3/FIGURES/8

Delibas, E., & Arslan, A. (2020). *DNA sequence similarity analysis using image texture analysis based on first-order statistics*. https://doi.org/10.1016/j.jmgm.2020.107603

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2010). *ImageNet: A large-scale hierarchical image database*. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

Desai, H. P., Parameshwaran, A. P., Sunderraman, R., & Weeks, M. (2020). Deep Ensemble Models for 16S Ribosomal Gene Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12304 LNBI*, 282-290. https://doi.org/10.1007/978-3-030-57821-3_25/COVER

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., & Fertit, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular biology and evolution*, *16*(10), 1391-1399. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A026048

Dixit, P., & Prajapati, G. I. (2015). Machine learning in bioinformatics: A novel approach for DNA sequencing. *International Conference on Advanced Computing and Communication Technologies, ACCT*, *2015-April*, 41-47. https://doi.org/10.1109/ACCT.2015.73

Dubey, A., & Rasool, A. (2021). Efficient technique of microarray missing data imputation using clustering and weighted nearest neighbour. *Scientific Reports 2021 11:1*, *11*(1), 1-12. https://doi.org/10.1038/s41598-021-03438-x

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797. https://doi.org/10.1093/NAR/GKH340

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *https://doi.org/10.1214/aos/1013203451*, *29*(5), 1189-1232. https://doi.org/10.1214/AOS/1013203451

Gates, M. A. (1985). Simpler DNA sequence representations. *Nature 1985 316:6025*, *316*(6025), 219-219. https://doi.org/10.1038/316219a0

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Mach Learn*. https://doi.org/10.1007/s10994-006-6226-1

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Nets*. http://www.github.com/goodfeli/adversarial

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual review of psychology*, *60*, 549-576. https://doi.org/10.1146/ANNUREV.PSYCH.58.110405.085530

Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Kanmani, S. D., Venkatesan, C., & Dhas, C. S. G. (2021). Analysis of DNA Sequence Classification Using CNN and Hybrid Models. *Computational and Mathematical Methods in Medicine*, *2021*. https://doi.org/10.1155/2021/1835056

Guo, X., Randic, M., & Basak, S. C. (2001). A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chemical Physics Letters*, *350*(1-2), 106-112. https://doi.org/10.1016/S0009-2614(01)01246-5

Hammad, M. S., Ghoneim, V. F., Mabrouk, M. S., & Al-Atabany, W. I. (2023). A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques. *Scientific Reports* |, *13*, 4003. https://doi.org/10.1038/s41598-023-30941-0

Hamori, E., & Ruskin J. (1983). H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Bioinformatics*, *5*(4), 263-269. https://doi.org/10.1093/bioinformatics/5.4.263

Hazra, D., Kim, M. R., & Byun, Y. C. (2022). Generative Adversarial Networks for Creating Synthetic Nucleic Acid Sequences of Cat Genome. *International Journal of Molecular Sciences 2022, Vol. 23, Page 3701*, *23*(7), 3701. https://doi.org/10.3390/IJMS23073701

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597

Hoang, T., Yin, C., & Yau, S. S. T. (2016). Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics*, *108*(3-4), 134-142. https://doi.org/10.1016/J.YGENO.2016.08.002

Hossain, S. N., Kabir, M. H., & Pal, A. (2021). Alignment Free Sequence Similarity Estimation using Local Binary Pattern on DNA Trajectory Images. *2021 Joint 10th International Conference on Informatics, Electronics and Vision, ICIEV 2021 and 2021 5th International Conference on Imaging, Vision and Pattern Recognition, icIVPR 2021*. https://doi.org/10.1109/ICIEVICIVPR52578.2021.9564141

Huang, G., Liao, B., Li, Y., & Liu, Z. (2008). H–L curve: A novel 2D graphical representation for DNA sequences. *Chemical Physics Letters*, *462*(1-3), 129-132. https://doi.org/10.1016/J.CPLETT.2008.07.046

Isawa, H., Asano, S., Sahara, K., Iizuka, T., & Bando, H. (1998). Analysis of genetic information of an insect picorna-like virus, infectious ¯acherie virus of silkworm:

evidence for evolutionary relationships among insect, mammalian and plant picorna(-like) viruses*. *Arch Virol*, *143*, 127-143.

Janes, D. E., Organ, C. L., Fujita, M. K., Shedlock, A. M., & Edwards, S. V. (2010). Genome Evolution in Reptilia, the Sister Group of Mammals. *Annu. Rev. Genomics Hum. Genet*, *11*, 239-264. https://doi.org/10.1146/annurev-genom-082509-141646

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic acids research*, *18*(8), 2163-2170. https://doi.org/10.1093/NAR/18.8.2163

Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, *4*(4), 320-330. https://doi.org/10.1007/S40484-016-0081-2/METRICS

Joseph, J., & Sasikumar, R. (2006). Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, *7*(1), 1-10. https://doi.org/10.1186/1471-2105-7-243/TABLES/2

Kalton, G. (1982). *IMPUTING FOR MISSING SURVEY RESPONSES*.

Kaya, V., Tuncer, S., & Baran, A. (2020). *Derin Öğrenme Yöntemleri Kullanılarak Nesne Tanıma*.

Keerin, P., Kurutach, W., & Boongoen, T. (2012). Cluster-based KNN missing value imputation for DNA microarray data. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 445-450. https://doi.org/10.1109/ICSMC.2012.6377764

Kim, H., Golub, G. H., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, *21*(2), 187-198. https://doi.org/10.1093/BIOINFORMATICS/BTH499

Kingma, D. P., & Lei Ba, J. (2015). *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION*.

Kobori, Y., & Mizuta, S. (2016). Similarity Estimation Between DNA Sequences Based on Local Pattern Histograms of Binary Images. *Genomics, Proteomics & Bioinformatics*, *14*(2), 103-112. https://doi.org/10.1016/J.GPB.2015.09.007

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 145-151. https://doi.org/10.1145/3383972.3383975

Kustin, T., & Stern, A. (2021). Biased Mutation and Selection in RNA Viruses. *Molecular Biology and Evolution*, *38*(2), 575-588. https://doi.org/10.1093/MOLBEV/MSAA247

Leong, P. M., & Morgenthaler, S. (1995). Random walk and gap plots of DNA sequences. *CABIOS*, *11*(5), 503-507. https://academic.oup.com/bioinformatics/article/11/5/503/236062

Li, H., & Sun, F. (2018). Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. *Scientific Reports*, *8*(1), 1-9. https://doi.org/10.1038/s41598-018-28308-x

Liang, Q., Bible, P. W., Liu, Y., Zou, B., & Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics*, *2*(1). https://doi.org/10.1093/NARGAB/LQAA009

Liao, B., Li, R., Zhu, W., & Xiang, X. (2007). On the similarity of DNA primary sequences based on 5-D representation. *Journal of Mathematical Chemistry*, *42*(1), 47-57. https://doi.org/10.1007/S10910-006-9091-Z/METRICS

Liao, B., & Wang, T. M. (2004). Analysis of Similarity/Dissimilarity of DNA Sequences Based on Nonoverlapping Triplets of Nucleotide Bases. *Journal of Chemical Information and Computer Sciences*, *44*(5), 1666-1670. https://doi.org/10.1021/CI034271F

Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). When should we ignore examples with missing values? *International Journal of Data Warehousing and Mining*, *13*(4), 53-63. https://doi.org/10.4018/IJDWM.2017100104

Liu, D., Tedbury, P. R., Lan, S., Huber, A. D., Puray-Chavez, M. N., Ji, J., Michailidis, E., Saeed, M., Ndongwe, T. P., Bassit, L. C., Schinazi, R. F., Ralston, R., Rice, C. M., & Sarafianos, S. G. (2019). Visualization of Positive and Negative Sense Viral RNA for Probing the Mechanism of Direct-Acting Antivirals against Hepatitis C Virus. *Viruses*, *11*(11). https://doi.org/10.3390/V11111039

Liu, X. Q., Dai, Q., Xiu, Z. L., & Wang, T. M. (2006). PNN-curve: A new 2D graphical representation of DNA sequences and its application. *Journal of Theoretical Biology*, *243*(4), 555-561. https://doi.org/10.1016/J.JTBI.2006.07.018

Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Mulders, D. G. J. C., Molenkamp, R., Perez-Romero, C. A., Claassen, E., Garssen, J., & Kraneveld, A. D. (2020). Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Scientific Reports* |, *11*, 947. https://doi.org/10.1038/s41598-020-80363-5

Löchel, H. F., & Heider, D. (2021). Chaos game representation and its applications in bioinformatics. *Computational and Structural Biotechnology Journal*, *19*, 6263-6271. https://doi.org/10.1016/J.CSBJ.2021.11.008

Lu, C. B., & Mei, Y. (2018). An Imputation Method for Missing Data Based on an Extreme Learning Machine Auto-Encoder. *IEEE Access*, *6*, 52930-52935. https://doi.org/10.1109/ACCESS.2018.2868729

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel

coronavirus: implications for virus origins and receptor binding. *The Lancet*, *395*(10224), 565-574. https://doi.org/10.1016/S0140-6736(20)30251-8

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature 2005 437:7057*, *437*(7057), 376-380. https://doi.org/10.1038/nature03959

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560-564. https://doi.org/10.1073/PNAS.74.2.560

Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., Iida, T., Yasunaga, T., Horii, T., Arakawa, K., Kasahara, M., & Nakamura, S. (2014). Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, *15*(1), 1-9. https://doi.org/10.1186/1471-2164-15-699/COMMENTS

Mock, F., Viehweger, A., Barth, E., & Marz, M. (2020). *VIDHOP, viral host prediction with deep learning*. https://doi.org/10.1093/bioinformatics/btaa705

Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S. Y., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G. A., & Georgakopoulos-Soares, I. (2024). A survey of k-mer methods and applications in bioinformatics. *Computational and Structural Biotechnology Journal*, *23*, 2289-2303. https://doi.org/10.1016/J.CSBJ.2024.05.025

Nandy A., N. P. (1995). Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. *Current Science*.

Newman, D. A. (2014). Missing Data. *https://doi.org/10.1177/1094428114548590*, *17*(4), 372-411. https://doi.org/10.1177/1094428114548590

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, *19*(16), 2088-2096. https://doi.org/10.1093/BIOINFORMATICS/BTG287

Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., & Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature*, *356*(6365), 168-170. https://doi.org/10.1038/356168A0

Pfeifer, B., Holzinger, A., & Schimek, M. G. (2022). Robust Random Forest-Based All-Relevant Feature Ranks for Trustworthy AI. *Studies in Health Technology and Informatics*, *294*, 137-138. https://doi.org/10.3233/SHTI220418

PhyVirus | adi-stern. (2021). https://www.sternadi.com/phyvirus

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C. N., Dietrich, J., Klem, E. B., & Scheuermann, R. H. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, *40*(D1), D593-D598. https://doi.org/10.1093/NAR/GKR859

Pigott, T. D. (2001). A Review of Methods for Missing Data. *Educational Research and Evaluation*, *21*(1), 353-383. https://doi.org/10.1076/EDRE.7.4.353.8937

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & Depristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology 2018 36:10*, *36*(10), 983-987. https://doi.org/10.1038/nbt.4235

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4), 7-9. https://doi.org/10.5120/IJCA2017915495

Qi, X. Q., Wen, J., & Qi, Z. H. (2007). New 3D graphical representation of DNA sequence based on dual nucleotides. *Journal of Theoretical Biology*, *249*(4), 681-690. https://doi.org/10.1016/J.JTBI.2007.08.025

Qi, Z.-H., & Fan, T.-R. (2007). *PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization*. https://doi.org/10.1016/j.cplett.2007.06.029

Qiu, Y. L., Zheng, H., & Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. *GigaScience*, *9*(8), 1-12. https://doi.org/10.1093/GIGASCIENCE/GIAA082

Remita, A. M., & Diallo, A. B. (2019). Statistical Linear Models in Virus Genomic Alignment-free Classification: Application to Hepatitis C Viruses. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 474-481. https://doi.org/10.1109/BIBM47256.2019.8983375

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2019). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, *8*(1), 64-77. https://doi.org/10.1007/s40484-019-0187-4

Rizzo, R., Fiannaca, A., La Rosa, M., & Urso, A. (2016). Classification experiments of DNA sequences by using a deep neural network and chaos game representation. *ACM International Conference Proceeding Series*, *1164*, 222-228. https://doi.org/10.1145/2983468.2983489

Saha, S., Bandopadhyay, S., Ghosh, A., & Dey, K. N. (2017). An ensemble based missing value estimation in DNA microarray using artificial neural network. *Proceedings - 2016 2nd IEEE International Conference on Research in Computational Intelligence*

and Communication Networks, ICRCICN 2016, 279-284. https://doi.org/10.1109/ICRCICN.2016.7813671

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, *26*(2), 544-548.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463-5467. https://doi.org/10.1073/PNAS.74.12.5463

Sanggaard, K. W., Bechsgaard, J. S., Fang, X., Duan, J., Dyrlund, T. F., Gupta, V., Jiang, X., Cheng, L., Fan, D., Feng, Y., Han, L., Huang, Z., Wu, Z., Liao, L., Settepani, V., Thøgersen, I. B., Vanthournout, B., Wang, T., Zhu, Y., … Wang, J. (2014). *ARTICLE Spider genomes provide insight into composition and evolution of venom and silk*. https://doi.org/10.1038/ncomms4765

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733-9748. https://doi.org/10.1128/JVI.00694-10

Santamaría, L. A., Zuñiga, S., Pineda, I. H., Somodevilla, M. J., & Rossainz, M. (2019). *DNA Sequence Recognition using Image Representation*.

Schmidt, B., & Hildebrandt, A. (2021). Deep learning in next-generation sequencing. *Drug Discovery Today*, *26*(1), 173. https://doi.org/10.1016/J.DRUDIS.2020.10.002

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature Methods 2008 5:1*, *5*(1), 16-18. https://doi.org/10.1038/nmeth1156

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., & Church, G. M. (2005). Molecular biology: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728-1732. https://doi.org/10.1126/SCIENCE.1117389/SUPPL_FILE/SHENDURE.SOM.PDF

Simmonds, P., Aiewsakun, P., & Katzourakis, A. (2018). *Prisoners of war — host adaptation and its constraints on virus evolution*. https://doi.org/10.1038/s41579-018-0120-2

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1). https://doi.org/10.1002/CPMB.59

Smith, M. (2017). DNA sequence analysis in clinical medicine, proceeding cautiously. *Frontiers in Molecular Biosciences*, *4*(MAY), 24. https://doi.org/10.3389/FMOLB.2017.00024/XML/NLM

Soliman, N. F., Abd-Alhalem, S. M., El-Shafai, W., Abdulrahman, S. E. S. E., Ismaiel, N., El-Rabaie, E. S. M., Algarni, A. D., Algarni, F., Alhussan, A. A., & El-Samie, F.

E. A. (2022). Hybrid Approach for Taxonomic Classification Based on Deep Learning. *Intelligent Automation and Soft Computing*, *32*(3), 1881-1891. https://doi.org/10.32604/IASC.2022.017683

Solis-Reyes, S., Avino, M., Poon, A., & Kari, L. (2018). An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLOS ONE*, *13*(11), e0206409. https://doi.org/10.1371/JOURNAL.PONE.0206409

Souza, L. C., Azevedo, K. S., de Souza, J. G., Barbosa, R. de M., & Fernandes, M. A. C. (2023). New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning. *BMC Bioinformatics*, *24*(1), 1-19. https://doi.org/10.1186/S12859-023-05188-1/TABLES/9

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631-643. https://doi.org/10.1093/BIOINFORMATICS/BTI033

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118. https://doi.org/10.1093/BIOINFORMATICS/BTR597

Sukhorukov, G., Khalili, M., Gascuel, O., Candresse, T., Marais-Colombel, A., & Nikolski, M. (2022). VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Frontiers in Bioinformatics*, *2*(May), 1-12. https://doi.org/10.3389/fbinf.2022.867111

Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947-1958. https://doi.org/10.1021/CI034160G/SUPPL_FILE/CI034160GSI20031008_041202.ZIP

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, *2020*(4), e270. https://doi.org/10.7717/PEERJ-CS.270/SUPP-2

Tampuu, A., Bzhalava, Z., Dillner, J., & Vicente, R. (2019). ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS ONE*, *14*(9), 1-17. https://doi.org/10.1371/journal.pone.0222271

Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics : TIG*, *30*(9), 418-426. https://doi.org/10.1016/j.tig.2014.07.001

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673. https://doi.org/10.1093/NAR/22.22.4673

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001a). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525. https://doi.org/10.1093/BIOINFORMATICS/17.6.520

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001b). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525. https://doi.org/10.1093/BIOINFORMATICS/17.6.520

Viñas, R., Azevedo, T., Gamazon, E. R., & Liò, P. (2020). Gene Expression Imputation with Generative Adversarial Imputation Nets. *bioRxiv*, 2020.06.09.141689. https://doi.org/10.1101/2020.06.09.141689

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 1096-1103. https://doi.org/10.1145/1390156.1390294

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737-738. https://doi.org/10.1038/171737A0

Watts, J. D., Powell, S. L., Lawrence, R. L., & Hilker, T. (2010). Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. *Remote Sensing of Environment*, *115*, 66-75. https://doi.org/10.1016/j.rse.2010.08.005

Wimmer, E., & Goldbach, R. (1996). Viral Genetics. *Current Opinion in Genetics and Development*, *2*(1), 59-60. https://doi.org/10.1016/S0959-437X(05)80322-3

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V, & Koonin, E. V. (2018). *Origins and Evolution of the Global RNA Virome*. https://doi.org/10.1128/mBio.02329-18

Wood, E. J. (1983). Molecular cloning. A laboratory manual by T Maniatis, E F Fritsch and J Sambrook. pp 545. Cold Spring Harbor Laboratory, New York. 1982. $48 ISBN 0-87969-136-0. *Biochemical Education*, *11*(2), 82-82. https://doi.org/10.1016/0307-4412(83)90068-7

Wu, Y., Wee, A., Liew, C., Yan, H., & Yang, M. (2003). *DB-Curve: a novel 2D method of DNA sequence visualization and representation*. www.elsevier.com/locate/cplett

Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160*, 104-118. https://doi.org/10.1016/J.KNOSYS.2018.06.012

Yoon, J., Jordon, J., & Van Der Schaar, M. (2018). *GAIN: Missing Data Imputation using Generative Adversarial Nets* (ss. 5689-5698). PMLR. https://proceedings.mlr.press/v80/yoon18a.html

Zhang, D. (2019). A New Numerical Method for DNA Sequence Analysis Based on 8-Dimensional Vector Representation. *Journal of Applied Mathematics and Physics*, *7*(12), 2941-2949. https://doi.org/10.4236/JAMP.2019.712204

Zhang, H., Hung, C. L., Liu, M., Hu, X., & Lin, Y. Y. (2019). NCNEt: Deep Learning Network Models for Predicting Function of Non-coding DNA. *Frontiers in Genetics*, *10*(MAY), 432. https://doi.org/10.3389/FGENE.2019.00432/BIBTEX

Zhang, L., Chen, F. X., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y. J., Hao, H. X., Yi, W., Li, M., & Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Frontiers in Microbiology*, *12*, 766364. https://doi.org/10.3389/FMICB.2021.766364/BIBTEX

Zhang, M., Yang, L., Ren, J., Ahlgren, N. A., Fuhrman, J. A., & Sun, F. (2017). *Prediction of virus-host infectious association by supervised learning methods*. https://doi.org/10.1186/s12859-017-1473-7

Zhang, X., Beinke, B., Kindhi, B. Al, & Wiering, M. (2021). *Comparing Machine Learning Algorithms with or without Feature Extraction for DNA Classification*. http://arxiv.org/abs/2011.00485

Zhang, X., Song, X., Wang, H., & Zhang, H. (2008). Sequential local least squares imputation estimating missing value of microarray data. *Computers in Biology and Medicine*, *38*(10), 1112-1120. https://doi.org/10.1016/J.COMPBIOMED.2008.08.006

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., MacKen, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., … Scheuermann, R. H. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, *45*(D1), D466-D474. https://doi.org/10.1093/NAR/GKW857

Zhang, Z.-J. (2009). *DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences*. *25*(9), 1112-1117. https://doi.org/10.1093/bioinformatics/btp130

Zhao, C., Chen, E., & Chen, T. (2023). Deep Learning Classification of Caries Based on DNA Sequences Transformed by Chaos Game Representation. *Proceedings - 2023 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023*, 4996-4998. https://doi.org/10.1109/BIBM58861.2023.10385889

Zhao, S. (2023). *ClueGAIN: Application of Transfer Learning On Generative Adversarial Imputation Nets (GAIN)*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2019). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555