T.C.
YALOVA ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

SARS-COV-2 PROTEİNİ İLE İNSAN PROTEİNİ ARASINDAKİ ETKİLEŞİMLERİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNİ

YÜKSEK LİSANS TEZİ

FİRDES GÜL KORKUT

DANIŞMAN: PROF. DR. MURAT GÖK

YALOVA
ŞUBAT 2022

T.C.
YALOVA ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

SARS-COV-2 PROTEİNİ İLE İNSAN PROTEİNİ ARASINDAKİ ETKİLEŞİMLERİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNİ

YÜKSEK LİSANS TEZİ

FİRDES GÜL KORKUT

DANIŞMAN: PROF. DR. MURAT GÖK

YALOVA
ŞUBAT 2022

ÖNSÖZ

Yüksek lisans eğitimim boyunca benden yardımlarını esirgemeyen, akademik anlamda bana her zaman destek olan ve doğru yolu gösteren sayın danışman hocam Prof. Dr. Murat Gök'e minnet ve teşekkürlerimi sunarım.

Bu çalışma süresince yanımda oldukları ve benden desteklerini hiçbir zaman esirgemedikleri için aileme; annem, babam, kardeşlerime ve arkadaşım Derya Mutlu'ya şükranlarımı sunarım.

Şubat – 2022 Firdes Gül KORKUT

İÇİNDEKİLER

TÜRKÇE – İNGİLİZCE TERİMLER

Alıcı Operatör Karakteristiği	: Receiver Operator Characteristic
Amfifilik Sözde Amino Asit Bileşimi	: Amphiphilic Pseudo-Amino Acid Composition
Amino Asit Birleşimi	: Amino Acid Composition
Aşırı Gradyan Artırma	: Extreme Gradient Boosting
Blok Değiştirme Matrisi	: Blocks Substitution Matrix
Çok Katmanlı Algılayıcılar	: Multilayer Perceptron
Derin Sinir Ağları	: Deep Neural Network
Destek Vektör Makineleri	: Support Vector Machines
Gruplandırılmış Ağırlığa Göre Kodlama	: Encoding Based on Grouped Weight
İnsan Protein Referans Veri tabanı	: Human Protein Reference Database
Kabul Edilmiş Nokta Mutasyon	: Point Accepted Mutation
Kompozisyon - Geçiş - dağılım	: Composition, Transition, Distribution
Kompozisyon Moment Vektörü	: Composition Moment Vector
k-En Yakın Komşuluğu	: k-Nearest Neighborhood
Matthew Correlation Karakteristliği	: Matthew Operating Characteristic
ROC Eğrisi Altında Kalan Alan	: Area Under the ROC Curve
Sözde Amino Asit Bileşimi	: Pseudo amino acid composition
Sözde Kod	: Pseudo Code
Temel Bileşen Analizi	: Principal Component Analysis
Topluluk Algoritması	: Ensemble Algorithm

KISALTMALAR

AAC	: Amino Acid Composition
APAAC	: Amphiphilic Pseudo-Amino Acid Composition
AUC	: Area Under the ROC Curve
BLOSUM	: Blocks Substitution Matrix
CMV	: Composition Moment Vector
CTD	: Composition, Transition, Distribution
CTDC	: Composition
CTDT	: Transition
ÇKA	: Multilayer Perceptron
DNN	: Deep Neural Network
DVM	: Support Vector Machines
EBGW	: Encoding Based on Grouped Weight
HPRD	: Human Protein Reference Database
k-EYK	: k-Nearest Neighborhood
MCC	: Matthews Correlation Coefficient
PAM	: Point Accepted Mutation
PseAAC	: Pseudo Amino Acid Composition
ROC	: Receiver Operator Characteristic
RVKDE	: Relaxed Variable Kernel Density Estimator
SKYM	: Sıklık Konum Yer değiştirme Matrisi
TBA	: Principal Component Analysis
XGBoost	: Extreme Gradient Boosting

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

SARS-COV-2 PROTEİNİ İLE İNSAN PROTEİNİ ARASINDAKİ ETKİLEŞİMLERİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNİ

ÖZET

Dünya genelinde insan sağlığını tehdit eden Kovid-19, 2020 yılında Dünya Sağlık Örgütü (WHO) tarafından salgın hastalık olarak ilan edilmiştir. Kovid-19 hastalığı yayılma hızı ve ölümcül etkisinden dolayı insanlık üzerinde yıkıcı bir etkiye sebep olmuştur. Kovid-19 hastalığı SARS-CoV-2 virüsünün neden olduğu yeni tip korona virüstür. Bulaşıcı hastalıklara karşı çalışmalarda patojen-konakçı etkileşimlerinin belirlenmesi, hastalığa karşı önlem alınmasında büyük önem taşır. Patojen proteinleri, konakçıyı istila etmek için konakçı proteinleri ile etkileşime girer. Konak proteinlerine normal fonksiyonlarını engeller hatta yanlış yönlendirerek ve zayıflatarak bir konağın bağışıklık sistemini tehlikeye atabilirler. Bunun için Patojen-konak arasında varsayılan protein-protein etkileşiminin tahmin edilmesi büyük önem taşımaktadır. Protein-protein etkileşimlerinin belirlenmesi ile virüs proteinlerinin nasıl çalıştığını, nasıl çoğaldıklarını ve hastalığa nasıl neden olduğunun bulunmasında yardımcı olur. Biz de bu çalışmada SARS-CoV-2 virüs proteinlerinin konakçı(insan) proteinler ile etkileşimleri makine öğrenmesi yöntemleri ile tahmin ettik. Çalışmamızda 30.046 negatif ve 20.365 pozitif veriden oluşan bir protein etkileşimli veri seti kullandık. Protein etkileşimleri tahmin etmek için protein dizilimindeki amino asitleri, protein öznitelik kodlama yöntemiyle kodladık. Sınıflandırma algoritmaları olarak Naif Bayes, Bayes Net, k-En Yakın Komşuluğu, Doğrusal Destek Vektör Makineleri, Radyal Destek Vektör Makineleri ve Çok Katmanlı Algılayıcı kullanılmıştır. Sınıflandırma algoritmalarına, öncelikle literatürde yer alan protein öznitelik kodlama yöntemleri ile elde edilen veriler verilmiştir. Daha sonra amino asitlerin sıklık, konum ve kimyasal özellikleri bilgisinin yer aldığı yeni bir yöntem geliştirildi. Geliştirilen yöntem ile literatürdeki yöntemlerin sınıflandırma sonuçları karşılaştırıldığında daha dengeli sonuçlar sağlamıştır. Elde ettiğimiz deneysel sonuçlara göre geliştirdiğimiz yöntem Naif Bayes ve Bayes Ağları sınıflandırma algoritmaları ile 0,513 doğruluk, 0,529 duyarlık, 0,471 özgünlük, 0.514 kesinlik, 0,514 F1-skor, 0,027 Kappa değeri vermiştir.

Anahtar Kelimeler: Kovid-19, SARS-CoV-2, Protein-Protein Etkileşimi, Protein çıkarım yöntemleri, Makine Öğrenmesi, Sınıflandırma Algoritmaları

PREDICTION OF INTERACTIONS BETWEEN SARS-COV-2 PROTEIN AND HUMAN PROTEIN USING MACHINE LEARNING METHODS

ABSTRACT

Covid-19, which threatens human health worldwide, was declared an epidemic disease by the World Health Organization (WHO) in 2020. Covid-19 disease has had a devastating effect on humanity due to its spread rate and deadly effect. Covid-19 disease is a new type of corona virus caused by the SARS-CoV-2 virus. In this study, we predicted the interactions of SARS-CoV-2 virus proteins with host (human) proteins using machine learning methods. Determination of pathogen-host interactions in studies against infectious diseases is of great importance in taking precautions against the disease. Pathogen proteins interact with host proteins to invade the host. They can interfere with the normal function of host proteins or even compromise a host's immune system by misdirecting and weakening it. For this, it is of great importance to predict the putative protein-protein interaction between the pathogen-host. Identifying protein-protein interactions helps to find out how virus proteins work, how they replicate, and how they cause disease. In our study, we used a protein interaction dataset consisting of 30,046 negative and 20,365 positive data. To predict protein interactions, we encoded amino acids in the protein sequence using the protein feature coding method. Naive Bayes, Bayes Net, k-Nearest Neighborhood, Linear Support Vector Machines, Radial Support Vector Machines and Multilayer Perceptron were used as classification algorithms. Firstly, the data obtained by protein feature coding methods in the literature are given to the classification algorithms. Then, a new method was developed that includes the frequency, location and chemical properties of amino acids. When the classification results of the developed method and the methods in the literature were compared, more balanced results were obtained. The method we developed according to the experimental results we obtained gave 0.513 accuracy, 0.529 sensitivity, 0.471 specificity, 0.514 precision, 0.514 F1-score, 0.027 Kappa values with Naive Bayes and Bayes Networks classification algorithms.

Keywords: Covid-19, SARS-CoV-2, Protein-Protein Interaction, Protein extraction methods, Machine Learning, Classification Algorithms

1. GİRİŞ

2019 yılında Çin'in Wuhan bölgesinde ortaya çıkan Kovid19, SARS-CoV-2 virüsünün neden olduğu yeni tip korona virüs hastalığıdır. 11 Mart 2020 tarihinde Dünya Sağlık Örgütü tarafından, Kovid-19 hastalığı salgın olarak ilan edilmiştir. Hastalık dünya çapında sağlık sistemlerinde ciddi bir baskı ve gerilim yaratmıştır. İlk vaka bildiriminin ardından Asya'dan, Avrupa ve Amerika'ya kadar ölümcül etkisi tüm dünyayı tehdit altına almıştır [1]. Our World In Data kaynağına göre, Kovid-19 çıktığı günden itibaren dünya genelinde 5 milyondan fazla insanın yaşamını kaybetmesine neden oldu [2]. SARS-CoV, MERS-CoV ve diğer birçok korona virüs gibi, SARS-CoV-2 de muhtemel olarak yarasalardan kaynaklandığını düşülmektedir, ancak SARS-CoV-2 zatürree enfeksiyonunun doğrudan yarasalardan mı yoksa bir ara konakçı yoluyla mı bulaştığının doğrulanması gerekmektedir.

Birçok virüs için ortaya çıkma sürecindeki kilit adımlardan biri hayvanlardan insanlara geçiştir. Bu nedenle virüs kaynağının belirlenmesi, yayılmasının kontrol altına alınmasına yardımcı olacaktır [3]. Hastaların erken izolasyonu, yakın temaslıların izlenmesi, hastaların karantinaya alınmasıyla bulaşmayı önlemek ve yayılmasını durdurmak için Kovid-19'un erken teşhisi önem arz etmektedir [4]. Kovid-19 hastalığı damlacık yoluyla bulaşır. Hastalığın bulaşıcılık süresi net olarak bilinmemektedir. Bulaşıcılık süresi, semptomlar başlamadan birkaç gün önce başlayıp semptomların yok olmasıyla bittiği düşünülmektedir [4].

Kovid-19 hastalarında hastalığın ilerlemesinin doğru bir şekilde gözlemlenmesi, hastalık yönetiminin kritik bir bileşenidir. Kovid-19'un en yaygın belirtileri, kuru öksürük, boğaz ağrısı, ateş, nefes darlığı, ishal, kas ağrısı, nefes darlığıdır. Bu belirtiler, septik şok, pulmoner ödem, akut solunum sıkıntısı, çoklu organ yetmezliği, zatürre gibi kritik komplikasyonlarla birlikte ciddi bir şekilde ilerleyebilir [4].

SARS-CoV-2 virüsüne karşı savaşmak için bu virüsünün genetik özellikleri iyi anlaşılmalıdır. SARS-CoV-2 virüsü 65 ile 125 nm arasında değişen parçacık boyutuna sahip yaklaşık 27-32 kb'den oluşan tek sarmallı bir RNA virüsüdür [5]. SARS-CoV-2, birçok yardımcı proteine ek olarak Zarf (envelope; E), Diken (spike; S), Membrane (M), Nükleokapsit (N) proteini olmak üzere dört ana yapısal protein içerir [6]. Korona virüsün dış yüzeyinde bulunan glikoprotein-spike (s) proteini ile konakçı hücreye

bağlanır ve korona virüs hücre içine girer. Akciğer epitel hücreleri virüs için birincil hedeftir [7]. Bu virüs proteinlerinin hayatta kalma ve üreme için konakçı hücrelerle nasıl etkileşime girdiğinin anlaşılması, bu hastalığa karşı çözüm bulunması için esastır. Bulaşıcı hastalıklara karşı çalışmalarda patojen-konakçı etkileşimlerinin belirlenmesi büyük önem taşır [8]. Patojen-konak etkileşimi (PKE) terimi, bir patojenin (virüs, bakteri, prion, mantar ve viroid) ev sahibi ile etkileşime girme yollarını ifade eder [9]. Patojenler hayatta kalmak ve bir konağa bulaşmak için alternatif yollar ararlar. Patojen proteinleri, konakçıyı istila etmek için konakçı proteinleri ile etkileşime girer. Bir patojenin proteinleri, konakçıya çeşitli şekillerde bağlanabilir. Konak proteinlerine doğrudan bağlanabilir ve konağın normal fonksiyonlarını engeller hatta yanlış yönlendirerek ve zayıflatarak bir konağın bağışıklık sistemini tehlikeye atabilirler. Bağışıklığı bastırılmış bir konağın ortam bileşenlerini bile kullanabilir [9].

PKE'nin belirlenmesi için patojen proteini ile konak proteini etkileşimi (Protein–protein etkileşimi, PPE) incelenir. Patojen-konak arasında varsayılan PPE tahmin edilmesi büyük önem taşımaktadır [9]. Virüs ve konakçı proteinler arasındaki PPE'lerin belirlenmesi, virüs proteinlerinin nasıl çalıştığını, nasıl çoğaldıklarını ve hastalığa nasıl neden olduklarını açıklamaya yardımcı olur. Geleneksel laboratuvar teknikleri zaman alıcı ve pahalıdır, bu da bir patojen ile konakçısı arasındaki olası tüm protein etkileşimi kombinasyonlarını değerlendirmeyi neredeyse imkansız hale getirir [10]. Bu nedenle, PKE'lerin tahmin etmek için hesaplama yaklaşımları kullanılır. Örneğin, eğer sadece 1000 konakçı proteinin 500 patojen protein ile olası etkileşimlerini bulmak istiyorsak, olası patojen-konak kombinasyonları bir milyon olarak çıkar. Bu nedenle, deneysel olarak doğrulanmış PKE verilerinin eksikliği vardır. Hesaplamalı çalışmalar, mevcut PPE verilerini artırmak ve nihayetinde ilaç tasarım araştırmalarının hızına katkıda bulunmak için hayati öneme sahiptir [10]. Biz bu tez çalışmamızda SARS-CoV-2 virüs proteinleri ile insan proteinlerinin etkileşimlerini hesaplama yaklaşımlar ile inceledik.

1.1. Tezin Amacı

Biz bu tez çalışmasında, biyolojik deneyler kullanılarak doğrulanmış yeni tip korona virüs - insan proteinleri arasındaki protein - protein etkileşimlerini tahmin edip potansiyel hedef insan proteinlerinin tanımlanmasını sağlamayı amaçladık. Makine öğrenmesi yöntemleri ile korona virüs – insan protein etkileşimleri daha iyi tahmin edilebilir.

1.2. Literatür Taraması

Makine öğrenmesi yöntemlerine dayalı patojen – konak etkileşimleri literatürde çoğunlukla protein dizilimi tabanlı gerçekleştirilmektedir. Bunun için protein dizilimleri protein çıkarım yöntemleri ile sayısallaştırılır ve elde edilen öznitelik vektörleri sınıflandırıcı algoritmaya verilir. Böylece patojen - konak etkileşimleri tahmini gerçekleştirilmektedir. Bulaşıcı hastalıklara karşı çalışmalarda patojen-konakçı etkileşimlerinin belirlenmesi büyük önem taşır [8].

Literatürde, Amino Asit Bileşimi (AAC) [11], N-Gram [12], Kompozisyon - Geçiş - Dağılım **(**CTD) [13], Kompozisyon moment vektör (CMV) [14] gibi protein çıkarım yöntemleri sıklıkla kullanılmaktadır.

Chen ve diğerleri [15], homolog olmayan protein içeren bir veri kümesi üzerinde AAC ve Sözde Amino Asit Bileşimi (PseAAC) yöntemlerini kullanılarak farklı sınıflandırma yöntemleri ile başarım metriklerini karşılaştırmışlardır. Başarılı sonuçlara ulaşmışlardır.

Aghajanbaglo ve diğerleri [16], proteinlerin dizi bilgilerini özellik vektörlerine dönüştürmek için N-Gram kodlama yaklaşımını ve etkileşimleri tahmin etmek için makine öğrenme aracı olarak Rahat Değişken Çekirdek Yoğunluğu Tahmincisi (RVKDE) kullandılar. Sonuç olarak, N-Gram kodlama yöntemleri arasında 2-Gram'ın üstün performansa sahip olduğunu ve tahmin sonuçlarını iyileştirdiğini belirlediler.

Barman ve diğerleri [17], bulaşıcı hastalıkla ilişkili konakçı genleri tanımlamak için AAC, Dipeptit Kompozisyon, PseAAC ve CTD yöntemlerini kullandılar. Makine öğrenimi tekniklerine dayalı bir sınıflandırma yaklaşımı olan Derin Sinir Ağları (DNN), Destek Vektör Makineleri (DVM), Naif Bayes ve Rastgele Orman gibi iyi

bilinen sınıflandırıcıları yöntemlerini kullandılar. Farklı yöntemler arasında, PseAAC ile DNN modeli ile en yüksek doğruluğu elde etti.

Gök ve diğerleri [18], iki virülent protein veri setinde CMV öznitelik kodlama yöntemini kullanarak k-En Yakın Komşuluğu (k-EYK), Rastgele Orman, Naif Bayes, C4.5 ve Ada Boost sınıflandırıcı algoritmaları ile bakteriyel virülent proteinleri tahmin ettiler. CMV yöntemi bir veri seti için k-EYK sınıflandırıcı yüksek doğruluk sağlarken diğer veri seti için Rotasyon Ormanı algoritması yüksek doğruluk sağlamıştır.

Hassan Mohabatkar ve diğerleri [19], alerjenik proteinleri tahmin etmek için sözde amino asit bileşimi (PseAAC) ve Destek Vektör Makineleri (DVM'ler) kullanılarak yüksek tahmin doğruluğu elde etmişlerdir.

Beltran ve diğerleri [20] yaptıkları çalışmada, farklı türde proteinden deneysel olarak tanımlanmış fiziksel protein-protein etkileşimi içeren veri setine AAC, Tripeptit Kompozisyon, Dipeptit Kompozisyon, PseAAC ve Otokovaryans olmak üzere protein etkileşimlerini temsil eden beş özellik çıkarma yöntemi kullanıldılar. Sınıflandırma algoritmaları olarak; Rastgele Orman, DVM ve Gradyan destekli karar ağaçlarını kullanan Aşırı Gradyan Artırma (XGBoost) yöntemini kullanarak sonuçlar elde edildi. Deneysel sonuçlara dayanarak, XGBoost, kullanılan ölçümlerin çoğu için DVM ve Rastgele Orman'dan daha iyi sonuçlar verdi.

Mahapatra ve diğerleri [21] eşit sayıda etkileşimli tür içi E.coli PPE veri setine standart dizi tabanlı özellikler çıkarım yöntemlerinden AAC, DC ve CTD yöntemleri uyguladı. DVM ile etkileşimleri tahmin etmişlerdir.

Tian ve diğerleri [22] yapmış olduğu çalışmada, H. pylori ve S. cerevisiae veri setlerindeki protein dizilerinin özelliklerini çıkarmak için PseAAC, AAC, Gruplandırılmış Ağırlığa Göre Kodlama (EBGW) yöntemleri ve bu yöntemlerin birleşimi ile geliştirdikleri yeni yöntemleri kullanılmışlardır. Sınıflandırıcı olarak, DVM, Naif Bayes, Lojistik Regresyon, Karar Ağaçları kullanılmıştır. DVM

algoritması her iki veri seti üzerinde diğer algoritmalara göre yüksek başarı elde etmiştir.

Du ve diğerleri [23], S. cerevisiae, H. pylori ve H. sapiens veri kümeleri için özellik çıkarma yöntemi olarak ACC, Dipeptit Kompozisyon, CTD ve Amfifilik Sözde Amino Asit Bileşimi (APAAC) kullanarak DNN ile tahmin yaptılar.

Göktepe ve diğerleri [24], protein veri tabanlarından topladıkları veriler ile kendi veri setini oluşturarak Bi-Gram, PseAAC ve farklı Bitişik Üçlü formlarından oluşan dizi tabanlı özellikler çıkarım yöntemleri kullanmışlardır. Temel bileşen analizi (TBA) ile boyutunu daralttıkları veriler üzerinden DVM ile sınıflandırma yapmışlardır.

Alguwaizani ve diğerleri [25], human papillomaviruses (HPV) – insan proteinleri arasındaki PPE tahmini için AAC ve kendi geliştirdikleri yöntemi kullanarak DVM ile sınıflandırma gerçekleştirmişlerdir.

Khorsand ve diğerleri [26] çalışmalarında, insan proteinleri ve Alphainfluenzavirüs proteinleri arasındaki protein-protein etkileşimlerini tahmini için AAC, Dipeptit Kompozisyon, Bitişik Üçlü protein dizisine dayalı yöntemler yanında çeşitli protein kodlama yöntemleri gerçekleştirmiş ve sınıflandırıcı olarak C5, Rastgele Orman, DVM, Naif Bayes, K-EYK gibi birkaç farklı sınıflandırıcıları birleştirerek topluluk sınıflandırma algoritması kullanıldılar. Analiz sonucunda yüksek doğruluk, özgünlük ve duyarlık elde ettiler.

Dey ve diğerleri [27], çalışmalarında insan ve dang virüsü arasındaki protein-protein etkileşimlerini AAC ve Bitişik Üçlü protein çıkarım yöntemleri ile tahmin etmişlerdir. Tahmin için üç iyi bilinen denetimli makine öğrenme yöntemi olan DVM, NAİF BAYES ve k-EYK yöntemlerini kullanmışlardır.

Dey ve diğerleri [28], SARS-CoV-2 virüs proteinleri ve insan proteinleri arasındaki PPE'leri tahmin etmek için çeşitli makine öğrenimi modelleri oluşturulmuşlardır. Sınıflandırma modelleri, AAC, PseAAC ve Bitişik Üçlü gibi insan proteinlerinin farklı dizi tabanlı özelliklerine dayalı olarak protein kodlama yöntemlerini kullanarak hesaplanmıştır. Modeli eğitmek için DVM, Rastgele Orman, k-EYK, Naif Bayes ve Derin Çok Katmanlı Algılayıcılar sınıflandırma algoritmalarını uygulayıp ve yeni etkileşim setlerinin tahmini için karşılaştırmalı analiz yapıldı. Topluluk sınıflandırma

(Ensemble Classification) algoritmasını kullanarak hedef insan proteinlerini belirlendi.

Pang ve diğerleri [29], ilk olarak AAC, Dipeptit Kompozisyon, k-Aralıklı Amino Asit Grup Çiftlerinin Bileşimi ve PseAAC protein kodlama yöntemlerini kullanarak protein dizilerini kodladılar. Sonrasında k-EYK, Rastgele Orman, Naif Bayes ve Bayes Ağı yöntemleri ile sınıflandırma gerçekleşmiştir. Elde ettikleri deneysel sonuçlara göre; AAC yöntemiyle kodlanan protein verileri üzerinde en iyi performansı Naif Bayes algoritması vermiştir.

1.3. Hipotez

Makine öğrenmesi sınıflandırma yöntemleri ile gerçekleştirilen SARS-CoV-2 - insan proteini etkileşimi tahminine ait başarım, amino asitlerin yer değiştirme özelliklerine dayanan yeni protein çıkarım yöntemleri geliştirilmesi ile arttırılabilir.

1.4. Tez İçeriği

Tez giriş bölümü ile birlikte beş bölümden oluşmaktadır.

İkinci bölümde protein etkileşim çalışmasında kullanılan yöntemler; protein çıkarım yöntemleri, amino asit yer değiştirme matrisleri, tezimizde kullandığımız yer değiştirme matrisi Blok Değiştirme Matrisi (BLOSUM) matris, veri ön işleme yöntemleri ve son olarak sınıflandırma algoritmaları anlatılmıştır.

Üçüncü bölümde; geliştirilen protein çıkarım yöntemi Sıklık Konum Yer değiştirme Matrisi (SKYM) olarak adlandırdığımız yöntem anlatılmıştır.

Dördüncü bölümde; veri seti ve deneysel sonuçlar anlatılmıştır.

Beşinci bölümde; kaynaklar ve öz geçmiş yer almaktadır.

2. YÖNTEMLER

2.1. Protein Çıkarım Yöntemleri

Bu bölümde biyoenformatik alınanda kullanılan protein çıkarım yöntemleri anlatılmıştır. Proteinler; amino asitlerin zincir halinde birbirine bağlanmasından oluşan büyük organik bileşikler olduğu için, dizilim bilgisinin sınıflandırıcıya doğrudan verilmesi mümkün değildir. Makine öğrenmesi algoritmalarında verinin kullanılması için proteinler sabit uzunlukta nitelik vektörleri şeklinde gösterilmesine ihtiyaç duyulmaktadır [28].

Protein sınıflandırması için oluşturulmuş hesaplama modeline dizi verilerinin girilmesi için bir ön işleme adımı olarak, verimli ve etkili verilerin modele dönüştürülmesi gereklidir. Protein dizi temsili algoritmaları, farklı vektör uzunluklarında hesaplama modeline mümkün olduğunca fazla bilgi sağlar. Dizi bilgisinin yüksek verimli teknoloji yoluyla elde edilmesi daha kolay olduğu için, öncelikle hem protein yapısı tahmini hem de etkileşim tahmini için kullanılır [30].

Proteinlerin dizi olarak kodlamasının amacı, protein dizisinin makine öğrenmesi algoritmaları tarafından daha anlamlı hale getirip performans ve doğruluk açısından yüksek başarı elde edilebilir hale getirmektir [11].

Tezde kullanılan yöntemlerin detaylı bilgileri aşağıda açıklanmıştır.

AAC yöntemi, protein dizilerinin gösteriminde kullanılan yöntemlerden biri amino asit bileşim yöntemidir. Bu yöntem, tüm proteinler, bilinen 20 çeşit amino asit olması göz önünde bulundurularak 20 boyutlu nitelik vektörleri ile göstermişlerdir. Her boyut ilgili amino asidin dizilim içerisinde bulunma sıklığıdır. Bu sıklık veri kümesi içinde bulunan her bir protein için ayrı ayrı hesaplanmıştır. Verilen dizilim bilgisi içerisinde ilgili amino asidin oranını bulunmuştur ve sınıflandırma için bu değer kullanılmıştır [29].

Bir proteinin AAC'si, o proteindeki her amino asidin normalleştirilmiş sıklığı olarak tanımlanır (Denk.2.1) [19].

$$AAC = [f_1, f_2, f_3, \ldots, f_{20}]^T \quad \textbf{(2.1)}$$

Dizi uzunluğu *L* olsun, her *i.* amino asidin *(i=1,2,...,20)* tekrarlanma sayısı n_i dersek;

$$f_i = \frac{n_i}{L} \quad \textbf{(2.2)}$$

Kompozisyon – Geçiş - Dağılım Yöntemleri (CTD): Etkileşen her protein çiftinin özellik vektörü, dizilerdeki her amino asidi kimyasal yapısına kodlanmış temsili kullanılarak oluşturulur [15]. Amino asitler kimyasal yapı olarak; hidrofobi, normalleştirilmiş Van der Waals hacmi, polarite, polarize edilebilirlik, yük, ikincil yapı, çözücü erişilebilirliği olmak üzere yedi özellik ve bu özelliklerin her biri için amino asitler üç farklı gruba ayrılır.

Bu çalışmada bu üçlü yöntemin Kompozisyon ve Geçiş yöntemleri kullanılmıştır.

Kompozisyon (CTDC) yöntemi, kimyasal özellikler baz alınarak amino asitlerin sıklık bilgileriyle hesaplanmış ve normalleştirilmiş, 21 boyutunda bir öznitelik vektörüdür [16].

Geçiş (CTDT) yöntemi, amino asitlerin ardı sıra farklı gruplara geçişlerinin sayılması ve normalleştirilmesi ile hesaplanır. Geçişlerin olası sayısı 21'dir ve bu sebeple öznitelik vektörü 21 boyutundadır [16].

N-gram yöntemi, protein dizilimindeki amino asitlerin n-gram tabanlı öznitelik çıkarma yöntemidir. Bir protein dizisindeki n-gramlar biyolojik olarak anlamlı bilgiler sunar, çünkü her n-gram bir protein dizisi motifini temsil eder [18]. Ayırt edici n-gramları tanımlamak amacıyla veri setindeki her protein dizisinden olası tüm n-gramlar çıkartılır [32].

Kompozisyon moment vektörü (CMV) yöntemi, amino asit dizilimindeki amino asitlerin konum ve frekans bilgilerinden faydalanarak hesaplanmıştır [16].

P veri setindeki bir protein; A_i, *P* proteininin *i.* proteini olsun.

Amino asitler: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y şeklinde sıralanır ve P'nin vektörel gösterimi $(x_1, x_2, \ldots, x_{20})$ şeklindedir.

N amino asit dizisinin uzunluğu, a_{ij} *i.* amino asidin j. konumu ve K_i dizideki *i.* amino asidin toplam sayısıdır. Bu bilgilerden yola çıkarak; bir $k \geq 0$ tamsayısı için, *k.* dereceden moment vektörünü $(x_1^{(k)}, x_2^{(k)}, \ldots, x_{20}^{(k)})$ Denk.2.3'deki gibi şekilde tanımlanır.

$$x_i^{(k)} = \frac{1}{N(N-1)\dots(N-k)} \sum_{j=1}^{K_i} a_{i\,j}{}^{k} \qquad \textbf{(2.3)}$$

CMV yönteminin vektör uzunluğu *K* değerine göre farklılık gösterdiği için Kx20 boyutunda olduğu söylenebilir.

İki proteinin, aynı moment matrisine sahip olmaları durumunda aynı oldukları gösterilebilir. Bu da matris ile ikincil yapı dizisi arasındaki fonksiyonel ilişkiyi garanti eder [22].

2.2. Amino Asit Yer Değiştirme Matrisi

Biyoenformatikteki temel sorunlardan biri, amino asitler dizilerinin arasındaki benzerlik veya farklılığı nicel bir ölçümü seçme problemidir, bu da proteinleri oluşturan amino asitlerin arasındaki benzerliklerin ölçülmesine dayanmaktadır [33].

Amino asit benzerliği geleneksel olarak bir yer değiştirme matrisine ile tanımlanır. Amino asit yer değiştirme matrisleri, protein veri tabanlarında benzer dizi araması ve çoklu dizi hizalaması için temel araçlardır. Çoğu yer değiştirme matrisi; protein aileleri veya ilgili proteinlerin gruplarını temsil eden bloklarda, her amino asit arasındaki istatistiksel yer değiştirme sıklığı kullanılarak geliştirilmiştir. Bununla birlikte, amino asitlerin yer değiştirmesi, her amino asidin yer değiştirme özelliklerinin benzerliğine dayanır [34]. Yer değiştirme matrislerinin içeriği; P_i ve P_j amino asitlerinin benzerlik oranı, i. satır ve j. sütununa yazılır [35]. Amino asitlerin bir kısmı kimyasal özelliklerinden dolayı birbirleri arasında yer değiştirme daha olasılığı yüksek iken, diğer bir kısmı için daha bu olasılık daha düşüktür. Amino asit hizalanmasında bu verileri desteklemek için puanlama matrislerine ihtiyaç duyulmuştur [36].

İlk yaygın olarak kullanılan amino asit yer değiştirme matrisi Kabul Edilmiş Nokta Mutasyon (PAM) matrisidir. Daha sonra BLOSUM serisi PAM matrisinin yerini aldı.

PAM matrisleri: 1978 yılında bilim insanları amino asitlerin birbiri yerine geçme olasılıklarını deneysel olarak hesaplayıp bir puanlama fonksiyonu sundular. Deneysel çalışmalarında, 71 protein ailesinden 1572 yakın akraba proteinini incelediler.

Bu proteinlerde meydana gelen nokta mutasyonları bağımsız olduğunu varsayarak amino asitlerin birbirinin yerine geçme olasılığını hesapladılar. PAM matrisleri

hesaplanan bu olasılıklarla oluşturulmuş 20x20 boyutunda matrislerin adıdır. Uzunlukları farklı mutasyon serileri için farklı PAM matrisleri vardır. Mutasyon serisi arttıkça amino asitlerin birbirinin yerine geçme olasılığını artar. Proteinin evrimsel uzaklıklarına göre hangi PAM matrisinin kullanılması gerektiği belirlenir [36].

2.2.1. Blok Değiştirme Matrisleri

BLOSUM yer değiştirme matrisleri bilim insanları tarafından 1992 yılında tanıtıldı [36]. Deneysel verilerle hazırlamış protein sekans hizalamaları için yaygın olarak kullanılan istatiksel bir yöntemdir. PAM serisi kısa mutasyonlar için uygun olsa da, mutasyon serileri uzadıkça mutasyonların birbirinden bağımsız olduğu varsayımı geçersiz olamaya başlar.

BLOSUM62 matris sayısal değerleri, dizi hizalamada %62 benzer olan iki DNA dizisi tarafından belirlenir. Matris değerleri, iki amino asidin birbirinin yerine ne kadar ikame edilebileceğini gösterir. Belirli bir çift için BLOSUM62 matris elemanı büyük bir pozitif değere sahipse, doğada daha sık meydana gelir ve protein işlevini etkileme olasılığı daha düşüktür. Örneğin, Glutamic ve Aspartic eşleşmesinin BLOSUM62 puanı 2'dir. Trytophan ve Trytophan yapıları aynı olduğu için 11 puan almıştır. Bazı amino asit çiftleri -4 veya -3 kadar düşük değerlere sahip olabilir, bu da bu ikamelerin doğada nadiren meydana geleceği ve meydana gelmeleri halinde muhtemelen protein fonksiyonunda önemli bir değişiklikle sonuçlanacağı anlamına gelir. Örneğin, Valine ve Triptofan -4 puana sahiptir; bu, Valin'in polar olmayan bir yan gruba sahip olması ve Triptofan'ın aromatik bir yan grubuna sahip olması nedeniyle yapıdaki bir farklılıktan kaynaklanmaktadır [37]. Tablo 2.1'de BLOSUM62 matrisi detaylı olarak verilmiştir.

Tablo 2.1. BLOSUM62 matrisi

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

2.3. Ön İşleme Yöntemleri

Veri ön işleme, veri analizi uygulamalarının en önemli aşamalarından biridir. Ham veriler genellikle çeşitli tutarsızlıklar, aralık dışı değerler, eksik değerler, gürültüler veya fazlalıklar gibi birçok kusurla birlikte gelir. Sonraki aşamalarda gerçekleştirilecek olan makine öğrenmesi algoritmalarının performansı bu düşük kaliteli veriler nedeniyle zayıflayacaktır. Bu nedenle ham verilerin çeşitli ön işleme aşamalarından geçirilerek kalitesinin artırılması gerekmektedir [38]. Veri analizi uygulamalarında veri seti için kullanılan veri ön işleme algoritmaları bu başlık altında incelenmektedir.

2.3.1. Sınıf Dengeleme

Verilerin sınırsız boyutu ve dengesizliği nedeniyle verilerin sınıflandırılması zorlaşır. Sınıf dengesizliği sorunu veri madenciliğinde en büyük sorun haline gelir [39]. Dengesizlik sorunu bir sınıf (azınlık sınıfı) diğer sınıflara (çoğunluk sınıfları) kıyasla eğitim verilerinde büyük ölçüde yetersiz temsil edilmesi sonucunda ortaya çıkar [40].

Standart makine öğrenmesi algoritmaları çoğunluk sınıfı tarafından boğulma ve azınlık sınıfı görmezden gelme eğilimindedir, çünkü sınıflandırma algoritmaları tüm örnekler üzerinden doğru bir performans ararlar [41]. Sınıflandırma algoritmaları çoğunluk sınıfına karşı ön yargılı olduğundan, sınıflandırma algoritmalarının doğruluğu daha fazla zarar görecektir. Böylece, yeni örnek sınıflandırılması için, sınıflandırıcının azınlık sınıfına yönelik tahmin doğruluğu daha düşük olduğundan çoğunluk sınıfı olarak sınıflandırılacaktır [42]. Bu tür bir sınıflandırma modeli kullanışlı değildir.

Dengesiz veri kümelerinin örnek seçiminin yaygın çözümleri; yetersiz temsil edilen sınıfın eğitim örneklerinin çoğaltılması (oversampling), aşırı temsil edilen sınıfın eğitim örneklerinin kaldırılması (undersampling) [40].

2.3.2. Ayrıklaştırma

Makine öğrenmesi algoritmalarındaki başarıyı arttırmak için veri setindeki değerleri ayrıklaştırma yöntemi kullanılır. Ayrıklaştırma yöntemi sürekli değerlerin kategorik değerlere dönüştürülmesidir. Bir öznitelik için sürekli değerlerin sayısı sonsuz sayıda olabilirken, ayrık değerlerin sayısı az veya sonludur [43]. Ayrıklaştırma veri setindeki gürültü ve doğrusal olamama durumunu azaltır [44]. Bu sebepten ayrıklaştırma, gelişmiş tahmin doğruluğuna sağlar.

Ayrıca literatürde bulunan birçok tümevarım algoritması ayrık öznitelikler gerektirir. Tüm bunlar, araştırmacıları bir makine öğrenimi görevi öncesinde sürekli özellikleri ayrıklaştırmaya yönlendirir. Literatürde çok sayıda ayrıklaştırma yöntemi mevcuttur [43].

2.4. Sınıflandırma Algoritmaları

Sınıflandırma, belirli bir girdiye dayalı olarak belirli bir sonucun tahminidir. Sonuçları tahmin etmek için mevcut olan verilerin sınıflandırıcıya verilmesi gerekir. Bu verilere dayanarak kayıtlar sınıflandırılır. Veri kaynakları eğitim seti ve test seti olarak ayrılır. Eğitim seti daha önce sınıflandırılmış ve sınıflandırma amacıyla referans olarak kullanılan verileri içermektedir. Özniteliklerin yardımıyla sonuçlar tahmin edilir. Daha sonra test verileri algoritmaya verilir. Bu veriler daha önce depolanan özniteliklere göre kontrol edilir ve bu olasılıklara dayalı olarak veriler sınıflandırılır. Algoritma verilen verileri analiz eder ve sonuçları tahmin eder [45].

Sınıflandırma, hedef sınıfı en yüksek hassasiyetle tahmin etmeye çalışır. Sınıflandırma algoritması, bir eğitim süreci olan bir model oluşturmak için girdi özniteliği ile çıktı özniteliği arasındaki ilişkiyi bulur [44].

Denetimli öğrenme teknikleri, gelecekteki olayları tahmin etmek için etiketlenmiş veriler yardımıyla önceki ve mevcut veri setini bağlayan sınıflandırma algoritmalarıdır.

Denetimsiz öğrenme, eğitim veri kümesi sınıflandırılmamış veya etiketlenmemiş olduğunda kullanılan sınıflandırma algoritmalarıdır [45].

Yarı denetimli öğrenme teknikleri, denetimli öğrenme teknikleri ile denetimsiz öğrenme teknikleri arasında kalan, eğitim sürecinde etiketli ve etiketsiz veri kümelerinin kullanıldığı öğrenme teknikleridir [45].

Tez çalışmasında kullanılan sınıflandırma algoritmaları Naif Bayes, Bayes Ağları, K-EYK, Rastgele Orman, DVM ve Çok Katmanlı Algılayıcılar (ÇKA) algoritmalarıdır.

2.4.1. Naif Bayes

Naif Bayes, Bayes teoremine dayanan olasılık tabanlı bir sınıflandırıcıdır, bu teori ilk kez 1763'te İngiliz bilim adamı Thomas Bayes tarafından yayınlanan bir makalede ortaya çıkmıştır. Naif Bayes sınıflandırıcısı, bir sınıftaki belirli bir öznitelik varlığının başka herhangi bir öznitelik varlığıyla ilgisi olmadığını varsayar. Böylece her özellik sonuca bağımsız ve eşit bir katkı sağlar [46]. Her özniteliğin sonuca olan etkileri olasılık olarak hesaplanır. Eğitim kümesi ile eğitilen sınıflandırıcı özniteliklerin her birinin sınıflarla olan ilişkisini olasılık oranı olarak hesaplar ve o değerleri içeren

modeli çıktı olarak verir. Daha sonra sınıflandırıcı, test örneklerini bu modeldeki öznitelik - sınıf olasılıklarını kullanarak, öznitelik bağımsızlığı varsayımıyla sınıflandırır.

Bayes teoremi matematiksel olarak aşağıdaki Denk.2.3'de gösterilmiştir.

P(B) ≠ 0, P(A|B) koşullu olasılıktır. B doğru olduğunda *A* olayının meydana gelme sayısı. *P(B|A)* da koşullu olasılıktır. A doğru olduğunda B olayının meydana gelme sayısı. *P(A)* ve *P(B)* marjinal olasılıktır, sırasıyla A ve B olasılıklarıdır.

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \tag{2.3}$$

Naif Bayes perspektifinde, *P(y|X), y* sınıf değişkenidir ve X_n boyutunda bağımlı bir özellik vektörüdür, burada $X = (x_1, x_2, x_3, \ldots\ldots, x_n)$

Bu nedenle, bir dizi bağımsız özellik için Naif Bayes denklemini Denk.2.4'deki gibi yeniden yazabiliriz:

$$P(y|x_1, x_2, x_3, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y) \ldots P(x_n|y)P(y)}{P(x_1)P(x_2) \ldots P(x_n)} \tag{2.4}$$

Sınıflandırıcı model oluşturulması için, y sınıf değişkeninin tüm olası değerleri için belirli bir girdi kümesinin olasılığını bulunmalı ve çıktıyı maksimum olasılıkla olarak almalıyız. Denk.2.5'de ifade edilir; denklemde *P(y)* sınıf olasılığı ve $P(x_i|y)$ koşullu olasılık olarak adlandırılır [27].

$$y = argmax_y\, P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.5}$$

2.4.2. Bayes Ağları

Bayes Ağları, veri incelemelerinde modeller oluşturmak için kullanılan bir Olasılık Grafik Model biçimidir. Bayes Ağları, rastgele değişkenler arasındaki hem koşullu hem de koşullu bağımsız etkileşimleri yakalar. Naif Bayes öznitelikler arasında güçlü bir bağımlılık olduğunda, sınıflandırma doğruluğu önemli ölçüde azalır. Bundan dolayı birçok araştırmacı Naif Bayes sınıflandırıcının performansını iyileştirmek ve öznitelik bağımsız varsayımının olumsuz etkisini azaltmak için Bayes Ağları sınıflandırma algoritmasını kullanırlar [47].

Dizi değişkenleri $V=m_1,m_2,..,m_n$ burada n ≥ 1'dir. Bayes ağı *G* ile gösterilir. Ağ yapısı G_S, Ağ olasılık tablosu G_P gösterilir. *pa(v)* G_S'deki *v*'nin ebeveyn kümesidir.

$$\mathrm{Gp}=\{\mathrm{p}(\mathrm{v}|\mathrm{pa}(\mathrm{v}))|\mathrm{v}\in\mathrm{V}) \quad \textbf{(2.6)}$$

Bayes Ağı olasılık dağılımı:

$$\mathrm{P}(\mathrm{V}) = \prod_{v\in V} p(v|pa(v)) \quad \textbf{(2.7)}$$

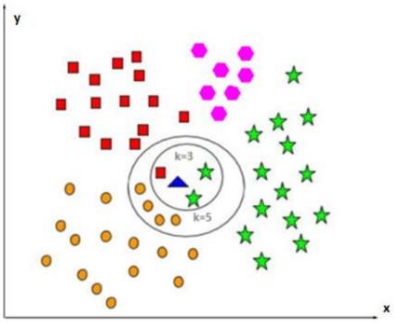
$m = m_1,m_2,..,m_n$ bir dizi öznitelik değişkeni olsun. Bir sınıflandırıcı için $q: m \rightarrow c$ tanımı, m örneğini c sınıfına eşleyen bir işlev. Bayes ağını sınıflandırıcı olarak kullanmak için *P(V)* dağılımı kullanılarak $argmax_c P(c|m)$ hesaplanır [48].

2.4.3. k – En Yakın Komşuluğu

k-EYK basit, etkili ve parametrik olmayan sınıflandırma yöntemidir. Sınıflandırma işlemine başlamadan önce en yakın kaç tane komşuya bakılacağının yani k sayısının belirlenmesi gerekir. Ancak, k-EYK'yi uygulamak için k için uygun bir değer seçmemiz gerekir ve sınıflandırmanın başarısı büyük ölçüde bu değere bağlıdır. Bir anlamda, k-EYK yöntemi k tarafından ön yargılıdır. k değerini seçmenin birçok yolu vardır, ancak basit olanı, algoritmayı farklı k değerleriyle birçok kez çalıştırmak ve en iyi performansa sahip olanı seçmektir [49]. Verinin sınıflandırılması için sınıfı bilinen örnekler seçilir.

Sınama örneğinin sınıfını belirlerken öğrenme örnekleri, kümesindeki örneklere olan uzaklıkları hesaplanır ve en yakın k tane örnek seçilir. Uzaklık hesaplanırken Öklid, Manhattan, Minkowski gibi farklı uzaklık ölçütlerinden yararlanmak mümkündür.

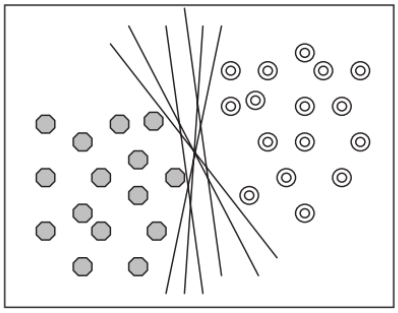
Daha sonra, seçilen k tane örnek arasında hangi sınıfa ait örnek sayısı en fazlaysa sınama örneği de bu sınıfa ait olduğu varsayılır. Şekil 2.1'de dört sınıflı veri setinin k-EYK ile sınıflandırması gösterilmektedir. Test verisi için önce k değeri 3, daha sonra k değeri 5 alınmıştır.



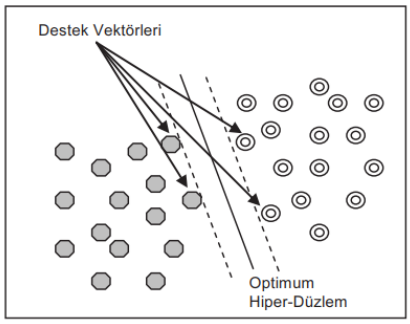
Şekil 2.1. Dört sınıflı bir veri setinin k-EYK sınıflandırılması

2.4.4. Destek Vektör Makineleri

DVM algoritması bilim insanları tarafından geliştirilen parametrik olmayan istatiksel öğrenme teorisine dayanan bir denetimli sınıflandırma yöntemidir [50]. DMV başlangıçta iki sınıflı doğrusal verilerin sınıflandırılması için tasarlanmış fakat yetersiz kalınca çok sınıflı ve doğrusal olmayan veriler için geliştirilmiştir [50]. Verilerin sınıflandırması en uygun karar fonksiyonunun (hiper düzlem) tanımlanması esasına dayanmaktadır. Hiper düzlemin belirlenmesinde en önemli olan en etkili ayrımı yapan düzlemin belirlenmesidir. Şekil 2.2'de gösterildiği gibi iki sınıflı verilerin ayrılmasında birden çok hiper düzlem çizilebilir. Sınıfların düzleme en yakın noktalarına destek vektörü denir. Bu en yakın noktalar arasındaki mesafesi maksimum olan düzlem, en uygun hiper düzlemdir. Şekil 2.3'de görüldüğü üzere en uygun düzlem elde edilmiştir.



Şekil 2.2. İki sınıflı verilerin ayrılmasında çizilen hiper düzlemler



Şekil 2.3. İki sınıflı verilerin ayrılmasında, en yakın noktalar arasındaki mesafesi maksimum olan düzlem

Bir hiper düzlem, n tane örnekten oluşan bir eğitim verisinin $\{ x_i , y_i \}, i = 1,.....,n$ olduğu kabul edilirse; Denk.2.8'deki gibi hesaplanır.

$$w. x_i + b = 0 \quad \textbf{(2.8)}$$

İki sınıflı veriler için doğrusal hiper düzlemler denklemi, Denk.2.9 ve Denk.2.10'da ki gibidir. x_i hiper düzlem üzerindeki noktayı, w hiper düzlemin normalini (ağırlık vektörü), b ise hiper düzlemin orijine olan uzaklığını (eğilim değeri) ifade etmektedir [50].

$$w \cdot x_i + b \geq +1 \text{ her } y = +1 \text{ için} \quad \textbf{(2.9)}$$

$$w \cdot x_i + b \leq +1 \text{ her } y = -1 \text{ için} \quad \textbf{(2.10)}$$

Bu eşitlik bir tek eşitlik haline getirilirse denklem 2.11 elde edilir.

$$y_i (w. x_i + b) - 1 \geq 0, \; x_i \in \{-1,+1\} \text{ ve } i = 1, \dots , n. \quad \textbf{(2.11)}$$

Doğrusal olarak ayrılabilen verilerde maksimum sınırlandırılmaları yapan en iyi hiper düzlem ξ değeri eklenerek sınıflandırmada doğrusal ayrılamama problemi çözümü için kullanılır. ξ değeri pozitiftir sınıflandırma hatalarını temsil eder. Eşitlik yeniden düzenlenirse Denk.2.12 elde edilir [50].

$$y_i (w. x_i + b) - 1 + \xi_i \geq 0 \quad \textbf{(2.12)}$$

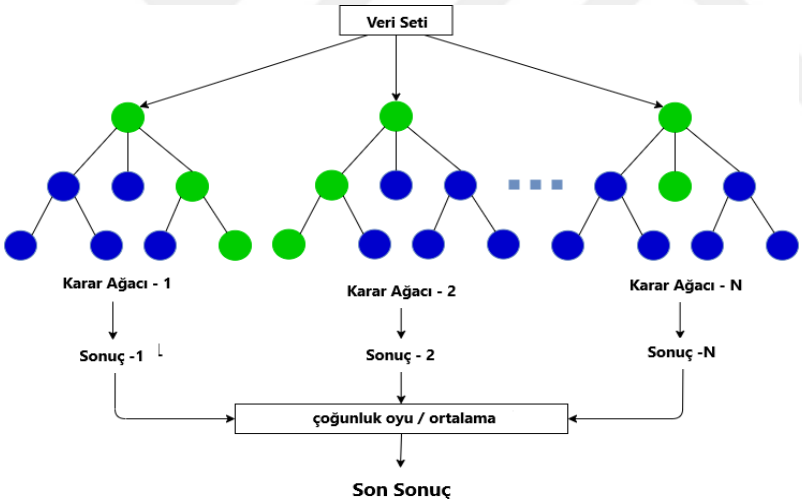
2.4.5. Rastgele Orman

Rastgele orman, 2001 yılında Leo Breiman [51] tarafından ortaya atılan bir yöntemdir. Rastgele Orman, karar ağacı sınıflandırıcılarının tahminlerini birleştiren bir topluluk öğrenme yöntemidir. Rastgele ormanlar, her ağacın farklı bir özellik alanına eşlendiği bir karar ağaçları ormanı oluşturur [52].

Rastgele ormanlar, her ağacın bağımsız olarak örneklenen rastgele bir vektörün değerlerine bağlıdır. Ormandaki tüm ağaçlar için aynı dağılıma sahip olduğu ağaç tahmin edicilerinin bir kombinasyonudur [51].

Biyoenformatik alanında, karar ağaçları topluluğunu içeren ve öğrenme sürecinde doğal olarak özellik seçimi ve etkileşimleri içeren Rastgele Orman tekniği popüler bir seçimdir [53].

Rastgele orman algoritması, doğru ve istikrarlı bir tahmin için birleştirilen karar ağaçlarının bir koleksiyonudur. Rastgele ormanlar, torbalama yönteminin kapsamına girer ve genel fikir, öğrenme modellerinin bir kombinasyonunun genel sınıflandırma sonucunu iyileştirmesidir. Şekil 2.4'de gibi rastgele orman algoritması adımları: N farklı ön yükleme eğitimi örneği oluşturulur, daha sonra algoritmayı her ön yüklenen örnek üzerinde ayrı ayrı eğitilir, son olarak tahminlerin ortalaması ya da çoğunluk oyunu alınır [18].

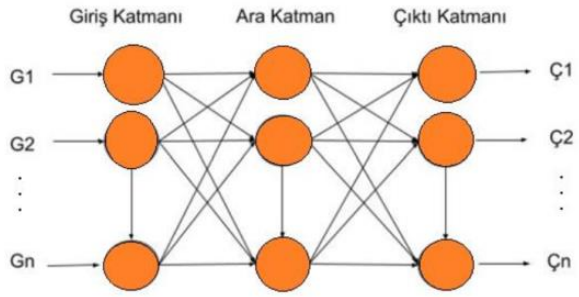


Şekil 2.4. Rastgele Orman algoritma yapısı

2.4.6. Çok Katmanlı Algılayıcılar

ÇKA, standart geri yayılım algoritması ile eğitilmiş ileri beslemeli sinir ağı modelleridir [54]. Eğitilmek istenen çıktı değerlerine ihtiyaç duyarlar, böylece girdi verilerini istenen çıktılara nasıl dönüştüreceklerini öğrenebilirler. Bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanı, ÇKA ağ bileşenleridir [55]. Değişken sayıda gizli katmanla, neredeyse tüm girdi-çıktı haritalarını tahmin

edebilirler. Zor problemlerde optimal istatistiksel sınıflandırıcıların performansını tahmin ettikleri gösterilmiştir. Şekil 2.5'de ÇKA'nın işleyişi gösterilmiştir.



Şekil 2.5. Çok Katmanlı Algılayıcı

Ağın oluşturduğu bileşenlerin her biri, kendilerine ait ağırlıklarla çarpılarak toplanır ve hesaplanan değerleri, çıktılarını üretmek için bir transfer fonksiyonundan geçirir. Birimler bu şekilde katmanlı bir ileri beslemeli topolojide tasarlanır. ÇKA ağları, fonksiyonların karmaşıklığını belirleyerek, gizli katmanların sayısı ve her katmandaki nöronların sayısı ile fonksiyonları modelleyebilir. Kullanılacak gizli katmanların sayısı, probleme ve modeller için kullanılan veri tipine bağlıdır [58].

Probleme göre; katman sayısı, her katmandaki algılayıcı sayısı, bias(b) katsayısının değeri, eğitimdeki iterasyon sayısı, en uygun aktivasyon fonksiyonu seçimi ÇKA için sınıflandırma başarısında farklılıklar sağlar. Birçok veri için en uygun katman sayısı üç olarak alınmaktadır [56]. Bias verisi içinde birçok araştırmada genel olarak en ideal değer 0.7 olduğu görülmüştür [56].

Transfer fonksiyonundan giriş vektörü x_i geçirilerek işlem elemanının çıktı değeri aşağıdaki Denk.2.13 ile elde edilir. Burada f aktivasyon fonksiyonunu, b eşik değerini, n girdi sayısını, w_{ji} i. işlem elemanından j. işlem elemanına olan ağırlık değerini göstermektedir. Aktivasyon fonksiyonların ikiye ayrılır doğrusal fonksiyonlar ve doğrusal olmayanlar olarak. Hiperbolik tanjant fonksiyonu ve sigmoid fonksiyonu en çok kullanılan doğrusal olmayan fonksiyonlardır [57].

$$y_i = f(\sum n\, w_{ji}\, x_i + b) \qquad \textbf{(2.13)}$$

3. GELİŞTİRİLEN PROTEİN ÇIKARIM YÖNTEMİ: SIKLIK KONUM YER DEĞİŞTİRME MATRİSİ

Biz bu tez çalışmasında, biyolojik deneyler kullanılarak doğrulanmış yeni tip korona virüs - insan proteinleri arasındaki PPE'leri makine öğrenmesi modelleri ile yüksek doğrulukla tahmin etmek için SKYM adını verdiğimiz yeni bir öznitelik çıkarım yöntemi geliştirdik. SKYM ile farklı uzunluktaki protein dizilimleri, sabit uzunlukta vektörlere dönüştürüldü. Daha sonra ön işleme adımı olarak sınıf dengeleme ve ayrıklaştırma yöntemleri uygulandı ve sınıflandırma algoritmaları ile tahmin gerçekleştirildi.

SKYM, amino asitlerin protein dizilimi içindeki konum, sıklık ve amino yer BLOSUM62 değiştirme matrisi bilgilerini kullanan bir protein dizilimi öznitelik çıkarım yöntemidir.

Bir protein dizilimi, SKYM ile 1x60 boyutunda vektörle ifade edilir. Vektörün, ilk 1x20 kısmı her bir amino asidin sıklığını; ikinci 1x20 kısmı konumunu; son 1x20 kısmı BLOSUM62 matrisinde yer alan köşegen bilgisini içerir.

İlk adımda dizilim içindeki amino asitlerin sıklık değerleri denklem 3.1'de görülen ifade ile hesaplanır.

Dizi uzunluğu L, her i. amino asidin ($i = 1,2,\ldots,20$) tekrarlanma sayısı n_i olmak üzere f_i sıklığı,

$$f_i = \frac{n_i}{L + (L + 1)} \tag{3.2}$$

ile hesaplanır.

Sıklık değerleri = [$f(f_1), f(f_2), \ldots f(f_{20})$] ilk 1x20 değer elde edilir.

İkinci olarak hesaplanan vektörün 1x20 konum değeri her i. amino asit a_i için; a_{ij} i. amino asidin j. konumu ve K_i, dizideki i. amino asidin toplam sayısı olmak üzere L değeri,

$$a_i = \frac{\sum_{j=1}^{K_i} a_{ij}}{L * (L+1)} \quad \textbf{(3.2)}$$

ile hesaplanır. Konum değerleri = [$f(a_1), f(a_2), \ldots f(a_{20})$] ikinci 1x20 değer elde edilir.

BLOSUM62 matrisinden her bir amino asidin köşegen değeri (amino asitlerin kendi yer değiştirme değerleri) b_i olmak üzere çarpım değeri, s_i,

$$s_i = \frac{b_i * n_i}{L * (L+1)} \quad \textbf{(3.33)}$$

işlevi ile elde edilir ve değerler 1x20 amino asit vektörüne yazılır. Son olarak, üç adımda elde edilen her bir 1x20 boyutundaki vektörler birleştirilir ve 1x60 boyutunda SKYM vektörü elde edilir. SKYM yönteminin sözde kodu aşağıda verilmiştir.

Algoritma: SKYM protein çıkarım yöntemi
Girdiler: Protein dizilimi ←('A','C',... ,'Y')
aminoasit sayısı ←20
d←Veri seti protein dizilimi
d_i←d proteinin i.amino asidi
a_i←P dizilimindeki i.amino asit
$count_i$←i.amino asit toplam değeri
B←Blosum62 matrisi
B_i←i.amino asitdin köşegen B değerleri

for each aa_P in P
  for each aa_d in d
    if(aa_P == aa_D)
      $count_{aap} \leftarrow$ *her x amino asidinin tekrarlanma sayısının toplamı*
  $s_{aap} \leftarrow \frac{count_{aap}}{len(d)+len(d)+1}$

for each aa_d in d
  $dkonum \leftarrow$ *d dizilimindeki her aa'ya sıralı değer ataması*
for each aa_P in P
  for each aa_d in d
    if(aa_P == aa_d)
      $dkonum_{aap} \leftarrow$ *her x amino asidinin konum değerileri toplamı*
  $k_{aap} \leftarrow \frac{dkonum_{aap}}{len(d)+len(d)+1}$

for each aa_B in B
  for each aa_d in d
    if(aa_P == aa_d)
      $B_{aap} \leftarrow$ *aaB değeri ile her aa'nın adet değeri çarpımı*
  $b_{aap} \leftarrow \frac{B_{aap}}{len(d)+len(d)+1}$

4. BULGULAR

4.1. Veri Setleri

Tez çalışmamızda, 30.046 SARS-CoV-2 virüsü ile etkileşime girmemiş insan proteini (negatif) ve 20.365 etkileşime girmiş insan proteini (pozitif) olmak üzere toplam 50.411 tane protein dizilimi içeren bir PPE veri seti [28] kullandık. Pozitif PPE verileri, 332 insan proteini ile dört yapısal ve ayrıca 20 yardımcı korona virüs proteini arasındaki 332 benzersiz etkileşimi içerir. Deneysel olarak doğrulanmış bu 332 insan proteini, çalışmamızda pozitif eğitim ve test veri kümeleri oluşturmak için kullanılmıştır [28]. PPE ağında negatif bir veri seti planlamak için "altın standart" yoktur. Bu nedenle, tahmin edilen PPE'lerin doğruluğunu olumsuz yönde etkileyebilecek negatif numuneler dikkatle seçilmelidir. Dey ve arkadaşları [28] negatif örnekleri, pozitif veri setinde bulunmayan ve insan PPE ağında düşük bir dereceye sahip olan Human Protein Reference Database (HPRD) veri tabanından insan proteinlerini seçerek hazırlamışlardır.

4.2. Deneysel Sonuçlar

Veri setini, sınıf dengeleme ve ayrıklaştırma yöntemleri ile ön işlem aşamasından geçirdik. Sınıf dengeleme yöntemi kullanmadaki amacımız, negatif verilerin sayıca pozitif verilerden çok fazla olmasıdır. Bu durum, sınıflandırmada negatif sınıfa doğru yanlama (bias) olmasına neden olur. Böylece pozitif sınıfın başarımı olması gerekenden düşük, negatif sınıfın başarımı yüksek çıkar. Bu durumu önlemek için, sınıf ağırlıklarını ayarlayarak pozitif ve negatif örnekler arasında dengeleme yapan sınıf dengeleme yöntemini kullandık.

Deneysel sonuçları elde ederken aşırı öğrenme probleminden kaçınmak ve rastlantısallığı en aza indirmek için veri seti üzerinde k-kat çapraz doğrulama yöntemi kullandık. k-kat çapraz doğrulama, mevcut tüm örnekleri eğitim ve test örnekleri olarak kullanan test protokol tekniğidir [58]. Bu çalışmada 10-kat çapraz doğrulama tekniği kullanılmıştır. 10-kat çapraz doğrulamada, veri seti 10 eşit kümeye bölünür. Dokuz küme eğitim verisi, bir küme test verisi için kullanır. Her bir küme bir defa test kümesi olacak şekilde 10 defa kaydırma ile seçilen sınıflandırıcı üzerinde başarım metrikleri elde edilir. 10 iterasyon sonunda elde edilen sınıflandırma sonuçlarının

ortalaması alınır ve böylece sınıflandırma başarım metriklerine veren karmaşıklık matrisi elde edilir.

Karmaşıklık matrisi, makine öğrenmesinde kullanılan bir değerlendirme tablosudur. Bir karmaşıklık matrisinin sütunları, tahmin sınıfı sonuçlarını temsil eder ve satırlar, gerçek sınıf sonuçlarını temsil eder. Bir sınıflandırma probleminin tüm olası durumlarını sıralar. Örnek olarak ikili sınıflandırma problemini alırsak, karmaşıklık matrisinin boyutu 2 × 2'dir [59].

Tablo 4.1'de görüldüğü gibi Karmaşıklık matrisinde, başarım sonuçlarının elde edilmesini sağlayan Doğru Pozitif (DP), Doğru Negatif (DN), Yanlış Negatif (YN), Yanlış Pozitif (YP) olmak üzere dört terim vardır.

Tablo 4.1. Karmaşıklık Matrisi

		Tahmin Edilen Sınıf	
		Pozitif	Negatif
Gerçek Sınıf	Pozitif	DP	YN
	Negatif	YP	DN

DP, doğru tahmin edilen pozitif; DN, doğru tahmin edilen negatif; YN, gerçekte pozitif veri olup negatif olarak tahmin edilen; YP, gerçekte negatif veri olup pozitif olarak tahmin edilen değerdir.

Doğruluk, doğru tahmin edilen pozitif ve negatif örneklerin oranını verir [60].

$$Doğruluk = \frac{DP + DN}{DP + YP + DN + YN} \quad \textbf{(4.4)}$$

Duyarlık, doğru tahmin edilen pozitif örneklerin oranını verir.

$$Duyarlık = \frac{DP}{DP + YN} \quad \textbf{(4.2)}$$

Özgünlük, doğru tahmin edilen negatif örneklerin oranını verir.

$$\text{Özgünlük} = \frac{DN}{DN + YP} \quad \textbf{(4.3)}$$

Kesinlik, tahmin edilen örnekler içinde doğru tahmin edilen pozitif örneklerin oranını verir.

$$Kesinlik = \frac{DP}{DP + YP} \tag{4.4}$$

MCC, [-1, 1] aralığında değerler alır. 1 tam bir uyum gösterirken -1 tam bir anlaşmazlık anlamına gelir [61].

$$MCC = \frac{(DP * DN) - (YP * YN)}{\sqrt{(DP + YP) * (DP + YN) * (DN + YP) * (DN + YN)}} \tag{4.5}$$

Duyarlık ve kesinlik değerlerinin harmonik ortalaması olan F1-skor, iki değer arasındaki dengeyi ölçer [62].

$$F1 - skor = 2 * \frac{Keskinlik * Duyarlılık}{Keskinlik + Duyarlılık} \tag{4.6}$$

Kappa, gözlemlenen doğruluk ve beklenen doğruluk temelinde çalışır. P_o, gözlemlenen doğruluğu, P_e ise beklenen doğruluğu gösterir. Standartlara göre Kappa değeri her zaman 1'den küçük veya eşittir. 0 veya 0'dan küçük Kappa değerleri, sınıflandırıcının iyi bir sonuç vermediğini gösterir [62].

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.7}$$

ROC, özgünlük ve duyarlık ile çizilen bir başarım grafiğidir [63]. AUC, ROC eğrisi altında kalan alanın sayısal karşılığıdır. AUC skorunun teorik aralığı 0 ile 1 arasındadır. AUC değeri ne kadar yüksekse model başarımı o kadar iyi demektir.

Çalışmamızda öncelikle literatürde öne çıkan protein diziliminden bağımsız AAC, CMV, 2-Gram, CTDC, CTDT ve geliştirdiğimiz SKYM protein çıkarım yöntemlerine göre veri setimizi kodladık. Böylece protein dizilimleri aynı boyuta getirilerek sınıflandırıcı algoritmalar için uygun hale getirildi. Sınıflandırmada, Naif Bayes, Bayes Ağları, Doğrusal DVM, Radyal DVM, Rastgele Orman, k-EYK ve ÇKA yöntemlerini kullandık. Tablo 4.2'de SKYM protein çıkarımı ile kodlanan verilerin sınıflandırma başarım metriklerine yer verilmiştir.

Tablo 4.2. SKYM protein çıkarım yöntemi uygulanmış verilerin sınıflanırma algoritmalarına göre başarım metrikleri

Sınıflandırma algoritmaları	Doğruluk	Kesinlik	Duyarlık	Özgünlük	AUC	F1 - Skor	MCC	Kappa
Naif Bayes	0,513	0,514	0,529	0,499	0,516	0,514	0,028	0,027
Bayes Ağları	0,513	0,514	0,529	0,499	0,516	0,514	0,028	0,027
k-EYK	0,512	0,513	0,516	0,509	0,514	0,513	0,025	0,025
Rastgele Orman	0,512	0,513	0,515	0,510	0,514	0,513	0,025	0,025
Doğrusal DVM	0,596	0,729	0,002	1	0,501	0,447	0,034	0,002
Radyal DVM	0,596	0,749	0,002	1	0,501	0,447	0,031	0,002
ÇKA	0,596	0,569	0,023	0,986	0,509	0,461	0,033	0,010

Elde edilen sonuçlara göre; SKYM öznitelik çıkarım yöntemi, Naif Bayes ve Bayes Ağları algoritmaları altında, 0,516 AUC, 0,514 F1-skor ve 0,529 duyarlık değerleri ile en iyi sonuçları vermiştir. Bununla birlikte ÇKA, Radyal DVM ve Doğrusal DVM algoritmaları 0,596 ile en yüksek doğruluğu vermiştir. Fakat düşük duyarlılık değerlerine sahiplerdir. Düşük duyarlılık, negatif sınıf yanlamalı olduğunu ve pozitif sınıf tahmininde başarılı olmadığını gösterir. Düşük Kappa ise elde edilen sonuçların rastlantısallıktan ne kadar uzak olduğunu yorumlamamızı sağlar. Kappa değeri, ne kadar 1'e yakınsa, elde edilen başarım o kadar rassal değildir.

Yeni koronovirüs - insan proteini etkileşimi tahmini problemi için geliştirdiğimiz SKYM öznitelik çıkarım yöntemimizi, literatürde öne çıkan öznitelik çıkarımı yöntemleri ile kıyasladık. Deneysel çalışmalarda, AAC, 2-Grams, CTDC, CTDC ve CMV öznitelik çıkarım yöntemlerini sınıflandırıcı algoritmalarla PPE veri setimiz üzerinde test ettik. Böylece her bir yöntem için en iyi sonucu veren sınıflandırıcı algoritmayı belirledik. Son olarak Tablo 4.3'de görüldüğü gibi SKYM ile test ettiğimiz öznitelik çıkarımı yöntemlerini kıyasladık.

Tablo 4.3. Protein çıkarım yöntemi uygulanmış verilerin sınıflanırma algoritmalarına göre başarım metrikleri

Çıkarım yöntemleri	Sınıflandırma algoritmaları	Doğruluk	Kesinlik	Duyarlık	Özgünlük	AUC	F1-Skor	MCC	Kappa
2-Grams	Naif Bayes	0,506	0,51	0,214	0,798	0,507	0,46	0,015	0,012
AAC	Naif Bayes	0,501	0,512	0,294	0,726	0,513	0,486	0,022	0,020
CTDT	ÇKA	0,596	0,559	0,008	0,994	0,506	0,451	0,017	0,003
CTDC	Naif Bayes	0,594	0,54	0,023	0,982	0,509	0,46	0,015	0,005
CMV	ÇKA	0,596	0,563	0,017	0,989	0,504	0,457	0,026	0,007
SKYM	Naif Bayes	0,513	0,514	0,529	0,499	0,516	0,514	0,028	0,027

Elde edilen deneysel sonuçlara göre; SKYM yöntemi, en yüksek AUC, duyarlık, F1-skor, MCC ve kappa değerlerinde en iyi sonucu vermiştir. SKYM, pozitif sınıfı en iyi tahmin ettiğini 0,529 en yüksek duyarlık performansı sergileyerek göstermektir. SKYM, diğer yöntemlere göre negatif sınıfa yanlama sorununu çözdüğünü 0,028 MCC ve 0,514 F1-skor değerleri ile göstermektedir. Ayrıca SKYM en yüksek kappa değeri elde etmesi diğer yöntemlere göre şans faktörüne daha az bağlı olduğunu gösterir.

0,008 en düşük duyarlık ve 0,994 en yüksek özgünlük değeri vermesi nedeniyle, CTDT'nin ÇKA ile 0,596 en yüksek doğruluk başarımı sergilemesi güvenilmezdir. Çünkü CTDT, gerçekte pozitif sınıf olmasına rağmen tüm değerleri baskın negatif sınıf olarak tahmin ettiği için doğruluk artmaktadır. Nitekim, CTDT'nin MCC değeri de 0'a çok yakın olması pozitif ve negatif sınıf arasında yanlama sorunu olduğunu göstermektedir.

5. SONUÇLAR

Virüs yoluyla bulaşan endemik ve/veya pandemik hastalıkların önlenmesinde patojen-konak etkileşimlerini ortaya çıkaracak çalışmalar ilaç ve aşı gelişim açısından çok önemlidir. Bu çalışmamızda; SARS-CoV-2 - insan proteinleri arasındaki PPE tahmini için SKYM adını verdiğimiz bir protein öznitelik çıkarımı yöntemi geliştirdik.

Protein verileri, sınıflandırma algoritmalarına verilmeden önce öznitelik kodlama teknikleri ile sınıflandırıcıya uygun hale getirilmiştir. Protein öznitelik çıkarımı yöntemleri, PPE etkileşimi incelenen protein dizilimleri ile hastalık arasında altta yatan işlevin sınıflandırıcı algoritma tarafından tanınması açısından önemlidirler. Sınıflandırıcı ne kadar bu işlevi anlarsa, hastalığın tahmininde o kadar yüksek başarım sergiler.

Amino asitlerin konum, sıklık ve yer değiştirme matrislerine dayanan SKYM yöntemi, SARS-CoV-2 - insan proteinleri arasındaki PPE tahmini için literatürde öne çıkan AAC, 2-Grams, CTDC, CTDC, CMV yöntemleri, sınıflandırma algoritmaları başarım metriklerine göre kıyasladık. PPE'ler; Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, Doğrusal DVM, Radyal DVM ve ÇKA sınıflandırma algoritmaları ile test edilmiştir. Elde ettiğimiz deneysel sonuçlara göre, SKYM en iyi başarımı göstermiştir. Sınıflandırma da tüm yöntemler için veri setimiz üzerinde sınıf dengeleme gerçekleştirdik.

Gelecekte çalışmalarda; SKYM protein çıkarım yöntemini değişik protein temelli sınıflandırma problemleri üzerinde, farklı PAM ve BLOSUM matrisleri ile geliştirilmeyi planlamaktayız. Ayrıca SKYM yöntemini uygulayan bir çevrim içi web sunucu geliştirmeyi hedeflenmekteyiz.

KAYNAKLAR

[1] F. Ucar ve D. Korkmaz, COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images, Medical Hypotheses, no. 140, 2020.

[2] Our World In Data, http://www.ourworldindata.org/, 2021

[3] D. Wu, T. Wu, Q. Liu ve Z. Yang, The SARS-CoV-2 outbreak: What we know, International Journal of Infectious Diseases, no. 94, pp. 44-48, 2020.

[4] S. H. Kassania, . P. H. Kassasni, M. J. Wesolowski, K. A. Schneider ve R. Deters, Automatic Detection of Coronavirus Disease (COVID-19) in, 2020.

[5] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir ve R. Siddique, COVID-19 infection: origin, transmission, and characteristics of human coronaviruses, Journal of Advanced Research, 2020.

[6] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu ve . K. Obernier1, A SARS-CoV-2 protein interaction map, Nature, no. 583, pp. 459-468, 2020.

[7] Ü. Abdulmecit, SARS-COV2: Genom Yapısı ve Yaşam Döngüsü, Yüksek İhtisas Üniversitesi Sağlık Bilimleri Dergisi, pp. 8-11, 2020.

[8] H. Lodish, A. Berk, C. A. Matsudaira, M. Krieger, M. P. Scott, L. Zipursky ve J. Darnell, Molecular Cell Biology, 5.

[9] R. Sen, L. Nayak ve R. K. DE, A review on host–pathogen interactions: classification, European Journal Clinical Microbiology Infectious Diseases, pp. 1581-1599, 2016.

[10] A. H. Basit, W. A. Abbasi, A. Asif, S. Gull ve F. U. A. A. Minhas, Training host-pathogen protein-protein interaction predictors, Journal of Bioinformatics and Computational Biology, 2018.

[11] H. Wang ve X. Hu, Accurate prediction of nuclear receptors, BMC Bioinformatics, 2015.

[12] S. M. Srinivasan, S. Vural, B. R. King ve C. Guda, Mining for class-specific motifs in protein, BMC Bioinformatics, 2013.

[13] S. L. Lo, C. Z. Cai, Y. Z. Chen ve M. C. M. Chung, Effect of training datasets on support vector machine, Proteomics, p. 876–884, 2005.

[14] İ. Kösesoy, Konak-Patojen Protein Etkileşiminin Hesaplamalı Yöntemler İle Tahmini, 2018.

[15] C. Chen, Y.-X. Tian, X.-Y. Zou, P.-X. Cai ve J.-Y. Mo, Using pseudo-amino acid composition and support vector machine to, Journal of Theoretical Biology, 2006.

[16] S. Aghajanbaglo, M. Rahgozar ve A. Rahimi, Predicting Protein-Protein Interactions Based on Sequence, Iranian Conference of Bioinformatics, 2014.

[17] R. K. Barman, A. Mukhopadhyay, U. Maulik ve S. Das, Identification of infectious disease-associated host genes using machine learning techniques, BMC Bioinformatics , 2019.

[18] M. Gök ve D. Herand, Prediction of Bacterial Virulent Proteins with Composition Moment Vector Feature Encoding Method, MATEC Web of Conferences, 2016.

[19] H. Mohabatkar, M. M. Beigi, K. Abdolahi ve S. Mohsenzadeh, Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach, Medicinal Chemistry, pp. 133-137, 2013.

[20] J. C. Beltran, P. Valdez ve P. Naval, Predicting Protein-Protein Interactions based on Biological Information using Extreme Gradient Boosting, IEEE , 2019.

[21] S. Mahapatra, A. Kumar, A. Sharma ve S. S. Sahu, Effect of Dimensionality Reduction on Classification Accuracy for Protein-Protein Interaction Prediction, Advanced Computing and Intelligent Engineering, 2020.

[22] B. Tian, X. Wu, C. Chen, W. Qiu, Q. Ma ve B. Yu, Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach, Journal of Theoretical Biology, 2018.

[23] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan ve . Y. Zhang, DeepPPI : Boosting Prediction of Protein-Protein Interactions with Deep Neural Networks, Journal of Chemical Information and Modeling, 2017.

[24] Y. E. Goktepe ve H. Kodaz, Prediction of Protein-Protein Interactions Using An Effective Sequence Based Combined Method, Neurocomputing, 2018.

[25] B. Kim, S. Alguwaizani, X. Zhou, D. S. Huang, B. Park ve K. Han, An improved method for predicting interactions between virus and human proteins, Journal of Bioinformatics and Computational Biology, 2017.

[26] B. Khorsand, A. Savadi, J. Zahiri ve M. Naghibzadeh, Alpha influenza virus infiltration prediction using virus-human protein–protein interaction network, Mathematical Biosciences and Engineering, 2020.

[27] L. Dey ve A. Mukhopadhyay, A Classification-based Approach to Prediction of Dengue Virus and Human Protein-Protein Interactions using Amino Acid Composition and Conjoint Triad Features, IEEE Region 10 Symposium , 2020.

[28] L. Dey, S. Chakraborty ve A. Mukhopadhyay, Machine learning techniques for sequence-based prediction of viralehost interactions between SARS-CoV-2 and human proteins, Biomedical Journal, 2020.

[29] Y. Pang, Z. Wang, J. H. Jhong ve T. Y. Lee, Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies, Briefings in Bioinformatics, 2021.

[30] H. Chen, W. Guo, J. Shen, L. Wang Ve . J. Song, Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey, IEEE Access, 2018.

[31] I. Kosesoy, M. Gok ve C. Oz, A New Sequence Based Encoding for Prediction of Host-Pathogen Protein Interactions, Computational Biology and Chemistry, 2018.

[32] H. Mohabatkar, S. Ebrahimi ve M. Moradi, Using Chou's Five-steps Rule to Classify and Predict Glutathione S-transferases with Different Machine Learning Algorithms and Pseudo Amino Acid Composition, International Journal of Peptide Research and Therapeutics , 2020.

[33] V. V. Sulimova, V. V. Mottl , C. A. Kulikowski ve I. B. Muchnik , Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families.

[34] Y. You, I. Jang, K. Lee, H. Kim ve K. Lee, An Approach for a Substitution Matrix Based on Protein Blocks and Physicochemical Properties of Amino Acids through PCA, Interdisciplinary Bio Central, 2014.

[35] M. Gök, HIV-1 Proteaz Eziminin Kesme Konumlarının Tespitinde Yeni Öznitelik Vektöleri, 2011.

[36] E. Aygün, Protein İşlev Kestiriminde Yapısal Bilginin Katkısı ve Dizi Geçiş Olasılıkları ile Peptit Sınıflandırılması, 2009.

[37] L. M. Aichinger, Evaluation of the Signature Molecular Descriptor with BLOSUM62 and an All-Atom Description for Use in Sequence Alignment of Proteins, 2015.

[38] V. Çetin ve O. Yıldız, A comprehensive review on data preprocessing techniques in data analysis Veri analizinde veri ön işleme teknikleri üzerine kapsamlı bir inceleme, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 2021.

[39] R. Longadge, S. S. Dongre ve L. Malik, Class Imbalance Problem in Data Mining: Review, International Journal of Computer Science and Network, 2013.

[40] S. B. Kotsiantis, D. Kanellopoulos ve P. Pintelas, Data Preprocessing for Supervised Leaning, Internatıonal Journal Of Computer Scıence, 2006.

[41] X. Guo, Y. Yin, C. Dong, G. Yang ve G. Zhou, On the Class Imbalance Problem, International Conference on Natural Computatin.

[42] Shwet Ketu ve Pramod Kumar Mishra, Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare, Complex & Intelligent Systems , 2021.

[43] H. Lıu, F. Hussaın, C. L. Tan ve M. Dash, Discretization: An Enabling Technique, p. 393–423, 2002.

[44] İ. Koç, Sınıflandırma Problemlerinde Meta-Sezgisel Optimizasyon Yöntemlerinin Özellik Seçimi ve Ayrıklaştırma Amacıyla Kullanımı, 2016.

[45] S.Umadevi ve K. Marseline, A Survey on Data Mining Classification Algorithms, International Conference on Signal Processing and Communication, 2017.

[46] F. Alzubaıdı, Detect Malware Url Usıng Naıve Bayes Algorıthm, 2021.

[47] Zhang Liwei, Predictive Analysis of Machine Learning Error Classification Based on Bayesian Network, Wireless Personal Communications, 2021.

[48] B. Abraham ve M. S. Nair, Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier, Biocybernetics and Biomedical Engineering, pp. 1436-1445.

[49] W. Xing ve Y. Bei, Medical Health Big Data Classification Based on KNN Classification Algorithm, IEEEAccess, 2017.

[50] H. Ç. Reis ve G. Yılancı, Destek vektör makineleri ve NDVI kullanarak pamuk ekili alanların tespiti: Harran ovası örneği, Türkiye Uzaktan Algılama Dergisi , pp. 29-41, 2020.

[51] L. Breıman, Random Forests, Machine Learning, p. 5–32, 2001.

[52] L. Vig, Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset, Open Access Library Journal, 2014.

[53] Y. Qi, Random Forest for Bioinformatics, 2012.

[54] M. C. Yüksel, Development Of A New Software For Network Traffıc Predıctıon And Forecastıng Usıng Tıme Serıes Multılayer Perceptron And Feedback Delays, 2019.

[55] Prediction of Output Energies for Broiler, Environmental Progress & Sustainable Energy Production Using Linear Regression, ANN (MLP, RBF), and ANFIS Models, 2016.

[56] F. Bulut Ve D. Bozyılan, Çok Katmanlı Algılayıcılar ile doğru meslek seçimi, Anadolu Üniversitesi Bilim ve Teknoloji Dergisi, pp. 97-109, 2016.

[57] B. Konakoğlu, Çok Katmanlı Algılayıcı Yapay Sinir Ağı ile Jeodezik Elipsoidal Koordinatların (φ, λ, h) 3 Boyutlu Global Kartezyen Koordinatlara (X, Y, Z) Dönüşümü, Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi, pp. 702-710, 2020.

[58] Y. Bengio ve Y. Grandvalet, No Unbiased Estimator of the Variance of K-Fold Cross-Validation, Journal of Machine Learning Research 5, p. 1089–1105, 2004.

[59] J. Xu, Y. Zhang ve D. Miao, Three-way confusion matrix for classification: A measure driven view, Information Sciences , 2019.

[60] C. Ferri, J. Hernández-Orallo ve R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognition Letters, pp. 27-38, 2009.

[61] S. Boughorbel, F. Jarray ve M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PLOS ONE, 2017.

[62] P. Bhat ve P. Malaganve, Metadata based Classification Techniques for Knowledge Discovery from Facebook Multimedia Database, I.J. Intelligent Systems and Applications, pp. 38-48 , 2021.

[63] R. Kannan ve V. Vasanthi, Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease, SpringerBriefs in Forensic and Medical Bioinformatics, 2019, pp. 63-72.

ÖZGEÇMİŞ

Ad Soyad: Firdes Gül KORKUT

Lisans: Bilgisayar Mühendisliği

Yayın ve Patent Listesi:

- Firdes Gül Korkut, Murat Gök, 2021. “Kompozisyon Moment Vektör Öznitelik Yöntemi ile Sars-Cov-2 ve İnsan Proteinleri Arasındaki Virüs-Konakçı Etkileşimlerinin İncelenmesi”, International Conference on Data