



**T.C.
YALOVA ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI**

**ANTİ-KORONAVİRÜS PEPTİTLERİNİN PROTEİN KODLAMA
YÖNTEMLERİ İLE TESPİTİ**

YÜKSEK LİSANS TEZİ

HASİBE CANDAN

DANIŞMAN: PROF. DR. MURAT GÖK

**YALOVA
ŞUBAT 2022**



T.C.
YALOVA ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

ANTİ-KORONAVİRÜS PEPTİTLERİNİN PROTEİN KODLAMA YÖNTEMLERİ
İLE TESPİTİ

YÜKSEK LİSANS TEZİ

HASİBE CANDAN
185105006

DANIŞMAN: PROF. DR. MURAT GÖK

YALOVA
ŞUBAT 2022

ÖNSÖZ

Yüksek lisans eğitimim boyunca bilgi ve tecrübelerinden yararlandığım, her zaman desteğini gördüğüm, benim için yeni bir ufuk çizmiş olan, saygı değer tez danışmanım Prof. Dr. Murat GÖK hocama minnet ve teşekkürlerimi sunarım.

Maddi ve manevi destekleri ile her koşulda yanımda olan sevgili aile üyelerime ayrı ayrı şükranlarımı sunarım.

Şubat – 2022

Hasibe CANDAN





İÇİNDEKİLER

ÖNSÖZ	i
İÇİNDEKİLER	iii
TÜRKÇE İNGİLİZCE TERİMLER ve KISALTMALAR LİSTESİ	vii
ÇİZELGE LİSTESİ.....	xi
ŞEKİLLER LİSTESİ	xiii
ANTI-KORONAVİRÜS PEPTİTLERİNİN PROTEİN KODLAMA YÖNTEMLERİ İLE TESPİTİ	xv
ÖZET.....	xv
ABSTRACT.....	xvii
1. GİRİŞ	1
1.1. Tezin Amacı	3
1.2. Literatür Çalışması	3
1.3. Hipotez	6
1.4. Tez İçeriği.....	6
2. YÖNTEMLER	7
2.1. Protein Kodlama Yöntemleri.....	7
2.1.1. Amino Asit Kompozisyonu.....	7
2.1.2. Amino Asit Çifti	8
2.1.3. Kompozisyon Moment Vektörü	8
2.1.4. Bitişik Üçlü.....	9
2.1.5. Kompozisyon Geçiş Dağılım.....	9
2.1.6. Digram	10
2.1.7. Amino Asitlerin Yer Değiştirme Matrisi	12
2.1.8. Taylor Venn Diyagramı	14
2.2. Önişleme Yöntemleri	15
2.2.1. Sınıf Dengeleme	15
2.2.2. Ayırıklaşma.....	16
2.3. Sınıflandırıcı Algoritmalar	16
2.3.1. Naif Bayes	17
2.3.2. Bayes Ağları	17
2.3.3. k-En Yakın Komşuluğu	17
2.3.4. Rastgele Orman	18
2.3.5. Çok Katmanlı Algılayıcı.....	18



2.3.6. Lojistik Regrasyon.....	19
2.3.7. Destek Vektör Makineleri.....	20
3. GELİŞTİRİLEN PROTEİN KODLAMA YÖNTEMLERİ.....	23
3.1. 2gBLO Yöntemi	23
3.2. 2gBLOTVD Yöntemi.....	25
4. BULGULAR	27
4.1. Veri Setleri	27
4.2. Deneysel Sonuçlar	28
5. SONUÇLAR	39
KAYNAKLAR	41
ÖZGEÇMİŞ	47



TÜRKÇE İNGİLİZCE TERİMLER ve KISALTMALAR LİSTESİ

Amino Asit Bileşimi	: Amino Acid Composition
Alıcı İşletim Karakteristiği	: Receiver Operating Characteristic
Amino Asit Çifti	: Amino Acid Pair
Bitişik Üçlü	: Conjoint Triad
Destek Vektör Makineleri	: Support Vector Machines
Dipeptit Kompozisyon	: Dipeptide Composition
Kompozisyon Moment Vektörü	: Composition Moment Vector
Kabul Edilen Noktasal Mutasyon	: Point Accepted Mutation
k-En Yakın Komşuluğu	: k-Nearest Neighborhood
Kompozisyon, Geçiş ve Dağılım	: Composition, Transition, Distribution
Matthew Corelation Katsayısı	: Matthew Correlation Coefficient
ROC Eğrisi Altında Kalan Alan	: ROC Area Under the Curve
Taylor'ın Venn Diyagramı	: Taylor's Venn Diagram
Yer Değiştirme Matrisi	: Block Substitution Matrix



KISALTMALAR

AAC	: Amino Asit Bileşimi
AAP	: Amino Asit Çifti
AMP	: Anti Mikrobiyal Peptit
Anti-CoV	: Koronavirüs Karşıtı Peptit
AUC	: ROC Eğrisi Altında Kalan Alan
AVP	: Anti Virüsal Peptit
CMV	: Kompozisyon Moment Vektörü
CT	: Bitişik Üçlü
CTD	: Kompozisyon, Geçiş ve Dağılım
CTDC	: Kompozisyon Yöntemi
CTDT	: Geçiş Yöntemi
DC	: Dipeptit Kompozisyon
DVM	: Destek Vektör Makineleri
DN	: Doğru Negatif
DP	: Doğru Pozitif
k-EYK	: k-En Yakın Komşuluğu
MKK	: Matthew Korelasyon Katsayısı
N	: Negatif
Non-AMP	: Anti Mikrobiyal Olmayan Peptit
Non-AVP	: Anti Virüsal Olmayan Peptit
P	: Pozitif
ROC	: Alıcı İşletim Karakteristiği
TVD	: Taylor'ın Venn Diyagramı
YN	: Yanlış Negatif
YP	: Yanlış Pozitif

ÇİZELGE LİSTESİ

Çizelge 2.1. Amino asitlerin kimyasal özelliklerine göre atanması.....	10
Çizelge 2.2. Digram tekrar sayısı.....	11
Çizelge 2.3. BLOSUM62 matrisi.....	14
Çizelge 3.1. Digram kodlama sonucu oluşan 20x20'lik matris.....	23
Çizelge 3.2. Digram ile BLOSUM62'nin 20x20'lik hibrit matrisi	25
Çizelge 3.3. Digram ile BLOSUM62 ve TVD'nin 20x20'lik hibrit matrisi	26
Çizelge 4.1. Veri setlerindeki örnek sayıları.....	27
Çizelge 4.2. 2gBLO'ya ait Anti-CoV ve AVP veri seti üzerindeki sınıflandırma sonuçları	30
Çizelge 4.3. 2gBLO'ya ait Anti-CoV ve non-AVP veri seti üzerindeki sınıflandırma sonuçları.....	31
Çizelge 4.4. 2gBLO'ya ait Anti-CoV ve non-AMP veri seti üzerindeki sınıflandırma sonuçları.....	32
Çizelge 4.5. 2gBLOTVD'ye ait Anti-CoV ve Antivirüs veri seti üzerindeki sınıflandırma sonuçları.....	33
Çizelge 4.6. 2gBLOTVD'ye ait Anti-CoV ve non-AVP veri seti üzerindeki sınıflandırma sonuçları.....	34
Çizelge 4.7. 2gBLOTVD'ye ait Anti-CoV ve non-AMP veri seti üzerindeki sınıflandırma sonuçları.....	35
Çizelge 4.8. Protein kodlama yöntemlerinin Anti-CoV ve non-AMP veri seti üzerinde en dengeli sonucu veren sınıflandırıcı sonuçları	36
Çizelge 4.9. Protein kodlama yöntemlerinin Anti-CoV ve non-AVP veri seti üzerinde en dengeli sonucu veren sınıflandırıcı sonuçları	37
Çizelge 4.10. Protein kodlama yöntemlerinin Anti-CoV ve AVP veri seti üzerinde en dengeli sonucu veren sınıflandırıcı sonuçları	38



ŞEKİLLER LİSTESİ

Şekil 2.1. Protein diziliminde digram	11
Şekil 2.2. Taylor'ın Venn diyagramı.....	15
Şekil 3.1. Sınıflandırma için oluşturulan model.....	23
Şekil 4.1. Karmaşıklık matrisi.....	28





ANTI-KORONAVİRÜS PEPTİTLERİNİN PROTEİN KODLAMA YÖNTEMLERİ İLE TESPİTİ

ÖZET

Patojenler canlılar üzerine yerleşir ve kendi proteinlerini canlının hücre çekirdeğine bırakarak protein yapısını bozar. Bozulan protein, fonksiyonlarını gerçekleştiremez ve canlının biyolojik işleyişi aksar. Patojenin canlı üzerine yerleşip çoğalmasıyla salgın hastalıklar ortaya çıkar. Günümüzde büyük bir salgına dönüşen Kovid-19 hastalığı, dünya çapında kitlesel can kayıplarına yol açmaktadır. Hastalığa karşı aşının keşfi Kovid-19 hastalığının önlenmesinde kesin sonuç verememiştir. İlaç geliştirme çalışmaları devam etmektedir. Biyoenformatik ve makine öğrenmesi temelli çalışmalar ilaç gelişimine büyük katkı sağlamaktadır.

Anti-koronavirüs (anti-CoV) peptitler, koronavirüsün inhibe edilmesini ve koronavirüsün zararlarının ortadan kaldırılmasını sağlayan bir tür terapötik ajandır. Anti-CoV peptit dizilimlerinin makine öğrenmesi yöntemleri ile tahmin edilmesi, koronavirüs hastalığının tedavisi ve ilaç gelişimi adımları için oldukça önemlidir. Ancak anti-CoV peptitleri diğer peptitlerden ayırt etme in vivo ya da in vitro ortamlarda çok daha fazla maliyet ve zaman gerektirir. Bu nedenle makine öğrenmesi temelli çalışmalar ile bilgisayar ortamında modellenerek tahmin edilmesi daha avantajlıdır. Bu tez çalışmasında, Kovid-19'a karşı terapötik ajan olan anti-CoV peptit dizilimlerinin yüksek doğrulukla tahmin edilebilmesi için protein verilerinin, protein kodlama yöntemleri ile sayısallaştırılması amaçlanmaktadır. Bu minvalde, 2gBLO ve 2gBLOTVD isimli iki yeni protein kodlama yöntemi geliştirilmiştir. Anti-CoV peptit veri seti pozitif sınıf, antimikrobiyal (AMP), antimikrobiyal olmayan (non-AMP) ve antivirüs olmayan (non-AVP) üç adet negatif sınıf peptit veri seti kullanılmıştır. Bu protein verileri pozitif anti-CoV olacak şekilde her bir negatif veri seti ile ayrı ayrı birleştirilerek üç adet negatif ve pozitif sınıfa sahip veri setleri elde edilmiştir. Veri setleri geliştirilen protein kodlama yöntemleriyle sayısallaştırılmıştır. 2gBLO yöntemi, digram metodu kullanılarak kodlanmakta olan her bir dizilimdeki amino asit ikilisinin BLOSUM62 matrisindeki yer değiştirme değeri ile çarpılarak güncellenmesi sonucu elde edilmektedir. 2gBLOTVD yöntemi ise, Taylor'ın Venn diyagramındaki ortak fiziko kimyasal özellik sayısı ile 2gBLO yönteminden elde edilen değer çarpılmaktadır. Pozitif sınıf olan anti-CoV peptit veri setindeki örnek azlığı sebebiyle, dengesiz olan veri setleri, başarımlarını yükseltmek amacıyla, sınıf dengeleme işlemine tabi tutulmuştur. Ayrıca, sınıflandırma sonuçlarının etkisini artırmak için, ayrıklaştırma yöntemi uygulanarak başarımların metrikleri, Naif Bayes, Bayes Ağları, Rastgele Orman, Doğrusal Destek Vektör Makineleri (Doğrusal DVM), Radyal Destek Vektör Makineleri (Radyal DVM), k-En Yakın Komşuluk (k-EYK), Çok Katmanlı Algılayıcı (ÇKA) ve Lojistik Regresyon sınıflandırıcı algoritmaları ile elde edilmiştir.

Elde edilen deneysel sonuçlara göre; sınıflandırıcı algoritmalar üzerinde geliştirilen protein kodlama yöntemlerinin, literatürde öne çıkan protein kodlama yöntemlerinden daha yüksek başarımlar verdiğini görülmektedir.

Anahtar Kelimeler: Kovid-19, Terapötik ajan, Öznitelik kodlama, Sınıflandırma, Makine öğrenmesi.



DETERMINATION OF ANTI-CORONAVIRUS PEPTIDES BY PROTEIN CODING METHODS

ABSTRACT

Pathogens settle on living beings and leave their own proteins in the cell nucleus of the host, disrupting the protein structure. The deteriorated protein cannot perform its functions and the biological functioning of the living being is disrupted. Epidemics occur when the pathogen settles and multiplies on the living being. Today, the Covid-19 disease, which has turned into a major epidemic, causes mass casualties around the world. The discovery of the vaccine against coronavirus could not give definite results in the prevention of the Covid-19 disease. Hence, drug development studies go on all over the world. Bioinformatics and machine learning-based studies make a great contribution to drug development for the Covid-19.

Anti-coronavirus (anti-CoV) peptides are a kind of therapeutic agent that inhibits the coronavirus and eliminates the harms of the coronavirus. Prediction of anti-CoV peptide sequences with machine learning methods is very important for the treatment of coronavirus disease and drug development steps. However, distinguishing anti-CoV peptides from other peptides requires much more cost and time consuming in vivo and in vitro. For this reason, it is more advantageous to model and predict in silico environment with machine learning-based studies. In this thesis, it is aimed to digitize protein data with protein coding methods so that anti-CoV peptide sequences, which are therapeutic agents against Covid-19, can be predicted with high accuracy. In this regard, two new protein coding methods named 2gBLO and 2gBLOTVD have been developed. Anti-CoV peptide dataset positive class, antimicrobial (AMP), non-antimicrobial (non-AMP) and non-antiviral (non-AVP) three negative class peptide datasets were used. These protein data were combined with each negative dataset separately to be positive anti-CoV, and datasets with three negative and positive classes were obtained. Data sets were digitized with the developed protein coding methods. The 2gBLO method is obtained by updating the amino acid pair in each coded sequence using the digram method by multiplying the displacement value in the BLOSUM62 matrix. In the 2gBLOTVD method, the number of common physicochemical properties in Taylor's Venn diagram is multiplied by the value obtained from the 2gBLO method. Due to the small sample size in the positive class anti-CoV peptide dataset, unstable datasets were subjected to class balancing in order to increase their performance values. In addition, in order to increase the effect of classification results, discretization method is applied to performance metrics, Naive Bayes, Bayesian Networks, Random Forest, Linear Support Vector Machines (Linear SVM), Radial Support Vector Machines (Radial SVM), k-Nearest Neighborhood (k-NN), Multi-Layer Perceptron (MLA) and Logistic Regression classifier algorithms.

According to the experimental results obtained; protein coding methods developed on classifier algorithms give higher performance than hitherto protein coding methods that are prominent.

Keywords: Covid-19, Therapeutic agent, Feature encoding, Classification, Machine learning.

1. GİRİŞ

Patojenlerin, hücre içindeki proteinlere ait fonksiyonlara ve biyolojik işleyişe nasıl etki ettiğinin anlaşılması hastalıkların önlenmesi açısından önemlidir [1]. Patojenler diğer canlılar üzerine herhangi bir şekilde yerleşip kendi proteinlerini canlının hücre çekirdeğinden içeri bırakarak proteinlerin yapısını bozarlar. Bozulan protein fonksiyonları canlının biyolojik işleyişinin aksamasına yol açmaktadır. Patojenlerin bir konak üzerine yerleşip çoğalması sonucunda oluşan hastalıklara salgın hastalık denmektedir [2]. Salgın hastalıklar, Kovid-19 dahil olmak üzere, insanlık tarihinde kitlesel ölümlere yol açmıştır.

Geçmiş zamanlarda, bilinen en büyük salgın kara veba salgınıdır. 14. yüzyılın ortalarında bir bakterinin neden olduğu kara veba salgını, Asya'da başlayıp, sıçanlar ve pireler aracılığı ile dünyanın farklı bölgelerine yayılmıştır. Kara veba, dünyada yaklaşık 200 milyon insanın ölümüne neden olduğu tahmin edilmektedir [3].

Tarihte büyük ölümlere neden olan diğer salgın ise çiçek hastalığı salgınıdır. 16. yüzyılda Meksika'da yayılmaya başlamış ve altı ay kadar kısa sürede ülkenin tamamını kaplamıştır. 22 milyon olan Meksika nüfusu salgın sonrasında 14 milyona düşmüştür [4].

Birinci dünya savaşı sırasında ortaya çıkan İspanyol gribi de büyük bir salgına neden olmuş ve birkaç ay içinde dünya nüfusunun üçte birine bulaşmıştır. İspanyol gribi salgını dünya genelinde yaklaşık 100 milyon insanın ölümüne yol açtığı tahmin edilmektedir [5].

Günümüzde de salgınlar hızını kesmeden devam etmektedir. Nitekim, 2002 Kasım ayında Çin'de ortaya çıkan SARS salgınından 8455 kişi etkilenmiştir. Bu kişilerin 790'ı vefat ile sonuçlanmıştır. MERS ise 2012 Eylül ayında Arap yarımadasında ortaya çıkmıştır. MERS virüsü, 1698 kişinin hasta olmasına ve 23 Mart 2016 itibarıyla 609 kişinin ölmesine neden olmuştur [6].

2019 Aralık ayında, Çin'in Wuhan kentinde ortaya çıkan koronavirüs salgını geçmişteki salgın örneklerinde olduğu gibi hızla dünyaya yayılmıştır. Koronavirüsün neden olduğu Kovid-19 hastalığına Kasım 2021 ayı itibarıyla 250 milyon kişi yakalanmış ve beş milyon kişi bu hastalıktan hayatını kaybetmiştir. Ülkemizde ise dokuz milyon kişi bu hastalığa yakalanmış ve 72 bin kişi hayatını kaybetmiştir [7].

Dünya’da Kovid-19’a yakalanan hastalar, hastalığa dair belirti göstermeyebileceği gibi mekanik ventilasyon gerektiren solunum yetmezliğine kadar varan ağır tıbbi durumlarla karşı karşıya kalabilmektedir. İleri vakalarda, çoklu organ yetmezliği, sepsis ve septik gibi ciddi durumlar sonucu hastalar ölümcül durumlarla karşılaşılabilir. Günümüzde tedavi gören hastaların, yaş gruplarına göre değişim göstermekle birlikte genel ortalama %0,5’i hayatını kaybetmektedir. Yaş arttıkça, ölüm oranı da artmaktadır, 70-79 yaş aralığında %5 iken, daha üst yaş gruplarında bu oran, %8’e çıkabilmektedir [8]. Günümüzde, aşının bulunmasıyla birlikte kesin olarak Kovid-19 tedavi edilememiştir ve dünya çapında salgının önüne geçilememiştir. İlaç geliştirme çalışmaları devam etmektedir. Biyoenformatik ve makine öğrenmesi temelli çalışmalar da ilaç gelişiminde büyük katkı sağlamaktadır. İşlevsel aktiviteye sahip anti-koronavirüs (anti-CoV) peptitlerinin makine öğrenmesi yöntemleri ile tespiti, ilaç geliştirmede büyük önem taşımaktadır.

Proteinlerin en küçük yapı taşına amino asit denir. Peptitler vücudun tüm hücrelerinde bulunan küçük amino asit zinciridir. Değerli mikro besin maddeleri olarak peptitler, vitamin ve mineraller gibi vücudun denge durumunda kalmasını sağladıkları gibi, diğer mikro besinleri optimize ederek vücut fonksiyonlarını yerine getirme, büyüme ve sağlıklı yaşam gibi genetik işlevleri de yerine getirirler. Peptitler, amino asitlerin peptit bağı ile bağlanması sonucunda oluşmaktadır. 20 farklı aminoasidin farklı sayıda ve sırada moleküler olarak bir araya gelmesiyle oluşmuş moleküllerdir [9]. Son yıllarda birçok tedavide umut vaat eden peptit tabanlı stratejiler kullanılmaktadır. Peptitler genellikle kısa dizilimlerden meydana gelmelerine rağmen koronavirüs karşıtı peptitleri diğer peptitlerden ayırt etmek maliyetli bir işlemdir.

Anti-CoV peptitler koronavirüsün inhibe edilmesini ve güçlü etkileri olan koronavirüsün zararlarının ortadan kaldırılmasını sağlamaktadır. Bu nedenle anti-CoV dizilimlerinin makine öğrenmesi yöntemleri ile tahmin edilmesi oldukça önemlidir [10]. Koronavirüsün teşhisinde ve tedavisinde dört şekilde araştırmaya ihtiyaç duyulmaktadır. Bunlardan; birincisi, moleküler seviyede anti-CoV ile koronavirüsün etkileşiminin tespiti, ikincisi proteinler arası etkileşimler incelenerek etkileşim ağlarının tespiti, üçüncü olarak hücresel işlemlerin tespiti ve son olarak ise dokusal seviyedeki etkilerin tespiti. Tezde önerilen yöntemler ile birinci aşamaya katkı

sunularak moleküler seviyede anti-CoV ile koronavirüsün etkileşim tahmini yapılacaktır.

1.1. Tezin Amacı

Makine öğrenmesinde, sınıflandırıcı algoritma, uygulandığı veri altında yatan işlevi ne kadar iyi tanırsa, o kadar iyi başarımlar verir.

Biyoformatikte makine öğrenmesi yöntemleriyle gerçekleştirilen protein temelli tahmin modellerinde protein diziliminin kodlanması verinin altında yatan işlevin sınıflandırıcı algoritma tarafından tanınmasında önemli rol oynamaktadır. Bu tez çalışmada, Kovid-19 hastalığına karşı ilaç geliştirilmesinde kullanılan anti-CoV peptitlerin sınıflandırıcı algoritmalar tarafından yüksek sınıf doğruluk oranı ile tahmin edilebilmeleri için protein dizilimi kodlama yöntemleri geliştirilmiştir.

1.2. Literatür Çalışması

Günümüzde, protein diziliminden protein kodlama yöntemleri ve sınıflandırıcı algoritmalar ile terapötik ajanların tespiti önemli bir konudur [11], [12], [13]. Kovid-19 salgınının yayılmasını engellemek için patojenlere karşı olan terapötik ajanlar belirli moleküler hedeflere tanımlanmaktadır. Hedeflediği patojeni inhibe edici özellikte olan terapötik ajanlar, reseptörlere bağlanarak patojenin etkisini yok etmektedir. Kovid-19 hastalığının ortadan kaldırılması açısından terapötik ajanların keşfedilmesi çok önemlidir [14]. Kovid-19 ve diğer hastalıklara karşı anti-CoV ve anti mikrobiyal peptit (AMP) gibi terapötik yöntemler deneysel (in vivo, in vitro) ve hesaplamalı (in silico) olmak üzere iki başlık altında incelenebilir.

Abid, Nishant ve diğerleri [15], in silico ortamda, makine öğrenmesi yöntemleri ile AVPdb isimli online bir veri tabanı geliştirmişlerdir. Bu veri tabanı 60'tan fazla, tıbbi açıdan önemli virüsleri hedef alan, deneysel olarak doğrulanmış antiviral peptitlerin veritabanını oluşturmaktadır. Ayrıca her bir peptit için fizikokimyasal özellikleride içermektedir.

Thakur ve diğerleri [16], AMP verilerine dayalı olarak, farklı fonksiyonel AMP'leri tanımlamak için geliştirilmiş AVPpred, antiviral peptitlerin toplanması ve tahmini konusuna dayanmaktadır. Çalışmada amino asit bileşimi (AAC), fizikokimyasal

özellikler ve destek vektör makinesi kullanılarak, antivirüs peptitlerinin tahmini üzerine çalışılmıştır.

Veltri ve diğerleri [17], antimikrobiyal aktiviteyi tanımak için derin öğrenme tabanlı bir peptit tarayıcı tasarlanmışlardır. Geliştirdikleri ourDNN yöntemi ile en yüksek sınıf doğruluğu, Matthew Korelasyon Katsayısı (MMC) ve Alıcı İşletim Karakteristiği (ROC) eğrisi değerlerini elde etmişlerdir.

Gök ve diğerleri [18], 2012 yılında, majör histo uyumluluk moleküllerine bağlı özgül peptit komplekslerini tahmin etmek için yeni bir amino asit kodlama yöntemi önermişlerdir. Yazarlar, ortonormal kodlama yöntemi ile Taylor Venn diyagramını birleştirmiş ve doğrusal destek vektör makinelerini ile sınıflandırmışlardır.

Protein kodlamada kaos teoresinden faydalanarak yapılmış bir çalışma da bulunmaktadır. Yazarlar bu çalışmada, amino asit dizilimlerinin düzensiz bölgelerini tahmin etme üzerine çalışmışlardır. İlgili protein dizisinin kaotik yapısına göre seçilen peptitlerin fizikokimyasal özelliklerinden oluşturulmuş yeni bir öznitelik kodlama tekniği sunmuşlar ve veri seti üzerinde çeşitli sınıflandırıcı algoritmalar ile test etmişlerdir. Önerdikleri yöntem sınıf doğruluğu, duyarlık ve özgüllük açısından iyi başarımlar göstermiştir. Lyapunov üsleri kullanarak yapılan bu çalışma, düzensiz proteinlerin tahmin edilmesinde oldukça etkili sonuçlar ortaya koymuştur [19].

Jain ve diğerleri [20], B hücresi veri setinden faydalanarak, makine öğrenmesi yöntemlerini kullanarak Kovid-19 tahmini yapmışlardır. İlgili çalışmada, tahminlerde bulunurken faydalı olduğu ortaya çıkan topluluk öğrenme stratejileri önerilmiştir. Destek Vektör Makineleri (DVM), Naif Bayes, k-En Yakın Komşu (k-EYK), AdaBoost, Gradyan Artırma, XGBoost, Rastgele Orman ve Yapay Sinir Ağları ile tahmin yapmışlardır.

Jain ve diğerleri [21], 2018 yılında, geliştirdikleri protein kodlama yönteminde, sınıflandırıcı algoritmalar ile protein fonksiyonel sınıflarının, protein-protein etkileşimlerinin, hücre altı yerleşiminin ve spesifik özelliklere sahip peptitlerin tahminine yönelik bir çalışma yapmışlardır. Kendi geliştirdikleri yöntemin yanı sıra literatürde öne çıkan protein kodlama yöntemlerini içeren PROSES adlı bir çevrimiçi sunucu geliştirmişlerdir.

Devi ve diğerleri [22], Kovid-19 ve insan proteinleri arasındaki viral konak etkileşimlerinin dizi tabanlı tahmini için amino asit kompozisyonu ve bitişik üçlü yöntemi ile verileri sayısallaştırmış ve Polinomial DVM, Radyal DVM, Rastgele Orman algoritmaları ile sınıflandırmışlardır.

Pang ve diğerleri [23], Kovid-19'un tüm genomundan tek tek protein yapılarını belirlemek ve modellemek için makine öğrenmesi modeli tasarlamışlardır. Yazarlar, Kabasura Kudineer adlı poli-bitkisel bir Siddha ilacı olan bir baharatın, koronavirüsü engelleyecek bir ilaç olarak kullanabileceğini ortaya koymuşlardır.

Ge ve diğerleri [24], Kovid-19'a karşı olarak patojeni hedef alan potansiyel bir terapötik ajan keşfetmişlerdir. Hem in silico hem in vitro ortamda çalışmalar yaparak, bir poli-ADP-riboz polimeraz-1 inhibitörü olan CVL218'in Kovid-19'u Tedavi etmek için kullanılabileceğini göstermişlerdir.

Arghiani ve diğerleri [25], Kovid-19'a karşı potansiyel terapötik faydaları olan spesifik miRNA'ları bulmak için miRNA modülatörlerinin, Kovid-19 hastalarının tedavisinde yardımcı olabileceğini göstermişlerdir.

Pang ve diğerleri [10], Anti-CoV peptitlerini tanımlamak için tahmin modeli oluşturmuşlardır. Anti-CoV ile birlikte antimikrobiyal, antiviral veri setlerinden faydalanarak, protein kodlama yöntemleri ile sayısallaştırılmış veriler üzerinden, sınıf dengeleme yaparak, Rastgele orman algoritması kullanmışlardır.

Bu tez çalışmasında, in silico ortamda koronavirüsü inhibe eden Anti-CoV peptitlerin tahmini için iki yeni protein kodlama yöntemi geliştirildi. Geliştirilen bu yöntemler Anti-CoV, mikrobiyal peptit (AMP), anti mikrobiyal olmayan peptit (non-AMP) ve anti virüsal olmayan peptit (non-AVP) peptitlerinin bir araya getirilmesiyle oluşturulmuş veri seti üzerinde makine öğrenmesi sınıflandırıcı algoritmaları ile test edildi. Elde edilen deneysel sonuçlar literatürde öne çıkan protein kodlama yöntemleri ile Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, Çok Katmanlı Algılayıcı (ÇKA), Lojistik Regresyon ve DVM sınıflandırıcıları ile kıyaslandı.

1.3. Hipotez

Kovid-19 hastalığının tedavisinde, virüsün inhibe edilmesi ile ilaç geliştirme çalışmalarında in silico analizler önemli yer tutmaktadır. In silico ortamda, Anti-Cov peptitler, diğer benzeri peptitlerden makine öğrenmesi yöntemleri ile ayırt edilebilmektedir. Protein dizilimi kodlama yöntemleri geliştirilmesi makine öğrenmesi yöntemlerinin ayırt etme başarımını artırılabilir. Böylece Anti-CoV peptitlerin yüksek sınıf doğruluğu ile tahminine bilimsel olarak katkı sağlayacaktır.

1.4. Tez İçeriği

Tez giriş kısmı ile birlikte toplamda beş bölümden oluşmaktadır.

Giriş kısmında koronavirüsün etkileri, koronavirüs karşıtı peptitler, tezin amacı, literatür çalışması ve hipotez konularından bahsedilmiştir.

İkinci kısımda anti-CoV tespiti için kullanılan protein kodlama yöntemleri, Amino Asit Yer Değiştirme Matrisi, Taylor'ın Venn Diyagramı, ön işleme yöntemlerinden sınıf dengeleme ve ayırıştırma yöntemi anlatılmıştır. Devamında kullanılan sınıflandırma yöntemlerine değinilmiştir.

Üçüncü kısımda anti-CoV tespiti için geliştirilen protein kodlama yöntemi olan 2gBLO ve 2gBLOTVD modellerinden bahsedilmiştir.

Dördüncü kısımda geliştirilen modelden, kullanılan veri setlerinden, başarımlerinden ve modelin deneysel sonuçlarından tablolar halinde bahsedilmiştir.

Tezin beşinci bölümünde geliştirilen model hakkında değerlendirmeler yapıp sonuçlar karşılaştırılarak özetlenmiştir.

2. YÖNTEMLER

2.1. Protein Kodlama Yöntemleri

Makine öğrenmesi algoritmalarının kullanımı için bir öznitelik çıkarımı olan protein kodlama, protein dizilimlerinin sayısallaştırılması için kullanılmaktadır. Makine öğrenmesi, biyoenformatik, örüntü tanıma ve görüntü işleme alanlarında kullanılan protein kodlama, girdi olan verilerden türetilmiş değerler elde eder [26]. Protein kodlama yöntemlerine, Amino asit (AAC), Amino asit çifti (AAP), Bitişik üçlü (CT), Dipeptit kompozisyon (DC) ve Kompozisyon, geçiş ve dağılım (CTD) yöntemleri örnek verilebilir. CTD yöntemleri olarak CTDC ve CTDT öznitelikleri, proteinlerin amino asitlere ait kimyasal ve yapısal özelliklerini göstermektedir. Bu yöntemler protein dizi uzunluğundan bağımsız, eşit uzunlukta öznitelik vektörleri üretmektedir. Bunlardan ayrı olarak, Kompozisyon moment vektör (CMV) yönteminde, M kaçınıcı derece moment alınacağını belirten parametre olarak, öznitelik vektör boyutu $20 \times M$ olmaktadır [25]. Bu çalışmada kullanılmış olan protein kodlama yöntemlerinin açıklamaları ayrı ayrı aşağıda açıklanmaktadır.

2.1.1. Amino Asit Kompozisyonu

Amino Acid Composition (AAC) yöntemi, uzunluğu 20 olan basit bir peptit tanımlayıcıdır. 20 farklı amino asidin protein dizilimindeki frekansına göre öznitelik vektörü çıkartılır [26]. Protein dizilimi içerisinde her bir amino asidin tekrar sayısı hesaplanır ve toplam dizi uzunluğuna bölünür. Böylece, öznitelik vektörü içindeki sayısal değer hesaplanır.

Amino asit dizi uzunluğundan bağımsız olarak 1×20 boyutunda bir öznitelik vektörü oluşmaktadır. Örnek bir “DILSRLDKVEAEVQIDRLIT” amino asit dizisinde, alaninin (A) normalleştirilmiş oluşumunda A sayısı 1 ve dizi uzunluğu 20 olduğuna göre, (A Sayısı) / (Dizi Uzunluğu) hesaplandığında 0,05 değerini alacaktır.

Dizi uzunluğuna N denilirse, öznitelik değeri hesaplanmak istenen i. aminoasidin tekrar sayısına da n_i denilirse, bu proteine ait öznitelik vektörü denklemi 2.1’e göre oluşturulur. Denklem 2.1’de bulunan öznitelik vektörü sadece proteinde yer alan amino asitlerin kompozisyon bilgisini vermektedir.

$$F_{AAC} = \left[\frac{n_1}{N}, \frac{n_2}{N}, \dots, \frac{n_{20}}{N} \right] \quad (2.1)$$

2.1.2. Amino Asit Çifti

Amino Acid Pair (AAP) yöntemi, Chen ve diğerleri tarafından ortaya atılmıştır [27]. Bu yöntemde bir dipeptitin protein diziliminde var olma sıklığı diğer dipeptitlerinkiyle karşılaştırılır. Örnek olarak, “MKEAAMT” protein dizilimi, MK, KE, EA, AA, AM, MT şeklinde dipeptitlerine ayrılır. Tüm dipeptitler göz önünde bulundurulduğunda dizi uzunluğundan bağımsız olarak 20 farklı aminoasit için 20×20 , yani 400 boyutunda öznitelik vektörü oluşmaktadır. Denklem 2.2’de görüldüğü gibi, bir dipeptitin dizi içerisinde geçme sayısının, f_{AAP}^+ , diğer tüm dipeptitlerin sayısına, f_{AAP}^- , oranı ile hesaplanmaktadır [28].

$$R = \log \left(\frac{f_{AAP}^+}{f_{AAP}^-} \right) \quad (2.2)$$

Her dipeptit için R değeri hesaplandıktan sonra çıkan değerler denklem 2.3’deki gibi $[-1, +1]$ aralığına normalize edilir. Normalizasyon yapılmasının sebebi öğrenme aşamasında özniteliklerin birbirlerini baskılamasına engel olmaktır. Denklem 2.3’de min ve max olarak belirtilenler en düşük ve en yüksek R değerini ifade etmektedir.

$$R_{AAP} = 2 \left(\frac{R_{AAP} - \min}{\max - \min} \right) - 1 \quad (2.3)$$

2.1.3. Kompozisyon Moment Vektörü

Composition Moment Vector (CMV) yöntemi, proteinlerin amino asit dizilimine bakarak, amino asitlerin frekans ve konum bilgileri çıkarabilir. CMV kodlamaya göre oluşturulan öznitelik vektör boyutu K parametresine göre değişkenlik gösterip $K \times 20$ boyutunda olmaktadır [29]. AAC ve AAP kodlama yöntemlerinden farklı olarak, bu yöntemde sadece frekans bilgisi kullanılmaz, bunun yanında konum bilgisinde kullanılmaktadır. Bu yöntem, proteinlerin fasta formatındaki aminoasit dizilim yapısından hem konumunu hemde frekans bilgisini kullanarak öznitelik vektörünü elde eder. P herhangi bir protein, A_i bu proteinin aminoasit dizisi içerisinde geçen i. amino asidi olduğunda, 20 amino asitlerin tümü A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y şeklinde sıralandığını ve her birinin sırasıyla ($x_1, x_2, \dots, x_{20}$) değişkenlerine karşılık geldiği düşünüldüğünde N boyutundaki bir aminoasit dizisine sahip bir proteinin, $i = 1, 2, \dots, 20$ ve $k \geq 0$ için k’ıncı dereceden moment vektörü denklemi,

2.4'e göre hesaplanır. Denklemden geçen $n_{i,j}$, i. amino asidin j. konumuna, K_i ise i. aminoasidin dizi içindeki toplam sayısına karşılık gelmektedir.

$$x_i^{(k)} = \frac{1}{N(N-1)\dots(N-k)} \sum_{j=1}^{K_i} n_{i,j}^k \quad (2.4)$$

$k = 0$ için oluşan CMV vektörü, sadece frekans bilgisi olup AAC kodlama ile aynı sonucu vermektedir. K 'nın maksimum değeri için P proteinine ait moment matrisi denklem 2.5'e göre hesaplanır.

$$A_{K+1} = \begin{bmatrix} x_1^{(0)} & \dots & x_{20}^{(0)} \\ \vdots & \ddots & \vdots \\ x_1^{(K)} & \dots & x_{20}^{(K)} \end{bmatrix} \quad (2.5)$$

Herhangi iki protein için oluşturulan moment matrisinin aynı olması bu iki proteinin aynı olduğunu göstermektedir [30].

2.1.4. Bitişik Üçlü

Conjoint Triad (CT), kodlama yöntemi proteinlerin öznitelik vektörlerini amino asitlerin frekanslarına ve fiziko kimyasal özelliklerine bakarak çıkarmaktadır [31]. Boyutu uzunluğu sabit olup 343'tür. Amino asitlerin özelliklerine göre hangi fiziko kimyasal özellik sınıfına aitse ona atanmaktadır.

Daha sonra protein dizisindeki her bir amino asit kendisinden sonra gelen iki amino asit ile üçlü birimlere ayrılır. Bir proteinin öznitelik vektörü, üçlü yapılara ait frekansları hesaplanıp normalize edilmesi ile çıkarılır. Protein dizisi içinde geçen i. üçlü gruba ait tekrar sayısına f_i denilirse, bu proteine ait R öznitelik vektöründeki r_i değeri denklem 2.6'ya göre hesaplanmaktadır.

$$r_i = \frac{f_i - \min\{f_1, f_2, \dots, f_{343}\}}{\max\{f_1, f_2, \dots, f_{343}\}} \quad (2.6)$$

2.1.5. Kompozisyon Geçiş Dağılımı

Composition, Transition, Distribution (CTD), kompozisyon, geçiş ve dağılımın kısaltılması ile ifade edilmiştir. Bu yöntemde öznitelikler proteinlerin amino asitlere ait yapısal ve kimyasal özelliklerini göstermektedir. Bu kimyasal özellikler hidrofobiklik, normalize edilmiş Van der Waals hacmi, polarite, polarize edilebilirlik, yük, ikincil yapılar, çözücü erişilebilirliği şeklindedir. Bu özellikler ayrı ayrı ya da bir tek vektör olarak kullanılabilir [32]. Öznitelik vektörü bu yöntem için iki aşamada

oluşturulur. Birinci aşamada protein dizisi amino asitlerin bu yedi farklı kimyasal özelliğe göre sınıflandırılır. Bu özelliklerin her biri önceden belirlenmiş üç farklı sınıfa atanır [33]. Çizelge 2.1.'de verilmiş olan tüm gruplamalardan, örnek olarak polar özelliği; pozitif, negatif ve nötr olarak üç şekilde gruplanmaktadır.

Kompozisyon yöntemi CTDC olarak ifade edilir. Bu yöntemde kimyasal özelliklerine göre etiketlenmiş amino asitlerin frekansları hesaplanır ve normalize edilir. Öznitelik vektör uzunluğu yedi özellik ve üç sınıf olduğundan 21 olmaktadır.

Geçiş yöntemi CTDT olarak ifade edilir. Bu yöntemde ise öznitelik vektörü, etiketlenen her bir amino asidin kendisinden sonra gelen amino asit sınıfına ve grubuna geçişinin sayılıp normalize edilmesi ile bulunur. Ayrıca CTDC'de olduğu gibi öznitelik vektör boyutu da 21'dir [34].

Çizelge 2.1. Amino asitlerin kimyasal özelliklerine göre atanması

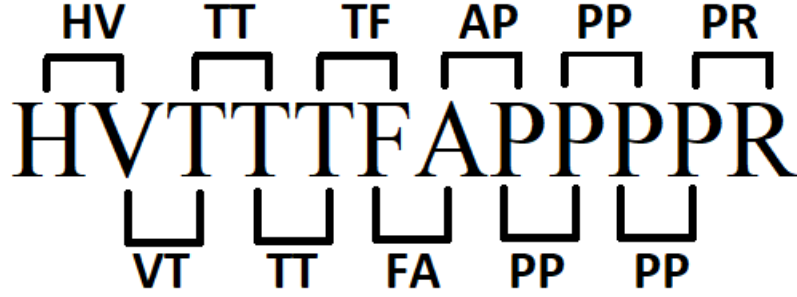
Fizikokimyasal Özellikler	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Hidrofobik	1	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
Pozitif	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0
Negatif	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Polar	0	1	1	1	0	0	1	0	1	0	0	1	0	1	1	1	1	0	1	1
Çok küçük	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Küçük	1	1	1	0	0	1	0	0	0	0	0	1	1	0	0	1	1	1	0	0
Aromatik	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1
Alifatik	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
Şarjlı	0	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0
Prolin	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2.1.6. Digram

N-gram, bir dizinin bir sonraki elemanını ardışık olarak, Markov zincirinden faydalanarak tahmin etmek için kullanılan bir istatistiksel dil modelidir. Doğal dil işleme, hesaplamalı dilbilim, istatistik gibi birçok alanda sıklıkla kullanılmaktadır. Aynı amaç ile kullanılan diğer algoritmalar ve modellerle karşılaştırıldığında, daha kolay olduğundan dolayı popülerdir. Ngram algoritması, ardışık bir dizi içinde tekrar

sayısını bulmak amacıyla kullanılmaktadır. n ile ifade edilen değişken, tekrarın kontrol edildiği değeri ifade ederken, gram da tekrar edilen bu değer dizideki ağırlığına karşılık gelmektedir. Bu tez çalışmasında, ngram yönteminin, n değeri iki olarak belirlenmiştir. n değeri üç ve daha fazla seçildiğinde protein dizilim uzunlukları kısa olduğundan protein dizilimi içinde rastlanacak örüntü miktarı azalacaktır.

Örnek olarak “HVTTFAPPPR” şeklinde bir dizilimde, digram değerlerine bakılırsa, dizilimdeki ardışık iki elemanlı olacak şekilde aminoasit bloklarına bölünür. Tekrar sayıları ile, protein dizilimleri, sayısallaştırılmış ve kodlanmıştır. Dizilimin, $n = 2$ ile hesaplanmış ngram değerleri Şekil 2.1’deki gibi olmaktadır. Tekrar sayıları ise Çizelge 2.2’de verilmiştir.



Şekil 2.1- Protein diziliminde digram

Çizelge 2.2. Digram tekrar sayısı

İkili AA	Tekrar
HV	1
VT	1
TT	2
TF	1
FA	1
AP	1
PP	3
PR	1

Her bir dizilim ayrı ayrı ele alınarak, dizideki ikili sayısına göre kodlanmaktadır. Çizelge 2.2’de tekrar sayıları verildiği gibi bu dizilimin kodlanmış hali, tekrar sayılarının yazılması ile olmaktadır.

2.1.7. Amino Asitlerin Yer Değiştirme Matrisi

Hizalama, en az iki genin ve protein dizisinde aynı hizada bulunan dizi elemanlarının karşılıklı olarak nasıl hizalandığının araştırıldığı bir biyoenformatik konusudur. Hizalama iki dizi ile gerekirse boşluklarda kullanılarak eşit uzunlukta olacak biçimde hizalanmasına global hizalama, dizilerin sadece belirli bir kısmının hizalanmasına ise yerel hizalama denmektedir.

Proteinlerin yapı taşı olan amino asitlerin hizalanmasında farklı kimyasal, fiziksel ve evrimsel şartlar vardır.

Hizalama, en az iki genin ve protein dizisinde aynı hizada bulunan dizi elemanlarının karşılıklı olarak nasıl hizalandığının araştırıldığı bir biyoenformatik konusudur. Hizalama iki dizi ile gerekirse boşluklarda kullanılarak eşit uzunlukta olacak biçimde hizalanmasına global hizalama, dizilerin sadece belirli bir kısmının hizalanmasına ise yerel hizalama denmektedir.

Proteinlerin yapı taşı olan amino asitlerin hizalanmasında farklı kimyasal, fiziksel ve evrimsel şartlar vardır.

Örnek olarak alanin amino asiti ile hangi amino asit hizalanmalıdır sorusuna valin mi yoksa lizin mi dersek, alanin ve valin küçük yapılı ve hidrofobiktir, lizin büyük ve suyu çeken yapıda olmasına göre, doğada daha çok ilk durum ile karşılaşmaktadır.

Bu görülme sıklıklarından yola çıkarak skorlama matrisleri tanımlanmaktadır. PAM (Percent Accepted Mutation) ve BLOSUM (Block Substitution Matrix) bu işlem için en çok kullanılan matrislerdir [35].

PAM global hizalama temelli olup az miktarda veri ile oluşturulmuştur ve global hizalamalarda ve evrimsel olarak benzer olan dizilimlerde daha iyi sonuç vermektedir. BLOSUM, yerel hizalama temellidir ve büyük miktarda veri ile oluşturulmuştur ve evrimsel olarak daha az benzeyen dizilimlerde daha iyi sonuç vermektedir. PAM ve BLOSUM'un ikisi aynı puanlama sonucunu verir, ancak farklı metodoloji kullanmaktadırlar [36]. Amino asitler için PAM matrisleri ve genetik olarak kodlanmış amino asitler için kullanılan tek harfli kısaltmalar Dayhoff tarafından geliştirilmiş ve ilk olarak 1978'de yayımlanmışlardır. Bir genin nükleotid diziliminin rastgele mutasyonları, buna karşılık gelen proteinin amino asit dizisini değiştirebilir. Bu mutasyonların bazıları proteinin yapısını çok değiştirmezken, bazıları ise proteinde

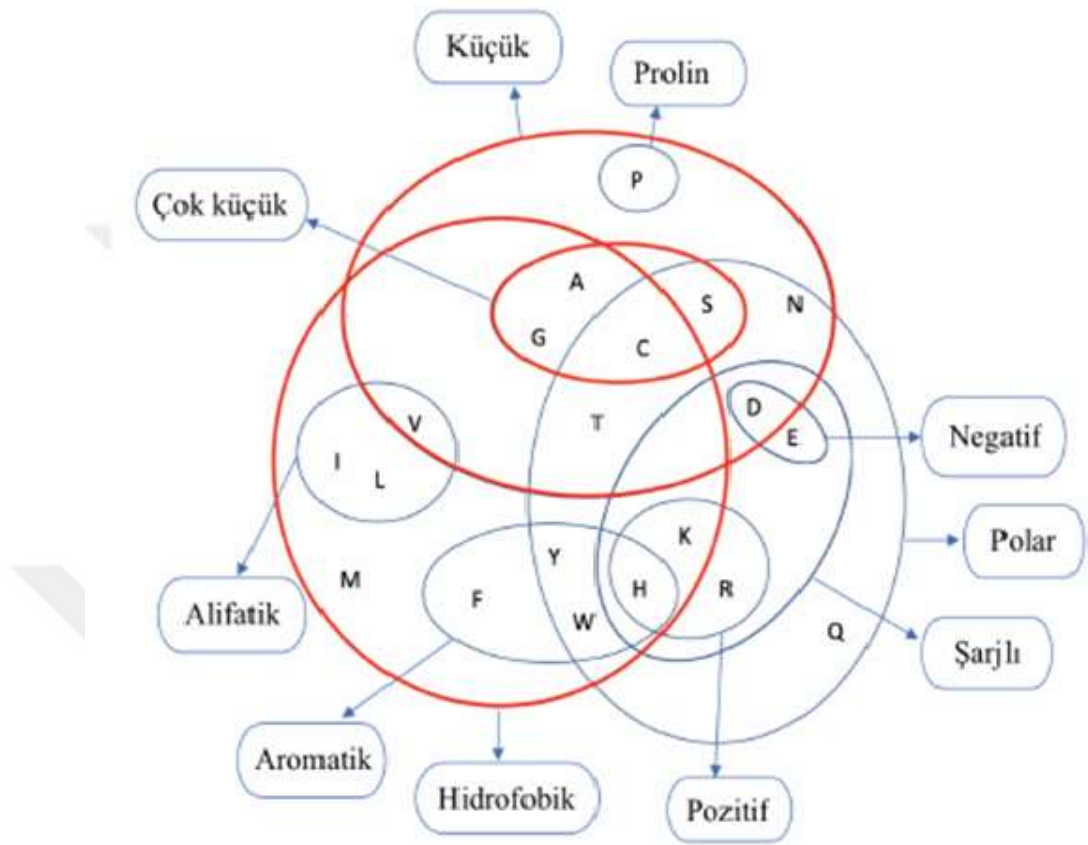
işlev kaybına bile sebep olmaktadır [37]. PAM matrisi çok benzer proteinlerin hizalanmasında tanımlanmaktadır. Daha az benzeyen proteinler için daha fazla mutasyon gerekebilir. PAM, tüm dizi çiftlerine doğru matris olarak hangisini seçeceğini önceden belirleyemez. Bunun için, BLOSUM62 gibi daha kullanışlı puanlama matrisi kullanmaya başlanmıştır. BLOSUM matrisleri, evrimsel olarak farklı protein dizileri arasındaki hizalamaları puanlamak amacıyla kullanılmaktadır [38]. BLOSUM62 matrisine blok ikame matrisi denilir. Steven ve Henikoff tarafından 1992’de ele alınan, BLOSUM matrisi, puanlamayı yaparken 2 tabanlı logaritmik olasılık kullanır. BLOSUM62 matrisi %62 ya da daha fazla amino asit kimliğine sahip bir hizalamada tüm proteinleri tek bir dizide birleştirmektedir [39]. Her blok içinde, ikili ortalama dizi kimlikleri bazı eşiklerden daha yüksek olduğunda, çoklu diziler kümelenir. Eşik, BLOSUM80 matrisi için %80, BLOSUM62 için %62, BLOSUM50 için %50 şeklindedir ve bunlardan fazla olmayan ortalama dipeptit kimliklerine sahip olacağı anlamına gelir. BLOSUM62 matrisi literatürde varsayılan (default) matris olduğu için seçilmiştir. Yer değiştirme matrislerinde, P_i ve P_j amino asitlerinin birbirlerine benzerlik oranı, i. satır ve j. sütunun kesişiminden elde edilmektedir [40]. Çizelge 2.3’ de BLOSUM62 amino asit yer değiştirme matrisi görülmektedir.

Çizelge 2.3. BLOSUM62 matrisi

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-2	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	-1	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	1	0	-3	0	1	-2	-1	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

2.1.8. Taylor Venn Diyagramı

Taylor Venn Diyagramı (TVD), amino asitlerin fizikokimyasal özelliklerini içeren ve proteinleri kodlayan bir yöntemdir [41]. TVD’de ilk olarak amino asitlerinin fizikokimyasal özellikleri belirlenir. Şekil 2.2’de Taylor venn diyagramında aminoasitler verilmiş fizikokimyasal özellikleri şekil halinde gösterilmiştir. Bu fizikokimyasal özellikler çok minik, alifatik, prolin, küçük, kutupsal, şarjlı, pozitif, negatif, aromatik ve hidrofobiktir.



Şekil 2.2. Taylor'ın Venn diyagramı [41]

Diyagramda kesişme ve kapsama durumları vardır ve bir amino asit birden fazla fizikokimyasal özelliğe sahip olabilir. Amino asitlerin sahip oldukları özellikleri sayısal formatta göstermek amacıyla 10×1 boyutunda vektör tanımlanır. Bu vektörün her bir elemanı bir fizikokimyasal özelliğe eşittir. Kodlanmak istenen aminoasidin sahip olduğu özelliklerin vektörde karşılık geldiği yerler 1, diğer yerler 0 olarak tanımlanır [42].

2.2. Önışleme Yöntemleri

2.2.1. Sınıf Dengeleme

Dengesiz veriler, veri madenciliğinde en zorlu araştırma probleminden biridir [43]. Sınıflandırmada, sınıf dengesindeki değişiklikler sıklıkla gözlemlenmektedir. Tüm sınıfların eşit olabilmesi için veri kümesindeki ağırlıklar yeniden atanarak sınıf dengeleme algoritması uygulanmaktadır. İki sınıflı veri kümesi için, genellikle hedef sınıf olan veriler, diğer verilerden çok daha az örneğe sahip olduğunda verilerde dengesizlik oluşur. Bu durumda, çoğunluk sınıf tarafından veriler, sınıflandırıcı

algoritmasını yanlış yönlendirir. Başka bir deyişle, sınıflandırıcılar aşırı uyum sağlama ve özellikle azınlık sınıfı üzerinde düşük performans gösterme eğilimindedir ve azınlık sınıf, çoğunluk sınıfı daha yanlış sınıflandırmaktadır [44].

Sorunu çözmek için, her sınıfın toplam ağırlığı ve ayrıca tüm örneklerdeki toplam ağırlıklar korunurken örnekler yeniden ağırlıklandırılır [45]. Bu tez çalışmasında da koronavirüse ait anti-CoV peptitlerin sayısı, diğer peptitlere göre daha az olduğu için sınıf dengeleme yapılmıştır.

2.2.2. Ayırıklaşma

Veri kümesindeki bir dizi sayısal özniteliği nominal özniteliklere ayıran bir örnek filtresidir. Ayırıklaştırma, Fayyad ve Irani'ni yöntemiyle yapılmaktadır [46]. Ayırıklaştırma, öznitelik aralığını aralıklara bölerek belirli sürekli öznitelikte değer sayısını azaltmak için kullanılan bir veri ön işleme sürecidir [47]. Ayırık özellikler, sürekli özelliklere kıyasla genellikle daha yoğun, daha kısa ve daha doğrudur. Bir ayırıklaştırma işlemi özellikle dört adımdan oluşur. İlk ayırıklaştırılacak özelliğin tüm sürekli değerleri sıralanır. İkinci olarak sürekli değerleri aralıklara bölmek için bir kesme noktası seçilir. Üçüncü olarak değerlerin sürekli aralıkları ayrılır veya birleştirilir, son olarak ayırıklaştırma işleminin durdurma kriterleri seçilir [48].

2.3. Sınıflandırıcı Algoritmalar

Bu bölümde, başarı değerlendirmesinde kullanılan sınıflandırma yöntemleri açıklanmıştır.

Sınıflandırma, verileri benzerliklerine göre gruplandırma işlemidir. Sınıflandırıcı algoritmalar, verilen eğitim kümesi ile dağılım şeklini öğrenip, sınıfının belirli olmadığı test verileri üzerinde sınıflandırma yapmaktadır.

Sınıflandırma algoritmaları olarak, Rastgele Orman, Naif Bayes, Bayes Ağları, k-En Yakın Komşu (k-EYK), Çok Katmanlı Algılayıcı (ÇKA) ve Destek Vektör Makineleri (DVM) gibi literatürde farklı problemlerin çözümünde sıkça kullanılan makine öğrenmesi tabanlı sınıflandırma yöntemleri kullanılmıştır. Anti-CoV peptitlerinin tespit edilebilmesi problemine uygunluğu elde edilen sonuçlara göre değerlendirilmiştir.

2.3.1. Naif Bayes

Naif Bayes, Bayes tabanlı bir istatistiksel sınıflandırma tekniğidir. Bu sınıflandırıcı, belirli bir sınıfta bir özellik değerinin etkisinin olduğunu tahminler ve diğer özelliklerin değerlerinden bağımsız olarak karar verir [49]. Naif Bayes sınıflandırıcı temelde bütün koşullu olasılıkların çarpımıdır. C verilen hedef ve F öznitelikler olmak üzere Naif Bayes'in olasılık hesaplaması Denklem 2.8'de verilmiştir.

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)} \quad (2.8)$$

2.3.2. Bayes Ağları

Bayes Ağları, rastgele değişkenler arasındaki istatistiksel ilişkilerin tanımlandığı graf tabanlı bir olasılıksal analiz yöntemidir. Olasılıksal tahmin yöntemleri, mevcut gözlemin birikmiş hipotezlerinden birine karşılık geldiğini belirlemek için kullanılır [50].

Bayes ağı, $G = (V, E)$, köşeleri olmak üzere, $V = \{V_1, V_2, \dots, V_n\}$, rastgele değişkenlere karşılık gelen ve yönlendirilmiş kenarlar E olmak üzere arasındaki koşullu olasılıkları temsil eden yönlendirilmiş, döngüsel olmayan bir grafiktir. P, bir dizi yerel olasılık dağılımı ile birlikte rastgele değişkenlere denir. V_i 'den V_j 'ye yönlendirilmiş bir kenar varsa, bu, V_j 'nin olasılıksal olarak V_i 'ye bağlı olduğu anlamına gelir. Bu durumda, yerel olasılık dağılımı, P, V_i üzerinde koşullandırılmış V_j düğümü olan $p(V_i|V_j)$ olarak hesaplanır [51].

2.3.3. k-En Yakın Komşuluğu

k- En Yakın Komşuluğu (k-EYK), çok etiketli öğrenmedir ve uygulaması, en yakın k tane komşunun sınıfına göre ilgili örneğin sınıfı ile ilişkilendirilmiş örneklerden oluşur. Bilinen etiket kümeleriyle eğitim örneklerini analiz etmekte ve sınıfı belli olmayan örneklerin etiket kümelerini tahmin etmektedir [52]. En yakın komşular, denklem 2.9'da verilen Öklid mesafesi formülü ile hesaplanır [53]. Öklid mesafesi, iki nokta arasındaki doğrusal uzaklıktır. N boyutlu Öklid uzayında $P = (p_1, p_2, \dots, p_n)$ ve $Q = (q_1, q_2, \dots, q_n)$ noktaları arasındaki Öklid uzaklığı:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2.9)$$

k sayısının belirlenmesinde matematiksel bir çerçeve yoktur, sezgisel olarak belirlenmektedir. Bu tez çalışmasında k değeri 1-30 arası değerler denenmiş olup, en başarılı sonuç k=1 olarak verdiği için seçilmiştir.

2.3.4. Rastgele Orman

Rastgele orman, bir topluluk öğrenme algoritması olup, büyük veri kümelerinde hızlı bir şekilde çalışılan, hesaplama olarak verimli bir sınıflandırma algoritmasıdır. Birçok araştırma projesinde ve farklı alanlardaki uygulamalarında kullanılmaktadır [54]. Yeniden örnekleme düşüncesini kullanan algoritma, önyükleme toplama, verilen bir veri kümesinden değiştirme ile belirli sayıda örnek seçer. Önyükleme toplama işleminin gerçekleştirilmesi için, seçilen her örnek bir temel karar ağacı sınıflandırıcısı oluşturmak için kullanılır. Son olarak, en iyi performansa sahip sınıflandırıcı oy çokluğu ile seçilir. Önyükleme toplamaya dayalı olarak, rastgele orman ayrıca karar ağacındaki düğümlerin parça özniteliklerini alt kümeler olarak rastgele seçer ve ardından sınıflandırma için alt kümelerden en iyi öznitelikleri seçer. Rastgele orman sınıflandırıcısı, birkaç temel sınıflandırıcının birleştirilmesine dayalı olarak kurulduğundan, girişime karşı önemli bir sağlamlığa sahiptir. Temel tahmin edicilerin sayısı arttıkça, sınıflandırıcı düşük bir tahmin hatasına yakınsama eğilimindedir [55].

2.3.5. Çok Katmanlı Algılayıcı

Çok Katmanlı Algılayıcı (ÇKA), nöron olarak adlandırılan, birbirine bağlı yapay sinir hücrelerinden oluşmaktadır. Yapay sinir hücrelerinin bağlantılarından bilgi akışı ağ içerisinde sağlanır. Doğrusal olmayan problemlerde başarılı olmaktadır [56]. ÇKA, girdi, ara ve çıktı katmanlarından oluşmaktadır. Çok katmanlı algılayıcılar, bilgilerin sisteme girişine girdi, bir ya da daha çok gizli ve son olmak üzere çıkış olan çıktı katmanlarından oluşur. ÇKA sistemi, içerisinde ileri ve geri yayılım gösteren geçişler vardır. İleri yayılımda, ağın ürettiği çıktı ve oluşan hata değeri hesaplanmaktadır. Geri yayılımdaysa, hesaplanan hata değerinin en düşük haline getirilmesi için katmanlar arasındaki bağlantı ağırlık değerleri güncellenmektedir [57]. Genellikle geri yayılım algoritması kullanılmaktadır. İlk adımda girdilere rastgele ağırlıklar üretilmektedir. Girişte girdiler elde edilir ve ara katmana iletilir. Ara katman bir ya da daha çok olup

her katmanda farklı sayıda işlem elemanı olabilmektedir. Ara katmanda ağırlıklar, girdiler ve eşik değeri ile toplam değer elde edilir. Daha sonra aktivasyon fonksiyonundan geçirilir ve net değer oluşur. Net değer, sonra gelen ara katmana ya da çıktı katmanına iletilir. Ara katmanda bulunan elemanlar aktivasyon fonksiyonu sonucu belirlenir.

Başlangıçtaki girdilere ait çıktılar ile ağırlık ürettiği çıktılar karşılaştırılır ve aradaki fark hata olarak kabul edilir. Elde edilen hata değeri modelin ağırlıklarına dağıtılır ve sonraki iterasyonda hatanın azaltılması hedeftir. Çıktı katmanındaki her işlem elemanına ait hata değerini azaltmak için hatalar ağırlıklara dağıtılır. Ağırlıkların güncellenmesi girdi katmanına kadar geri yayılım olarak devam ederek çıktı katmanındaki hata en aza indirilir.

2.3.6. Lojistik Regrasyon

Lojistik, bir veya daha fazla tahmin değişkenine bağlı kategorik olarak bağımlı değişkenlerin sonuçlarını tahmin etmek için kullanılan bir tür regresyon analizidir. Analiz etmek için basit ve çoklu regresyon analizleri kullanılmaktadır. Lojistik regresyon, bağımlı değişkenler ile bağımsız değişkenler arasındaki sebep-sonuç ilişkisinin analizi için önemlidir [58]. Bir ya da daha fazla değişkenin iki ya da daha fazla bağımlı değişken ile ilişkisini modellemek için kullanılan bir istatistiki yaklaşımdır. Bağımlı değişkenin belirli bir aralıkta veya ortamdaki değerini tahmin eder [59]. Lojistik regresyon modeli eğitilirken sınıflandırmaya girecek örüntüye ait özniteliklerle, belirlenmiş ağırlıkların çarpımının toplamına bias değeri eklenmektedir. Bu değer aktivasyon fonksiyonu olan, denklem 2.10'da görülen sigmoid fonksiyonuna girer.

$$y = \frac{y}{1+e^{-x}} \quad (2.10)$$

Sonraki aşamada, fonksiyondan elde edilen sonuç ve beklenen çıktı değerleri karşılaştırılır. Karşılaştırma sonrasında hata değeri ağırlıklar ve biasa dağıtılır. Bu döngü belirlenen sayıda tekrarlanır. Test edilirken, karar sigmoid fonksiyonundaki çıktı 0,5 eşik değerinden büyük ise sonuç 1, küçük ise 0 olmaktadır.

2.3.7. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), örüntü tanıma, sınıflandırma problemleri çözümü için 1979'da geliştirilmiştir [60]. DVM iki sınıfı birbirinden en uygun biçimde ayırması için bir hiper düzlemin tanımlanmasına dayanır.

Bu tez çalışmasında, doğrusal ve radyal temel fonksiyon çekirdek fonksiyonlarına sahip DVM algoritmaları kullanılmıştır.

Örnek uzayındaki, tüm örnekleri başarılı sınıflandıracak şekilde ayıran birçok doğru bulunabilir. Bu doğrulardan hangisinin diğerlerinden daha uygun olduğunu belirlerken, doğru bütün noktalara yakın olmamalıdır. DVM'nin çalışma mekanizması, eğitim verilerinden kendine en yakın 27 örnek için en uzak mesafeye sahip ayırıcı hiperdüzlemi bulmaktır.

Sınıflandırma için bir düzlemde bulunan iki grup arasına bir sınır çizilerek iki grubu ayırmak mümkündür. DVM algoritması ait olduğu sınıfın sınırını belirler [61]. Destek vektörlerinin üzerinde bulunduğu ve noktalarla gösterilmiş düzlemlere sınır düzlemleri denir. Sınır düzlemlerinin tam ortasından geçen ve her iki düzleme de eşit uzaklıkta bulunan düzlem ise hiperdüzlem olarak ifade edilir [62].

İki sınıftan oluşan, r tane örnek için, birbirinden ayrılabilir $(x_1, y_1), \dots, (x_r, y_r)$ şeklinde tanımlanmış bir eğitim seti üzerinde; $x \in \mathbb{R}^N$ olup N boyutlu bir uzayı, $y \in \{+1, -1\}$ olmak üzere sınıf etiketlerini belirtmektedir. Bu iki sınıf $N-1$ boyutlu hiperdüzlemler tarafından ayrılabilir [63]. Bir hiperdüzlem denklem 2.11'deki gibi tanımlanmaktadır.

$$w \cdot x_i + b = 0 \quad (2.11)$$

x_i hiperdüzlem üzerindeki noktayı belirtir, w hiperdüzlemin normalini belirtirken, b ise hiperdüzlemin orijine olan uzaklığı olarak ifade edilen biası (eğilim değeri) ifade etmektedir.

İki sınıfı doğrusal olarak ayrılabilen hiperdüzlem için;

$$w \cdot x_i + b \geq +1, \text{ her } y = +1 \text{ için} \quad (2.12)$$

$$w \cdot x_i + b \leq -1, \text{ her } y = -1 \text{ için} \quad (2.13)$$

olarak tanımlanabilir.

Denklem 2.12 ve 2.13'teki iki eşitsizlik tek bir eşitsizlik haline getirilirse;

$$y_i (w \cdot x_i + b) - 1 \geq 0, y_i \in \{1, -1\} \quad (2.14)$$

Denklem 2.14' deki gibi olur.

Optimum hiperdüzleme paralel ve $w \cdot x_i + b = \pm 1$ eşitliği ile tanımlanan ayırım yapabilen iki hiper düzlem üzerinde bulunan ve sınır genişliğini belirleyen noktalar destek vektörleri olarak adlandırılır [64], [65], [66]. Bu iki hiperdüzlem arasındaki sınır genişliği (marjin)dir.

$$\frac{2}{\|w\|} \quad (2.15)$$

İki sınıfı maksimum sınır genişliği ile ayıran optimum hiperdüzlem, denklem 2.15'teki gibi sınırlara bağlı olarak minimum hale getirilerek bulunabilir, denklem 2.16'da gösterilmiştir.

$$\min \left[\frac{1}{2} \|w\|^2 \right] \quad (2.16)$$

Bu problem Lagrange denklemleri kullanılarak çözülebilecek bir optimizasyon problemidir ve çözüm için karar fonksiyonu;

$$f(x) = \text{sign}(\sum_{i=1}^r \tau_i \cdot y_i (x \cdot x_i) + b) \quad (2.17)$$

şeklinde olur.

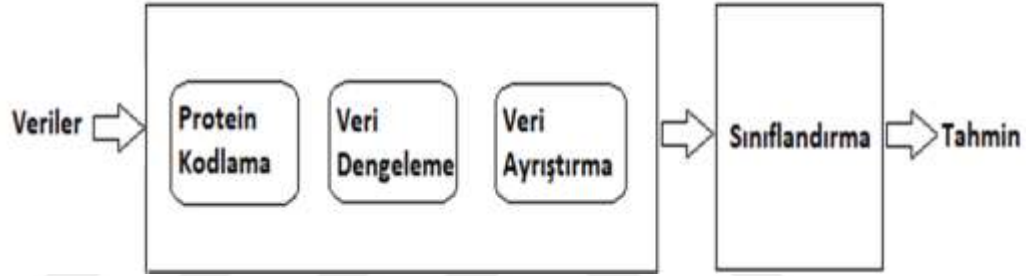
Burada, $\tau_i = 1, 2, \dots, r$, Lagrange çarpanlarını temsil etmektedir [67].



3. GELİŞTİRİLEN PROTEİN KODLAMA YÖNTEMLERİ

Protein kodlama yöntemleri geliştirilmesi ve sınıflandırılması aşamalarını gösteren modelin çizimi, Şekil 3.1’de gösterilmektedir. Veriler ilk olarak protein kodlama yöntemleri ile sayısallaştırılmıştır.

Sonra veri dengeleme yapılmıştır ve verilere, veri ayrıklaştırma yöntemi uygulanmıştır. Daha sonra sınıflandırma algoritmaları ile sınıflandırma sonuçları elde edilmiştir. Bu bölümün devamında geliştirilmiş protein kodlama yöntemleri açıklanmaktadır. Geliştirilen 2gBLO ve 2gBLOTVD protein kodlama yöntemleri, model çiziminde protein kodlama kısmında yer almaktadır.



Şekil 3.1. Sınıflandırma için oluşturulan model

3.1. 2gBLO Yöntemi

2gBLO, bir protein kodlama yöntemidir. Bu protein kodlama yönteminde n-Gram metodunda n değeri iki alınarak, digram uygulanmıştır. Daha sonra BLOSUM62 matrisi kullanılarak verilere karşılık gelen değerler ile çarpılmıştır. Veriler Python programlama dili kullanılarak sayısallaştırılmıştır. Böylece yeni hibrit bir protein kodlama yöntemi geliştirilmiştir.

İlk adımda 20 adet aminoasit içeren farklı uzunluktaki protein dizilimleri digram yöntemiyle, ikili olarak bölünmüş, ikili grubun tekrar sayılarına göre kodlanmıştır. Kodlama aşamasında ilgili ikiliye ait tekrar sayısı ve BLOSUM62 matrisinde ilgili ikiliye karşılık gelen puan ile çarpılmaktadır. Bu şekilde çarpılarak, her ikili amino asit değeri için bir değer bulunmaktadır. Bulunan bu değer için, 20×20 boyutunda bir matris oluşturulmaktadır. Kodlanmış kısımlara karşılık gelen yer dışındaki kısımlara sıfır değeri yazılmıştır. Veri setinde yer alan her peptit dizilimi için, aynı şekilde 20×20 boyutunda matris hesaplaması yapılmaktadır. Her işlem aşamasında, bu 20×20’lik matris, 1×400’lük vektöre çevrilmektedir. Sonucunda, örnek sayısı n olmak üzere, n

kez bu işlem tekrarlanarak, $n \times 400$ 'lük bir matris elde edilmektedir. Böylece, kodlanmış yeni veri seti $n \times 400$ boyutunda olmaktadır.

HVTTTFAPPPPR dizilimi örnek verilirse, ilk olarak dizilimdeki ikili tekrar sayıları bulunur.

Daha sonra, sıfır matrisine, 20×20 boyutundaki digram yöntemine göre tekrar sayıları yazılır. Yani, TT için T ve T kesişimine 2, P ve P kesişimine 3 geri kalan ikililer için HV, VT, TF, FA, AP ve PR için ise 1 değeri yazılır. Geri kalan değerler 0 olarak kalır. Elde edilen matris Çizelge 3.1'de gösterilmiştir.

Çizelge 3.1. Digram kodlama sonucu oluşan 20×20 'lik matris

	A	C	D	E	F	G	H	I	K	L	M	N	Q	R	S	W	V	P	T
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
F	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
T	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Çizelge 3.1'de verilmiş, digram olarak kodlanmış matris, Çizelge 2.3' de verilmiş, BLOSUM62 matrisindeki kesişimlere karşılık gelen değerler ile çarpılır ve Çizelge 3.2'de gösterildiği gibi 20×20 boyutunda bir matris oluşur.

Çizelge 3.2. Digram ile BLOSUM62'nin 20x20'lik hibrit matrisi

	A	C	D	E	F	G	H	I	K	L	M	N	Q	R	S	W	V	P	T
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
F	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
T	0	0	0	0	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	8
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Son olarak her 20×20'lik bu matrisler, 1×400'lük vektöre dönüştürülür ve tüm dizilimler için aynı işlem tekrarlanır ve verinin örnek sayısına n denildiğinde, $n \times 400$ boyutunda veri seti elde edilir.

3.2. 2gBLOTVD Yöntemi

2gBLOTVD, bir protein kodlama yöntemidir. Bu protein kodlama yönteminde 2gBLO'ya ek Taylor'ın Venn diyagramındaki fizikokimyasal özelliklerden faydalanılmıştır. 2gBLOTVD yönteminde, 20 amino asitten oluşan farklı uzunluklardaki veri seti üzerinde python ortamında digram yöntemine göre, aminoasitlerin ikili kombinasyon tekrar sayısı ile BLOSUM62 matrisinin ikiliye karşılık gelen kesişimi çarpılıp, buna ek olarak Taylor'ın Venn diyagramındaki her bir aminoasitin fizikokimyasal özelliklerinin kesişim sayısında çarpılması ile 1×400 boyutunda bir özellik vektör elde edilir. Böylece bu üç yöntem çarpılmış ve yeni hibrit bir öznitelik yöntemi elde edilmiştir. Örnek olarak, AC ikilisi için, A ve C amino asitleri, 20×20 boyutuna sahip özellik vektöründeki değerini, sahip oldukları fizikokimyasal özelliklerin ortak özellik sayısı ile alır. A amino asidi, çok küçük, hidrofobik ve küçük özelliklerine sahipken, C amino asidi polar, çok küçük, küçük ve hidrofobik özelliklerine sahiptir. Her iki amino asidin kesişim kümesi incelendiğinde çok küçük, küçük ve hidrofobik özelliklerinin ortak olduğu görülmektedir. İşlem yapılan peptit dizilimde, özellik vektöründeki AC sütunun değeri üç ortak özellik olduğu için üç ile ifade edilmektedir.

Çizelge 3.3. Digram ile BLOSUM62 ve TVD'nin 20x20'lik hibrit matrisi

	A	C	D	E	F	G	H	I	K	L	M	N	Q	R	S	W	V	P	T
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
F	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
T	0	0	0	0	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	24
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Çizelge 3.3'e göre, Digram, BLOSUM62 ve TVD değerlerinin çarpılması sonucundaki değerler gösterilmiştir. Örnek olarak TT ikilisinin, T ve T kesişimine bakıldığında, üç ortak değeri olduğu için Çizelge 3.2'deki TT değeride 8 olduğu için, 8×3 olmak üzere 24 olmaktadır. Tüm aminoasit ikililerine daha sonra dizilimlere aynı işlem uygulanmaktadır. Daha sonra, elde edilen 20×20 'lik matris, 1×400 'lük matrise çevrilir. Tüm aşamalar tamamlandığında elde edilen her 1×400 'lük vektör için, n dizilim için yeni veri seti, $n \times 400$ boyutunda bir matris oluşur.

4. BULGULAR

4.1. Veri Setleri

Veri setleri, Anti-CoV peptitleri, antivirüs peptitleri olan AVP seti, antivirüs aktiviteleri olmayan non-AVP peptitleri ve antimikrobiyal aktiviteleri olmayan non-AMP peptitleri olmak üzere dört veri seti bulunmaktadır. Negatif olanlar AVP, nonAMP ve nonAVP protein verileridir.

Pozitif olan anti-CoV verileri ile birleştirilerek üç farklı veri seti elde edilmiştir.

Veri setlerinin veri sayıları Çizelge 4.1’de ifade edilmiştir. Anti-CoV veri seti için 137 dizi kaydı bulunmaktadır. Anti-CoV veri seti, diğer veri setlerindeki değerlerden az sayıda veri içerdiği için ayrı ayrı birleştirildi. Böylece, her bir veri setinde pozitif ve negatif sınıflara ait proteinler elde edildi.

Anti-CoV verilerinin, 99 tanesi AVPdb veritabanından alınmıştır [68]. Bu veriler, P9, Mucroporin-M1 ve HR2P gibi bildirilen güçlü koronavirüs inhibitörlerini içermektedir. Geri kalan protein dizilimleri ise, koronavirüse karşı varsayılan anti-CoV peptitlerdir [69]. AVP veri seti 1999 protein dizilimi, non-AVP veri seti, 5217 protein dizilimi içermektedir [70], [71], [72], [73]. Non-AMP veri kümesi, 4979 protein diziliminden oluşmaktadır. Tüm veri setleri uzunluğu 100'e eşit veya daha az olan kısa peptitler içermektedir.

Çizelge 4.1. Veri setlerindeki örnek sayıları

	Veri Setleri		
	Anti-CoV+AVP	Anti-CoV+nonAMP	Anti-CoV+nonAVP
Örnek Sayısı	2136	5116	5354

4.2. Deneysel Sonuçlar

Makine öğrenmesi modelleri ile gerçekleştirilen, Anti-CoV peptitlerin tahmini probleminde, deneysel sonuçlar 10-kat çapraz doğrulama test tekniği ile elde edilmiştir. Bu teknikte, veri seti 10 eşit kümeye ayrılır, dokuzu eğitim, kalan biri ise test için kullanılır ve böylece toplamda 10 test gerçekleştirilir. Bu testlerin sonucunda elde edilen performans değerlerinin ortalaması alınır.

Anti-CoV peptit tahmininde sınıflandırıcı performansı, sınıf doğruluğu, ROC eğrisi altında kalan alan (AUC), Kesinlik, Duyarlık, Kappa ve Mathew'sın Korelasyon Katsayısı (MKK) metrikleri ile değerlendirilmiştir. Bu metrikler, pozitif (P), negatif (N), doğru pozitif (DP), doğru negatif (DN), yanlış pozitif (YP) ve yanlış negatif (YN) olmak üzere, Şekil 4.1'de görülen karmaşıklık matrisinden elde edilmektedir [75].

Şekil 4.1. Karmaşıklık matrisi

	Tahmin	
	P	N
Gerçek		
P	DP	YN
N	YP	DN

Sınıf doğruluğu, pozitif ve negatif doğru tahmin edilen değerlerin oranını vermektedir.

$$Sınıf\ Doğruluğu = \frac{DP + DN}{P + N} \quad (4.1)$$

Denklem 4.2'de gösterilmiş kesinlik değeri, tahmin edilecek veriler için modelin ne kadar kararlı sonuç verdiğini göstermektedir ve dengesiz veri setlerinde kesinlik değeri önemlidir.

$$Kesinlik = \frac{DP}{DP + YN} * (DP + YN) \quad (4.2)$$

Duyarlık denklem 4.3' de gösterildiği şekilde hedef sınıfın (DP) doğru tahmin edilme oranını verir.

$$Duyarlılık = \frac{DP}{DP + YN} \quad (4.3)$$

Denklem 4.4'de verilen, AUC değeri, duyarlığın 1-özgünlüğe oranını gösteren ROC eğrisi grafiğidir. Bu eğri altında kalan alan AUC olarak tanımlanır. 0-1 aralığında ifade edilen AUC değeri, 1'e ne kadar yakınsa modelin performansı o kadar iyi demektir.

$$AUC = \frac{1}{2} * \left(\frac{DP}{DP + YN} + \frac{DN}{DN + YP} \right) \quad (4.4)$$

MKK, ikili sınıflandırma için gözlenen ve öngörülen çıktılar arasındaki korelasyon katsayısıdır. İki sınıfı olan problemlerde modelin gücünü belirten bir ölçüttür [76]. MKK değeri -1 ile +1 arasında değişen performans metriğidir. Değer 1'e yaklaştıkça iyi tahmin yapıldığını, 0 şansa bağlı bir tahmin yapıldığını, -1'e yaklaştıkça ters tahmin yapıldığını gösterir.

$$MKK = \frac{1}{2} * \left(\frac{(DP*DN)-(YP*YN)}{\sqrt{(DP+YP)*(DP+YN)*(DN+YP)*(DN+YN)}} \right) \quad (4.5)$$

Kappa değeri, tahmin edilen değer in şans eseri mi bulunduğunu verir. Po kabul edilen oran, Pc beklenen oran olmak üzere denklem 4.6'da verilmiştir.

$$Kappa = \frac{P_o - P_c}{1 - P_c} \quad (4.6)$$

F-Skoru, denklem 4.7'de belirtildiği gibi, duyarlık ve kesinlik değerlerinin harmonik ortalaması hesaplanarak, iki metrik arasındaki dengenin sağlanması ile elde edilir. F-Skor değerine bakarak sınıflandırmayı daha sağlıklı yorumlamak mümkündür.

$$F - Skor = \frac{2 * Duyarlilik * Kesinlik}{Duyarlilik + Kesinlik} \quad (4.7)$$

Tüm test sonucunda, performans metriklerinin ortalaması alınmaktadır.

Geliştirilen 2gBLO öznitelik kodlama yöntemi, Bayes Ağları, Naif Bayes, Rastgele Orman, k-EYK, ÇKA, Doğrusal DVM, Radyal DVM ve Lojistik Regresyon sınıflandırma algoritmaları ile Anti-CoV ve Antivirüs veri setleri üzerinde test edilmiştir.

Çizelge 4.2'de 2gBLO yönteminin Anti-CoV ve AVP veri setleri üzerindeki başarımların değerleri görülmektedir.

Çizelge 4.2. 2gBLO'ya ait Anti-CoV ve AVP veri seti üzerindeki sınıflandırma sonuçları

2gBLO Yöntemi - Anti-CoV + AVP Veri Seti							
Sınıflandırıcılar	Doğruluk(%)	Duyarlılık	F-Skor	AUC	Kappa	MKK	Kesinlik
Bayes Ağları	70,98	0,563	0,703	0,822	0,420	0,439	0,729
Naif Bayes	71,03	0,710	0,704	0,821	0,421	0,440	0,730
k-EYK(1)	74,40	0,521	0,731	0,763	0,488	0,545	0,804
Rastgele Orman	71,98	0,720	0,698	0,881	0,440	0,522	0,809
ÇKA	56,20	0,923	0,497	0,577	0,124	0,179	0,629
Doğrusal DVM	92,0306	0,549	0,926	0,747	0,4264	0,432	0,933
Radyal DVM	94,4169	0,127	0,225	0,563	0,2137	0,346	0,947
Lojistik Regresyon	78,983	0,606	0,782	0,918	0,5797	0,624	0,835

Elde edilen deneysel sonuçlara göre; 2gBLO yöntemi, Radyal DVM algoritması, %94,41 sınıf doğruluğu ve 0,947 kesinlik ile en iyi sonuçları vermiştir. Bununla birlikte, Doğrusal DVM algoritması, 0,926 F-Skor ile en iyi değeri vermiştir. En düşük sınıf doğruluğunu %56,20 ile veren ÇKA yöntemi, aynı zamanda 0,923 ile en yüksek duyarlık değerini vermiştir. Bunun anlamı, ÇKA yöntemi pozitif sınıfın tahmininde çok başarılı iken, negatif sınıfın tahmininde bu başarımları göstermemektedir.

Aynı sınıflandırıcılar ile Çizelge 4.3'te, 2gBLO yönteminin Anti-CoV ve non-AVP veri setleri üzerindeki performans değerleri verilmiştir.

Çizelge 4.3. 2gBLO'ya ait Anti-CoV ve non-AVP veri seti üzerindeki sınıflandırma sonuçları

2gBLO Yöntemi - Anti-CoV + non-AVP Veri Seti							
Sınıflandırıcılar	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
Bayes Ağları	79,60	0,634	0,790	0,904	0,592	0,626	0,831
Naif Bayes	79,12	0,634	0,786	0,897	0,582	0,614	0,823
k-EYK(1)	78,69	0,606	0,780	0,791	0,574	0,616	0,830
Rastgele Orman	71,52	0,437	0,691	0,918	0,430	0,518	0,812
ÇKA	68,22	0,965	0,655	0,694	0,365	0,442	0,768
Doğrusal DVM	92,0306	0,549	0,926	0,747	0,4264	0,432	0,933
Radyal DVM	94,4169	0,127	0,923	0,563	0,2137	0,346	0,947
Lojistik Regresyon	78,98	0,606	0,782	0,918	0,5797	0,624	0,835

Elde edilen sonuçlardan en yüksek değerleri sınıf doğruluğu %94,41 ve kesinlik 0,947 olarak Radyal DVM vermiştir. En düşük sınıf doğruluğunu %68,22 ile ÇKA verirken en yüksek duyarlık değerinde 0,965 ile ÇKA vermektedir.

Aynı sınıflandırıcılar ve Anti-CoV ile non-AMP birleşiminden oluşmuş veri seti üzerinde tahminlenmiş olan tablo, Çizelge 4.4'te, 2gBLO yönteminin performans değerleri verilmiştir.

Çizelge 4.4. 2gBLO'ya ait Anti-CoV ve non-AMP veri seti üzerindeki sınıflandırma sonuçları

2gBLO Yöntemi - Anti-CoV + non-AMP Veri Seti							
Sınıflandırıcılar	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
Bayes Ağları	98,28	1,000	0,983	0,989	0,966	0,966	0,983
Naif Bayes	98,28	1,000	0,983	0,988	0,966	0,966	0,983
k-EYK(1)	89,33	0,789	0,892	0,986	0,787	0,804	0,911
Rastgele Orman	93,22	0,932	0,932	0,999	0,864	0,872	0,940
ÇKA	50	1	NA	0,646	0	NA	NA
Doğrusal DVM	99,6455	0,824	0,996	0,911	0,8222	0,822	0,996
Radyal DVM	99,2201	0,225	0,99	0,613	0,3655	0,473	0,992
Lojistik Regresyon	88,64	0,775	0,885	0,925	0,7733	0,793	0,907

Elde edilen sonuçlardan ÇKA hariç hepsi iyi sonuçlar vermesi il birlikte, en yüksek değerleri sınıf doğruluğu %99,64 ve kesinlik 0,996 olarak Doğrusal DVM vermiştir. Bunun yanında duyarlık değeri 1 ve kappa değeri ile MKK değeri 0,966 olmasıyla en dengeli olarak başarılı değerleri Bayes Ağları vermiştir. En düşük sınıf doğruluğunu hesaplanamayan manasına gelen not applicable (NA) değerleri ile doğruluk değeri %50 olmasıyla ÇKA vermiştir. Bu modelde ÇKA değerleri hesaplanamamaktadır.

Geliştirilen diğer öznitelik kodlama yöntemi, 2gBLOTVD için, aynı sınıflandırma algoritmaları kullanılmıştır. Çizelge 4.5' de Anti-CoV ve Antivirüs veri setleri üzerinde test edilmiş 2gBLOTVD yönteminin performans değerleri görülmektedir

Çizelge 4.5. 2gBLOTVD'ye ait Anti-CoV ve Antivirüs veri seti üzerindeki sınıflandırma sonuçları

2gBLOTVD Yöntemi - Anti-CoV + Antivirüs Veri Seti							
Sınıflandırıcılar	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
Bayes Ağları	77,63	0,652	0,773	0,878	0,553	0,570	0,794
Naif Bayes	77,63	0,652	0,773	0,878	0,553	0,570	0,794
k-EYK(1)	76,92	0,546	0,757	0,736	0,538	0,602	0,836
Rastgele Orman	69,64	0,397	0,667	0,919	0,393	0,490	0,806
ÇKA	50	1	NA	0,598	0	NA	NA
Doğrusal DVM	90,3201	0,532	0,917	0,728	0,324	0,343	0,285
Radyal DVM	94,7027	0,014	0,922	0,507	0,026	0,116	0,95
Lojistik Regresyon	71,02	0,489	0,695	0,693	0,420	0,469	0,761

Elde edilen sonuçlardan en yüksek değerleri sınıf doğruluğu %94,70 ve kesinlik 0,95 olarak Radyal DVM vermiştir. Bu modelde hesaplanamamış olan ÇKA sınıflandırıcısı, birçok NA değeri vermiştir. Dengeli başarımların değerlendirilmesinde, en kararlı sınıflandırıcının Bayes temelli olan Bayes ağları ve Naif Bayes sınıflandırıcılar olduğu görülmektedir.

Aynı sınıflandırıcılar ve Anti-CoV ile non-AVP birleşiminden oluşmuş veri seti üzerinde tahminlenmiş olan tablo, Çizelge 4.6' da, 2gBLOTVD yönteminin performans değerleri verilmiştir.

Çizelge 4.6. 2gBLOTVD'ye ait Anti-CoV ve non-AVP veri seti üzerindeki sınıflandırma sonuçları

2gBLOTVD Yöntemi - Anti-CoV + non-AVP Veri Seti							
Sınıflandırıcılar	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
Bayes Ağları	85,44	0,709	0,851	0,966	0,709	0,741	0,887
Naif Bayes	85,77	0,999	0,855	0,969	0,715	0,746	0,889
k-EYK(1)	74,42	0,496	0,727	0,799	0,488	0,562	0,824
Rastgele Orman	67,02	0,340	0,630	0,946	0,340	0,453	0,801
ÇKA	50,00	1,000	NA	0,793	0,000	NA	NA
Doğrusal DVM	98,18	0,440	0,982	0,715	0,443	0,444	0,981
Radyal DVM	98,67	0,227	0,983	0,613	0,122	0,473	0,987
Lojistik Regresyon	71,26	0,454	0,692	0,809	0,425	0,497	0,790

Elde edilen sonuçlardan en yüksek değerleri sınıf doğruluğu %98,67 ve kesinlik 0,987 olarak Radyal DVM vermiştir. Bunun yanında duyarlık değeri 0,999, kappa değeri 0,715 ve MKK değeri 0,746 olmasıyla en dengeli olarak başarılı sonuçları Naif Bayes sınıflandırıcısı vermiştir

Aynı sınıflandırıcılar kullanılarak, 2gBLOTVD kodlama yöntemi için, Anti-CoV ile non-AMP birleşiminden oluşmuş veri seti üzerinde yapılan analiz sonuçları, Çizelge 4.7'de, verilmiştir.

Çizelge 4.7. 2gBLOTVD'ye ait Anti-CoV ve non-AMP veri seti üzerindeki sınıflandırma sonuçları

2gBLOTVD Yöntemi - Anti-CoV + non-AMP Veri Seti							
Sınıflandırıcılar	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
Bayes Ağları	80,10	0,603	0,793	0,930	0,602	0,656	0,857
Naif Bayes	82,57	0,652	0,820	0,937	0,652	0,695	0,870
k-EYK(1)	74,79	0,504	0,732	0,710	0,496	0,568	0,826
Rastgele Orman	64,54	0,291	0,594	0,929	0,291	0,412	0,793
ÇKA	50,36	1,000	0,341	0,543	0,007	0,060	0,751
Doğrusal DVM	99,0154	0,014	0,985	0,507	0,0277	0,119	0,99
Radyal DVM	97,6059	0,976	0,980	0,672	0,2212	0,238	0,985
Lojistik Regresyon	80,6221	0,617	0,799	0,885	0,6124	0,662	0,857

Elde edilen sonuçlardan en yüksek değerleri sınıf doğruluğu %99,01 ve kesinlik 0,99 olarak Doğrusal DVM vermiştir. Bunun yanı sıra en dengeli başarımları veren olan sınıflandırıcı ise en yüksek 0,937 AUC, 0,652 kappa ve 0,695 MKK değerleri ile Naif Bayes'tir.

2gBLO ve 2gBLOTVD protein kodlama yönteminden en dengeli sonuçları veren sınıflandırıcı ise Naif Bayes'tir.

Bu tez çalışmasında geliştirilen, 2gBLO ve 2gBLOTVD öznelik kodlama yöntemleriyle birlikte, literatürde yer alan AAC, CT, AAP, DC, CMV, CTDT, CTDC ve Digram yöntemleri yer almaktadır. Veri setlerinin üzerinde kullanılmış tüm bu kodlama teknikleri ve sınıflandırma algoritmalarının en dengeli değerlerini veren sonuçlar Anti-Cov ve non-AMP veri seti için Çizelge 4.8.'de, Anti-CoV ve non-AVP veri seti için Çizelge 4.9 'da, Anti-CoV ve AVP veri seti için ise Çizelge 4.10'da gösterilmiştir.

Çizelge 4.8. Protein kodlama yöntemlerinin Anti-CoV ve non-AMP veri seti üzerinde en dengeli sonucu veren sınıflandırıcı sonuçları

Kodlayıcı	Sınıflandırıcı	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
AAC	Bayes Ağları	75,86	0,664	0,756	0,836	0,517	0,527	0,768
CT	Bayes Ağları	79,8	0,779	0,798	0,891	0,596	0,596	0,798
AAP	Bayes Ağları	83,98	0,81	0,84	0,902	0,68	0,681	0,841
DC	Rastgele orman	78,1	0,562	0,72	0,995	0,562	0,625	1
CMV	Rastgele orman	78,8	0,577	0,731	0,99	0,576	0,636	0,999
CTDT	Bayes Ağları	79,33	0,599	0,743	0,958	0,587	0,637	0,981
CTDC	ÇKA	86,72	0,818	0,867	0,913	0,735	0,738	0,871
Digram	Rastgele orman	74,71	0,504	0,731	0,756	0,494	0,566	0,824
2gBLO	Naif Bayes	98,28	1	0,983	0,989	0,966	0,966	0,983
2gBLOTVD	Naif Bayes	82,57	0,652	0,82	0,937	0,652	0,695	0,870

Anti-CoV ve non-AMP veri setinde uygulanmış protein kodlama yöntemlerine bakıldığında, AAC, CT, AAP, CTDT için Bayes Ağları, DC, CMV ve Digram için Rastgele Orman, CTDC için ÇKA, 2gBLO ve 2gBLOTVD için Naif Bayes en dengeli ve başarılı sonuçları vermektedir. Tüm değerlere bakıldığında, en başarılı sonuçları 2gBLO kodlayıcı yöntemi Bayes Ağları sınıflandırıcısı, doğruluk %98,28, duyarlık 1, f-skor 0,983, AUC 0,989, kappa ve MKK 0,966 ve kesinlik 0,983 olmak üzere değerleri ile vermiştir.

Diğer geliştirilmiş 2gBLOTVD yöntemi ise, 82,57 doğruluk, 0,652 duyarlık, 0,820 f-skor, 0,937 AUC, 0,652 Kappa, 0,695 MKK ve 0,870 kesinlik değerleri ile diğer en dengeli ve en başarılı sonucu vermiştir.

Çizelge 4.8.'e benzer şekilde aynı protein kodlama yöntemleri, Anti-CoV ve non-AMP veri seti üzerinde kullanılmış ve sınıflandırma algoritmalarının en dengeli değerlerini veren sonuçlar Çizelge 4.9.'da gösterilmiştir.

Çizelge 4.9. Protein kodlama yöntemlerinin Anti-CoV ve non-AVP veri seti üzerinde en dengeli sonucu veren sınıflandırıcı sonuçları

Kodlayıcı	Sınıflandırıcı	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
AAC	Rastgele Orman	84,926	0,701	0,846	0,984	0,699	0,732	0,883
CT	k-EYK(1)	70,862	0,431	0,684	0,702	0,417	0,502	0,802
AAP	k-EYK(1)	74,017	0,496	0,724	0,755	0,480	0,550	0,815
DC	k-EYK(1)	76,411	0,533	0,693	0,797	0,528	0,596	0,991
CMV	k-EYK(1)	72,683	0,460	0,627	0,819	0,454	0,537	0,987
CTDT	k-EYK(1)	72,916	0,467	0,633	0,860	0,458	0,538	0,981
CTDC	k-EYK(1)	70,281	0,413	0,582	0,874	0,406	0,498	0,982
Digram	k-EYK(1)	87,534	0,953	0,875	0,881	0,751	0,760	0,885
2gBLO	Doğrusal DVM	92,031	0,549	0,926	0,747	0,426	0,432	0,933
2gBLOTVD	Naif Bayes	85,766	0,999	0,855	0,969	0,715	0,746	0,889

Anti-CoV ve non-AVP veri setinde uygulanmış protein kodlama yöntemlerine bakıldığında, AAC yöntemi Rastgele Orman, CT, AAP, DC, CMV, CTDT, CTDC ve Digram için k değeri bir olmak üzere k-EYK, 2gBLO yöntemi ile Doğrusal DVM ve 2gBLOTVD yöntemiyle Naif Bayes en dengeli ve başarılı sonuçları vermektedir. Tüm değerlere bakıldığında, en başarılı sonuçları 2gBLO kodlayıcı yöntemi, Doğrusal DVM sınıflandırıcısı, doğruluk %92,031, duyarlık 0,549 , f-skor 0,926, AUC 0,747, kappa 0,426, MKK 0,432 ve kesinlik 0,933 değerleri ile vermiştir. 2gBLOTVD yöntemi ise doğruluk açısından diğer yöntemleri geçememektedir. Ancak, pozitif sınıfın doğru tahmini için kullanılan duyarlık değeri ile en başarılı pozitif sınıf tahminini Naif Bayes ile 2gBLOTVD yöntemi sağlamıştır.

Çizelge 4.8. ve Çizelge 4.9.'a benzer şekilde bir aynı protein kodlama yöntemleri, bir diğer veri seti olan Anti-CoV ve AVP ile oluşturulmuş veriler üzerinde kullanılmış ve sınıflandırma algoritmalarının en dengeli sonuçları Çizelge 4.10.'da verilmiştir.

Çizelge 4.10. Protein kodlama yöntemlerinin Anti-CoV ve AVP veri seti üzerinde en dengeli sonucu veren sınıflandırıcı sonuçları

Kodlayıcı	Sınıflandırıcı	Doğruluk(%)	Duyarlık	F-Skor	AUC	Kappa	MKK	Kesinlik
AAC	Bayes Ağları	75,858	0,664	0,756	0,836	0,517	0,527	0,768
CT	Bayes Ağları	78,6313	0,401	0,768	0,932	0,679	0,709	0,84
AAP	k-EYK(1)	75,4019	0,526	0,74	0,763	0,508	0,571	0,821
DC	k-EYK(1)	76,7515	0,562	0,707	0,77	0,535	0,587	0,954
CMV	k-EYK(1)	74,3366	0,518	0,669	0,749	0,486	0,545	0,943
CTDT	k-EYK(1)	71,7114	0,482	0,63	0,755	0,434	0,492	0,91
CTDC	k-EYK(1)	70,7018	0,453	0,607	0,736	0,414	0,481	0,922
Digram	Doğrusal DVM	91,8999	0,962	0,918	0,873	0,770	0,772	0,917
2gBLO	Doğrusal DVM	92,0306	0,549	0,926	0,747	0,426	0,432	0,933
2gBLOTVD	Doğrusal DVM	90,3201	0,532	0,917	0,728	0,324	0,343	0,285

Anti-CoV ve AVP veri setinde uygulanmış protein kodlama yöntemlerine bakıldığında, AAC, CT için Bayes Ağları, AAP, DC, CMV, CTDT ve CTDC için k-EYK, Digram, 2gBLO ve 2gBLOTVD için Doğrusal DVM en dengeli ve başarılı sonuçları vermektedir. Tüm değerlere bakıldığında, en başarılı sonuçları 2gBLO kodlayıcı yöntemi Doğrusal DVM sınıflandırıcısı ile doğruluk %92,0306, duyarlık 0,549 , f-skor 0,926, AUC 0,747, kappa 0,426, MKK 0,432 ve kesinlik 0,933 olarak vermiştir.

Diğer en dengeli sonucu Digram yöntemi ile Doğrusal DVM sınıflandırıcısı vermektedir. Yüksek olan 0,962 duyarlık değeri ile pozitif sınıf tahminini iyi yaptığı görülmektedir.

5. SONUÇLAR

İnsan vücudunda patojenler, hücre içindeki proteinlere ait fonksiyonlara ve biyolojik işleyişe etki etmektedir. Patojenler canlı üzerine yerleşir ve kendi proteinlerini canlının hücre çekirdeğinden içeri bırakarak canlının protein yapısını bozarlar. Bozulan protein, canlının biyolojik işleyişinin aksamasına yol açmaktadır. Buradan yola çıkarak, patojenlere karşı gelen ve patojenlerin yapısını bozan peptit dizilimlerinin tespit edilmesi, ilgili patojenin inhibe edilmesi açısından önem kazanmaktadır. İlaç geliştirme çalışmaları bu düşünce ile yürütülmektedir.

Biyoenformatik ve makine öğrenmesi temelinde, patojeni inhibe eden peptitin analiz edilebilmesi için bu peptit dizilimlerinin kodlanması gerekmektedir. Bu tez çalışmasında, Kovid-19 karşısında işlevsel aktiviteye sahip anti-CoV peptitlerinin tahmini için üç farklı protein veri seti üzerinde makine öğrenmesi yöntemleri ile sınıflandırma yapılmıştır. Bu minvalde, sınıflandırıcının başarımını artırmak için protein dizilimleri 2gBLO ve 2gBLOTVD adı verilen iki yeni protein kodlama yöntemi geliştirilmiştir.

2gBLO yönteminde, digram ile kodlanan her bir dizilimin, her ikili aminoasit karşılıkları BLOSUM62 matrisindeki değer ile çarpılmıştır. Tüm veri setindeki dizilimler bu yöntemle sayısallaştırıldıktan sonra, ayırıklaştırma metodu uygulanarak sınıflandırıcı algoritmalara verilmiştir. 2gBLOTVD yönteminde ise, 2gBLO yöntemine benzer, digram ve BLOSUM62'nin yanında TVD ortak özellik değerleri kullanılmıştır. Buna göre her bir dizilimdeki ikili aminoasitlere karşılık gelen TVD tablosundaki ortak özellik sayısı bulunmuştur. Sonrasında, Digram ile kodlanmış dizilim, BLOSUM62 matrisi ve TVD ortak özellik sayısının çarpılması sonucu yeni protein dizilimi kodlaması gerçekleştirilmiştir. Bu şekilde kodlanmış veri setine, sınıf dengeleme yapılarak, ayırıklaştırma yöntemi uygulanarak sınıflandırıcı algoritmalara verilmiştir.

Geliştirilen her iki protein kodlama yöntemi, Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, ÇKA, Doğrusal DVM, Radyal DVM ve Lojistik Regresyon makine öğrenmesi sınıflandırıcı algoritmaları ile test edilmiştir. Elde edilen sonuçlara göre, 2gBLOTVD ve 2gBLO protein kodlama yöntemleri Bayes tabanlı sınıflandırıcılardan Doğrusal DVM ve Naif Bayes ile iyi sonuçlar vermişlerdir.

Gelecekte yapılacak çalışmalarda; başka protein temelli sınıflandırma problemleri üzerinde, 2gBLO ve 2gBLOTVD protein dizilimi kodlama yöntemlerinin, farklı BLOSUM ve PAM matrisleri ile gerçekleştirmeyi planlamaktayız. Ayrıca, geliştirdiğimiz 2gBLO ve 2gBLOTVD yöntemlerini de kapsayan çevrim içi bir web sunucu geliştirmeyi hedeflemekteyiz.



KAYNAKLAR

- [1] Al-Tawfiq, J. A., Al-Homoud, A. H., & Memish, Z. A. (2020). Remdesivir as a possible therapeutic option for the COVID-19. *Travel medicine and infectious disease*, 34, 101615.p. 288, 2009.
- [2] Bates, J. H., & Stead, W. W. (1993). The history of tuberculosis as a global epidemic. *The Medical clinics of North America*, 77(6), 1205-1217.
- [3] A. W. Crosby, *The Columbian Exchange: Biological and Cultural Consequences of 1492*. 1972.
- [4] R. Acuna-Soto, D. W. Stahle, M. K. Cleaveland, and M. D. Therrell, "Megadrought and megadeath in 16th century Mexico," *Emerging Infectious Diseases*, vol. 8, no. 4. pp. 360–362, 2002.
- [5] Sen, S., Thejas, B. K., Pranitha, B. L., & Amrita, I. (2021). Analysis, Visualization and Prediction of COVID-19 Pandemic Spread Using Machine Learning. In *Innovations in Computer Science and Engineering* (pp. 597-603). Springer, Singapore.
- [6] Gao, H., Yao, H., Yang, S., & Li, L. (2016). From SARS to MERS: evidence and speculation. *Frontiers of medicine*, 10(4), 377-382.
- [7] WHO Coronavirus Dashboard. <https://covid19.who.int/>, 2021
- [8] Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... & Ghani, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.
- [9] Eidhammer, I., Jonassen, I., & Taylor, W. R. (2004). *Protein Bioinformatics: An algorithmic approach to sequence and structure analysis* (Vol. 1). Chichester: Wiley.
- [10] Pang, Y., Wang, Z., Jhong, J. H., & Lee, T. Y. (2021). Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. *Briefings in bioinformatics*, 22(2), 1085-1095.
- [11] S. Mei, "Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins," *PLoS One*, vol. 8, no. 11, p.e79606, 2013.
- [12] Arslan, I., Akgul, H., & Kara, M. (2021). Saporin, a Polynucleotide–Adenosine Nucleosidase, May Be an Efficacious Therapeutic Agent for SARS-CoV-2 Infection. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 26(3), 330-335.
- [13] Debnath, N., & Yadav, A. K. (2021). Potential Correlation Between Homeostasis Control and Tumor Microenvironment Regulation of Probiotic as

- a Therapeutic Agent to Manage Gastrointestinal Cancer. In *Probiotic Research in Therapeutics* (pp. 167-189). Springer, Singapore.
- [14] Ghosh, S. K., & Weinberg, A. (2021). Ramping up AMPs against SARS-CoV-2. *Frontiers in Molecular Biosciences*, 8, 547.
 - [15] Qureshi, A., Thakur, N., Tandon, H., & Kumar, M. (2014). AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic acids research*, 42(D1), D1147-D1153.
 - [16] Thakur, N., Qureshi, A., & Kumar, M. (2012). AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic acids research*, 40(W1), W199-W204.
 - [17] Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16), 2740-2747.
 - [18] Gök, M., & Özcerit, A. T. (2012). OETMAP: a new feature encoding scheme for MHC class I binding prediction. *Molecular and cellular biochemistry*, 359(1), 67-72.
 - [19] Gök, M., Koçal, O. H., & Genç, S. (2016). Prediction of disordered regions in proteins using physicochemical properties of amino acids. *International Journal of Peptide Research and Therapeutics*, 22(1), 31-36.
 - [20] Jain, N., Jhunthra, S., Garg, H., Gupta, V., Mohan, S., Ahmadian, A., ... & Ferrara, M. (2021). Prediction modelling of COVID using machine learning methods from B-cell dataset. *Results in physics*, 21, 103813.
 - [21] Kösesoy, İ., Gök, M., & Öz, C. (2018). PROSES: a web server for sequence-based protein encoding. *Journal of Computational Biology*, 25(10), 1120-1122.
 - [22] Devi, M. S., Sathiyarajeswaran, P., Kumar, D. T., Kumar, S. U., Siva, R., Doss, G. P., & Kanakavalli, K. (2022). Siddha Medicine and Computer Modeling: A Treasure for SARS-CoV-2 Treatment. In *Modeling, Control and Drug Development for COVID-19 Outbreak Prevention* (pp. 521-541). Springer, Cham.
 - [23] Pang, Y., Wang, Z., Jhong, J. H., & Lee, T. Y. (2021). Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. *Briefings in bioinformatics*, 22(2), 1085-1095.
 - [24] Ge, Y., Tian, T., Huang, S., Wan, F., Li, J., Li, S., ... & Zeng, J. (2020). A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. *BioRxiv*.
 - [25] Arghiani, N., Nissan, T., & Matin, M. M. (2021). Role of microRNAs in COVID-19 with implications for therapeutics. *Biomedicine & Pharmacotherapy*, 112247.
 - [26] Alpaydin, E. (2010). *Introduction to machine learning*: London.

- [27] Islam, M. M., Alam, M., Ahmed, F. F., Hasan, M., & Mollah, M. N. (2021). Improved prediction of protein-protein interaction mapping on Homo sapiens by using amino acid sequence features in a supervised learning framework. *Protein and Peptide Letters*, 28(1), 74-83.
- [28] Kösesoy, İ. (2018). Konak-patojen protein etkileşiminin hesaplamalı yöntemler ile tahmini.
- [29] J. Chen, H. Liu, J. Yang, and K. C. Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," *Amino Acids*, vol. 33, no. 3, pp. 423–428, 2007.
- [30] Usman, M., & Lee, J. A. (2019, October). Afp-cksaap: Prediction of antifreeze proteins using composition of k-spaced amino acid pairs with deep neural network. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 38-43). IEEE.
- [31] Ruan, J., Wang, K., Yang, J., Kurgan, L. A., & Cios, K. (2005). Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. *Artificial Intelligence in Medicine*, 35(1-2), 19-35.
- [32] Gök, M., & Herand, D. (2016). Prediction of bacterial virulent proteins with composition moment vector feature encoding method. In *MATEC Web of Conferences* (Vol. 49, p. 07001). EDP Sciences.
- [33] J. Shen, J. Zhang, X. Luo, and W. Zhu, "Predicting protein–protein interactions based only on sequences information," *Proc. ...*, vol. 104, no. 11, pp. 4337–4341, 2007.
- [34] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. M. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions," *Proteomics*, vol. 5, no. 4, pp. 876–884, 2005.
- [35] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. H. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins Struct. Funct. Genet.*, vol. 35, no. 4, pp. 401–407, 1999.
- [36] M. Bhasin and G. P. S. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *J. Biol. Chem.*, vol. 279, no. 22, pp. 23262–23266, 2004.
- [37] NANNI L., LUMINI A., (2006). A reliable method for HIV-1 protease cleavage site prediction methods. *Neuro Computing* 69: pp. 838- 841.
- [38] Pevsner, J. (2015). *Bioinformatics and functional genomics*. John Wiley & Sons.
- [39] Jones, N. C., Pevzner, P. A., & Pevzner, P. (2004). *An introduction to bioinformatics algorithms*. MIT press.
- [40] Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". *PNAS*. 89 (22): 10915–10919.

Bibcode:1992PNAS...8910915H. doi:10.1073/pnas.89.22.10915. PMC 50453. PMID 1438297.

- [41] Polyanovsky, V. O., Roytberg, M. A., & Tumanyan, V. G. (2011). Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms for molecular biology*, 6(1), 25.
- [42] Gök, M., Hiv-1 Proteaz Enziminin Kesme Konumlarının Tespitinde Yeni Öznitelik Vektörleri, Sakarya Üniversitesi . Sakarya, Türkiye.
- [43] Lamy, J. B. (2018, July). A new diagram for amino acids: User study comparing rainbow boxes to Venn/Euler diagram. In 2018 22nd International Conference Information Visualisation (IV) (pp. 561-566). IEEE.
- [44] Taylor, W. R. (1986). The classification of amino acid conservation. *Journal of theoretical Biology*, 119(2), 205-218.
- [45] Yang Q, Wu X (2006) 10 Challenging problems in data mining research. *Int J Inf Technol Decis Mak IJITDM* 5(4):597–604.
- [46] Sandberg K (2000) The Haar wavelet transform. University of Colorado at Boulder, Boulder.
- [47] Elkan, C. (2001, August). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.
- [48] Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning.
- [49] Gök, M. (2019). A novel machine learning model to predict autism spectrum disorders risk gene. *Neural Computing and Applications*, 31(10), 6711-6717.
- [50] CANDAN, H., DURMUŞ, A., & HARMAN, G. Genetik Algoritma Ve Sınıflandırıcı Yöntemler İle Kanser Tahmini. *Veri Bilimi*, 2(1), 30-34.
- [51] Morales, Dinora A., et al. "Predicting dementia development in Parkinson's disease using Bayesian classifiers." *Psychiatry network Research: NeuroImaging* 213.2 (2013): 92-98.
- [52] Valdes, A. D. J., Fong, M. W., & Porras, P. A. (2008). U.S. Patent No. 7,379,993. Washington, DC: U.S. Patent and Trademark Office.
- [53] Kaderali L, Radde N (2008) Inferring gene regulatory networks from expression data. *Comput Intell Bioinform. Springer, Berlin*, pp 33–74.
- [54] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7), 2038-2048.
- [55] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." *Pattern recognition* 40.7 (2007): 2038-2048.

- [56] Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a Rotasyon Ormanı?. In International Workshop on Machine Learning and Data Mining in Pattern Recognition(pp. 154-168). Springer, Berlin, Heidelberg.
- [57] Leo B. Random forests. Mach Learn 2001; 45(1): 5–32.
- [58] Budak, H., & Erpolat, S. (2012). Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması. AJIT-e: Online Academic Journal of Information Technology, 3(9), 23-30.
- [59] Arı, A., & BERBERLER, M. E. Yapay Sinir Ağları ile Yaklaşım ve Sınıflandırma Problemlerinin Çözümü İçin Arayüz Tasarımı. Acta INFOLOGICA, 1(2), 55-73.
- [60] V. Vapnik, (1995) .The nature of statistical learning theory, Springer, New York.
- [61] Bhattacharya, I., & Bhatia, M. P. S. (2010). SVM classification to distinguish Parkinson disease patients. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (pp. 1-6).
- [62] A.Sevgi. (2013). Kaba Küme ve Destek Vektör Makineleri Kullanılarak Nitelik İndirgeme ve Sınıflandırma Problemlerinin Çözümü İçin Bütünleşik Bir Yaklaşım. Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü.Doktora Tezi.
- [63] Huang, C., Davis L. S., vd. (2002). "An assessment of support vector machines for land cover classification." International Journal of Remote Sensing 23(4): 725-749.
- [64] Mathur, A., Foody, G.M. (2008). "Multiclass and binary SVM classification: Implications for training and classification users." IEEE Geoscience and Remote Sensing Letters (5):241–245.
- [65] Kavzoglu, T. ve Colkesen, I. (2009). "A kernel functions analysis for support vector machines for land cover classification." International Journal of Applied Earth Observation and Geoinformation 11(5): 352-359.
- [66] Mathur, A. ve Foody G. M. (2008). "Crop classification by support vector machine with intelligently selected training data for an operational application." International Journal of Remote Sensing 29(8): 2227- 2240.
- [67] Song, X., Duan Z., at al. (2011). "Comparison of artificial neural networks and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image." International Journal of Remote Sensing 33(10): 3301-3320.
- [68] Qureshi A, Thakur N, Tandon H, et al. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. Nucleic Acids Res 2014; 42(D1): D1147–53.

- [69] Mustafa S, et al. Peptide-protein interaction studies of antimicrobial peptides targeting middle east respiratory syndrome coronavirus spike protein: an in silico approach. *Adv Bioinformat* 2019; 2019:1–16.
- [70] Jhong J-H, et al. dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Res* 2018; 47:D1285–97. 23.
- [71] Kang X, Dong F, Shi C, et al. Dramp 2.0, an updated data repository of antimicrobial peptides. *Scientific Data* 2019; 6(1): 148–8.
- [72] Tyagi A, Tuknait A, Anand P, et al. CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* 2015; 43(D1): D837–43.
- [73] Agrawal P, et al. In silico approach for prediction of antifungal peptides. *Front Microbiol* 2019; 9:323–3.
- [74] H. Candan, M. Cetin, Makine Öğrenmesi Algoritmaları ile Parkinson Hastalığının Teşhisi, 2nd International Conference on Data Science and Applications, 2019, Balıkesir, Türkiye.
- [75] M. Eser, "Kanser Karşıtı Peptitlerin Tahmininde Yeni Öznitelik Kodlama Yöntemleri Geliştirilmesi", Yalova Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2019.
- [76] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).
- [77] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

ÖZGEÇMİŞ

İlk, orta ve lise öğrenimimi Bursa’da tamamladım. 2014 yılında girdiğim Yalova Üniversitesi Bilgisayar Mühendisliği Bölümünden 2018 yılında mezun oldum. 2019 yılında Yalova Üniversitesi Bilgisayar Mühendisliği bölümünde Yüksek Lisans eğitimime başladım. 2019-2020 yılları arasında bir yıl özel sektör deneyimim oldu. 2022 yılı itibariyle Bursa Teknik Üniversitesinde araştırma görevlisi olmaya hak kazandım.

TEZDEN TÜRETİLEN YAYIN VE ESERLER

ICONDATA 2021, Prediction of Anti-CoV Protein Sequences Using Machine Learning Methods – Hasibe Candan, Murat Gök

DiĞER YAYIN VE ESERLER

CANDAN, H., DURMUŞ, A., & HARMAN, G. Genetik Algoritma ve Sınıflandırıcı Yöntemler ile Kanser Tahmini. Veri Bilimi, 2(1), 30-34.

ICONDATA 2018, Genetik Algoritma ve Sınıflandırıcı Yöntemler ile Kanser Tahmini– (eISBN:978-605-68882-5-0) - Hasibe Candan, Arzu Durmuş, Güneş Harman

ICONDATA 2019, Anomaly Detection in Computer Networks by Machine Learning Algorithms – (eISBN:978-605-031-662-9) - Hasibe Candan, Murat Gök

ICONDATA 2019, Makine Öğrenmesi Algoritmaları ile Parkinson Hastalığının Teşhisi- (eISBN: 978-605-031-662-9) - Hasibe Candan, Müfit Çetin