YALOVA ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KANSER KARŞITI PEPTİTLERİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ GELİŞTİRİLMESİ

YÜKSEK LİSANS TEZİ

Murat ESER

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

HAZİRAN 2019

YALOVA ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KANSER KARŞITI PEPTİTLERİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ GELİŞTİRİLMESİ

YÜKSEK LİSANS TEZİ

Murat ESER
(155105007)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Murat GÖK

HAZİRAN 2019

YALOVA Üniversitesi Fen Bilimleri Enstitüsü'nün 155105007 numaralı Yüksek Lisans Öğrencisi **Murat ESER**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"KANSER KARŞITI PEPTİTLERİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ GELİŞTİRİLMESİ"** başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmustur.

Tez Danışmanı : **Doç. Dr. Murat GÖK** .....
Yalova Üniversitesi

Jüri Üyeleri : **Doç. Dr. Murat GÖK** .....
Yalova Üniversitesi

Dr.Öğr. Üyesi Murat OKKALIOĞLU .....
Yalova Üniversitesi

Dr.Öğr. Üyesi Emre DANDIL .....
Bilecik Şeyh Edebali Üniversitesi

Teslim Tarihi : **16 Mayıs 2019**
Savunma Tarihi : **21 Haziran 2019**

Eğitim hayatımda desteklerini esirgemeyen aileme ve eşime,

ÖNSÖZ

Yüksek lisans eğitimi boyunca bana yol gösteren, fikir veren, deneyimlerinden faydalandığım Sayın Doç. Dr. Murat GÖK'e, iş arkadaşlarım Sayın Emrah GÜRSAÇ, Sayın Ömer AKAY, Sayın Varol YETİŞ ve Sayın Hasan ŞİMŞEK'e, eğitime verdiği önem ile personeline daima katkıda bulunan Sayın Salih ÇAKIROĞLU'na, beni destekleyen, yol gösteren Sayın Dr. Öğr. Üyesi Metin BİLGİN'e, desteğini esirgemeyen aileme ve hayat arkadaşım Fatma ESER'e sonsuz teşekkürlerimi sunarım.

Haziran 2019 Murat Eser

İÇİNDEKİLER

Sayfa

KISALTMALAR

DNA : Deoksiribo Nükleik Asit
AA : Amino Asit
AKV : Anlık Komposizyon Vektör
ÇD : Çapraz Doğrulama
İDBA : İkili Düzensiz ve Birleştirilmiş Ağaç
ÇKA : Çok Katmanlı Algılayıcı
DVM : Destek Vektör Makineleri
k-EYK : k En Yakın Komşuluk
NB : Naive Bayes
BA : Bayes Ağları
RO : Rastgele Orman
LR : Lojistik Regresyon
DDVM : Doğrusal Destek Vektör Makineleri
RTFDVM : Radyal Temelli Fonksiyon Destek Vektör Makineleri
DP : Doğru Pozitif
DN : Doğru Negatif
YP : Yanlış Negatif
YN : Yanlış Negatif

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

KANSER KARŞITI PEPTİTLERİN TAHMİNİNDE YENİ ÖZNİTELİK KODLAMA YÖNTEMLERİ GELİŞTİRİLMESİ

ÖZET

Kanser, ölüme sebebiyet verme olasılığı yüksek olan hastalıkların başında gelmektedir. Bilinen kanser tedavilerinde kullanılan yöntemler, tümörlü hücreleri yok ederken tümör bulaşmamış hücreler de tedavi süresince etkilenmektedir. Son yıllarda çeşitli tümör tedavilerinde umut vaat eden peptid tabanlı stratejiler kullanılmaktadır. Bu doğrultuda kanser karşıtı peptidler gelişme sürecindedir. Kanser karşıtı peptidlerin ortaya çıkması ile sağlıklı hücrelere zarar verilmeden sadece tümörlü hücreler yok edilebilir. Şöyle ki, kanser karşıtı peptidler doğada katyonik olduklarından, kanser hücrelerinin anyonik hücre zar bileşenleri ile etkileşip özellikle kanser hücrelerini ortadan kaldırabilirler. Ayrıca kanser karşıtı peptidler vücuda fiziksel olarak zarar vermezler ve bu özellikleri ile yapay ilaçlardan daha etkili ve güvenilirdirler. Kanser karşıtı peptidlerin tespiti hastalıkların tedavisi ve ilaç geliştirilmesi açısından önemli bir adımdır. Ancak kanser karşıtı peptidleri diğer peptidlerden ayırt etmek maliyetli ve zor bir işlemdir. Peptid dizilimlerin istatistiki yöntemler ile tahmin edilmesi yerinde olacaktır. Bu nedenle makine öğrenmesi temelli çalışmalar ile bilgisayar ortamında hastalığın modellenerek tahmin edilmesi daha avantajlıdır.

Kanser karşıtı peptidler üzerinde yapılan çalışmalar incelendiğinde dizilim tabanlı metotların daha etkili sonuçlar verdiği anlaşılmaktadır. Bizim önerdiğimiz yeni yöntemde Evrensel Protein Kaynağı tarafından yayınlanan veri seti üzerinde 2-grams özellik çıkarım yöntemi ve Taylor Venn Diyagramı kullanılarak peptidlere ait özellikler çıkarılmış ve çıkarılan özelliklerin değerleri Blosum 30 yer değiştirme matrisi kullanılarak güncellenmiştir. Çalışma sonuçlarının etkisini artırmak için boyut daraltma yöntemlerinden Fisher doğrusal ayırtaç analizi yöntemini kullanılarak özelliklerin boyutu daraltılıp İkili Düzensiz ve Birleştirilmiş Ağaç, Çok Katmanlı Algılayıcı, Rastgele Orman, Naive Bayes, Bayes Ağları, doğrusal Destek Vektör Makineleri, radyal tabanlı fonksiyon Destek Vektör Makineleri, k-En Yakın Komşuluk, Adaboost, Bagging, k-Yıldız ve Lojistik Regresyon sınıflandırıcı algoritmalar ile sınıflandırılmıştır.

Yapılan deneysel çalışmalarla önerilen yöntemin var olan yöntemlerle karşılaştırılıp analiz edilmiştir. Analiz sonucunda kanser karşıtı peptidlerin tespiti için geliştirdiğimiz yöntem, literatürde aynı veri seti üzerinde gerçekleştirilen çalışmalara göre en yüksek performansı göstermiştir.

DEVELOPMENT OF NEW FEATURE ENCODING METHODS IN PREDICTION OF ANTICANCER PEPTIDES

SUMMARY

Cancer is one of the diseases that are likely to cause death. The methods used in known cancer therapies destroy tumor cells while tumor uninfected cells are also affected during treatment. In recent years, promising peptide-based strategies have been used in various tumor therapies. In this respect, anti-cancer peptides are in the process of development. With the emergence of anti-cancer peptides, only tumor cells can be destroyed without damaging healthy cells. Thus, as anti-cancer peptides are cationic in nature, they can interfere with the anionic cell membrane components of cancer cells, and in particular eliminate cancer cells. In addition, anti-cancer peptides do not physically damage the body and are more effective and reliable than artificial drugs. Detection of anti-cancer peptides is an important step in the treatment of diseases and drug development. However, distinguishing anti-cancer peptides from other peptides is costly and difficult. It is appropriate to estimate peptide sequences by statistical methods. For this reason, it is more advantageous to estimate the disease in computer environment with machine learning based studies.

When studies on anti-cancer peptides are examined, it can be seen that sequence based methods give more effective results. In our proposed new method, 2-grams feature extraction method and Taylor Venn Diagram were used to extract the properties of the peptides on the dataset published by Universal Protein Resource and the values of the extracted properties were updated using Blosum 30 matrix. In order to increase the effect of the results of the study, Fisher's multiple class linear discriminant analysis method was used to reduce the size of the features and classified them using Dual Perturb and Combine Tree, Multilayer Perceptron, Random Forest, Naive Bayes, Bayes Network Support Vector Machines, K-Nearest Neighbor, Adaboost, Bagging, KStar and Logistic classifier algorithms.

The experimental method has been compared and analyzed with the existing methods. As a result of the analysis, the method we developed for the detection of anti-cancer peptides showed the highest performance in the literature compared to the studies performed on the same dataset.

1. GİRİŞ

Kanser, dünya genelinde milyonlarca insanın ölümüne yol açan son zamanlardaki en büyük sağlık problemlerinden biridir [1,2]. Kanser tedavisinin zor olmasının yanı sıra tedavi maliyeti de yüksek bir hastalıktır. Geleneksel tedavi yöntemlerinde hasta hücreler ile mücadele edilirken hasta olmayan hücreler de zarar görebilmektedir [3,4]. Bu nedenle farklı tedavi yöntemlerinin bulunmasına şiddetle ihtiyaç duyulmaktadır. Peptid, amino asitlerin peptid bağı ile bağlanması ile oluşur. Son yıllarda çeşitli tümör tedavilerinde umut vaat eden peptid tabanlı stratejiler kullanılmaktadır. Bu doğrultuda kanser karşıtı peptidler gelişme sürecindedir. Kanser karşıtı peptidlerin ortaya çıkmasıyla birlikte sağlıklı hücrelere zarar verilmeden sadece hasta hücreler ortadan kaldırılabilir. Şöyle ki, kanser karşıtı peptidler doğada katyonik olarak bulunduklarından, hasta hücrelerinin anyonik hücre zar bileşenleri ile etkileşip özellikle kanser hücrelerini yok edebilirler [5,6]. Bunun yanı sıra vücuda fiziksel olarak zarar vermezler, bu ayrıcalıkları ile yapay ilaçlardan daha etkili ve güvenilirdirler [7,8]. Bunun nedeni kanser karşıtı peptidlerin doğal olarak meydana gelen biyolojik maddeler olması ve bundan dolayı sentetik ilaçlardan daha güvenli, yüksek etkinlik, seçicilik ve özgünlüğe sahip olmasıdır [9]. Bu avantajlarına ek olarak peptidler genellikle kısa dizilimlerden meydana gelirler. Avantajlarına rağmen kanser karşıtı peptidleri diğer peptidlerden ayırt etmek maliyetli ve zor bir işlemdir. Bu nedenle dizilimlerin istatistiki yöntemler ile tahmin edilmesi yerinde olacaktır.

1.1 Tezin Amacı

Kanser hastalığının tedavisinde alternatif ve gelecek vaat eden kanser karşıtı peptitler, doğada biyolojik olarak meydana gelirler. Gelenekse tedavi yöntemlerine göre avantajları olmasına rağmen diğer peptidlerden laboratuvar ortamında ayırt etmek zor ve maliyetli bir işlemdir. Bu nedenle bilgisayar ortamında makine öğrenmesi algoritmaları kullanarak çözüm aramak daha uygundur.

Bu tez çalışmasında amacımız kanser karşıtı peptidleri belirlemek için yeni öznitelik kodlama yöntemleri geliştirmektedir.

1.2 Literatür Araştırması

Kanser karşıtı peptidleri belirlenmesi üzerine günümüze kadar laboratuvar ve bilgisayar ortamında bazı çalışmalar yapılmıştır. Yapılan çalışmalarda öznitelik çıkarımı, dizilim tabanlı, amino asitlerin kimyasal özellikleri ve fiziksel özellikleri temel alınarak oluşturulmuştur. Laboratuvar ortamında yapılan çalışmalar maliyet ve zaman açısından büyük handikaplar oluşturmaktadır. Bu nedenle, bilgisayar ortamında peptidlerin dizilimlerine ait istatistiki yöntemler kullanılarak kanser karşıtı peptidlerin yüksek doğruluk ve düşük zaman maliyeti ile belirlenmesi hedeflenmektedir. Kanser karşıtı peptidlerin sağlık açısından önemi ve henüz tam anlamıyla başarıya ulaşmamış olması yeni ve daha güçlü bir kanser karşıtı peptid tespit edici ihtiyacı doğurmaktadır.

Tyagi ve arkadaşları kanser karşıtı peptidleri ve mikrobiyal karşıtı peptidleri sınıflandırmak için, ikili peptid bileşimi, ikili profil ve amino asit bileşimi özellik çıkarımı yöntemlerini kullanarak destek vektör makineleri ile sınıflandırma işlemi gerçekleştirmişlerdir [10]. Gerçekleştirdikleri çalışmada farklı yöntemlere ait en yüksek %92,65 sınıf doğruluğu elde etmişlerdir. Hajisharifi ve arkadaşları yerel hizalama kernel metot tabanlı yeni bir yöntem önermişlerdir [11]. Önerdikleri modelde sözde amino asit bileşimi ile özellik çıkarımı yaptıktan sonra destek vektör makineleri ile sınıflandırma işlemi gerçekleştirmişlerdir. Gerçekleştirdikleri çalışmada %89,70 sınıf doğruluğu elde etmişlerdir. Chen ve arkadaşları kanser karşıtı peptidleri sınıflandırmak için, üzere dizilim tabanlı bir yöntem geliştirmişlerdir. Geliştirdikleri modelde g-gap ikili peptid mod ile sözde amino asit bileşimi özellik çıkarımı ve varyans analizi ile özellik seçimi yöntemini kullanarak destek vektör makineleri ile kanser karşıtı peptidleri belirlemişlerdir. Gerçekleştirdikleri çalışmada %95,06 sınıf doğrulu elde etmişlerdir [9]. Xu ve arkadaşları 400D ve g-gap ikili peptid bileşimi özellik çıkarımı, maksimum ilgi-maksimum uzaklık özellik seçimi yöntemini kullanarak destek vektör makineleri ile kanser karşıtı peptidleri belirlemişlerdir. Gerçekleştirdikleri çalışmada %91,86 sınıf doğrulu elde etmişlerdir [12]. Manavalan ve arkadaşları gerçekleştirdikleri çalışmada, amino asit bileşimi, ikili peptid bileşimi, atomik bileşim ve amino asitlerin fizikokimyasal özelliklerini kullanarak özellik çıkarımı yaptıktan sonra destek vektör makineleri ve rastgele orman sınıflandırma algoritmaları kullanarak %88,70 sınıf doğruluğu elde etmişlerdir [13].

1.3 Hipotez

Kanser karşıtı peptidlerin makine öğrenmesi yöntemleri ile tahmin edilmesinde; peptidde yer alan amino asitlerin konumu, fizikokimyasal özellikleri ve birbirlerine olan benzerlikleri ile ilgili peptidin kanser karşıtı olması arasında ilişki vardır.

2. KANSER KARŞITI PEPTİDLER

2.1 Amino Asit

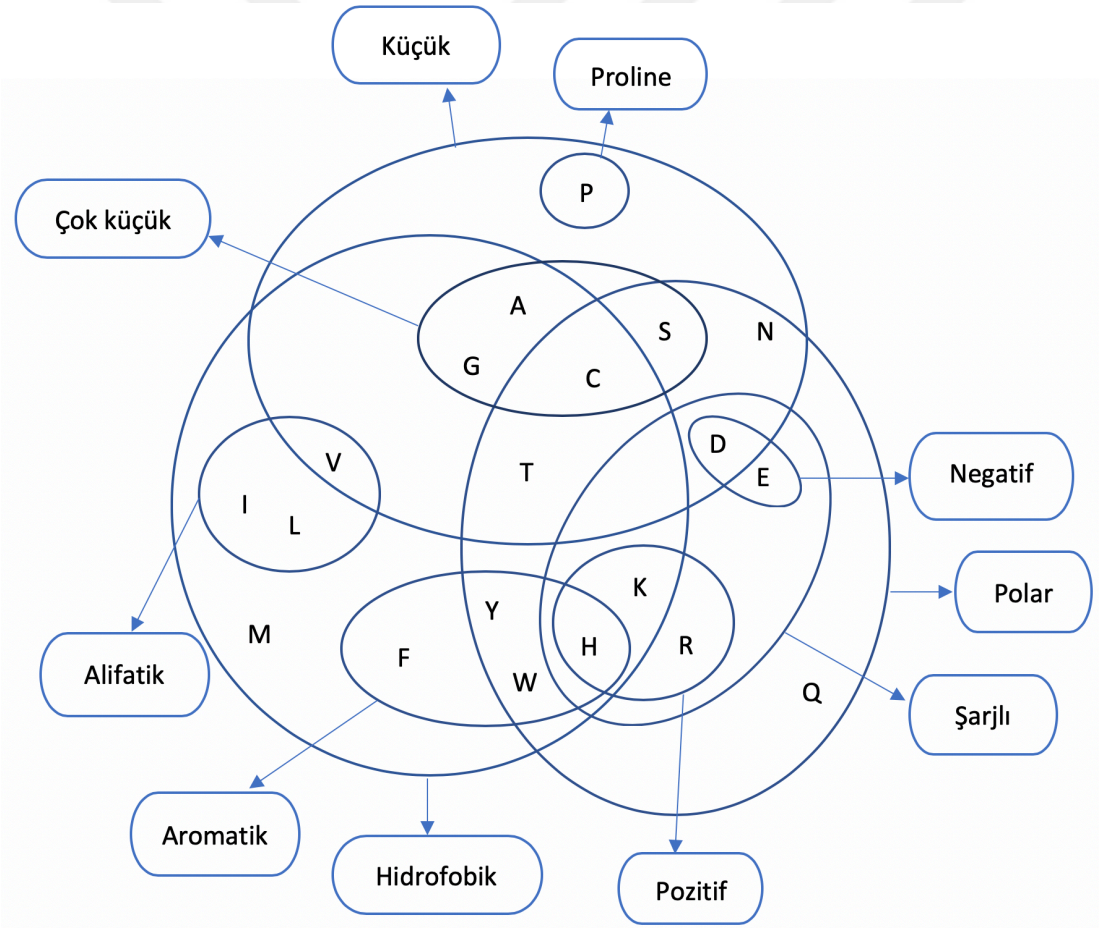
Yapılarında amino grubu ve karboksil grubu içeren bileşiklere amino asit (AA) denir. Çok sayıda amino asit bir araya gelerek temel taşı oldukları proteinleri oluştururlar. Doğada 300'den fazla bulunan amino asitler, proteinlerde farklı sayıda ve dizilimde bulunarak proteinlerin işlevinde rol oynar. Proteinlerde 20 çeşit amino asit bulunmaktadır. Doğada 300'den fazla amino asit bulunmasına rağmen insanlarda sadece 20 adet bulunur. İnsanlarda bulunan 20 amino asidin 12 tanesi insan vücudunda doğal olarak bulunurken kalan 8 tanesi dışarıdan alınır. Bu 20 amino asit DNA tarafından kodlanarak protein sentezinde kullanılır [14]. Çizelge 2.1'de 20 standart amino asit görülmektedir.

Çizelge 2.1: Standart 20 AA.

Amino Asit	3 Harf	1 Harf
Alanin	Ala	A
Arginin	Arg	R
Asparajin	Asn	N
Aspartik Asit	Asp	D
Sistenin	Cys	C
Glütamin	Gln	Q
Glütamik Asit	Glu	E
Glisin	Gly	G
Histidin	His	H
İzolösin	Ile	I
Losin	Leu	L
Lizin	Lys	K
Metiyonin	Met	M
Fenilalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Treonin	Thr	T
Triptofan	Trp	W
Trozin	Tyr	Y
Valin	Val	V

2.1.1 Amino asitlerin fizikokimyasal özellikleri

Doğada farklı karakteristikte bulunan 20 temel amino aside ait fizikokimyasal özellikler amino asitlerin etkinliklerinin incelenmesinde önemli rol oynamaktadır. Amino asitler 544 adet fizikokimyasal özelliğe ve her özellik sayısal değer indeks değerlerine sahiptir. Fizikokimyasal özellikler peptidlerin sınıflandırılmasında, peptidlerin özelliklerinin ve aktivite alanlarının tanımlanmasında yardımcı olmaktadır. Bu tez çalışmasında amino asitlere ait fizikokimyasal özellikler için Taylor Venn Diyagramındaki 10 adet özellik kullanılmıştır. Diyagramda bulunan özellikler, amino asitlerin fizikokimyasal özelliklerine göre küme oluşturmaktadır. Şekil 2.1'de Taylor Venn Diyagramı incelendiğinde amino asitler küçük, hidrofobik, çok küçük, polar, proline, negatif, pozitif, şarjlı, aromatik ve alifatik özelliklerine sahiptirler.



Şekil 2.1: Taylor Venn diyagramı.

2.1.2 Blosum yer değiştirme matrisleri

Protein dizilim karşılaştırmalarında ve protein hizalama problemlerinde kullanılır. Yer değiştirme matirisndeki sayılar, amino asitlerin birbirleri yerine geçebilme

durumlarına dair bilgi verir [14]. Yer değiştirme matrisleri, amino asitlerin birbirlerine benzerliklerini puanlar. İlk olarak Steven Henikoff ve Jorja Henikoff tarafından önerilmiştir [15].

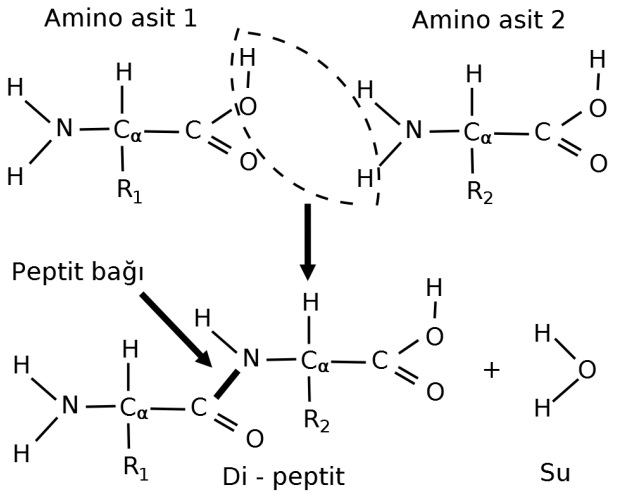
Yer değiştirme matrislerinde A_x ve A_y amino asitlerinin benzerlik oranı, x. satır ve y. sütunun kesişiminde yer alan değer ile elde edilir [14]. ACCYVLLYAC dizilimine sahip peptid, yer değiştirme matrisinde kodlarsak AC= -3, CC=17, CY= -6, YV= 1, VL= 1, LL= 4, LY= 3, YA= -4 değerlerini almaktadır. Şekil 2.2'de Blosum 30 yer değiştirme matrisi görülmektedir.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	0	0	-3	1	0	0	-2	0	-1	0	1	-2	-1	1	1	-5	-4	1
R	-1	8	-2	-1	-2	3	-1	-2	-1	-3	-2	1	0	-1	-1	-1	-3	0	0	-1
N	0	-2	8	1	-1	-1	-1	0	-1	0	-2	0	0	-1	-3	0	1	-7	-4	-2
D	0	-1	1	9	-3	-1	1	-1	-2	-4	-1	0	-3	-5	-1	0	-1	-4	-1	-2
C	-3	-2	-1	-3	17	-2	1	-4	-5	-2	0	-3	-2	-3	-3	-2	-2	-2	-6	-2
Q	1	3	-1	-1	-2	8	2	-2	0	-2	-2	0	-1	-3	0	-1	0	-1	-1	-3
E	0	-1	-1	1	1	2	6	-2	0	-3	-1	2	-1	-4	1	0	-2	-1	-2	-3
G	0	-2	0	-1	-4	-2	-2	8	-3	-1	-2	-1	-2	-3	-1	0	-2	1	-3	-3
H	-2	-1	-1	-2	-5	0	0	-3	14	-2	-1	-2	2	-3	1	-1	-2	-5	0	-3
I	0	-3	0	-4	-2	-2	-3	-1	-2	6	2	-2	1	0	-3	-1	0	-3	-1	4
L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4	-2	2	2	-3	-2	0	-2	3	1
K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4	2	-1	1	0	-1	-2	-1	-2
M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6	-2	-4	-2	0	-3	-1	0
F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10	-4	-1	-2	1	3	1
P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11	-1	0	-3	-2	-4
S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4	2	-3	-2	-1
T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5	-5	-1	1
W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20	5	-3
Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9	1
V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5

Şekil 2.2: Blosum 30 yer değiştirme matrisi.

2.2 Peptid

Doğada bulunan herhangi bir amino asit ile başka bir amino asidin, amino ve karboksil grubuna bağlanmasıyla peptid bağı ve peptid oluşur. Amino asitlerin peptid bağlar kurarak oluşturdukları bileşiklere peptid denir [14]. Şekil 2.3'te bağ oluşumu görülmektedir. İki amino asit molekülü bir amit bağı ile birbirlerine bağlandıklarında bir dipeptid oluşmaktadır. 3 amino asit bir araya gelerek tripeptidleri oluşturur. 20 ve daha fazla amino asitten oluşan peptidler polipeptid olarak isimlendirilirler.



Şekil 2.3: İki amino asidin peptid bağı oluşturması [16].

2.2.1 Kanser karşıtı peptidler

Kanser karşıtı peptidler doğada katyonik olduklarından, kanser hücrelerinin anyonik hücre zar bileşenleri ile etkileşime girip özellikle kanser hücrelerini yok edebilirler [5,6]. Bunun yanı sıra vücuda fiziksel olarak zarar vermezler, bu ayrıcalıkları ile yapay ilaçlardan daha etkili ve güvenilirdirler [7,8]. Bunun nedeni kanser karşıtı peptidlerin doğal olarak meydana gelen biyolojik maddeler olması ve bundan dolayı sentetik ilaçlardan daha güvenli, yüksek etkinlik, seçicilik ve özgünlüğe sahip olmasıdır [9]. Bu avantajlarına ek olarak peptidler genellikle kısa dizilimlerden meydana gelirler. Ancak kanser karşıtı peptidleri diğer doğal veya yapay peptidlerden ayırmak zordur. Deneysel metotları kullanarak kanser karşıtı peptidleri belirlemek oldukça pahalı ve zaman alıcıdır. Dahası bu yöntemlerden sadece bir kaç tanesi kliniklerde uygulanabilir [17].

2.3 Veri Seti

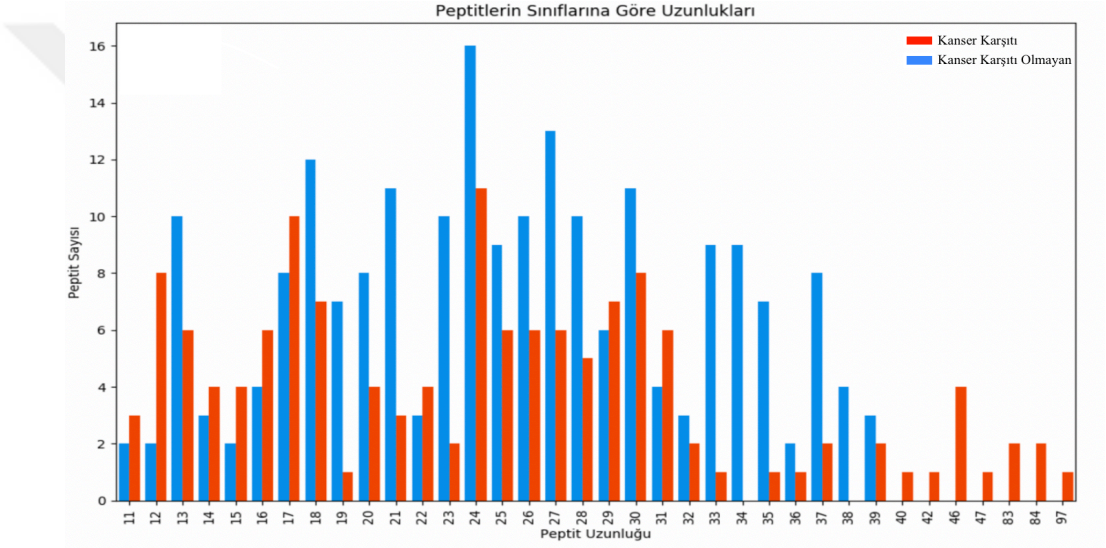
Bu tez çalışmasında, $P = P^+ \cup P^-$ ile formüle edilen, içerisinde pozitif ve negatif örnekler barındıran veri seti kullanılmıştır [18]. Veri setinde tekrarlı veriden kaçınmak ve biası düşürmek için orjinal veri seti üzerinde çalışmalar yapılarak 138 pozitif ve 206 negatif etiketine sahip peptid elde edilmiştir [9]. Veri setinde bulunan pozitif örnekler P^+, negatif örnekler P^- ile gösterilmektedir. Çalışmada 10 kat çapraz doğrulama yöntemi kullanılmasından dolayı kullanılan veri seti eğitim ve test veri seti olarak ayrılmamıştır. Veri seti incelendiğinde pozitif örnekler, negatif örnekler arasında bulunmamaktadır. Aynı kriter negatif örnekler içinde geçerlidir. Veri setinde bulunan pozitif yani kanser karşıtı peptidler incelendiğinde büyük çoğunluğunun 30

veya daha az amino asitten oluştuğu anlaşılmaktadır. Veri seti özellikleri Çizelge 2.2 de verilmiştir.

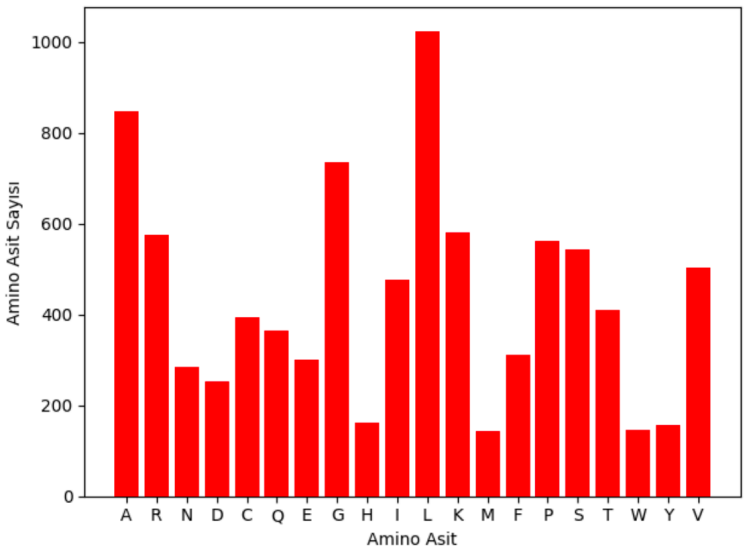
Çizelge 2.2: Veri seti özellikleri.

Peptid Sayısı	Kanser Karşıtı	Kanser Karşıtı Olmayan
344	138	206

Veri setinde bulunan peptidlerin uzunlukları birbirinden farklıdır. Her peptid içerisinde 20 adet temel amino asit bulunmaktadır. Veri setinde bulunan peptidlerin uzunluklarına ve amino asit sayılarını ilişkin veriler Şekil 2.4 ve Şekil 2.5'te gösterilmektedir.



Şekil 2.4: Peptidlerin uzunluklarına göre istatistikleri.

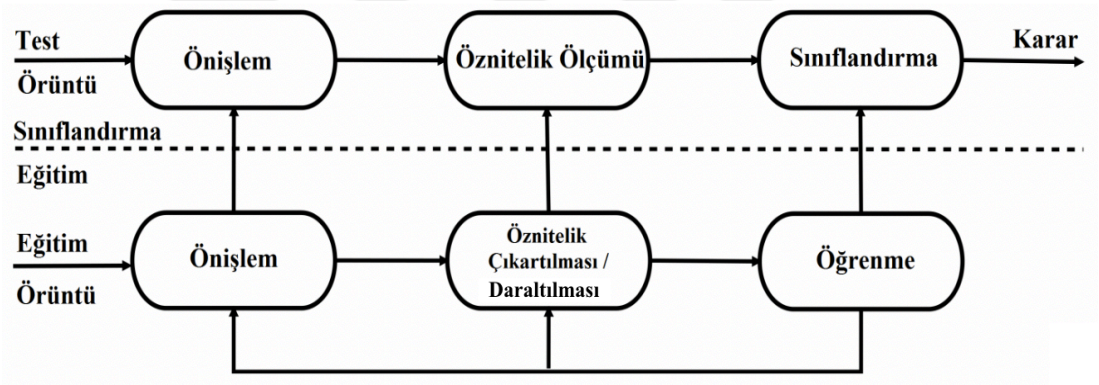


Şekil 2.5: Peptidlerde bulunan amino asit istatistikleri.

Veri setinde bulunan peptidlerin benzerlik analizi eşleşme skoru yöntemi ile yapılmıştır. Benzerlik analizi, sınıflandırma işleminde ezberlemenin önlenmesi için yapılmaktadır. Makine öğrenmesinde ezberleme, test verisinin eğitim sırasında sınıflandırması yapılan veriler ile benzerliğinin yüksek olması durumunda sınıflandırma işleminin yüksek doğrulukla yapılmasıdır. Bu durumda öğrenme değil ezberleme işlemi gerçekleştirilir. Veri seti ne kadar özgün olursa öğrenme işlemi o kadar başarılı olur.

3. MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİN

Makine öğrenmesinin kollarından biri olan örüntü tanıma, nesneleri sınıflara ve kategorilere ayırma amacına dayalı bir bilim dalıdır [14]. Nesneleri sınıflandırmak veya kategorilere ayırmak için nesneye ait öznitelikleri kullanır. Örüntü tanıma, örüntü içindeki sürekli tekrar eden veya yer alan ifadelerden anlamlar çıkararak örüntüyü tanımaya çalışmaktadır. Biyoinformatik, otonom araçlar, parmak izi tanıma, plaka tanıma gibi makine öğrenmesi konularında en uygun kararın verilmesinde önemli görevler üstlenmektedir [19]. Biyoinformatik alanında, teknolojik gelişmelere bağlı olarak canlılar üzerinde yapılması zor ve maliyetli uygulamalar, bilgisayar ortamında gerçekleşebilmektedir [14].



Şekil 3.1: Örüntü tanıma sistemi modeli [20].

Örüntü tanıma eğitim ve sınıflandırma olmak üzere iki bölümden oluşmaktadır. Şekil 3.1'de örüntü tanıma modeli görülmektedir. Sınıflandırılacak örüntü ön işlemden geçerek örüntüye ait öznitelikler çıkarılır. Özniteliklerin elde edilmesi ile işlenebilir hale gelmektedir. Öğrenme sistemine ait problem bağlanım, sınıflandırma veya kümeleme problemi olabilir [21]. Bağlanım iki veya daha fazla değişken arasındaki bağlantıyı belirlemek için kullanılan istatistiki yöntemdir [19]. En basit makine öğrenimi modelidir. Bu bağlantının düzeyini ölçmeye bağlanım analizi denir. Temel amaç verilere ait parametreler arasındaki ilişki fonksiyonunu elde etmektir. Fonksiyon elde edilerek yeni veriler üzerinde tahminler yapılabilir. Sınıflandırma örüntü tanıma sürecinde daha önceki deneyimlere dayanarak verinin tanımlanmış olan sınıflardan hangisine ait olduğunun belirlenmesidir. Sınıflandırma gözetimli öğrenme alanına

girmektedir. Sınıfların önceden belli olduğu ve sonucu parametrelerin etkilediği yöntemdir. Sonucu etkileyen parametrelerin önceden öğrenilip model oluşturulması ile daha önceden belirlenen sınıflardan biri üzerinde tahmin eden yöntemdir. Kümeleme, verilerin özelliklerine göre analiz edilerek sınıflandırma yapılması işlemidir. Gözetimsiz öğrenme yöntemleri arasındadır. Gözetimsiz öğrenmede, sınıflandırma işleminde sınıfların daha önceden bilinmeden tahmin edilmesi yöntemidir. Yani girdilere ait çıktılar yoktur.

Örüntü tanımada, örüntüye ait gözlemlenebilir ve ölçülebilir verilere dayanarak öznitelikler tanımlanmaktadır [14]. Şekil 3.1'de örüntüye ait verilerin eğitim süresinde özniteliklerin çıkarılması ve seçilmesi, ardından sınıflandırıcı tarafından sınıflandırma işlemi yapılmaktadır. Oluşturulan modelde, başarı alınan geri bildirim ile artırılabilir.

3.1 Öznitelik Çıkarım Yöntemleri

3.1.1 2-grams özellik çıkarım yöntemi

2-grams özellik çıkarımı ile peptid diziliminde 2'li ardışık kalıntı çiftleri aranır [22]. Böylece ardışık ikili aminoasit çiftlerinin sıklığı (frekansı) ortaya çıkarılır. Bu durumda her bir peptid için 2-grams öznitelik vektörü 20x20=400 uzunluğunda olur. ACCYVLLYAC diziliminde 2-grams özellik çıkarım yöntemi uygulanırsa 9 adet kalıntı çifti elde edilir. AC amino asit ikilisi dizilimde iki defa tekrar etmektedir. Bu şekilde dizilimlerde yer alan bütün amino asitlere uygulanır. Buna göre AC = 2, CC = 1, CY = 1, LL = 1, LY = 1, VL = 1, YA = 1, YV = 1 olur. Şekil 3.2'de ACCYVLLYAC peptidinin 2-grams özellik çıkarım yöntemine göre elde edilen 1×400 büyüklüğündeki özellik vektörü yer almaktadır.

AA	AC	AD	……	CC	CY	LL	LY	VL	YA	YV
0	2	0	……	1	1	1	1	1	1	1

Şekil 3.2: ACCYVLLYAC peptidinin 2-grams yöntemine göre özellik çıkarımı.

3.1.2 Anlık kompozisyon vektörü özellik çıkarım yöntemi

Anlık kompozisyon vektörü (AKV-Composition Moment Vector), amino asit dizilimi ile ilgili fonksiyonel bir ilişki sağlamaktadır. AKV öznitelik vektörü 1×40 boyutundadır. Vektörün ilk 1×20 kısmı, AA bileşimi ve frekansı ikinci 1×20 kısmı, amino asidin konumu hakkında bilgi içerir. Bununla birlikte, kompozisyon vektörü

pozisyon bilgisini tamamen reddeder. Bu nedenle Ruan ve arkadaşları tarafından, AA'nın hem kompozisyonu hem de pozisyonu hakkında bilgiler içeren yeni bir yöntem önerilmiştir. Kompozisyon vektörünün aksine, konum ve pozisyon bilgisi kullanarak dizilim ile ilgili fonksiyonel bir ilişki sağlar [23].

Anlık kompozisyon vektörü yönteminde amino asitlerin öznitelik vektöründeki değeri 3.1 denklemindeki formül ile elde edilir. Pozisyon ve kompozisyon bilgisi ile hesaplama yapıldığında her dizilim için 1×40 boyutunda öznitelik vektörü elde edilir. N peptid dizilim uzunluğu, i hesaplanacak amino asidin sırasını, j amino asidin tekrar sayısını, K amino asitin dizilim içindeki sayısını vermektedir.

$$x_i{}^{(k)} = \frac{1}{N(N-1)\dots(n-k)} \sum_{j=1}^{K_i} n_{ij}^{k} \; for\; i = 1,2,..,20 \qquad (\mathbf{3.1})$$

ACCYVLLYAC peptid diziliminin anlık kompozisyon vektörü öznitelik kodlama yöntemine göre hesaplaması 2.2 denklemindeki formül ile elde edilir. N=10, K_1 = 2(A amino asidinin tekrar sayısı), n_{11}= 1, n_{12} =9 (A amino asidinin pozisyonları) değişkenleri 2.1'de verilen formül ile A amino asidi için öznitelik vektöründeki pozisyon değeri 2.2, kompozisyon değeri 2.3'te yer alan formül ile hesaplanır.

$$x_i{}^{(k)} = \frac{1}{N(N-0)} \sum_{j=1}^{2} n_{1j} = \frac{1}{10 \times 10}(1+9) = \frac{10}{100} = 0{,}10 \qquad (\mathbf{3.2})$$

$$x_i{}^{(k)} = \frac{1}{N(N-1)} \sum_{j=1}^{2} n_{1j} = \frac{1}{10 \times 9}(1+9) = \frac{10}{90} = 0{,}11 \qquad (\mathbf{3.3})$$

ACCYVLLYAC diziliminde yer alan amino asitler için hesaplamalar yapıldığında öznitelik vektörü değeri sıfır olanlar hariç şekil 3.3'te verilmektedir.

A	C	Y	V	L	A'	C'	Y'	V'	L'
0.10000	0.15000	0.13000	0.12000	0.05000	0.11111	0.16667	0.14444	0.13333	0.05556

Şekil 3.3: ACCYVLLYAC peptidinin AKV yöntemine göre özellik çıkarımı.

3.2 Boyut Daraltma Yöntemleri

Makine öğrenmesinde problemin sınıflandırılma başarısı öznitelik vektörünün problemi ifade etmesine bağlıdır. Ancak öznitelik vektöründe yer alan özniteliklerin

sayısı fazla olursa problemin sınıflandırma başarısını olumsuz etkileyebilir. Bu nedenle varolan özniteliklerin daha anlamlı az sayıda bileşenleri ile ifade edilmeleri gerekebilir. Bu işlem, boyut daraltma yöntemleri ile gerçekleştirilir. Boyut daraltma yöntemlerinde temel bileşenler analizi, Fisher doğrusal ayırtaç analizi gibi yöntemler sıklıkla kullanılmaktadır. Biz kanser karşıtı peptidlerin tahmininde Fisher doğrusal ayırtaç analizi yöntemini kullandık.

3.2.1 Fisher doğrusal ayırtaç analizi

Genel olarak, sınıflar arası dağılma oranını en yüksek yaparken sınıf içindeki dağılmayı en küçük yapar. Yöntemin amacı, iki veya daha fazla sınıfı, kendilerini ifade eden niteliklerden doğrusal bir kombinasyon elde ederek ayırmaktır. Fisher doğrusal ayırtaç analizinde veri tek boyut ile ifade edilir. Eldeki veri iyi bir şekilde ayrılmış olsa da, bir doğru üzerine yapılan yansıtma sonunda karışık bir durum oluşabilir. Doğru üzerinde belirtilen örneklerin iç içe girmesi sonucunda ayrım zorlaşır, dolayısıyla örnekleri tanımada doğruluk düşer [24].

3.3 Sınıflandırma Yöntemleri

Makine öğrenmesi bilgisayar biliminin yapay zeka üzerinde istatistiki öğrenme ve model geliştirme çalışmalarına ait alt bilim dalıdır. Örnek veriler üzerinde öğrenme işleminden sonra örneklere ait çıkarım yapabilirler. Makine öğrenmesi, istenmeyen elektronik posta, karakter tanıma, ses tanıma, bilgisayarlı görme, örüntü tanıma gibi bir çok alanda uygulaması vardır [25]. Makine öğrenmesi tahmin etme aracı olarak sınıflandırıcı algoritmalar kullanıp, veri setinde sınıflandırma yaparak sonuçlar üretir. Sınıflandırma, veri seti üzerinde özniteliklere bağlı olarak istatistiki hesaplamalar sonucu verinin daha önceden belirlenmiş sınıflar arasında dağıtılmasıdır. Sınıflandırma algoritmaları, önceden öznitelikler ile eğitilen veri setini kullanarak, test verileri üzerinde sınıf tahmini yapar.

Makine öğrenmesinde temel olarak gözetimli ve gözetimsiz olmak üzere iki yöntem kullanılır. Gözetimli öğrenmede sınıflandırıcı algoritma giriş değerleri için çıkış değerleri üzerinde tahminde bulunur. Eğitim sırasında girdiler ile çıktılar arasında oluşturulan fonksiyon çıktılar kategorik ise sınıflandırma, nümerik ise bağlanım algoritmaları kullanılır. Gözetimsiz öğrenmede etiket bilgisine sahip olmayan veriler üzerinde çalışarak sınıf bilgisini tahmin etmek için sınıflandırıcılar kullanılır. Giriş

verisinin hangi sınıfa ait olduğu belirsizdir. Kümeleme, gözetimsiz sınıflandırma yöntemine örnektir [26] ve gözetimsiz öğrenmede en çok kullanılan tekniklerden biridir. Gözetimli öğrenme bir sınıf verisine ait olan giriş değerlerini en uygun sınıfa atarken gözetimsiz öğrenme sınıf verisi olmadan giriş değerini sınıflara ayırır.

Kanser karşıtı peptidlerin tahmin edilmesi için geliştirilen öznitelik kodlama yöntemleri İkili ve Düzensiz Birleştirilmiş Ağaç, Çok Katmanlı Algılayıcı, Rastgele Orman, Naive Bayes, Bayes ağları, Doğrusal Destek Vektör MAkineleri, Radyal Tabanlı Fonksiyon Destek Vektör Makineleri, k-En Yakın Komşuluk, Adaboost, Bagging, k-Yıldız ve Lojistik Regresyon sınıflandırıcı algoritmalar ile test edilmiştir. Geliştirilen öznitelik kodlama yöntemleri ile elde edilen öznitelik vektörleri üzerinde sınıflandırma yapılıp, sonuçlar incelenmiştir.

3.3.1 İkili düzensiz ve birleştirilmiş ağaç

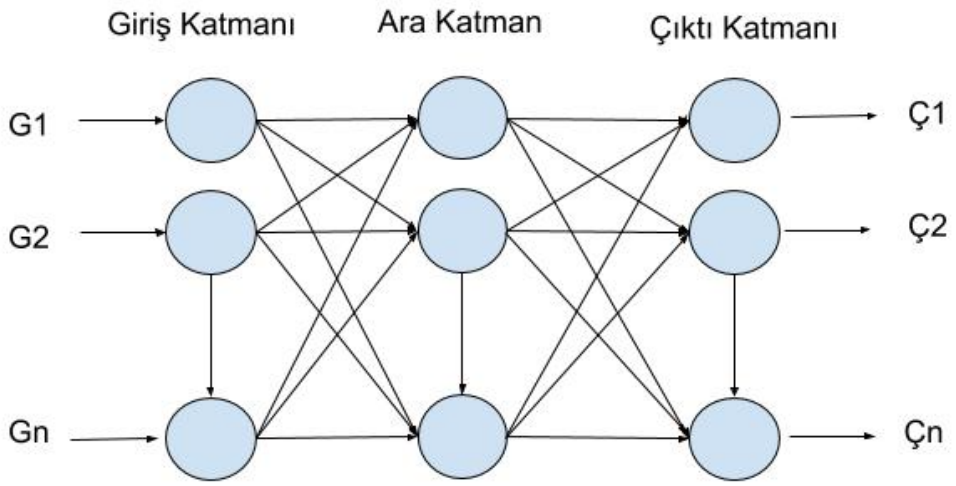
İkili Düzensiz ve Birleştirilmiş Ağaç (İDBA-Dual Perturb and Combine Tree) algoritması 2001 yılında Geurts tarafından tanıtılmıştır. Düzensiz ve birleştirilmiş algoritmadan farklı olarak, sadece bir model kullanır ve tahmin aşamasına bir test vakasına karşılık gelen öznitelik vektörünü bozarak çoklu tahminlerin oluşturulmasını geciktirir. İkili Düzensiz ve Birleştirilmiş Ağaçta sınıflandırma ve regresyon ağaçları paralel olarak ele alınmakta ve bozulma derecesini otomatik olarak ayarlamak için bir çapraz doğrulama tekniği önerilmektedir [27]. Kullandığı teknik ile hesaplama zamanı ve hafıza kullanımı açısından verimlidir. Ayrıca sinir ağları tarafından sergilenen varyansın azaltılmasına izin veren, aynı zamanda tahminlerin güven değerlerinin tahmin edilmesini sağlayan etkili bir yöntemdir. Yöntemde sırasıyla test örneği birçok kez karıştırılır, özellik değerlerine gürültü eklenir, test örneklerinin karıştırılmış her versiyonu tahmin edilir. Nihai olarak yapıalcak son tahmin, tüm tahminlerin toplanmasıyla elde edilir [28].

3.3.2 Çok katmanlı algılayıcı

Çok katmanlı algılayıcılar (ÇKA-Multilayer Perceptron), nöron olarak adlandırılan ve birbirine bağlı yapay sinir hücrelerinden oluşur. Yapay sinir hücrelerinin birbirleri ile olan bağlantılarıyla bilgi akışı ağ içerisinde sağlanır [29]. Doğrusal olmayan problemlerde başarılı olmaktadır. Çok katmanlı algılayıcılar, bilgilerin sisteme giriş yaptığı girdi, bir veya daha fazla gizli (ara) ve son olarak çıktı katmanlarından oluşur. Çok katmanlı algılayıcılarda katmanlar içerisinde ileri ve geri yayılım olarak

isimlendirilen geçişler vardır. İleri yayılım aşamasında, ağın ürettiği çıktı ve oluşan hata değeri hesaplanır. Geri yayılım aşamasında ise hesaplanan hata değerinin en düşük hale getirilmesi için katmanlar arası bağlantıda bulunan ağırlık değerleri güncellenir [30].

Çok katmanlı algılayıcı, bilgi girişinin yapıldığı, bilgilerin işlendiği, çıktıların elde edildiği girdi, ara(gizli) ve çıktı katmanlarından oluşmaktadır. Genellikle geri yayılım algoritması kullanılmaktadır. İlk aşamada girdilere ait rastgele ağırlıklar üretilmektedir. Giriş katmanında girdiler elde edilerek ara katmana iletilir. Ara katman bir veya daha fazla olup her katman farklı sayıda işlem elemanına sahip olabilir. Ara katmanda ağırlıklar, girdiler ve eşik değeri ile toplam değer elde edilir. Toplam değer aktivasyon fonksiyonundan geçirilerek net değer elde edilir. Net değer bir sonraki ara katmana veya çıktı katmanına iletilir. Ara katmanda bulunan elemanlar ve çıktı katmanında bulunan elemanların ürettikleri çıktı net girdinin hesaplanması ve aktivasyon fonksiyonundan geçirilmesi sonucu belirlenir.

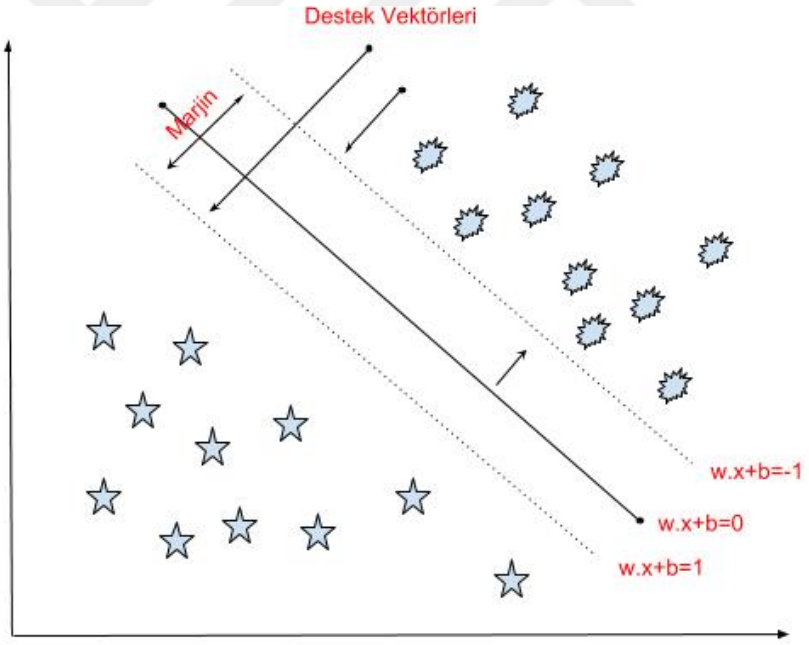


Şekil 3.4: Çok katmanlı algılayıcı.

Başlangıçta verilen girdilere ait çıktılar ile ağın ürettiği çıktılar karşılaştırılır. Aradaki fark hata olarak kabul edilir. Elde edilen hata değeri algılayıcı modelin ağırlıklarına dağıtılarak bir sonraki iterasyonda hatanın azaltılması hedeflenir. Çıktı katmanındaki her işlem elemanına ait hata değerini azaltmak için hatalar ağırlıklara dağıtılır. Ağırlıkların güncellenmesi girdi katmanına kadar devam eder. Bu şekilde geri yayılım izlenerek bir sonraki iterasyonda katmanlardaki ağırlıklar güncellenerek çıktı katmanındaki hata en aza indirilir. Şekil 3.4'de çok katmanlı algılayıcının gösterimi yapılmaktadır.

3.3.3 Destek vektör makineleri

Destek vektör makineleri (DVM-Support Vector Machines) Vapnik tarafından geliştirilmiştir [31]. Temeli istatistiksel öğrenme teoremine dayanan DVM algoritmasında özellik vektörü olarak isimlendirilen ve içerisinde en uygun aşırı düzlemlerin belirlendiği yüksek boyutlu bir vektörü haritalandırmaktadır. Haritalandırılan özellik, eğitim verilerinden elde edilerek oluşturulmaktadır [32]. DVM'nin temel amacı, eğitim verisinde yer alan sınıfların üyeleri arasında, bu sınıfları birbirinden ayıran sonsuz sayıdaki doğru içinden marjini en yüksek doğruyu seçerek sınıflandırmayı gerçekleştirmeye çalışmaktır. Oluşan marjin doğrusu, sınıf üyelerinin seçilen doğruya en yakın olan üyelerine paralel olmalıdır. Çizilen doğrulara üst düzlem adı verilmektedir [33]. En büyük marjine sahip üst düzlemi seçebilmek için her iki sınıfa yakın noktalardan geçen paralel doğrular çizilebilir. Paralel olarak çizilen ve destek vektörleri adı verilen bu doğrular arasındaki uzaklığa marjin adı verilmektedir [34]. Şekil 3.5'te marjin, destek vektörleri ve üst düzlem görülmektedir.



Şekil 3.5: İki sınıflı veri kümesinde DVM algoritması üst düzlemi.

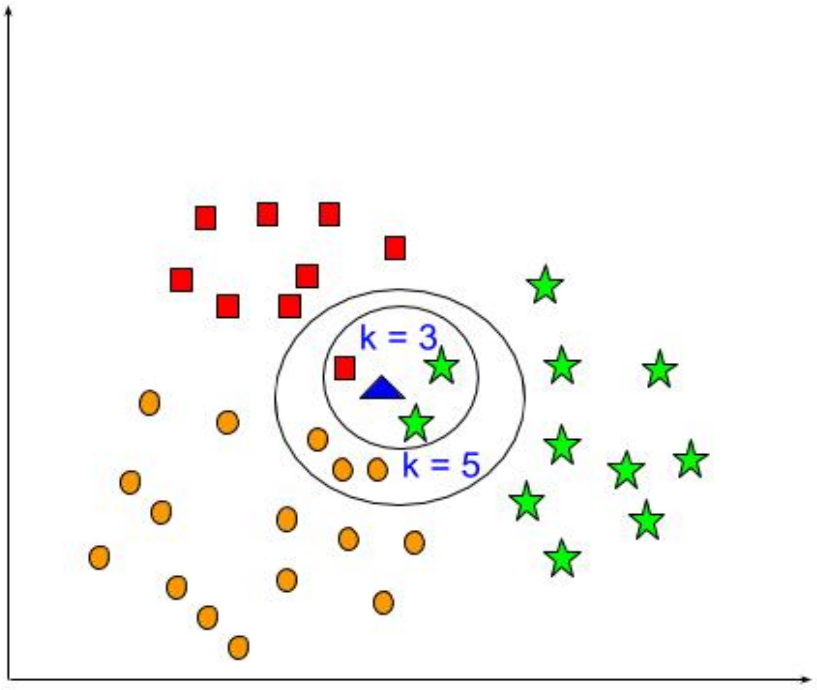
Şekil 3.3'te iki sınıfı birbirinden ayıran üst düzlem için $w \times x + b = 0$ iken destek vektörleri $w \times x + b = 1$ ve $w \times x + b = -1$ olmaktadır.

Çalışmada LIBSVM kütüphanesi kullanılmıştır [35]. DVM çekirdek fonksiyonu olarak lineer ve radyal tabanlı fonksiyon kullanılmıştır.

3.3.4 k-En yakın komşuluk

k-En Yakın Komşuluk (k-EYK- k-Nearest Neighborhood) algoritması, Cover ve Hart tarafından önerilmiştir. Algoritma, örnek veri noktasının örnek uzayında sınıfların ve en yakın komşularının, k değerine göre belirlendiği bir sınıflandırma yöntemidir [36]. Algoritmanın başarısını etkileyen en önemli faktör k değeridir. Parametre olarak k değeri büyük seçildiğinde farklı sınıflara ait örnekler aynı sınıfa gösterilebilir veya k değeri küçük seçildiğinde aynı sınıfa ait örnekler farklı sınıflara ayrılabilir [37].

Şekil 3.6'da üçgen ile gösterilen örnek veriye ait sınıf bilgisinin tespit edilmesi gösterilmiştir. k, 3 alındığında örnek veri yıldız sınıfa atanırken 5 alındığında ise yuvarlak sınıfa atanmaktadır.



Şekil 3.6: 3 sınıflı veri setinde k-EYK algoritması.

3.3.5 Naive bayes

Naive Bayes (NB), bayes teoremine dayanan olasılıksal sınıflandırma yöntemidir. Mevcut sınıflandırılmış örneklerde yer alan verilerden yola çıkarak, sınıflandırılmamış örneğe ait verileri kullanıp mevcut sınıflardan hangisine ait olduğunu hesaplayan bir yaklaşıma sahiptir. Örneklerin önem derecesi sıralaması yoktur. Naive Bayes, modeli öğrenilmesi sayesinde, her çıktının öğrenme kümesinde kaç kere meydana geldiğinin hesaplanmasını ve hesaplanan bu değerlerin öncelikli olasılık olarak adlandırılmasını

sağlar [38]. Sınıflandırıcının en temel dezavantajı örnek veri setinin boyutu yüksek ise sonuçların doğruluk oranı azalabilmektedir.

Bayes teoremi;

$$P(X|Y) = \frac{P(Y|X) * P(X)}{\mathrm{P(Y)}} \quad (\mathbf{3.4})$$

P(X) =>X olayının olasılığı (bağımsız)

P(Y)=>Y olayının olasılığı (bağımsız)

P(X|Y)=>Y olayının olduğu durumda X olayının olma ihtimali

P(Y|X)=>X olayının olduğu durumda Yolayının olma ihtimali

İşlevi ile ifade edilir.

Eşitlik 3.1 temel alınarak P(X|Y)'yi maksimum durumlar hesaplanarak NB algoritması çalıştırılmış olur. P(Y) bütün sınıflar için sabit olduğundan;

$$P(X|Y) = P(Y|X)P(X) \quad (\mathbf{3.5})$$

olasılığı için en büyük değer atanır.

3.3.6 Bayes ağları

Bayes ağları (BA-Bayesian Network), bilgisayar bilimlerinde verilerin açıklanması ve veriler arası durum geçişlerini ifade etmek için kullanılan yöntemlerden birisidir. Literatürde Bayes ağları veya inanç ağı olarak bilinen ağların özelliği istatistiksel ağlar olmaları ve düğümler arası geçiş yapan kolların istatistiki yöntemlere göre seçilmesidir. Bayes ağları yönlü dönüşsüz ağlardır ve her düğüm ayrı bir değişkeni ifade eder [39].

3.3.7 k-Yıldız

k-Yıldız algoritması örnek tabanlı sınıflandırıcıdır. İki nokta arasındaki uzaklığın veya benzerliğin belirlenmesinde kullanılmaktadır. Entropik uzaklık ölçüsü kullanılır [40]. Entropi rastgele değişkenler ile ilgili belirsizliğin ölçüsü olarak tanımlanmaktadır. İşlem yapılan veri setinde belirlenen K* noktasına en yakın olan örnekler sınıflandırılır, sonrasında belirlenen K* noktasını kendine yakın verilerin merkezine taşıyarak sınıflandırma işlemini gerçekleştirmiş olur [41]. k-Yıldız algoritması fonksiyonu eşitlik 3.6'daki gibi ifade edilir [42].

$$K^*(b/a) = -log_2 P^*(b/a) \quad (3.6)$$

3.3.8 Adaboost

En çok kullanılan başarı artırıcı algoritmalarından olan Adaboost, Freund ve arkadaşları tarafından geliştirilmiştir [43]. Sınıflandırma işlemi için kullanılan topluluk öğrenme algoritmasıdır. Topluluk öğrenme başarısı tekil öğrenmedeki sınıflandırma başarısına göre daha yüksektir. Topluluk öğreniminde, bir önceki sınıflandırıcı tarafından yanlış sınıflandırılan örnekleri sonraki sınıflandırıcıların kullanılması ile başarı artırımı hedeflenir. Yanlış sınıflandırılan eğitim örneklerinin göreceli ağırlığı artırılarak sonraki sınıflandırıcının güncellenmiş ağırlıklar ile tekrardan tahmin işlemi gerçekleştirir. İkili sınıflandırma işleminde başarılı sonuçlar vermektedir.

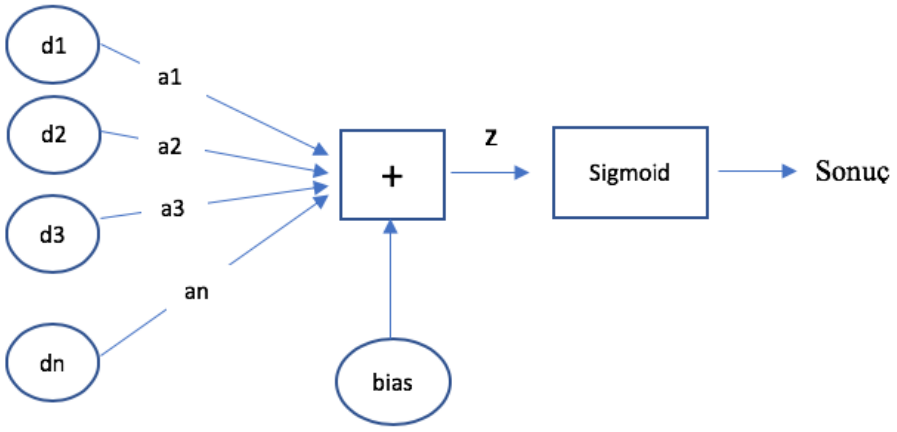
3.3.9 Bagging

Leo Breiman tarafından geliştirilen sınıflandırma ve regresyon düzenlerindeki tahminlerin doğruluğunu artırmak için önerilmiştir [44]. Bagging yönteminde var olan eğitim verisinden alt eğitim kümeleri oluşturulur. Sınıflandırma orjinal veri setinden rastgele seçilen alt küme bir sınıflandırma algoritması ile eğitilir. Alt kümelerdeki sınıflandırma işleminin ardından sonuçlar birleştirilir. Sonuç tahmini regresyon örneklerinde ortalama, sınıflandırma örneklerinde oylama yöntemi ile belirlenir. Yöntemin avantajı aşırı uyumu engeller.

3.3.10 Lojistik regresyon

Lojistik regresyon (LR-Logistic Regression) bağımlı değişkenler ile bağımsız değişkenler arasındaki sebep-sonuç ilişkisinin incelenmesinde önem taşımaktadır [45]. Bir veya daha fazla değişkeni, iki veya daha fazla bağımlı değişken arasındaki ilişkiyi modellemek için kullanılan istatistiki modelleme yaklaşımıdır [46]. Bağımlı değişkenin belirli bir aralıkta veya ortamdaki değerini tahmin eder [47]. Bağımlı değişken bir sınıf, kategori veya ölçülebilir bir sonuç olabilir. Lojistik regresyonda modeli eğitirken sınıflandırma yapılacak örüntüye ait özniteliklerle, belirlenen ağırlıkların çarpımının toplamına bias değeri eklenir. Elde edilen değer aktivasyon fonksiyonu olarak belirlenen sigmoid fonksiyonundan geçirilir ve fonksiyon sonucu ile beklenen çıktı değeri karşılaştırılır. Karşılaştırma sonucundaki oluşan hata değeri ağırlıklar ve biasa dağıtılır. Bu işlem belirlenen döngü sayısı kadar tekrarlanır. Test

işleminde sigmoid fonksiyonundan elde edilen çıktı 0.5 eşik değerinden büyükse sınıflandırma sonucu 1, küçükse 0 olur. Şekil 3.7'de lojistik regresyon uygulaması verilmektedir.

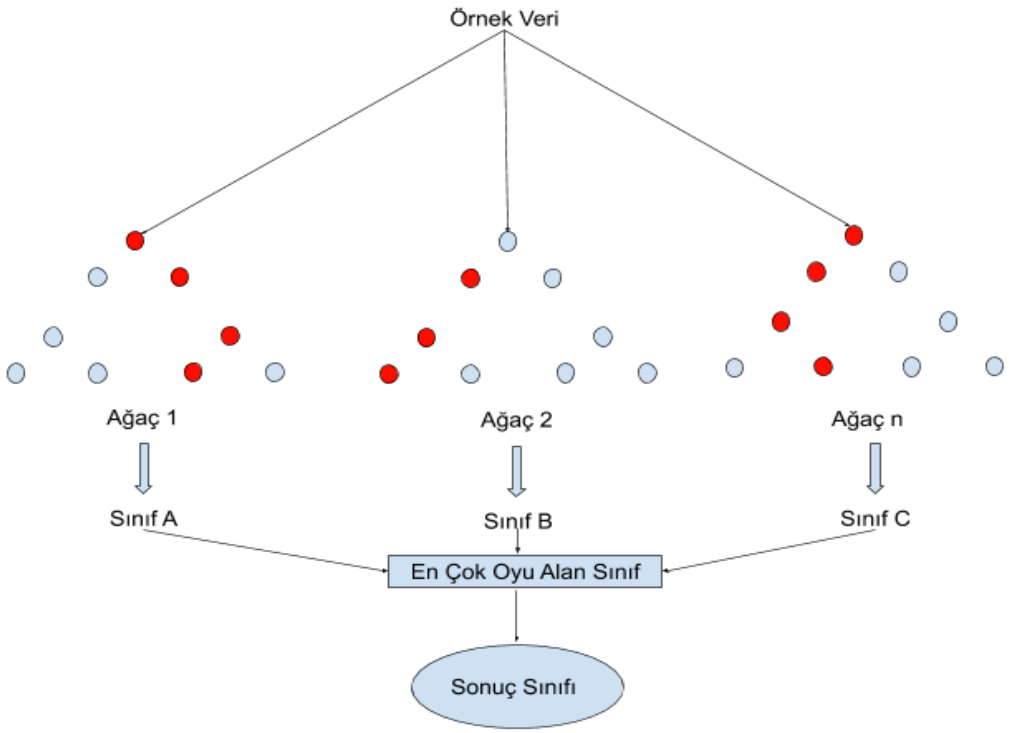


Şekil 3.7: Lojistik regresyon uygulaması.

3.3.11 Rastgele orman

Topluluk sınıflandırma yöntemleri birden fazla sınıflandırıcının ürettiği sonuçlar üzerinden örnek veriyi sınıflandırır. Rastgele orman (RO-Random Forest), topluluk sınıflandırma yöntemlerinden en yaygın olarak kullanılan sınıflandırıcı algoritmadır. İçerisinde çoklu ağaçlar barındırır. Örnek veriyi sınıflandırma için toplulukta bulunan her ağaca giriş vektörü verilir ve her ağaç bir çıktı üretir. RO algoritması sonuç olarak verilen kararlardan en çok oyu alan sınıfı seçer [33].

Rastgele orman, Breiman'ın geliştirdiği bu algoritmanın amacı, algoritma içinde bir adet karar ağacı bulundurmak yerine her biri ayrı eğitim kümelerinde eğitilmiş olan farklı sayıda ve birbirinden farklı karar ağacının ürettikleri sonucu birleştirmektir [44]. Farklı eğitim kümeleri oluştururken ön yükleme ve rastgele özellik seçimi kullanılır. Çok değişkenli karar ağaçları oluşturulurken sınıflandırma ve regresyon ağaçları [48] algoritması kullanılır. Her düzeydeki öznitelik belirlenirken öncelikle bütün ağaçlarda hesaplamalar sonucu nitelikler belirlenir, ardından algoritma içinde ağaçlardaki nitelikler birleştirilerek en çok kullanılan öznitelik seçilir. Seçilen nitelik ağaca dahil edilerek diğer seviyelerde aynı işlemler tekrarlanır. Şekil 3.8'de RO algoritması karar mekanizması ve çalışma şekli verilmektedir. Şekilde algoritma içinde yer alan birden fazla karar ağacı sınıflandırma işlemi yapmaktadır. Karar ağaçlarının yaptığı sınıflandırma sonucu oylanır ve en yüksek oyu alan sınıf sonuç olarak belirlenir.



Şekil 3.8: Rastgele orman algoritması.

3.4 Öznitelik Çıkarımı

Makine öğrenmesi ve alt alanlarında girdi olarak kullanılan örüntülerden anlamlı veriler üretmeye öznitelik çıkarımı denmektedir. Öznitelik çıkarımı bir örüntü tanıma modelinde veri önişleme adımında gerçekleştirilir. Örüntüye ait oluşturulan öznitelikler, mevcut çözüm uzayında bulunan örnekler üzerinde yüksek doğruluk ile sınıflandırma amacıyla kullanılır.

Özniteliklerin örüntüye ait bilgiyi doğru ve kesin ifade etmesi ile sınıflandırma başarısı arasında bağlantı bulunmaktadır. Bir örüntü kümesinin öznitelik çıkarımı yapılacaksa elde edilen öznitelikler bütününe öznitelik kümesi denir. Özniteliklerin içinde bulunduğu n boyutlu uzay ise öznitelik uzayı olarak isimlendirilir [49-50].

Peptidlere ait öznitelik vektörünün çıkarımında amino asitlerin dizilimleri kullanılabilir. Dizilim tabanlı öznitelik vektörleri sınıflandırıcı algoritmalar için örüntü tanımada önemli rol oynamaktadır. Peptidlere ait öznitelik vektörü oluşturmak için kullanılan bazı yöntemler şunlardır: Amino asit bileşimi [51], anlık kompozisyon vektörü [52] ve kalıntı çift modeli [53].

3.5 Başarım Analizi

Sınıflandırıcı Algoritmaların test edilecek veri üzerindeki başarımları doğruluk, duyarlılık, özgünlük F-ölçütü, Matthew Korelasyon Katsayısı (MCC) ve alıcı işletim karakteristiği (ROC) alanı ölçütleri ile değerlendirilmiştir. Başarım metrikleri test verisi sınıflandırıldıktan sonra çizelge 3.1'de görülen karmaşıklık matrisinden yararlanılarak hesaplanmaktadır.

Çizelge 3.1: Karmaşıklık matrisi.

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

DP; sınıflandırıcı tarafından pozitif tahmin edilen gerçek pozitifleri, DN; sınıflandırıcı tarafından negatif tahmin edilen gerçek negatifleri, YP; sınıflandırıcı tarafından negatif tahmin edilen gerçek pozitifleri, YN; sınıflandırıcı tarafından pozitif tahmin edilen gerçek negatifleri gösterir.

Doğruluk; kanser karşıtı peptidlerin ve kanser karşıtı olmayan peptidlerin ne kadarının sınıflandırıcı algoritma tarafından ayırt edilebildiğini gösterir.

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \quad (\mathbf{3.7})$$

Duyarlılık; kanser karşıtı peptidlerin ne kadarının sınıflandırıcı algoritma tarafından ayırt edilebildiğini gösterir.

$$Duyarlılık = \frac{DP}{DP + YP} \quad (\mathbf{3.8})$$

Özgünlük; kanser karşıtı olmayan peptidlerin ne kadarının sınıflandırıcı algoritma tarafından ayırt edilebildiğini gösterir.

$$Özgünlük = \frac{DN}{DN + YN} \quad (\mathbf{3.9})$$

MCC; ikili sınıflandırmada gözlenen ve öngörülen çıktılar arasında bir korelasyon katsayısıdır. İki sınıflı problemlerde yöntemin gücünü belirten ölçüttür. Sınıflardaki veri sayısı karışık olduğunda diğer performans metriklerine göre daha doğru sonuç verir [54]. MCC 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.

$$MCC = \frac{(DP * DN) - (YP * YN)}{\sqrt{(DP + YP) * (DP + YN) * (DN + YP) * (DN + YN)}} \qquad (\mathbf{3.10})$$

Kesinlik; sınıflandırıcı algoritmanın pozitif olarak tahmin ettiği pozitif örnek sayısının, pozitif tahmin ettiği örneklerin sayısının toplamına bölümüdür Kanser karşıtı peptid olarak sınıflandırılan örneklerden ne kadarının gerçekten kanser karşıtı peptid olduğunu gösterir.

$$Kesinlik = \frac{DP}{DP + YN} \qquad (\mathbf{3.11})$$

F-Ölçütü; kesinlik ve duyarlılık ölçütleri sonuçları karşılaştırmada yeterli değildir. Duyarlılık ve kesinlik değerlerinin harmonik ortalaması hesaplanarak iki metrik arasındaki denge sağlanır. Özgüllük ve duyarlılık değerlerinden birinin yüksek, diğerinin düşük olması durumunda ikisinin harmonik ortalamasını veren F-Ölçütü değerine bakarak sınıflandırmayı daha sağlıklı yorumlamak mümkündür

$$F - \text{Ölçütü} = \frac{2.Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \qquad (\mathbf{3.12})$$

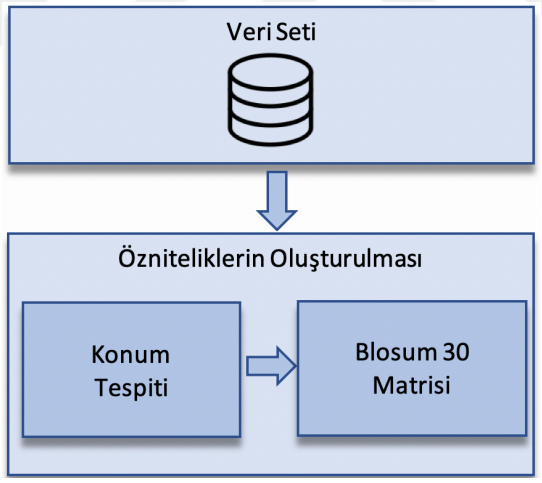
ROC alanı; gerçek pozitif oran, bir parametrenin farklı kesme noktaları için yanlış pozitif oranı işlevinde çizilir. ROC eğrisindeki her nokta belirli bir karar eşiğine karşılık gelir. ROC eğrisi altındaki alan, iki sınıfın ne kadar iyi ayırt edilebildiğini gösterir. ROC eğrisinde x ekseni gerçek pozitif oranını, y ekseni yanlış pozitif oranını göstermektedir [33].

4. GELİŞTİRİLEN ÖZNİTELİK KODLAMA YÖNTEMLERİ

Konum ve Taylor Venn Diyagramı temel alınarak iki adet öznitelik kodlama yöntemi geliştirilmiştir. İlk yöntemde peptid dizilimlerinde yer alan 20 amino asidin konumu ve peptid uzunluğu hesaplanmıştır. İkinci yöntemde 20 amino asidin Taylor Venn Diyagramında bulunan fizikokimyasal özelliklerine dayanarak hesaplama yapılmıştır.

4.1 AKBlo Yöntemi

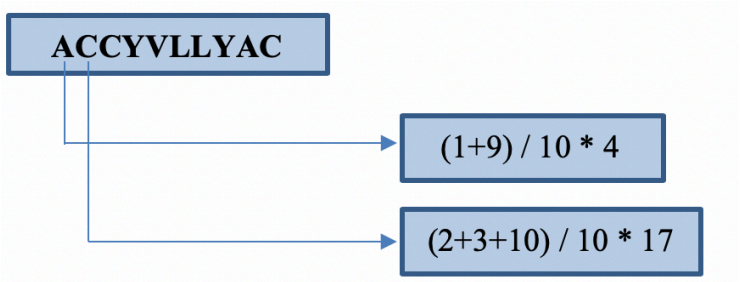
Dizilim tabanlı yöntemler peptidlere ait özelliklerin ortaya çıkmasında önemli rol oynamaktadır. AKBlo yöntemi, amino asitlerin peptid içerisinde nerede yer aldıkları bilgisini kullanmaktadır. Şekil 4.1'de konum tabanlı öznitelik kodlama yöntemi AKBlo verilmektedir.



Şekil 4.1: AKBlo öznitelik kodlama yöntemi.

İlk aşamada 20 amino asit, peptid dizilimlerinde yer aldığı konum bilgisi toplamı dizilim uzunluğuna bölünmektedir. Daha sonra, Blosum 30 yer değiştirme matrisinde amino asidin karşısında bulunan puan ile çarpılmaktadır. Her bir amino asit değeri bulunduktan sonra 1×20 boyutunda özellik vektörü oluşturulup, amino asidin karşısına daha önce hesaplanan konum değerlerinin peptid uzunluğuna bölümünün sayısal değeri yazılmaktadır. Veri setinde yer alan her peptid dizilimi için 1×20 boyutunda aynı vektör hesaplaması yapılmaktadır. Hesaplama işlemi sonucunda

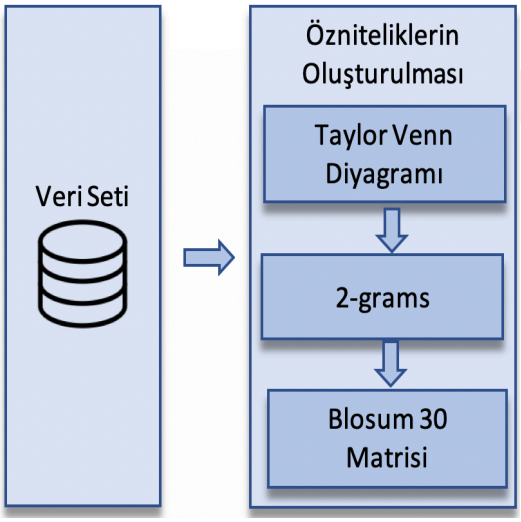
344×20 boyutunda öznitelik vektörü elde edilmektedir. Şekil 4.2'de AKBlo yöntemine göre ACCYVLLYAC dizilimine ait özellik çıkarımı görülmektedir.



Şekil 4.2: AKBlo yöntemi ile özellik çıkarımı

4.2 ngTBlo Yöntemi

ngTBlo yönteminde, 20 amino asit temel alınarak amino asitlerin ikili kombinasyonu sonucunda her peptid için 1×400 boyutunda özellik vektörü elde edilir. Şekil 4.3'te konum tabanlı öznitelik kodlama yöntemi görülmektedir.

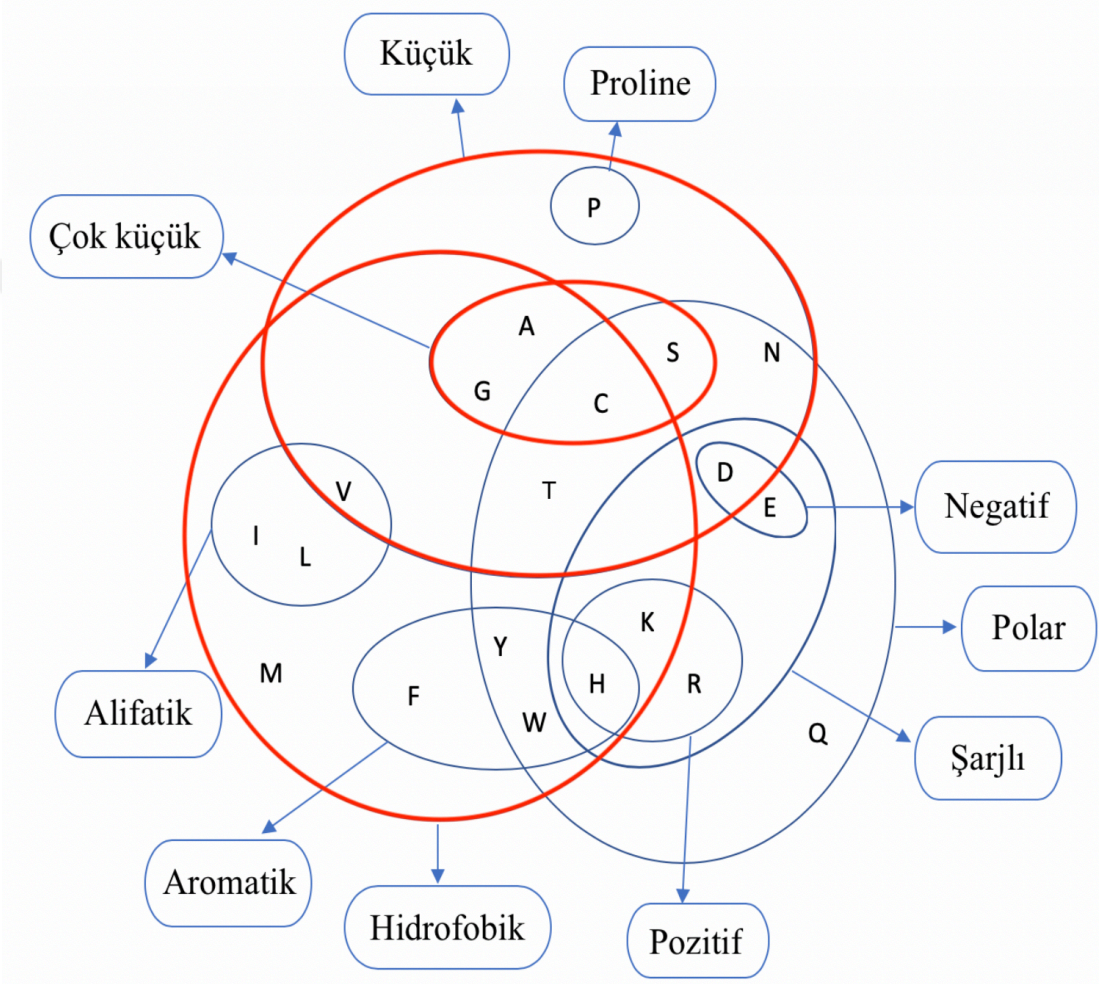


Şekil 4.3: ngTBlo öznitelik kodlama yöntemi.

Peptid dizilimlerinde bulunan amino asitler fizikokimyasal özelliklerine göre sınıflandırmaktadır. Diyagramda bir amino asit birden fazla özelliği barındırabilir. Geliştirilen yöntemde amino asitlerin sahip olduğu fizikokimyasal özellikler kullanılmıştır. Şekil 2.1'de Taylor Venn Diyagramı yer almaktadır.

Şekil 4.4'te gösterilen diyagramda A ve C amino asitlerinin 1×400 boyutuna sahip özellik vektöründeki değeri, sahip oldukları fizikokimyasal özelliklerin kesişim sayısı ile hesaplanmaktadır. A amino asidi; çok küçük, hidrofobik, küçük özelliklerine sahipken C amino asidi polar, çok küçük, küçük, hidrofobik özelliklerine sahiptir. Her

iki amino asidin kesişim kümesi incelendiğinde çok küçük, küçük, hidrofobik özelliklerinin ortak olduğu görülmektedir. İşlem yapılan peptid dizilimin özellik vektöründeki AC sütunun değeri 3 ile ifade edilmektedir. Peptid dizilimde yer alan bütün amino asit ikililerine işlem uygulanmaktadır. Şekil 4.4'te Taylor Venn Diyagramına göre A ve C amino asitlerin kesişim değerlerinin belirlenmesi verilmektedir.



Şekil 4.4: Özellik vektöründe amino asit kesişim değerleri.

İlk aşamada elde edilen 1×400'lük vektör, peptid diziliminde yer alan amino asit ikililerinin frekansı ile güncellenmektedir. ACCYVLLYAC diziliminde AC amino asit ikilisi dizilimde iki defa tekrar etmektedir. Hesaplanan sıklık değeri ile vektörde yer alan AC özelliğindeki değer, çarpma işlemine tabi tutulmaktadır. Bir sonraki CC amino asit ikilisi, incelendiğinde sıklık değeri 1 olarak hesaplanmaktadır. Özellik vektöründe bulunan CC özelliği değeri ile sıklık değeri çarpılır. Bu yöntem ile tüm dizilimin amino asitleri tamamlanır.

Özellik vektörünün oluşturulmasındaki son adım, vektörde yer alan ağırlıkların Blosum 30 matrisi ile çarpılmasıdır. Peptid diziliminde yer alan amino asit ikilileri Blosum 30 matrisinde ikiliye karşılık gelen değer ile çarpılmaktadır. Dizilimde yer alan bütün amino asitler için işlem tekrarlanır. 3 aşamalı özellik vektörü oluşumunda işlemler veri setinde yer alan tüm peptidler için gerçekleştirildikten sonra 344×400 boyutunda özellik vektörü elde edilmiştir. Şekil 4.5'te ACCYVLLYAC dizilimine ait ngTBlo yöntemine göre hesaplama yapılmaktadır.

AC	CC	CY	YV	VL	LL	LY	YA
2×3×-3	1×0×17	1×2×-6	1×1×1	1×2×1	1×0×4	1×1×3	1×1×-4

Şekil 4.5: ngTBlo yöntemi ile özellik çıkarımı.

Sınıflandırma başarısının artması için boyut daraltma yöntemi olan Fisher doğrusal ayırtaç analizi özellik vektörüne uygulanmıştır.

4.3 Bulgular ve Analiz

Kanser karşıtı peptidlerin tahmin tespiti probleminde, testler 10 kat çapraz doğrulama gerçekleştirilmiştir. 10 kat çapraz doğrulamada veri seti 10 eşit bölüme ayrılır. Bölümlerden 9 tanesi eğitim 1 tanesi test için kullanılıp sonuçları saklanır. İlk döngüde 1. bölüm test için alındığında kalan 9 bölüm eğitim için kullanılır ve sonuçlar saklanır. Sonraki döngüde 2. bölüm test, diğer 9 bölüm eğitim için kullanılır. Bu işlem her bölümün test için kullanılması işlemi ile sona erer. Nihai sonuçlar için her döngüde elde edilen sonuçların ortalaması alınır. Bu yöntem ile örnek uzayın tüm elemanları eğitim ve test işlemlerinde kullanılır.

Kanser karşıtı peptidleri belirlemek için oluşturulan modelin başarısının rastlantı sonucu olmadığını kanıtlayabilmek için Chi-square [55] hesaplama yöntemi kullanılarak gerçek değerler ve tahmin edilen sonuçlara dair p-value hesaplanmıştır. Elde edilen 0,001 değeri, eşik değeri 0,05 in altındadır.

Öznitelik kodlama yöntemleri python programlama dilinde, sınıflandırma testleri macOS Mojave işletim sistemine sahip 3,2 GHz i5 işlemcili bilgisayarda bulunan Weka [56] yazılımı ortamında gerçekleştirilmiştir. Çalışmada kullanılan sınıflandırıcı algoritmaların parameter değerleri Çizelge 4.1'de görülmektedir.

Çizelge 4.1: Sınıflandırma algoritmalarının parametre değerleri.

Sınıflandırma Algoritması	Parametre
İDBA	Yumuşatma miktarı = 0.2
k-Yıldız	Genel karışım=20, kayıp modu=’Ortalama Sütun Entropi Eğrileri’
ÇKA	Öğrenme katsayısı=0.3, momentum=0.2, eğitim süresi=500
DDVM	Maliyet=1.0, kayıp=0.1, iptal kriter toleransı=0.001, nu=0.5, derece=3
RTFDVM	Maliyet=1.0, kayıp=0.1, iptal kriter toleransı=0.001, nu=0.5, derece=3
k-EYK	k=3, arama algoritması= ‘düz arama’, mesafe ölçümü= ‘öklid mesafesi’
RO	Eğitim set büyüklüğü yüzdesi=100, model çıktı sayısı=2, iterasyon sayısı=100
Adaboost	Sınıflandırıcı=’RO’, iterasyon sayısı=10, model çıktı sayısı=2, budama ağırlık eşiği=100
Bagging	Sınıflandırıcı=’RO’, iterasyon sayısı=10, model çıktı sayısı=2
LR	Yükselti=0.00000001, model çıktı sayısı=4, en fazla iterasyon=-1
NB	Çekirdek kestirimi kullanımı = ‘Hayır’
BA	Tahmin edici="Basit Tahmin Edici", arama algoritması=’K2’

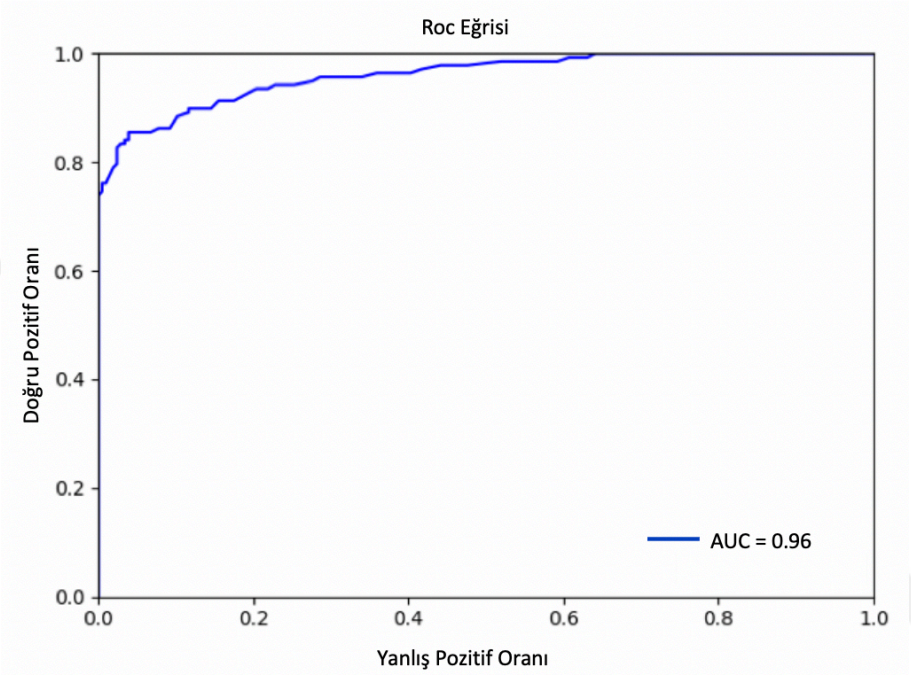
Veri seti üzerinde AKBlo yöntemi kullanılarak oluşturulan öznitelik vektörü İDBA ve ÇKA, RO, NB, BA, DDVM, RTFDVM, k-EYK, AD, BG, k-Yıldız ve LR sınıflandırıcı algoritmalar ile sınıflandırılmıştır. Sınıflandırma işleminde 10 kat çapraz doğrulama yöntemi kullanılarak başarım sonuçları elde edilmiştir. Elde edilen başarım sonuçları Çizelge 4.2'de görülmektedir.

Çizelge 4.2: Sınıflandırma algoritmalarının AKBlo yöntemine göre başarım metrikleri.

Sınıflandırma Algoritması	Doğruluk (%)	Duyarlılık	Özgünlük	Kesinlik	ROC Alanı	MCC	F-Ölçütü
İDBA	40,11	1,00	0,00	0,40	0,58	0,00	0,57
k-Yıldız	61,62	0,79	0,49	0,51	0,75	0,29	0,62
ÇKA	60,46	0,14	0,91	0,52	0,53	0,00	0,22
DDVM	49,41	0,66	0,37	0,41	052	0,00	0,51
RTFDVM	59,88	0,00	1,00	0,00	0,50	0,00	0,00
k-EYK	69,18	0,65	0,71	0,60	0,75	0,36	0,62
RO	90,69	0,81	0,96	0,94	0,95	0,80	0,87
Adaboost	91,57	0,84	0,96	0,94	0,96	0,82	0,88
Bagging	90,69	0,81	0,97	094	0,96	0,80	0,87
LR	74,70	0,46	0,93	0,83	0,78	0,47	0,59
NB	61,62	0,06	0,98	0,75	0,53	0,13	0,12
BA	86,04	0,80	0,89	0,84	0,91	0,70	0,82

AKBlo yöntemi ile kanser karşıtı peptidler, RO algoritması kullanarak %90,69 doğruluk, 0,81 duyarlılık, 0,96 özgünlük, 0,94 kesinlik, 0,95 ROC alanı, 0,80 MCC ve 0,87 F-Ölçütü değerleri elde edilmiştir. Adaboost algoritması kullanarak %91,57 doğruluk, 0,84 duyarlılık, 0,96 özgünlük, 0,94 kesinlik, 0,96 ROC alanı, 0,82 MCC ve 0,88 F-Ölçütü değerleri elde edilmiştir. Adaboost algoritması ile elde edilen sonuçlar DP=116, YN=22, DN=199 ve YP=7 karmaşıklık matrisi değerlerine sahiptir. Bagging algoritması kullanarak %90,69 doğruluk, 0,81 duyarlılık, 0,97 özgünlük, 0.94 kesinlik, 0,96 ROC alanı, 0,80 MCC ve 0,87 F-Ölçütü değerleri elde edilmiştir. BA algoritması

kullanarak %86,04 doğruluk, 0,80 duyarlılık, 0,89 özgünlük, 0,84 kesinlik, 0,91 ROC alanı, 0,70 MCC ve 0,82 F-Ölçütü değerleri elde edilmiştir. Sonuçlar incelendiğinde en yüksek ROC alanı değerini 0,96 ile Adaboost ve Bagging sınıflandırıcı algoritmaları elde etmiştir. En yüksek ROC alanı Şekil 4.6'da görülmektedir. Şekil 4.6 incelendiğinde kanser karşıtı ve kanser karşıtı olmayan peptidlerin yüksek başarı ile ayırt edildiği görülmektedir.



Şekil 4.6: Adaboost ve Bagging ROC eğrisi.

Veri seti üzerinde ngTBlo yöntemi kullanılarak oluşturulan öznitelik vektörü, fisher çoklu sınıf doğrusal ayırt edici analizi filtresi uygulanıp İDBA ve ÇKA, RO, NB, BA, DDVM, RTFDVM, k-EYK, Adaboost, Bagging, k-Yıldız ve LR sınıflandırıcı algoritmalar ile sınıflandırılmıştır. Sınıflandırma işleminde 10 katlı çapraz doğrulama yöntemi kullanılarak başarım sonuçları elde edilmiştir. Elde edilen başarım sonuçları Çizelge 4.3'te görülmektedir.

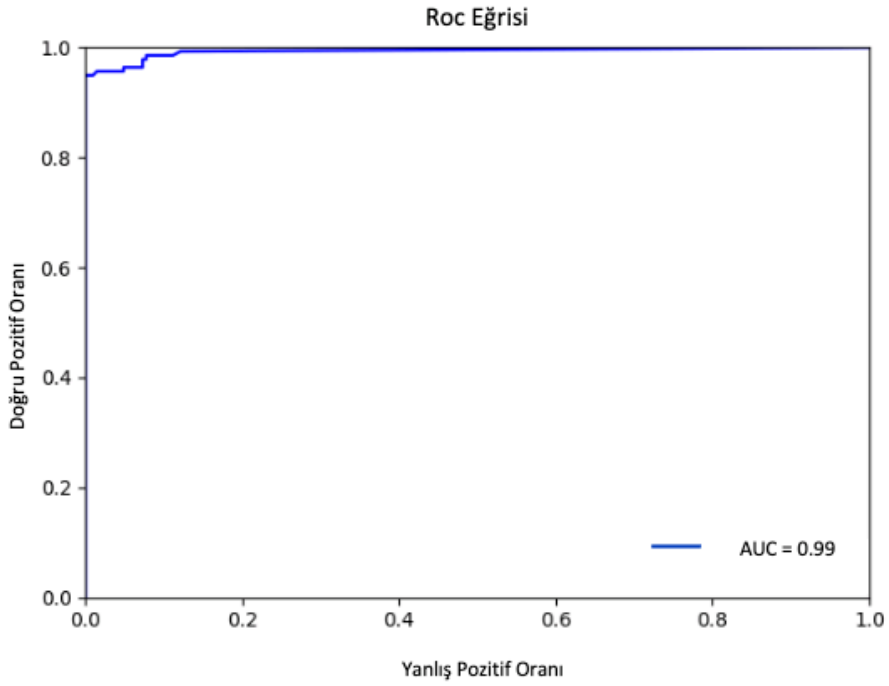
Çizelge 4.3: Sınıflandırma algoritmalarının ngTBlo yöntemine göre başarım metrikleri.

Sınıflandırma Algoritması	Doğruluk (%)	Duyarlılık (%)	Özgünlük	Kesinlik	ROC Alanı	MCC	F-Ölçütü
İDBA	97,09	0,95	0,98	0,97	0.99	0,93	0,96
k-Yıldız	97,38	0,95	0,98	0,97	0.99	0,94	0,96
ÇKA	97,09	0,95	0,98	0,97	0.99	0,93	0,96
DDVM	97,97	0,95	0,99	0,99	0.97	0,95	0,97
RTFDVM	97,09	0,95	0,98	0,97	0.97	0,93	0,96
k-EYK	97,38	0,94	0,99	0,99	0.98	0,94	0,96
RO	97,09	0,94	0,98	0,97	0.98	0,93	0,96
Adaboost	97,09	0,94	0,98	0,97	0.98	0,93	0,96
Bagging	97,38	0,95	0,98	0,97	0.99	0,94	0,96
LR	97,38	0,96	0,98	0,97	0.99	0,94	0,96
NB	97,38	0,96	0,98	0,97	0.99	0,94	0,96
BA	97,97	0,94	1,00	1,00	0.98	0,95	0,97

ngTBlo yöntemi ile kanser karşıtı peptidler, BA algoritması kullanarak %97,97 doğruluk, 0,94 duyarlılık, 1,00 özgünlük, 1,00 kesinlik, 0,99 ROC alanı, 0,95 MCC ve 0,77 F-Ölçütü değerleri elde edilmiştir. BA algoritması ile elde edilen sonuçlar DP=131, YN=7, DN=206 ve YP=0 karmaşıklık matrisi değerlerine sahiptir. DDVM algoritması kullanarak %97,97 doğruluk, 0,95 duyarlılık, 0,99 özgünlük, 0,99 kesinlik, 0,97 ROC alanı, 0,95 MCC ve 0,97 F-Ölçütü değerleri elde edilmiştir. DDVM algoritması ile elde edilen sonuçlar DP=132, YN=6, DN=205 ve YP=1 karmaşıklık matrisi değerlerine sahiptir.

Duyarlılık başarım metriğinde LR ve NB algoritmaları 0,96 değeri ile en yüksek puanı elde etmiştir. Özgünlük başarım metriğinde BA algoritması 1,00 değeri ile en yüksek puanı elde etmiştir. Kesinlik başarım metriğinde BA algoritması 1,00 değeri ile en yüksek puanı elde etmiştir. MCC başarım metriğinde BA ve DDVM algoritmaları 0,95 değeri ile en iyi sonucu elde etmiştir. F-Ölçütü başarım metriğinde BA ve DDVM algoritmaları 0,97 değeri ile en yüksek puanı elde etmiştir.

Sonuçlar incelendiğinde en yüksek ROC alanı değerini 0,99 ile NB, İDBA, k-Yıldız, ÇKA, LR ve Bagging sınıflandırıcı algoritmaları elde etmiştir. En yüksek ROC alanı Şekil 4.7'de görülmektedir. Şekil 4.6 incelendiğinde roc eğrisi altında kalan alan 1'e çok yakın olduğundan kanser karşıtı ve kanser karşıtı olmayan peptidlerin çok yüksek başarı ile ayırt edildiği görülmektedir.



Şekil 4.7: NB, İDBA, k-Yıldız, ÇKA, LR ve Bagging ROC eğrisi.

Veri seti üzerinde literatürde var olan n-grams öznitelik çıkarımı yöntemi kullanılarak elde edilen öznitelik vektörü İDBA ve ÇKA, RO, NB, BA, DDVM, RTFDVM, k-EYK, AD, BG, k-Yıldız ve LR sınıflandırıcı algoritmalar ile sınıflandırılmıştır. Sınıflandırma işleminde 10 katlı çapraz doğrulama yöntemi kullanılarak başarım sonuçları elde edilmiştir. Elde edilen başarım sonuçları Çizelge 4.4'te görülmektedir.

Çizelge 4.4: 2-grams başarım metrikleri.

Sınıflandırma Algoritması	Doğruluk (%)	Duyarlılık	Özgünlük	Kesinlik	ROC Alanı	MCC	F-Ölçütü
İDBA	79,06	0,73	0,83	0,74	0.80	0,56	0,74
k-Yıldız	77,03	0,86	0,71	0,66	0.86	0,56	0,75
ÇKA	80,00	0,75	0,83	0,75	0.87	0,58	0,75
DDVM	82,26	0,78	0,85	0,78	0.81	0,63	0,78
RTFDVM	75,58	0,39	1,00	1,00	0.69	0,53	0,56
k-EYK	77,03	0,86	0,71	0,66	0.86	0,56	0,75
RO	89,24	0,80	0,96	0,92	0.95	0,78	0,86
Adaboost	87,50	0,78	0,94	0,90	0.94	0,74	0,83
Bagging	86,62	0,78	0,96	0,93	0.94	0,77	0,85
LR	81,10	0,78	0,83	0,76	0.87	0,61	0,77
NB	84,88	0,78	0,90	0,84	0.89	0,68	0,80
BA	79,65	0,62	0,92	0,83	0.85	0,57	0,71

2-grams yöntemi ile kanser karşıtı peptidler, RO algoritması kullanılarak %89,24 doğruluk, 0,95 değeri ile ROC alanı, 0,78 değeri ile MCC ve 0,86 değeri ile F-Ölçütü başarım metriklerinde en iyi sonucu vermiştir. k-Yıldız ve k-EYK algoritmaları

kullanılarak 0,86 değeri ile duyarlılık, RTFDVM algoritması kullanılarak 1,00 değeri ile özgünlük ve kesinlik başarım metrikleri elde edilmiştir.

Veri seti üzerinde literatürde var olan anlık kompozisyon vektörü öznitelik çıkarımı yöntemi kullanılarak elde edilen öznitelik vektörü İDBA ve ÇKA, RO, NB, BA, DDVM, RTFDVM, k-EYK, AD, BG, k-Yıldız ve LR sınıflandırıcı algoritmalar ile sınıflandırılmıştır. Sınıflandırma işleminde 10 katlı çapraz doğrulama yöntemi kullanılarak başarım sonuçları elde edilmiştir. Elde edilen başarım sonuçları Çizelge 4.5'te görülmektedir.

Çizelge 4.5: Anlık kompozisyon vektörü başarım metrikleri.

Sınıflandırma Algoritması	Doğruluk (%)	Duyarlılık	Özgünlük	Kesinlik	ROC Alanı	MCC	F-Ölçütü
İDBA	85,17	0,81	0,88	0,82	0,89	0,69	0,81
k-Yıldız	86,92	0,86	0,88	0,83	0,94	0,73	0,84
ÇKA	90,70	0,84	0,95	0,92	0,80	0,81	0,88
DDVM	85,47	0,83	0,87	0,81	0,85	0,70	0,82
RTFDVM	91,86	0,85	0,97	0,94	0,90	0,83	0,89
k-EYK	88,66	0,82	0,93	0,89	0,91	0,76	0,85
RO	90,70	0,83	0,96	0,93	0,96	0,81	0,88
Adaboost	90,99	0,83	0,96	0,93	0,96	0,81	0,88
Bagging	91,86	0,83	0,98	0,97	0,96	0,83	0,89
LR	87,50	0,85	0,89	0,84	0,91	0,74	0,84
NB	87,79	0,84	0,90	0,85	0,91	0,75	0,85
BA	88,37	0,84	0,91	0,87	0,93	0,76	0,85

Anlık kompozisyon vektörü yöntemi ile kanser karşıtı peptidler, RTFDVM ve Bagging algoritmaları kullanılarak %91,86 değeri ile doğruluk, 0,83 değeri ile MCC ve 0,89 değeri ile F-Ölçütü başarım metriğinde en yüksek değer elde edilmiştir. k-Yıldız algoritması 0.86 değeri ile duyarlılık başarım metriğinde en yüksek değer elde edilmiştir. RTFDVM algoritması özgünlük başarım metriğinde 0,97 değeri ile en yüksek değer elde edilmiştir. Bagging algoritması kesinlik başarım metriğinde 0,97 ile en yüksek değer elde edilmiştir.

Geliştirilen öznitelik kodlama yöntemi ve literatürde var olan yöntemlerin aynı veri seti üzerinde elde ettiği sonuçlar Çizelge 4.6'da görülmektedir.

Çizelge 4.6: ngTBlo yöntemi ile diğer çalışmaların karşılaştırılması.

Model	Doğruluk	Duyarlılık	Özgünlük	MCC
ngTBlo	97,97	94,92	1,00	0.95
iACP [9]	95,06	89,86	98,54	0.89
SAP [12]	91,86	86,23	95,63	0.83

Gerçekleştirilen makine öğrenmesi modeli ngTBlo yöntemi ile BA sınıflandırıcı algoritması, İACP yöntemine göre doğruluk metriğinde %2,37, duyarlılık metriğinde %5,06, özgünlük metriğinde %1,46 ve MCC metriğinde 0,06 yaklaşık değerleri ile daha başarılı sonuçlar elde ettik. ngTBlo yöntemi ile BA sınıflandırıcı algoritması ile SAP yöntemine göre doğruluk metriğinde %6,11, duyarlılık metriğinde %8,69, özgünlük metriğinde %4,37 ve MCC metriğinde 0,12 ile daha başarılı sonuçlar elde edilmiştir. Gerçekleştirdiğimiz çalışmada yöntemimiz ile doğruluk, duyarlılık, özgünlük ve MCC değerlerinde karşılaştırdığımız metotlara göre daha iyi sonuçlar elde edilmiştir.

5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, kanser karşıtı peptitleri belirlemek için konum tabanlı AKBlo ve amino asitlerin fizikokimyasal özelliklerini kullanan Taylor Venn Diyagramına dayanan ngTBlo yöntemi isimli iki yeni öznitelik kodlama yöntemi geliştirilmiştir. Bu öznitelik kodlama yöntemleri geliştirilirken Blosum 30 yer değiştirme matrisinden faydalanılmıştır.

AKBlo yöntemi, amino asitlerin peptid içindeki konumları ve doğadaki yer değiştirme sıklıklarını temel alan bir yöntemdir.

ngTBlo yönteminde Taylor Venn Diyagramında bulunan amino asitlerin fizikokimyasal özelliklerinin kesişim değerleri skorlanıp ardından 2-grams yönteminden elde edilen değerlerle çarpılmıştır. Özellik vektöründe yer alan değerler, Blosum 30 yer değiştirme matrisinde karşılık gelen değerler ile çarpılıp özellik vektörünün son değeri elde edilmiştir.

AKBlo yöntemi ile Adaboost algoritmasını kullanarak%91,57 sınıf doğruluğu, 0,94 duyarlılık, 0,96 özgünlük, 0,81 kesinlik, 0,95 ROC alanı, 0,80 MCC, 0,87 F-Ölçütü performans değerleri elde edilmiştir.

ngTBlo yöntemi ile Bayes ağları algoritmasını kullanarak üzerinde %97,97 sınıf doğruluğu, 1,0 duyarlılık, 1,0 özgünlük, 0,94 kesinlik, 0,98 ROC alanı, 0,95 MCC, 0,97 F-Ölçütü performans değerleri elde edilmiştir. AKBlo yöntemine göre, NgTBlo yöntemi ile Taylor Venn diyagramı kanser karşıtı peptidlerin belirlenmesinde başarılı olduğu görülmektedir.

ngTBlo yöntemi ile geliştirilen öznitelik kodlama yönteminin kanser karşıtı peptidlerin belirlenmesi probleminde etkili bir yöntem ve model olduğu görülmüştür. Yöntem içerisinde kullanılan çeşitli sınıflandırıcılar birbirlerine yakın sonuçlar vermektedir. Ayrıca yöntem içerisinde kullanılan sınıflandırıcılar literatürde gerçekleştirilen çalışmalar ile karşılaştırıldığında kanser karşıtı peptidleri belirlemede yüksek performans göstermektedirler. Çeşitli sınıflandırıcılardan elde edilen yüksek

performans geliştirilen dizilim tabanlı öznitelik kodlama yönteminin başarılı olduğunu göstermektedir.

Geliştirilen öznitelik kodlama yöntemlerinin kullanılması için peptid dizilimlerinden öznitelik vektörü elde edilen çevrimiçi bir sunucu geliştirilmiştir. Çevrimiçi öznitelik kodlama yöntemlerine proses.yalova.edu.tr [57] adresinden erişilerek tekli, çoklu dizilim veya fasta dosya formatında peptid dizilimlerine ait öznitelik vektörleri elde edilebilir.

İleride yapılacak çalışmalarda, her iki yöntemin farklı peptit sınıflandırma problemlerinde birleştirilmiş sınıflandırıcılar ile kullanılması planlanmaktadır.

KAYNAKLAR

[1] **Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., & Chou, K. C.** (2017). iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget, 8(3), 4208.

[2] **Al-Benna, S., Shai, Y., Jacobsen, F., & Steinstraesser, L.** (2011). Oncolytic activities of host defense peptides. International journal of molecular sciences, 12(11), 8027-8051.

[3] **Kalyanaraman, B. J. S. S. E. S., Joseph, J., Kalivendi, S., Wang, S., Konorev, E., & Kotamraju, S.** (2002). Doxorubicin-induced apoptosis: implications in cardiotoxicity. Molecular and cellular biochemistry, 234(1), 119-124.

[4] **Maliepaard, M., Scheffer, G. L., Faneyte, I. F., van Gastelen, M. A., Pijnenborg, A. C., Schinkel, A. H., ... & Schellens, J. H.** (2001). Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues. Cancer research, 61(8), 3458-3464.

[5] **Mader, J. S., & Hoskin, D. W.** (2006). Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment. Expert opinion on investigational drugs, 15(8), 933-946. [6] Hoskin DW, Ramamoorthy A. Studies on anticancer activities of antimicrobial peptides. Biochimica et biophysica acta. 2008; 1778:357–375.

[7] **Gaspar, D., Veiga, A. S., & Castanho, M. A. (**2013). From antimicrobial to anticancer peptides. A review. Frontiers in microbiology, 4, 294.

[8] **Huang, Y., Feng, Q. I., Yan, Q., Hao, X., & Chen, Y.** (2015). Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs. Mini reviews in medicinal chemistry, 15(1), 73-81.

[9] **Chen, W., Ding, H., Feng, P., Lin, H., & Chou, K. C.** (2016). iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget, 7(13), 16895.

[10] **Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., & Raghava, G. P. S.** (2013). In silico models for designing and discovering novel anticancer peptides. Scientific reports, 3, 2984.

[11] **Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M., & Mohabatkar, H.** (2014). Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology, 341, 34-40.

[12] **Xu, L., Liang, G., Wang, L., & Liao, C.** (2018). A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. Genes, 9(3), 158.

[13] **Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., & Lee, G.** (2017). MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget, 8(44), 77121.

[14] **Gök, M.,** "HIV-1 Proteaz Enziminin Kesme Konumlarının Tespitinde Yeni Öznitelik Vektörleri", Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2011.

[15] **Henikoff, S., & Henikoff, J. G.** (1992). Amino acid substitution matrices from protein blocks. PROCeedings of the National Academy of Sciences, 89(22), 10915-10919.

[16] **Genç, S.,** "Proteinlerdeki düzensiz bölgelerin tespiti için kaotik ve fizikokimyasal özellikler tabanlı yeni öznitelik kodlama yöntemleri geliştirilmesi", Yalova Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2015.

[17] **Thundimadathil**, J. (2012). Cancer treatment using peptides: current therapies and future prospects. *Journal of amino acids*, *2012*.

[18] **Wang, Z., & Wang, G.** (2004). APD: the antimicrobial peptide database. Nucleic acids research, 32(suppl_1), D590-D592.

[19] **Babur, S.**, "Proteinlerin düzensiz bölgelerinin tahmininde yeni öznitelik kodlama yöntemleri", Yalova Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2016.

[20] **Murphy, K.** (2001). The bayes net toolbox for matlab. Computing science and statistics, 33(2), 1024-1034.

[21] **Wlodawer**, A., & Erickson, J. W. (1993). Structure-based inhibitors of HIV-1 protease. *Annual review of biochemistry*, *62*(1), 543-585.

[22] **Wu, C., Whitson, G., McLarty, J., Ermongkonchai, A., & Chang, T. C.** (1992). Protein classification artificial neural system. *Protein Science*, *1*(5), 667-677.

[23] **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.

[24] **Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K. R.** (1999, August). Fisher discriminant analysis with kernels. In Neural networks for signal pROCessing IX: PROCeedings of the 1999 IEEE signal pROCessing society workshop (cat. no. 98th8468) (pp. 41-48). Ieee.

[25] **Wernick, M. N., Yang, Y., Brankov, J. G., Yourganov, G., & Strother, S. C.** (2010). Machine learning in medical imaging. IEEE signal processing magazine, 27(4), 25-38.

[26] **Alpaydin, E.** (2009). Introduction to machine learning. MIT press.

[27] **Geurts, P.** (2001). Dual perturb and combine algorithm. In PROCeedings of AISTATS 2001, Eighth International Workshop on Artificial Intelligence and Statistics (pp. 196-201). Key-West, Florida.

[28] **Rodrigues, P. P., Gama, J., & Bosnic**, Z. (2008, December). Online reliability estimates for individual predictions in data streams. In *2008 IEEE International Conference on Data Mining Workshops* (pp. 36-45). IEEE.

[29] **Bal, Ç.**, "Yapay sinir ağlarında en iyi mimari seçimi için kullanılan kriterlerin incelenmesi", Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2018.

[30] **Arı, A., & BERBERLER, M. E.** Yapay Sinir Ağları ile Yaklaşım ve Sınıflandırma Problemlerinin Çözümü İçin Arayüz Tasarımı. Acta INFOLOGICA, 1(2), 55-73.

[31] **Cortes, C., & Vapnik, V.** (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[32] **Zhi-qiang, J., Hang-guang, F., & Ling-jun, L. I.** (2005). Support vector machine for mechanical faults classification. Journal of Zhejiang University-SCIENCE A, 6(5), 433-439.

[33] **Bilgin, M.** (2018). Makine Öğrenmesi: Teori ve Algoritmaları. İstanbul: Papatya Bilim

[34] **Akarsu, C.** (2016) Twitter Verileri ile Türk Televizyonları İzleneme Oranı Sıralamaları Tahmini, Tez, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü.

[35] **Chang, C. C., & Lin, C. J.** (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.

[36] **Cover, T., & Hart, P.** (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[37] **Mitchell, T.** (1997) Machine Learning. McGraw Hill. ISBN

[38] **Burhan, Y.,** "Elektronik Postaların Ayrıştırılmasında Naive Bayes ve Bulanık Mantık Yöntemlerinin Karşılaştırılması", Gazi Ünversitesi Bilişim Enstitüsü, Yüksek Lisans Tezi, 2011.

[39] **Şeker, Ş. E**. (2019) Bayes Ağları. http://bilgisayarkavramlari.sadievrenseker.com/2008/12/21/ bayes-aglari- bayesian-network/

[40] **Cleary, J. G., & Trigg, L. E.** (1995). K*: An instance-based learner using an entropic distance measure. In *Machine Learning PROCeedings 1995* (pp. 108-114). Morgan Kaufmann.

[41] **Nalçakan, Y., Bayramoğlu, Ş. S., & Tuna, S.** (2015). *Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu AnaliziÇalışması*. Technical Report.

[42] **Çölkesen, İ., & Kavzoğlu, T.** Örnek Tabanlı K-Star Algoritması İle Uzaktan Algılanmış Görüntülerin Sınıflandırılması.

[43] **Freund, Y., Schapire, R., & Abe, N.** (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

[44] **Breiman, L.** (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[45] **Agresti, A.** (1996). An Introduction to Categorical Data Analysis. 3rd Edition. Wiley.744.

[46] **Çolak, E.**, "Koşullu ve sınırlandırılmış lojistik regresyon yöntemlerinin karşılaştırılması ve bir uygulama", *Osmangazi Üniversitesi,* Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2002.

[47] **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.

[48] **Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.** (1984) Classification and Regression Trees.

[49] **Cevikalp, H.** (2005). Feature extraction techniques in high-dimensional spaces: Linear and nonlinear approaches.

[50] **Kuncheva, L. I.** (2004). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

[51] **Bhasin, M., & Raghava, G. P.** (2004). Classification of nuclear receptors based on amino acid composition and dipeptide composition. Journal of Biological Chemistry, 279(22), 23262-23266.

[52] **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.

[53] **Guo, J., Lin, Y., & Sun, Z.** (2005). A novel method for protein subcellular localization: Combining residue-couple model and SVM. In PROCeedings of the 3rd Asia-Pacific bioinformatics conference (pp. 117-129).

[54] **Çınar Efe, İ.**, "Yardımcı T Hücreleri / Büyük Doku Uygunluk Kompleksi Molekülleri Bağlanma Yerlerinin Tespitinde Öznitelik Kodlama Yöntemleri Geliştirilmesi", Yalova Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 2016.

[55] **Plackett, R. L.** (1983). Karl Pearson and the chi-squared test. International Statistical Review/Revue Internationale de Statistique, 59-72.

[56] **Witten**, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[57] **Kösesoy, İ., Gök, M., & Öz, C.** (2018). PROSES: A Web Server for Sequence-Based Protein Encoding. *Journal of Computational Biology*, *25*(10), 1120-1122.

ÖZGEÇMİŞ

Ad Soyad:	Murat ESER
Doğum Yeri ve Tarihi:	Düzce / 08.07.1990
Adres:	ÇÖMÜ Bilgi İşlem Daire Başkanlığı
E-Posta:	meser@comu.edu.tr
Lisans:	Yalova Üniversitesi, Bilgisayar Mühendisliği

Mesleki Deneyim ve Ödüller:

2013 - 2018	Yalova Üniversitesi Bilgi İşlem Daire Başkanlığı
2018 - Halen	ÇÖMÜ Bilgi İşlem Daire Başkanlığı

Yayın ve Patent Listesi:

- Eser, M., Gök, M., "Makine Öğrenmesi Sınıflandırıcı Algoritmalar ile Kanser Karşıtı Peptidlerin Tahmin Edilmesi". *International Conference on Data Science and Applications,* Yalova - Türkiye, 4-7 Ekim 2018.
- Eser, M., Bilgin, M., "Microbiyal Karşıtı Peptidlerin Belirlenmesinde Bir Makine Öğrenmesi Modeli Kullanılması". *International Conference on Data Science and Applications,* Yalova - Türkiye, 4-7 Ekim 2018.
- Eser, M., Gök, M., " Mikrobiyal Karşıtı Peptidlerin Sınıflandırıcı Algoritmalar ile Tahmin Edilmesi". *International Academic Research Congress,* Alanya - Türkiye, 17-21 Nisan 2019.
- Eser, M., Gök, M., "Antimikrobiyal Peptidlerin Makine Öğrenmesi Yöntemler ile Tahmin Edilmesi". *Uluslararası Mardin Artuklu Multidisipliner Çalışmalar Kongresi,* Mardin - Türkiye, 19-21 Nisan 2019