



**KODLAMA YAPMAYAN UZUN VE KISA BOYUTLU
RNA'LAR İLE HASTALIKLAR VE ÇEVRESEL
FAKTÖRLER ARASINDAKİ BAĞLANTILARIN TAHMİNİ**

Hüseyin VURAL

DOKTORA TEZİ

Anabilim Dalı: Bilgisayar Mühendisliği Bölümü

Programı: Kuramsal Temeller

Danışman: Prof. Dr. Mehmet KAYA

MAYIS-2019

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**KODLAMA YAPMAYAN KISA VE UZUN BOYUTLU RNA'LAR İLE
ÇEVRESEL FAKTÖRLER VE HASTALIKLAR ARASINDAKİ
BAĞLANTILARIN TAHMİNİ**

DOKTORA TEZİ

Hüseyin VURAL

(142129204)

Tezin Enstitüye Verildiği Tarih : 2 Nisan 2019

Tezin Savunulduğu Tarih : 2 Mayıs 2019

Tez Danışmanı : Prof.Dr. Mehmet KAYA (F. Ü.)

Diğer Jüri Üyeleri : Prof.Dr. Reda ALHAJJ (İ. M. Ü.)

Prof.Dr. Ali KARCI (İ. Ö.)

Prof.Dr. İbrahim TÜRKOĞLU (F. Ü.)

Doç.Dr. Burhan ERGEN (F. Ü.)

MAYIS-2019

ÖNSÖZ

Tezin ortaya çıkması ve çalışma süresince her zaman desteğini sunan, yol gösterici olan değerli danışman hocam Sayın Prof. Dr. Mehmet KAYA'ya teşekkürlerimi sunarım. Uzun tez çalışma sürecinde her zaman destek olan aileme ve özellikle kardeşim Giray'a teşekkür ederim.

Hüseyin VURAL
ELAZIĞ-2019

İÇİNDEKİLER

Sayfa No

ÖNSÖZ	II
İÇİNDEKİLER	III
ÖZET	VI
SUMMARY	VII
ŞEKİLLER LİSTESİ	VIII
TABLolar LİSTESİ	X
KISALTMALAR	XI
1. GİRİŞ	1
1.1. Tezin Amacı ve Katkıları	2
1.2. Tezin Yapısı	3
2. İNSAN GENOM PROJESİ	4
2.1. Protein Kodlaması Yapmayan RNA'lar	4
2.1.1. LncRNA'lar	6
2.1.2. MikroRNA'lar	7
2.1.3. Çevresel Faktörler	9
2.2. Etkileşimler	10
2.2.1. LncRNA'ların Hastalıklar Üzerindeki Etkileri	10
2.2.2. MiRNA-Çevresel Faktör Etkileşimlerinin Hastalıklar Üzerindeki Etkileri	11
2.2.3. Çevresel Faktörlerin Hastalıklar Üzerindeki Etkileri	12
2.2.4. LncRNA-Çevresel Faktör Etkileşimlerinin Hastalıklar Üzerindeki Etkileri	13
3. VERİTABANLARI VE HESAPLAMALI MODELLER	14
3.1. LncRNA Veritabanları	14
3.1.1. NONCODE Veritabanı	14
3.1.2. lncRNome Veritabanı	14
3.1.3. lncRNAdb Veritabanı	14
3.1.4. LncRNADisease Veritabanı	15
3.1.5. Lnc2Cancer Veritabanı	15
3.2. MiRNA Veritabanları	15
3.2.1. MiREnvironment	15

3.2.2.	PEMDAM	15
3.3.	LncRNA Çevresel Faktörler Veritabanları	16
3.3.1.	DLREFD	16
3.4.	LncRNA-Hastalık Modelleri	16
3.4.1.	LncRNA-Hastalık İlişkileri için Düzenlenmiş Laplas En Küçük Kareler Yöntemi ..	16
3.4.2.	LRLSLDA-LNCRNA Fonksiyonel Benzerlik Hesaplama Modeli	17
3.4.3.	Naif Bayes Sınıflandırıcı	18
3.4.4.	RWRlncD	19
3.4.5.	İki Parçalı lncRNA-PCG Ağı Üzerinde Yeniden Başlatma ile Rastgele Yürüyüş....	19
3.4.6.	Hastalık Genleri Temel Alınarak lncRNA-Hastalık Tahmin Modeli	19
3.5.	LncRNA-Çevresel Faktörler Modelleri	20
3.6.	MiRNA-Çevresel Faktörler Modelleri	20
4.	MATERYAL VE METOD	21
4.1.	KATZ	21
4.2.	Gauss Etkileşim Profili Çekirdek Benzerliği	21
4.3.	Spearman Korelasyon Katsayısı	23
4.4.	Hastalık Semantik Benzerliği	23
4.4.1.	Bir Hastalığın Semantik Değeri	24
4.4.2.	İki Hastalığın Semantik Benzerliği	24
4.5.	Yapay Sinir Ağları	25
4.5.1.	Derin Sinir Ağları.....	26
4.5.1.1.	Kayıp Fonksiyon	29
4.5.1.2.	Optimizasyon Algoritmaları	30
4.5.1.3.	Gradyan İnişi Optimizasyon Algoritmaları	31
4.5.1.4.	İletim Sönümü	34
4.6.	İstif Edilmiş OtoKodlayıcılar	35
5.	ÖNERİLEN MODELLER.....	39
5.1.	MiRNA Çevresel Faktör Etkileşim Modeli	39
5.1.1.	Veri Seti	40
5.1.2.	MikroRNA’lar için Gauss Etkileşim Profili Çekirdek Benzerliği	40

5.1.3.	Çevresel Faktörler için Gauss Etkileşim Profili Çekirdek Benzerliği	41
5.1.4.	KATZ	42
5.1.5.	Değerlendirme	43
5.1.6.	Global Tek Çıkarımlı Çapraz Doğrulama Yöntemi	44
5.1.7.	K-katlı Çapraz Doğrulama Yöntemi	44
5.1.8.	Sonuçlar	47
5.2.	LncRNA Çevresel Faktör Etkileşim Modeli	47
5.2.1.	Veritabanı	48
5.2.2.	LncRNA'lar için Gauss Etkileşim Profili Çekirdek Benzerliği	48
5.2.3.	Çevresel Faktörler için Gauss Etkileşim Profili Çekirdek Benzerliği	49
5.2.4.	KATZ	50
5.2.5.	Değerlendirme	51
5.2.6.	Global Tek Çıkarımlı Çapraz Doğrulama Yöntemi	51
5.2.7.	K-katlı Çapraz Doğrulama Yöntemi	52
5.2.8.	Diğer Metodlar ile Karşılaştırma	55
5.2.9.	Örnek Çalışmalar	56
5.2.10.	Sonuçlar	57
5.3.	OtoKodlayıcı ve Derin Öğrenme Tabanlı lncRNA-Hastalık Modeli	57
5.3.1.	Veritabanı	59
5.3.2.	LncRNA Ekspresyon Benzerliği	60
5.3.3.	Hastalık Semantik Benzerliği	60
5.3.4.	LncRNA'lar için Gauss Etkileşim Profili Çekirdek Benzerliği	61
5.3.5.	Hastalıklar için Gauss Etkileşim Profili Çekirdek Benzerliği	62
5.3.6.	LncRNA'lar için Entegre Benzerlik Matrisi	62
5.3.7.	Hastalıklar için Entegre Benzerlik Matrisi	63
5.3.8.	Metod	63
5.3.9.	5-katlı Çapraz Doğrulama Yöntemi	66
6.	SONUÇLAR VE ÖNERİLER	69
	KAYNAKLAR	71
	ÖZGEÇMİŞ	81

ÖZET

İnsan vücudunda bulunan lncRNA, miRNA vb. moleküller birbirleriyle, genlerle ve çevresel faktörlerle etkileşim halinde bulunmaktadır. Etkileşimler bir ağ yapısı şeklinde tarif edilebilmektedir. Ağ yapısında bulunan bağlantıların her hangi bir nedenden dolayı kopması veya ağda yeni bağlantıların oluşması çeşitli hastalıklara neden olabilmektedir. Bir molekül ile hastalık arasında veya çevresel faktör arasında bağlantı olup olmadığının laboratuvar ortamında test edilmesi oldukça zaman gerektiren ve yoğun insan kaynağına ihtiyaç duyan bir işlemdir. Daha önce aralarında bağlantı olmayan moleküller ile çevresel faktörler ve hastalıklar arasında olabilecek potansiyel bağlantıların geliştirilecek yöntemler ile tahmin edilmesinin bu işlemi hızlandırmaya katkı sunacağı düşünülmektedir.

Tezimizde bu amaçla üç yöntem önerilmiştir. LncRNA-çevresel faktörler ve miRNA-çevresel faktörler arasında potansiyel yeni bağlantıların tahmin edilmesi amacıyla KATZ tabanlı iki model geliştirilmiştir. Diğer bir model ise lncRNA'lar ile hastalıklar arasında potansiyel yeni bağlantıların tahmin edilmesi amacıyla istiflenmiş otokodlayıcılar ve derin öğrenme temel alınarak geliştirilmiştir. Geliştirilen modellerin test sonuçlarının literatürdeki modellere karşı daha başarılı AUC sonuçları elde ettiği gösterilmiştir.

Anahtar Kelimeler: Bağlantı Tahmini, lncRNA, miRNA, çevresel faktör, hastalık.

SUMMARY

Prediction of Associations Between Diseases and Environmental Factors of Long and Short Non-coding RNAs

The lncRNA, miRNA, etc. molecules in the human body interact with each other, genes and environmental factors. The interactions can be described as a network structure. The interruption of the connections in the network structure for any reason or the formation of new connections in the network can cause various diseases. Testing in a laboratory whether there is an association between a molecule and a disease or an environmental factor is a time-consuming process that requires intensive human resources. It is thought that prediction of potential associations between non-linking molecules and environmental factors and diseases by methods to be developed will contribute to this process.

Three methods have been proposed for this purpose. Two KATZ-based models have been developed in order to estimate the possible associations between lncRNA-environmental factors and miRNA-environmental factors. Another model has been developed based on stacked autoencoders and deep learning to estimate potential new associations between lncRNAs and diseases. The test results of the developed models show that they achieve more successful results in terms of AUC values against the models in the literature.

Keywords: Link Prediction, lncRNA, miRNA, environmental factor, disease.

ŞEKİLLER LİSTESİ

Sayfa No

Şekil 1.1.	Farklı bilim dallarının biyoinformatik bilim dalına katkısı	2
Şekil 2.1.	Kodlama Yapmayan RNA'ların sınıflandırılması	6
Şekil 3.1.	LRLSLDA'nın akış diyagramı	17
Şekil 3.2.	İki lncRNA arasındaki fonksiyonel benzerliğin hesaplanmasını temel alan LNCSIM modeli	18
Şekil 4.1.	İlaç-hedef ağından etkileşim profillerinin oluşturulması	22
Şekil 4.2.	Spearman Korelasyon Katsayısı	23
Şekil 4.3.	Bir nöronun temel yapısı	26
Şekil 4.4.	Derin Sinir Ağları Örneği	27
Şekil 4.5.	Sigmoid fonksiyon	28
Şekil 4.6.	Tanh fonksiyon	28
Şekil 4.7.	ReLU fonksiyon	29
Şekil 4.8.	Derin sinir ağında iletim sönümü eklenmiş gösterim	35
Şekil 4.9.	Birden fazla otokodlayıcıdan oluşan istif edilmiş otokodlayıcı örneği	37
Şekil 5.1.	MiRNA-Çevresel Faktörler modelinin akış diyagramı	40
Şekil 5.2.	Gauss kullanılarak komşuluk matrisinden benzerlik matrislerinin elde edilmesi... ..	42
Şekil 5.3.	k parametresinin 2 seçilmesi halindeki AUC sonuçları	45
Şekil 5.4.	k parametresinin 3 seçilmesi halindeki AUC sonuçları	46
Şekil 5.5.	k parametresinin 4 seçilmesi halindeki AUC sonuçları	46
Şekil 5.6.	LncRNA-Çevresel Faktörler modelinin akış diyagramı	48
Şekil 5.7.	k parametresinin 2 seçilmesi halindeki AUC sonuçları	53
Şekil 5.8.	k parametresinin 3 seçilmesi halindeki AUC sonuçları	54
Şekil 5.9.	k parametresinin 4 seçilmesi halindeki AUC sonuçları	54

Şekil 5.10. Modelimiz ile diğer yöntemlerin AUC değerlerinin karşılaştırılması	56
Şekil 5.11. Hastalıkların yönlü çevrimsiz çizgesel olarak gösterilmesi	61
Şekil 5.12. İstif edilmiş otokodlayıcılarda geriye-yayılım ve parametre güncellemeleri.....	64
Şekil 5.13. İstif edilmiş otokodlayıcılardan elde edilen öznitelik vektörlerinin derin sinir ağlarına aktarılması ve derin sinir ağlarından lncRNA-hastalık tahminlerinin elde edilmesi.....	65
Şekil 5.14. Modelimizin 5-katlı çapraz doğrulama yönteminde elde ettiği AUC değeri	67
Şekil 5.15. Modelimiz ile diğer yöntemlerin AUC değerlerinin karşılaştırılması.....	68

TABLÖLER LİSTESİ

	<u>Sayfa No</u>
Tablo 2.1. LncRNA veri örnekleri	7
Tablo 2.2. MiRNA veri örnekleri	8
Tablo 2.3. Çevresel Faktör örnekleri.....	9
Tablo 5.1. LncRNA-Çevresel Faktör bağlantı örnekleri.....	52
Tablo 5.2. LncRNA-Çevresel Faktör bağlantı örnekleri üzerinde KATZ uygulama sonuçları	52
Tablo 5.3. Derin Öğrenme'nin toplam yayınlanan makale sayısı ile Derin Öğrenme'nin Biyoinformatik alanındaki makalele sayısı	58
Tablo 5.4. Derin Sinir Ağları ve İstif Edilmiş Otokodlayıcıların Biyoinformatik uygulamalarında kullanım alanları	59

KISALTMALAR

LncRNA	:Kodlama Yapmayan Uzun Boyutlu RNA'lar
MiRNA	:Kodlama Yapmayan Mikro RNA
Non-coding RNA	:Kodlama Yapmayan RNA'lar
ÇF	:Çevresel Faktörler
YÇÇ	:Yönlü Çevrimsiz Çizgesel
YSA	:Yapay Sinir Ağları
ReLU	:Düzeltilmiş Doğrusal Birim
MSE	:Ortalama Karesel Hata
SAE	:İstif Edilmiş Otokodlayıcılar
AUC	:Eğrinin altında kalan alan

1. GİRİŞ

Bilgisayarlar'ın tıp bilimlerinde kullanılması 1960'lı yıllara kadar uzanmaktadır. O yıllarda elde edilen tıp verilerinin analiz edilmesi ve yorumlanması ihtiyacı biyoinformatik biliminin doğmasına neden olmuştur. Biyoinformatik; hesaplama araçlarının kullanılarak biyolojik verinin yorumlanması ve analiz edilmesi şeklinde tanımlanmakta ayrıca bilgisayar bilimi, matematik, biyoloji/genetik ile istatistik bilimlerinden yararlanan disiplinlerarası bir alandır [1] ve Şekil 1.1.'de gösterilmiştir. 2003 yılında tamamlanan insan genom projesi ile ortaya çıkan çok büyük miktardaki biyolojik verinin yorumlanması ve işlenmesi ihtiyacı bilgisayarların tıp bilimlerinde kullanılmasını yaygınlaştırmıştır [2]. İnsan Genom Projesinin [3] tamamlanmasından bir kaç yıl sonra çok sayıda protein kodlaması yapmayan uzun boyutlu RNA (lncRNA) ve mikroRNA (miRNA) keşfedilmiştir. LncRNA'ların hastalıklar üzerindeki etkisi [4] nedeniyle lncRNA-hastalık bağlantılarının bilinmesi önem teşkil etmektedir. Aynı zamanda lncRNA'ların ve miRNA'ların çevresel faktörler ile olan bağlantılarının hastalıklar üzerinde etkili olması [5,6] nedeniyle çevresel faktörler ile olabilecek potansiyel yeni bağlantıların bilinmesi de önem teşkil etmektedir. Bu tezde lncRNA'lar ile hastalıklar ve çevresel faktörler arasında yeni potansiyel bağlantıların bulunması amacıyla iki model geliştirilmiştir. Ek olarak miRNA'lar ile çevresel faktörler arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla bir model geliştirilmiştir.



Şekil 1.1. Farklı bilim dallarının biyoinformatik bilim dalına katkısı

1.1. Tezin Amacı ve Katkıları

Çevresel faktörler bazı lncRNA'ların ve miRNA'ların ekspresyon düzeylerini değiştirebilmekte ve ekspresyon düzeyi değişen lncRNA'lar ve miRNA'ların hastalıkların ortaya çıkmasında etkin olduğu yapılan çalışmalarda gösterilmiştir [5,6]. Bu nedenle, lncRNA-çevresel faktörler ve miRNA-çevresel faktörler arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla iki model ve lncRNA-hastalıklar arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla üçüncü bir model geliştirilmiştir. Geliştirilen birinci model, KATZ ölçütünü [7] temel alarak lncRNA ve çevresel faktörler arasında farklı yürüyüş yollarını hesaplayarak yeni potansiyel bağlantıların tahmin edilmesi amacıyla geliştirilmiştir. Geliştirilen model AUC sonuçları açısından literatür çalışmalarından daha yüksek sonuçlar elde etmiştir. Çalışmada geliştirilen ikinci model lncRNA ve hastalıklar arasında yeni potansiyel bağlantıların bulunmasını amaçlamaktadır. Bu model; hastalıklar ve lncRNA benzerlik matrislerinden istif edilmiş otokodlayıcılar kullanılarak özneteliklerin ortaya çıkarılması ve bu özneteliklerin derin sinir ağlarına aktarılıp potansiyel bağlantıların tahmin edilmesi temelinde geliştirilmiştir. Geliştirilen model daha önce geliştirilen modellerden AUC sonuçları açısından daha yüksek sonuçlar elde etmiştir. Çalışmada geliştirilen üçüncü model

KATZ ölçütünü temel alarak miRNA ve çevresel faktörler arasında farklı yürüyüş yollarını hesaplayarak yeni potansiyel bağlantıların tahmin edilmesi amacıyla geliştirilmiştir. LncRNA-hastalık, lncRNA-çevresel faktör ve miRNA-çevresel faktör arasında yeni bağlantıların laboratuvar ortamında elde edilmesi çok fazla insan kaynağı ve zaman gerektirmektedir. Örneğin insanlarda görülen bir hastalığın lncRNA'lar ile bağlantıya sahip olup olmadığının laboratuvar ortamında test edilmesi, insan vücudunda bulunan lncRNA'ların çok sayıda (~280.000) olması nedeniyle çok fazla zaman ve insan kaynağı gerektirecektir. Geliştirilen lncRNA-hastalık modeli ile bu hastalığın bağlantılı olabileceği lncRNA'lar bağlantı değerlerinin büyüklüğüne göre sıralanabilmektedir. Belirlenen hastalığın, modelden elde edilen lncRNA sıralama değerleri dikkate alınarak laboratuvar ortamında test edilmesi ile zamandan ve insan kaynağından tasarruf edilmesine katkı sunulması amaçlanmaktadır.

1.2.Tezin Yapısı

Tezin sonraki aşamaları aşağıdaki şekilde hazırlanmıştır:

- Bölüm 2’de lncRNA’lar, miRNA’lar ve çevresel faktörler hakkında bilgiler verilmiştir ek olarak lncRNA’ların, miRNA-çevresel faktör etkileşimlerinin ve lncRNA-çevresel faktör etkileşimlerinin hastalıklar üzerindeki etkilerine değinilmiştir.
- Bölüm 3’te literatür çalışmaları ve veritabanı bilgileri verilmiştir.
- Bölüm 4’te materyal ve metod bilgileri açıklanmıştır.
- Bölüm 5’te lncRNA-çevresel faktör, miRNA-çevresel faktör ve lncRNA-hastalık arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla geliştirilen modellere yer verilmiştir.
- Bölüm 6’da elde edilen sonuçlar değerlendirilmiştir.

2. İNSAN GENOM PROJESİ

İnsan genomunun bütün DNA diziliminin belirlenmesi amacıyla 1990 yılında başlanan İnsan Genom Projesinin 2003 yılında tamamlanmasıyla insanlık tarihinde çok önemli bir eşik atlanmıştır. Biyolojik ve tıbbi araştırma topluluklarının büyük bilime attıkları ilk büyük atılım İnsan Genom Projesidir. Bu proje ile biyolojinin ilk büyük ölçekli projesi tamamlanmış ve çok önemli çıktılar olmuştur. Bu çıktılar; dizilemede yeni teknolojilerin geliştirilmesi, çeşitli organizmaların genomlarından oldukça faydalı genetik, fiziksel ve transkript haritalarının çıkartılması, bilimsel bir araştırma programı ile paralel bir biyoetik araştırma programının birleştirilmesi, insan genomunun bütünsel dizisinin herkes için ücretsiz ve kolayca erişilebilir olması şeklinde sıralanabilmektedir [3].

İnsan genom projesi ile insan vücudunda protein sentezinden sorumlu ve sorumlu olmayan genlerin sayısı belirlenmiştir. Bu genler insan vücudunun her bir fonksiyonunun düzgün bir şekilde işleyişini sağlamak amacıyla eşgüdüm içerisinde çalışmaktadır. Genler, farklı fonksiyonlar yardımıyla protein sentezini sağlamakta ve ayrıca çeşitli kompleks prosedürler ile organize ve düzgün çalışan bir insan vücudunu oluşturmaktadır [8].

İnsan Genom Projesi çalışmalarında daha çok protein kodlayan genlerin keşfedilmesi üzerine çaba harcanılmıştır çünkü daha önce protein sentezinde görev almayan genlerin insan genomunun fonksiyonel özelliklerinin çok küçük bir kısmını oluşturduğu varsayılmıştır. Bununla birlikte, İnsan Genom Projesi'nin tamamlanmasıyla birlikte protein sentezinde görev alan genlerin oranının yaklaşık olarak %1.5 olduğu belirlenmiştir [9] ve protein sentezinde görev almayan çok sayıda uzun kodlama yapmayan RNA geni keşfedilmiştir [10,11]. İnsan Genom Projesinden sonraki yapılan çalışmalar ile protein sentezlemeyen genlerinde çok sayıda önemli fonksiyonel özelliklere sahip olduğu belirlenmiş ve ayrıca protein sentezlemeyen genlerinde protein sentezleyen genler kadar insan biyolojisi için yaşamsal önemde oldukları keşfedilmiştir [12].

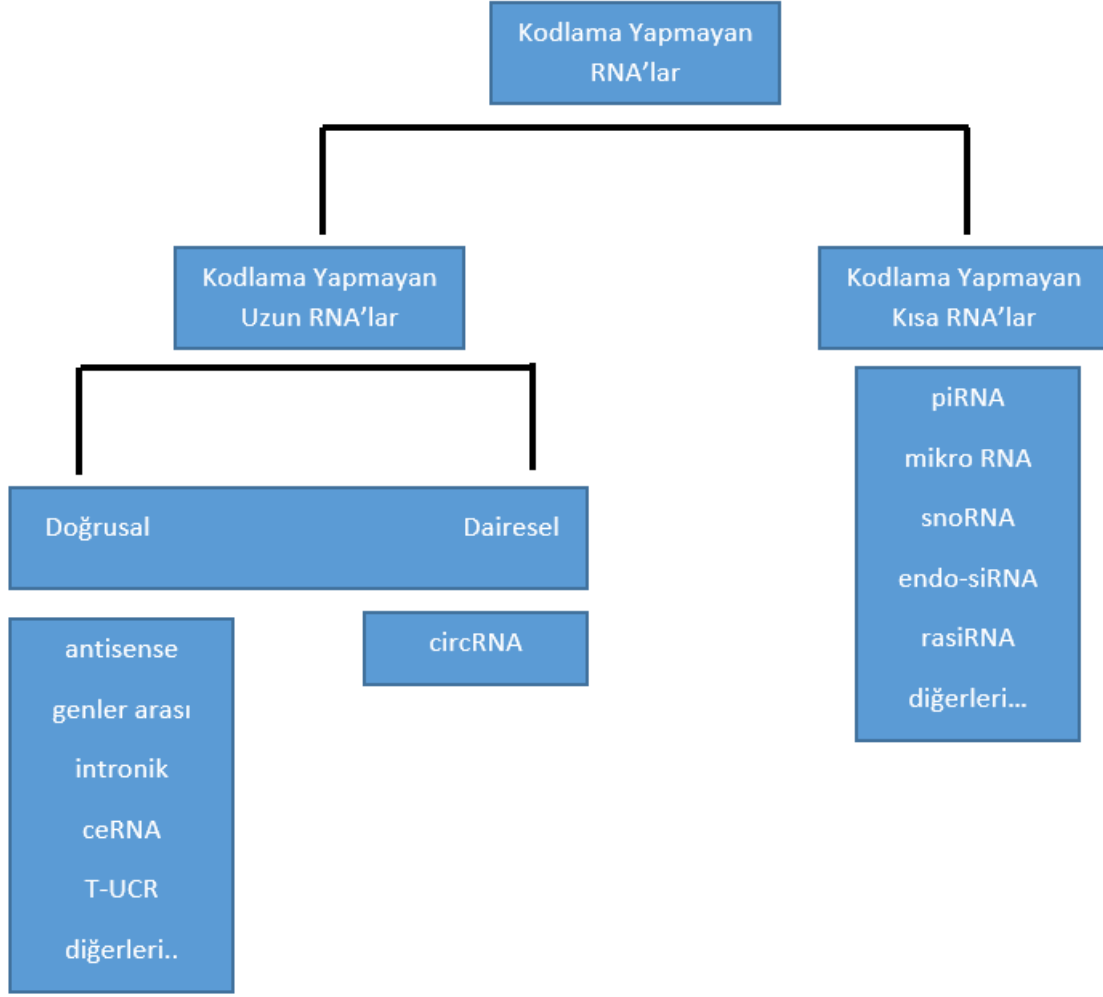
2.1. Protein Kodlaması Yapmayan RNA'lar

Organizmaların karmaşıklığı ve kompleksliği sahip oldukları gen sayısı ile orantılıdır varsayımı, karmaşık ve kompleks bir organizmaya sahip olan insan ile basit bir ökaryot olan *Caenorhabditis Elegans* yuvarlak solucanının yaklaşık olarak aynı sayıda protein sentezleyen gene sahip olmasının keşfedilmesinin ardından kabul görmemeye başlamıştır [13].

Organizmaların karmaşıklığının ve kompleksliğinin neye bağlı olarak oluştuğunu açıklamak için iki yeni teori öne sürülmüştür:

- Biyolojik karmaşıklık, genellikle genomun protein kodlaması yapmayan oranı ile ilişkilidir [14].
- Memeli genomunun yaklaşık olarak %1.5 arasında değişen bir kısmı mRNA'ları kodlarken genomun büyük çoğunluğu kısa ve uzun protein kodlaması yapmayan RNA'lar (ncRNAs) olarak transkripsiyon edilmektedir [2-10].

Bu teoriler RNA hakkındaki sadece DNA ve protein sentezi arasında aracı olduğu fikrini değiştirmekte ve genomun büyük çoğunluğunun, karmaşık organizmaların gelişimini yöneten işlevsel RNA türlerini şifrelediğini göstermektedir [15,16]. Yapılan çalışmalar ile RNA sinyallerinin ökaryotlardaki tüm kompleks genetik olguların merkezinde yer aldığı gösterilmiştir. Protein sentezine katılmayan RNA'lar (non-coding RNAs) Şekil 2.1'de gösterildiği üzere uzunluklarına bağlı olarak uzun veya kısa şeklinde sınıflandırılmakta ve genomik pozisyonlarına, yapılarına ve fonksiyonlarına göre alt gruplara ayrılmaktadır.



Şekil 2.1. Kodlama Yapmayan RNA'ların sınıflandırılması [17].

2.1.1. LncRNA'lar

LncRNA'lar genel olarak 200 nükleotidden uzun kodlama yapmayan RNA molekülleri olarak tanımlanmaktadır [17]. Hücrelerde ekspresyon edilen transkriptlerin haritalandırılması çalışmalarından elde edilen bilgiler büyük sayıda lncRNA varlığını göstermektedir [18,19]. Proteinlere sentezinde görev almamalarına rağmen lncRNA'lar fonksiyonel moleküllerdir. Giderek artan sayıda kanıt; gen baskılanması [20], hücre farklılaşması ve gelişimi [21], antiviral cevap [22] ve bitkilerde vernalizasyon [23] gibi birçok hücresel süreçte lncRNA fonksiyonlarının etkili olduğunu göstermektedir. Ek olarak; lncRNA'ların epigenetik düzenleme, kromatinin yeniden modellenmesi, gen transkripsiyonu, protein taşınımı, organ veya doku gelişimi, hücresel taşınım ve metabolik işlemler gibi bir çok biyolojiksel işlemde

önemli kritik rollere sahip olduğunu gösteren kanıtların sayısında giderek artmaktadır [24-27]. Önemli fonksiyonel özelliklere sahip olan lncRNA'lar da meydana gelen mutasyonların ve düzensizliklerin farklı kompleks insan hastalıklarının ortaya çıkması ve gelişimi ile ilişkili olduğunu gösteren kanıtların sayısında da artış gözlemlenmektedir [28-30]. Tüm bu çalışmalar, lncRNA'ların oldukça heterojen oldukları ve uzun RNA molekülleri olarak farklı yapılara ve moleküler etkileşimlere uyma yeteneklerine dayanan önemli bir fonksiyonel yönlülüğe sahip olduğunu göstermektedir. Tablo 2.1'de lncRNA veri örnekleri gösterilmektedir.

Tablo 2.1. lncRNA veri örnekleri

ABCC6P1	AC019109.1	AC098973.2	AK022798	AK094457
AC005682.8	AC022596.2	AC144449.1	AK027193	AK095147
AC006050.3	AC022596.6	ADAMTS9-AS2	AK028326	AK125699
AC006369.2	AC022816.2	AF085935	AK055347	AK126393
AC007879.5	AC023115.3	AF086415	AK056098	AK126698
AC012668.2	AC074363.1	AF118081	AK056155	AK293020
AC015818.3	AC083843.1	AFAP1-AS1	AK058003	AK294004
AC018647.3	AC092835.2	AK001796	AK093733	AL049354.1

2.1.2. MikroRNA'lar

MikroRNA'lar (miRNA) transkrip uzunlukları 21-25 arası olan nükleotidlerdir ve RNA'ların küçük bir alt sınıfıdır. [31]. Deney ve hesaplama yaklaşımları kullanılarak miRNA'ların işlevlerini anlama konusunda hızlı ilerleme sağlanmış olsa da miRNA'ların insan hastalıklarında ki etki tarzını yöneten genel ilkeleri belirleme konusunda hala önemli bir çaba gösterilmesi gerekmektedir. MiRNA'ların hareket mekanizmaları ve biyo oluşumları hakkındaki bilgiler son on yılda artış göstererek, miRNA'ların gen ekspresyonunun düzenlenmesi üzerindeki rolleri keşfedilmiştir. Bu sayede miRNA'ların bir çok patolojik

işlem üzerinde etkin olduğu belirlenmiştir. Bu patolojik işlemler; kanser, kalp hastalığı, diyabet, zihinsel hastalığı ve virüs enfeksiyonu gibi bir çok hastalığı içermektedir [32]. Yüksek çıktılı genomik ve biyoinformatik çalışmaları sayesinde, miRNA araştırmaları laboratuvar çalışmalarından klinik çalışmalara geçişi sağlayacak şekilde konumlandırılmakta ve bu da miRNA'ların çok sayıda hastalık için önemli tedavi edici hedefler olarak kullanılabileceğini göstermektedir [33]. MiRNA'lar, protein kodlayan genlerin protein sentezlemesini, translasyonu engelleyerek veya mRNA'yı baskılayarak önleyebilmektedir [34]. Bununla birlikte, her bir miRNA'nın gen düzenleme mekanizmasına etkisi bilinmemektedir. MiRNA'ların hedeflerini belirleme ve miRNA'ların ekspresyon seviyesinin onlarca veya yüzlerce mRNA'nın işleyişi üzerindeki kontrol mekanizması ile ilgili olarak birçok deneysel yaklaşım mevcuttur. Yapılan çalışmalar, miRNA'ların memeli transkriptomunun önemli bir bölümünü düzenlemekte olduğunu ve miRNA'ların düzenleyici ağlarda etkin bir şekilde organize olduklarını göstermektedir [35]. Bu karmaşık gen düzenleyici ağların ve bunların normal ve patolojik durumlardaki rollerinin araştırılması konusunda daha fazla sayıda çalışma yapılmaktadır. Tüm bu yapılan çalışmalar; miRNA'ların hücre büyümesi, farklılaşması, çoğalması, sonlanması, sinyal iletim yolu vb. fizyolojik süreçlerde önemli rollere sahip olduklarını göstermektedir [36]. Tablo 2.2.'de miRNA vary örnekleri gösterilmektedir.

Tablo 2.2. MiRNA veri örnekleri

miR-21	miR-100	miR-186	miR-30e-5p	miR-186
miR-212	miR-101	miR-197	miR-101	miR-197
miR-451	miR-101	miR-19a	miR-101	miR-19a
miR-20a	miR-122a	miR-20a	miR-122a	miR-20a
miR-21	miR-141	miR-20b	miR-141	miR-20b
miR-26a	miR-15b	miR-29b	miR-15b	miR-29b
miR-26a	miR-16	miR-29b	miR-16	miR-29b
miR-34a	miR-16	miR-30b	miR-16	miR-30b

2.1.3. Çevresel Faktörler

Birçok hastalığın kökeninin sadece genetik veya çevresel faktörlerle açıklanamayacağı aksine ikisi arasındaki etkileşimlerin bilinmesinin hastalıkların kökeninin anlaşılması için gerekli olduğunu gösteren çalışmaların sayısında sürekli artış gözlemlenmektedir. Sigara dumanı, radyasyon, çevre kirliliği, ilaçlar gibi fiziksel, biyolojik ve kimyasal çevresel faktörlere maruz kalmanın birçok hastalığın oluşumu üzerinde etkili olduğunu gösteren çalışmalar, çevresel faktörlerin hastalıklar üzerinde etkili olduğunu göstermektedir [37]. Örneğin, insan vücudunda ve üzerinde bulunan 10^{14} bakteri, yaklaşık 1 milyon genlik bir metagenom barındırarak, mikrobiyomu oluşturmaktadır. Mikrobiyom, insan bağışıklık sisteminin gelişimi ve sürdürülebilirliği açısından önemlidir. Yapılan çalışmalar; mikrobiyomun obezite, damar tıkanıklığı ve iltihaplı hastalıklar ile arasında ilişki olduğunu göstermektedir [38]. Başka bir çevresel faktör olan psikososyal stres; hormonal dengeyi değiştirerek oksidatif hasar vermekte, hücrelerin yaşlanmasını hızlandırmakta, koroner kalp hastalığı riskini arttırmaktadır [39]. Tüm bu yapılan çalışmalar çevresel faktörlerin hastalıklar üzerinde etkili olduğunu göstermektedir. Tablo 2.3’de bazı çevresel faktör örnekleri gösterilmiştir.

Tablo 2.3. Çevresel Faktör örnekleri

Sisplatin	Kurkumin	Oksijen yetmezliği	Gemsitabin	Hava kirliliği
İnterlökin-1 β (IL-1 β)	Radyasyon	Doksorubisin	ısı şok	Gefitinib
Dihidrotestosteron	BPDE	mitomisin C	Hepatit C Virüs	Dosetaksel
Deksametazon	Resveratrol (antioksidan madde)	5-florourasil	Prednizolon	İnsülin
Fötal serum	klopidogrel	paklitaksel	fentanil	HPV16
trombosit transfüzyonu	Yüksek glikoz	Glikoz	Düşük doz iyonlaştırıcı radyasyon	β -Elemene
Hepatit B Virus	AKT durdurucu (GDC-0068)	Platin	Arsenik	Oxaliplatin
5-Aza-2-deoksisitidin	PI3K durdurucu (LY294002)	γ -ray ışın	jenistein	karboplatin

2.2. Etkileşimler

2.2.1. LncRNA'ların Hastalıklar Üzerindeki Etkileri

Giderek artan sayıda çalışma; lncRNA'ların hastalık teşhisi, tedavisi ve seyri için potansiyel biyobelirteç ve aynı zamanda ilaç keşfi ve klinik tedavi için potansiyel ilaç hedefi işlevi olarak kullanılabileceğini göstermektedir. LncRNA'ların hastalıkların üzerindeki etkilerini içeren literatür çalışmalarının bir kısmı aşağıda gösterilmiştir.

Göğüs Kanseri: Göğüs kanseri dünya genelinde kadınlar arasında görülen en yaygın kanser türlerinden bir tanesidir [40]. Yapılan çalışmalar, moleküler değişikliklerin göğüs kanserinin nedenleri arasında yer alabileceğini göstermektedir. Örneğin bazı lncRNA'ların ekspresyon düzeylerinin yükselişi göğüs kanseri hücrelerinde karsinojenite artışına neden olabilmektedir [41]. Yapılan bir çalışma [42] ile H19 lncRNA'sının birincil göğüs kanserlerinde büyük etkilere sahip olduğu belirlenmiştir. BC200 lncRNA'sının göğüs kanserinde ekspresyon edildiği keşfedildikten sonra bu lncRNA'nın ekspresyon düzeyi tümör gelişimini tahmin etmekte kullanılmaktadır [43]. Ek olarak, CDKN2B-AS1 lncRNA'sı insan göğüs tümörlerinde p14/ARF ile birlikte kümelenmiş olarak ekspresyon edilmektedir [44]. GAS5 lncRNA'sının normal göğüs dokusu ile karşılaştırıldığında göğüs kanserindeki transkripsiyon seviyesinin düşük olduğu belirlenmiştir [45]. XIST lncRNA'sının göğüs kanseri ile yakından ilişkili olduğu yapılan başka bir çalışmada [46] gösterilmiştir.

Akciğer Kanseri: Akciğer kanseri dünyada kansere bağlı ölüm oranları arasında ilk sıradadır [47]. BCYRN1 lncRNA'sının akciğer, göğüs, serviks, yemek borusu, yumurtalık ve dil kanseri dokularında ekspresyon edildiği ancak karşılık gelen normal dokularda ekspresyon edilmediği yapılan bir çalışma [48] ile göstermiştir. H19 lncRNA'sının akciğer kanseri ile ilişkili olduğu saptanmıştır [49]. Ek olarak, tümör baskılayıcı bir lncRNA çeşidi olan GAS5'in akciğer kanseri dokularında azaldığı saptanmıştır [50].

Hepatosellüler Kanseri: Dünyada kanser ölümlerinin üçüncü önde gelen türü hepatosellüler kanserdir [51]. Bu hastalığın nedenleri arasında hepatit B virüsü (HBV) veya hepatit C virüsü (HCV) ile enfeksiyon, aflatoksin B1 alımı, alkol tüketimi, alkolsüz yağlı karaciğer hastalığı ve bazı kalıtsal hastalıklar olduğu bilinmektedir [52]. Son yıllarda yapılan çalışmalar, lncRNA'ların karaciğer kanserine neden olan etmenler arasında olduğunu göstermektedir [53]. HOTAIR, LALR ve HULC lncRNA'ları hepatosellüler kanserdeki farklı yolların

düzenleyicilerini hedef alarak yayılımını etkileyebilmektedir [54]. Özellikle HULC lncRNA'sının ekspresyon seviyesinin azalması hepatosellüler kanser ile ilişkili birkaç gende önemli anormallik artışına neden olmaktadır [55]. Ayrıca, ATB lncRNA'sının hepatosellüler kanser hastalığının seyrinin ölçülmesi ile ilişkili olduğu saptanmıştır [56]. Ek olarak, MEG3 lncRNA'sının normal hücrelere kıyasla insanlardaki dört farklı hepatosellüler kanser türünde azaldığı keşfedilmiştir [57].

Alzheimer hastalığı: Yapılan araştırmalar Alzheimer hastalığının dünyada giderek arttığını göstermektedir [58]. Alzheimer, hipokampusun CA1 bölgesi gibi spesifik beyin bölgelerindeki sinaps ve nöronların kaybindan kaynaklanan kronik ilerleyen bir nörodejeneratif hastalıktır [59]. Birçok araştırma, BACE1-AS ve BC200 gibi lncRNA'ların Alzheimer hastalığı ile yakından ilişkili olduğunu göstermektedir. Örneğin, BACE1-AS lncRNA'sının Alzheimer hastalığında beta sekretazı hızlı bir şekilde düzenleyebildiği belirlenmiştir [60]. Ayrıca, yaşla eşleştirilen normal beyinlerle karşılaştırıldığında Alzheimer hastalarının beyin bölgelerinde BC200 lncRNA'sının anlamlı şekilde arttığı gösterilmiştir. Aynı zamanda, Alzheimer hastalığının ilerlemesi ile BC200 lncRNA'sının ekspresyon düzeyinde de artma eğiliminde olduğu belirlenmiştir [61].

Kalp Krizi: Kalp krizi, yüksek ölüm oranına neden olan karmaşık bir klinik olgudur [62]. Yapılan çalışmalarda, bazı lncRNA çeşitlerinin kalp krizi ile ilişkili olduğu ve aynı zamanda kritik fonksiyonlara sahip olduğu gösterilmiştir [63]. Örneğin, dokuya özgü bir lncRNA çeşidi olan Fendrr, kalp gelişiminin temel bir düzenleyicisidir [64]. Ayrıca, bir lncRNA çeşidi olan Nkx2-5 kalp fonksiyonelliğin bozulması üzerinde etkili olan miyotonik kas bozukluğu RNA toksisitesinin genetik düzenleyicisidir [65]. Ek olarak, bir lncRNA çeşidi olan LIPCAR kalp krizinin ilk aşamasından sonra azalmakta ancak sonraki aşamalarında artış göstermektedir. Bu nedenle, LIPCAR kalp krizi geçiren hastalarda hastalığın seyrinin tahmininde kullanılabilmektedir [66].

2.2.2. MiRNA-Çevresel Faktör Etkileşimlerinin Hastalıklar Üzerindeki Etkileri

Hastalıkların ortaya çıkmasında bir çok risk faktörü vardır. Çevresel faktörler ve miRNA'lar bu riskler arasındadır. Çevresel faktörler ile miRNA'ların ayrı ayrı risk faktörü taşımalarının yanında ikisi arasındaki etkileşimlerde risk faktörü taşımaktadır çünkü insan dokularındaki miRNA'ların ekspresyon düzeylerinin çevresel faktörler tarafından değiştirilebilmesi nedeniyle hastalıkların altında yatan mekanizmaların moleküler düzeyde

anlaşılabilmesi açısından önem taşımaktadır [67]. Artan sayıda çalışma, miRNA'ların ilaç, virüs, alkol, sigara dumanı, stres, radyasyon gibi çevresel faktörlerin büyük bir kısmı ile fonksiyonel olarak ilişkili olduğunu göstermektedir [68]. Sonuç olarak, miRNA ile çevresel faktörler arasındaki etkileşimler anormal fenotiplerin ve hastalıkların oluşmasında kritik rollere sahiptir [69]. Yapılan bazı çalışmalar ile bu olgu gösterilmiştir. Örneğin, Kulshreshtha [70] yaptığı çalışmada oksijen yetmezliği ile bazı spesifik miRNA grupları arasında bir bağlantı olduğunu ve bunun da bazı tip insan kanserlerinde ekspresyon düzeyini artırdığını ve bu durumun tümör oluşumu üzerinde etkili olduğunu belirtmiştir. Yoğun sigara dumanına maruz kalmanın miRNA'lar ve çevresel faktörler arasındaki ilişkiyi araştırmak için yapılan bir çalışmada normal epitel ve akciğer kanseri hücrelerinde miR-487b'nin baskılanması üzerinde bir etkisinin olduğu gösterilmiştir [71].

2.2.3. Çevresel Faktörlerin Hastalıklar Üzerindeki Etkileri

Bazı çevresel faktörlere maruz kalmak, hastalıkların meydana gelmesinde önemli bir risk faktörü teşkil etmektedir. Çevresel faktörler ile hastalıklar arasında ilişki olduğunu gösteren çok sayıda çalışma yapılmıştır. Örneğin, yapılan bir çalışmada [72] bazı çevresel faktörlere maruz kalmanın kardiyovasküler hastalığın gelişmesi ve ilerlemesi ile ilişkili olduğu gösterilmiştir. Çevresel bir faktör olan alkol'ün bazı kanser çeşitlerinin gelişiminde risk teşkil ettiği yapılan başka bir çalışmada gösterilmiştir [73]. Virüsler ve bakteriler de bir çok kanser türü ile ilişkilidir. Örneğin, yapılan bir çalışmada bir bakteri türü olan *Helicobacter pylori* bakterisinin mide ülserlerine neden olduğu ve bununda mide ve yemek borusu kanserlerine sebep olduğu bildirilmiştir [74]. Kendler & Eaves [75] yaptıkları çalışmada bir çok psikiyatrik rahatsızlığın nedenini eksiksiz olarak anlamak için ilgili genetik risk faktörlerinin, ilgili çevresel risk faktörlerinin ve bu risk faktörlerinin hangi yollar ile etkileşimde olduğunun bilinmesine ihtiyaç olduğunu bildirmişlerdir. Andreassi yaptığı çalışmada [76] hem çevresel hem de genetik faktörlerin etkileşimlerinin metabolik sendrom ve tip 2 diyabet gelişimini ciddi şekilde etkileyebildiğini güçlü kanıtlarla göstermiştir. Kardiyovasküler hastalığın altında yatan kavramsal çerçeveler, hastalık koşulların çok sayıda genetik ve çevresel risk faktörünün varlığında ortaya çıktığını, bu faktörlerin arasındaki etkileşimlerin çeşitli düzeylerde gerçekleştiğini ve bu etkileşimlerin, hastalıkların ortaya çıkma ihtimali üzerinde önemli birincil etkiler oluşturduğu başka bir çalışmada [77] belirtilmiştir.

2.2.4. LncRNA-Çevresel Faktör Etkileşimlerinin Hastalıklar Üzerindeki Etkileri

LncRNA'lar çevresel kimyasallar, sigara dumanı ve hava kirliliği gibi bir çok çevresel faktör ile etkileşime girebilmektedir [78,79]. Çevresel faktörler nedeniyle ortaya çıkan bazı hastalıklar ile lncRNA'larında aynı zamanda ilişkili oldukları yapılan çalışmalarda gösterilmiştir [80]. Örneğin, PRINS lncRNA'sının ekspresyon düzeyinin viral enfeksiyon ile artabileceği ve bunun da sedef hastalığının ortaya çıkmasına katkı sunabileceği yapılan bir çalışmada gösterilmiştir [81]. Başka bir çalışmada, HOTAIR lncRNA'sının göğüs kanseri hücrelerinde (MCF7) bizfenol A konsantrasyonuna maruz kalmasının bu lncRNA'nın ekspresyon düzeyinde artış olabileceğini göstermektedir [82]. Yapılan başka bir çalışmada ise HSR1 lncRNA'sının ısı şokuna cevap olarak yapısal bir değişikliğe uğradığını ve bu nedenle ısı şoku tepkisi sürecini aktive ettiğini göstermektedir [83].

3. VERİTABANLARI VE HESAPLAMALI MODELLER

3.1. LncRNA Veritabanları

LncRNA ekspresyon bilgileri ve lncRNA-hastalık bağlantılarını içeren çok sayıda veritabanı oluşturulmuştur. Bunlar arasından bazıları aşağıda gösterilmiştir.

3.1.1. NONCODE Veritabanı

NONCODE veritabanı, hemen hemen tüm geleneksel ncRNA sınıflarını (tRNA'lar ve rRNA'lar hariç) içeren entegre bir veritabanıdır ayrıca lncRNA ekspresyon profillerini ve fonksiyonlarını da içermektedir. NONCODE 2016 sürümünde lncRNA'ların toplam sayısı 527336'dır. İnsanlar ve fareler hakkında sırasıyla 167150 ve 130558 lncRNA bilgisi içermektedir. NONCODE veritabanı, lncRNA-hastalık bağlantılarını da içermektedir [84].

3.1.2. lncRNome Veritabanı

lncRNome insanlar özelinde 17000'in üzerinde lncRNA kaynağı içermektedir. Her bir lncRNA için bir kaç farklı özellik bulundurmaktadır. Bu özellikler; kromozom lokasyonu, türü, biyolojik fonksiyon tanımı ve hastalık ilişkisi olarak sıralanmaktadır. Kullanıcılar istenilen lncRNA'nın ismini girerek o lncRNA'ya karşılık gelen bilgilere ulaşabilmektedir. Ek olarak, metilasyon ve histon modifikasyonu, tek nükleotid polimorfizmleri, miRNA bağlanma alanları ve entegre lncRNA-protein etkileşim bilgilerini içermektedir [85].

3.1.3. lncRNAdb Veritabanı

lncRNAdb veritabanı kolay erişilebilir ve aranabilir bir formatta, ökaryot organizmaların lncRNA bilgilerini özetlemektedir. Her bir lncRNA'nın nükleotid dizileri, genomik bağlam, gen ekspresyonu verileri, yapısal bilgileri, hücre altı lokalizasyonu, literatür referanslı fonksiyon bilgilerini içermektedir [86].

3.1.4. LncRNADisease Veritabanı

LncRNADisease veritabanı, 500 yayından elde edilmiş 221 hastalık ve 321 lncRNA bilgisine ek olarak 1000 lncRNA-hastalık kaydı ile 475 lncRNA etkileşim kayıtlarını içermektedir. Etkileşim kayıtlarında; lncRNA'ların protein, RNA, DNA ve miRNA'lar ile farklı düzeylerde olan etkileşim bilgilerini içermektedir. LncRNADisease veritabanı; kullanıcıların lncRNA-hastalık bilgileri ve lncRNA etkileşim bilgileri üzerinde arama yapmalarına, verileri kişisel bilgisayarlarına aktarmalarına ve yeni kayıtlar yüklemelerine olanak vermektedir. Ek olarak, kullanıcıların literatürde olan hesaplamalı modellerine dayanarak lncRNA-hastalık etkileşim tahminlerini elde edebilmelerine olanak sağlamaktadır [87].

3.1.5. Lnc2Cancer Veritabanı

Lnc2Cancer manuel olarak oluşturulmuş bir veritabanıdır. Kanser ile alakalı lncRNA'ların fonksiyonlarını ve mekanizmalarını keşfetme amacıyla bilgi sağlayan entegre bir veritabanıdır. 1500'den fazla literatür çalışmasından elde edilmiş 579 insan lncRNA'sı ile insanlarda görülen 93 kanser türü arasındaki bağlantı bilgilerini içermektedir. lncRNA ve kanser isimleri, lncRNA ekspresyon yapısı, deneysel teknikler, kısa fonksiyonel tanımlar, orijinal referans bilgileri Lnc2Cancer veritabanı tarafından sağlanmaktadır [88].

3.2. MiRNA Veritabanları

3.2.1. MiREnvironment

MiREnvironment veritabanı miRNA'lar ile çevresel faktörler arasındaki bağlantıları içermektedir. MiREnvironment veritabanının son sürümünde 557 tane literatür çalışmasından 24 tür için elde edilmiş 1242 miRNA, 394 çevresel faktör ve 304 tane fenotip kaydı bulunmaktadır [89].

3.2.2. PEMDAM

MiRNA'lar ile çevresel faktörler arasındaki bağlantılar hastalıkların moleküler düzeydeki etkileşimleri nedeniyle önem arz etmektedir. PEMDAM veritabanı, üç ikili bağlantı ağını

kurarak çevresel faktörler ile miRNA'lar arasındaki bağlantı bilgilerini elde etmektedir. Bu üç ikili ağ: Çevresel faktörler-miRNA ağı, Çevresel faktörler-hastalık ağı ve miRNA-hastalık ağıdır. Çevresel faktörler ile miRNA'lar arasındaki bağlantılar toksikoloji mekanizmaları ve hastalık etiyolojilerini anlamayı kolaylaştırabilmektedir [90].

3.3. LncRNA Çevresel Faktör Veritabanları

3.3.1. DLREFD

DLREFD veritabanı çevresel faktörler ile lncRNA'lar arasındaki bağlantıların kayıtlarını içermektedir. 475 lncRNA ile 153 çevresel faktör arasında 835 bağlantı kaydı bulunmaktadır [91]. Her kayıt lncRNA ile ilgili bilgiler içermektedir.

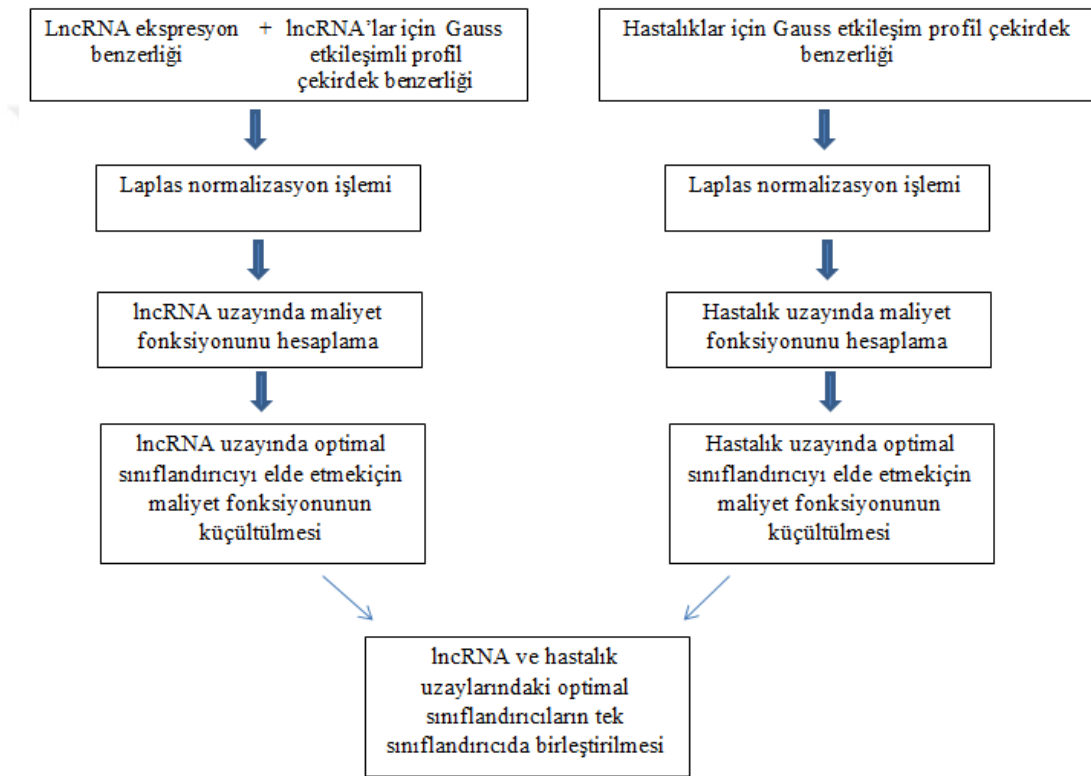
3.4. LncRNA-Hastalık Modelleri

LncRNA'lar üzerinde meydana gelen mutasyonların ve düzensizliklerin çeşitli hastalıklar ile yakın bir ilişkide olduğunu gösteren çalışmaların sayısında artış görülmektedir dolayısıyla lncRNA'ların fonksiyonlarının ve hastalıklar ile olan bağlantılarının anlaşılması konusu daha fazla önem teşkil etmektedir [92-94]. LncRNA'ların fonksiyonlarının anlaşılmasında ve lncRNA-hastalık bağlantılarının ortaya çıkarılmasında özellikle hesaplamalı modeller daha fazla etkili olabilmektedir. LncRNA-hastalık ilişkilerini hesaplamak için geliştirilen bazı hesaplama modellerinin bilgileri sonraki kısımda verilmiştir.

3.4.1. LncRNA-Hastalık İlişkileri için Düzenlenmiş Laplas En Küçük Kareler Yöntemi

Chen, potansiyel lncRNA-hastalık ilişkilerinin tahmin edilmesi amacıyla Yarı Gözetimli Öğrenme yapısına dayalı bir hesaplama modeli olan LRLSLDA'yı geliştirmiştir. LRLSLDA; benzer hastalıkların fonksiyonel olarak benzer lncRNA'lar ile bağlantılı olacağı varsayımına dayanan ve hastalık-lncRNA bağlantılarını tahmin etmek için geliştirilen bir modeldir. LRLSLDA modelinde, hastalık-lncRNA arasındaki potansiyel bağlantıların hesaplanabilmesi amacıyla bilinen hastalık-lncRNA bağlantıları ile lncRNA ekspresyon özellikleri birbirine entegre edilmektedir. Şekil 3.1.'de LRLSLDA modelinin akış şeması gösterilmektedir. LRLSLDA tekli çıkarımlı çapraz doğrulama (leave-one-out cross validation) tekniğinde

0.7760 AUC değerini elde ederek benzer hesaplamalı biyolojik problemlerin çözümünde kullanılan önceki metodların performansı ile karşılaştırıldığında daha yüksek bir başarı göstermiştir. Ek olarak, LRLSLDA negatif örneklerin bilgisine ihtiyaç duymamaktadır. LRLSLDA modelinin eksik yönleri; modelin çok fazla parametre içermesi ve bu parametlerin hangi kriterlere göre seçileceğinin açık bir şekilde çözümünün gösterilememesi ve ayrıca aynı hastalık-lncRNA çifti için lncRNA ve hastalık uzayında iki farklı skor elde edilebilmesi olarak sıralanmaktadır [95].

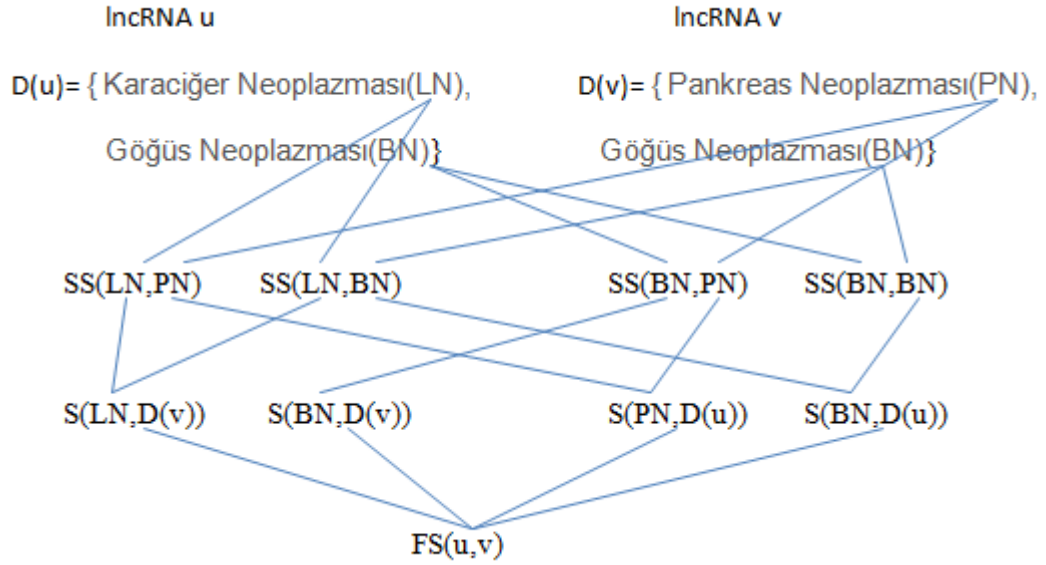


Şekil 3.1. LRLSLDA'nın akış diyagramı

3.4.2. LRLSLDA-LNCRNA Fonksiyonel Benzerlik Hesaplama Modeli

Fonksiyonel olarak benzerlik gösteren lncRNA'ların genel olarak benzer hastalıklar ile bağlantılı olduğu varsayımını temel alan Chen, hastalık gruplarının aralarındaki ilişkilerinin semantik benzerliklerini hesaplamaya dayalı iki yeni model geliştirmiştir. Bu iki model (LNCSIM1 VE LNCSIM2) arasındaki farklılık, hastalıklar üzerindeki yönlü çevrimsiz

çizgesel'i temel alarak, hastalıkların semantik benzerliklerinin hesaplanmasında ki teknik farklılıklardan kaynaklanmaktadır. Şekil 3.2.'de LNCSIM modelinin akış şeması gösterilmiştir [96].



Şekil 3.2. İki lncRNA arasındaki fonksiyonel benzerliğin hesaplanmasını temel alan LNCSIM modeli: lncRNA u ve v ile aralarındaki fonksiyonel benzerliğin hesaplanması.

3.4.3. Naif Bayes Sınıflandırıcı

Zhao, kanser ile bağlantıları bilinen lncRNA'ları kullanarak kanser ile bağlantılı olabilecek yeni lncRNA'lar keşfetmek amacıyla genomik, çoklu-omik ve transkriptom verilerinin entegrasyonuna dayalı bir Naif Bayes (Naïve Bayes) sınıflandırıcı modeli geliştirmiştir. Geliştirilen model etkili bir performans göstererek kanser ile ilişkili olabilecek 707 yeni lncRNA bilgisini ortaya çıkartmıştır. Geliştirilen modelde kullanılan denetimli öğretici (Naif Bayes Sınıflandırıcı) negatif örneklerin bilgilerine ihtiyaç duymaktadır ancak bu bilinmemektedir. Bu nedenle kullanılan veri bilgileri arasında negatif örnek verileri yerine rastgele etiketsiz lncRNA-hastalık çiftleri arasından seçilen veriler kullanılmaktadır ve bu da tahmin performansını önemli ölçüde etkilemektedir [97].

3.4.4. RWRIncD

Fenotipik olarak benzer hastalıklar ile fonksiyonel olarak benzer lncRNA'ların birbirleri ile ilişki kurma eğiliminde olduğu varsayımı temel alınarak, Sun tarafından lncRNA-lncRNA fonksiyonel benzerlik ağına dayanan global ağ-tabanlı bir hesaplama modeli olan RWRIncD geliştirilmiştir. RWRIncD modeli; lncRNA-hastalık bağlantı ağı, hastalık benzerlik ağı ve lncRNA fonksiyonel benzerlik ağı oluşturulduktan sonra lncRNA fonksiyonel benzerlik ağı üzerinde yeniden başlatmalı rassal yürüyüş (RWR) algoritmasının uygulanması ile potansiyel lncRNA-hastalık bağlantılarının ortaya çıkarılabileceği fikrine dayanmaktadır. RWRIncD, lncRNA'lar ile hastalıklar arasında, deneysel olarak kanıtlanmış bağlantılı veriler üzerinde tekli çıkarımlı çapraz doğrulamada 0.822 AUC değerini elde etmektedir. Bununla beraber, RWRIncD modelinde lncRNA'lar ile bağlantıları bilinmeyen hastalıklar üzerinde uygulama yapılamamaktadır. RWRIncD'in tahmin performansının, gelecekte, lncRNA-hastalık bağlantı sayısının artış göstermesi ve lncRNA fonksiyonel benzerlik ölçümlerinin daha kesin bulunabilmesi ile artacağı tahmin edilmektedir [98].

3.4.5. İki Parçalı lncRNA-PCG Ağı Üzerinde Yeniden Başlatma ile Rastgele Yürüyüş

Liu, protein etkileşim veri kümeleri ve prostat kanserindeki PCG (protein kodlayan genler) ekspresyon profillerini temel alarak iki parçalı PCG-lncRNA ağı oluşturmuştur. Ek olarak RWR temel alınarak kanser ile bağlantılı olabilecek lncRNA'lar tahmin edilmeye çalışılmaktadır. Bununla beraber, bu model eksik protein etkileşim veri kümelerinden ciddi bir biçimde etkilenmekte ve zayıf bir performans göstermektedir [99].

3.4.6. Hastalık Genleri Temel Alınarak lncRNA-Hastalık Tahmin Modeli

lncRNA ve genlerin ekspresyon düzeyleri ile hastalık genlerinin entegre edilmesi ile bilinen lncRNA-hastalık bağlantılarına bağlı kalmayan ilk hesaplamalı model Liu tarafından geliştirilmiştir. Bu modelde, lncRNA'lar iki kısma ayrılmaktadır: dokuya özgü ve dokuya özgü olmayan lncRNA'lar. Farklı dokulardaki lncRNA'ların ekspresyon düzeyleri temel alınarak dokuya özgün skorlar hesaplanmaktadır. Daha sonra, dokuya özgü lncRNA'lar için, bu lncRNA'lar ile bağlantılı olabilecek hastalıklar hesaplanmaktadır. Ek olarak, dokuya özgün olmayan lncRNA'lar ile hastalıklar arasındaki potansiyel ilişki tahminleri hastalık-gen bağlantıları ve gen-lncRNA ilişkileri temel alınarak hesaplanmaktadır. Bu model gen

bağlantısı olmayan hastalıklar için lncRNA ilişki tahmini yapamamaktadır. Model, LOOCV’de 0.7645 AUC değeri elde etmektedir [100].

3.5. LncRNA-Çevresel Faktörler Modelleri

LncRNA’lar ile çevresel faktörler arasında yeni potansiyel bağlantıların hesaplanması amacıyla sınırlı sayıda hesaplama modeli geliştirilmiştir. Bilinen ilk model, yeniden başlamaya dayalı rastgele yürüyüş temel alınarak geliştirilen RWREFD modelidir [101]. Bu model; eğer bir lncRNA ve çevresel faktör ortak bir şekilde bir miRNA ile ilişkili ise o zaman bu lncRNA ile çevresel faktörün bağlantılı olma ihtimali yüksektir hipotezine dayanmaktadır. Bu model ile tek-çıkışlı çapraz doğrulama yönteminde 0.71 AUC değeri elde edilmiştir. Diğer bir çalışma, iki parçalı ağlar üzerinde kaynak aktarma tekniği (resource transfer technique) kullanılarak gerçekleştirilmiştir [102]. Bu çalışmada, lncRNA’lar ve çevresel faktörler arasında yeni potansiyel bağlantılar tahmin edilerek, bu bağlantıları içeren bir veritabanı oluşturulmuştur ve daha önce oluşturulmuş başka bir veritabanı ile karşılaştırılmıştır. Zenginleştirme analiz tekniği (enrichment analysis technique) temel alınarak modellerinin önceki modele [101] göre daha başarılı sonuçlar elde ettiği belirtilmiş ancak bu çalışmada sonuçların değerlendirilmesi için AUC değeri kullanılmamıştır.

3.6. MiRNA-Çevresel Faktörler Modelleri

MiRNA’lar ile çevresel faktörler arasında yeni potansiyel bağlantıları bulmaya yönelik bazı hesaplamalı modeller geliştirilmiştir. Geliştirilen miREFRWR modelinde, miRNA ile çevresel faktörler arasındaki yeni potansiyel bağlantıların bulunması amacıyla miRNA-miRNA benzerlik ağı, çevresel faktör-çevresel faktör benzerlik ağı ve miRNA-çevresel faktör ağı olmak üzere üç tane ağ oluşturulmaktadır. İkinci olarak, miRNA benzerlik ağı ve çevresel faktör benzerlik ağı üzerinde sırasıyla rastgele yürüyüşler uygulanmaktadır. Üçüncü olarak, tahmin sonuçlarını elde etmek için miRNA benzerlik ağı ve çevresel faktör benzerlik ağı üzerinde rastgele yürüyüşe dayalı tahmin sonuçları birleştirilmektedir. Son olarak, yeni potansiyel miRNA-çevresel faktör bağlantı çiftleri rastgele yürüyüşün olasılık sabitine göre hesaplanmaktadır [103]. Başka bir çalışmada, 23 kanser içerisinde küçük molekül-miRNA etkileşim ağı kurularak daha sonra hipergometrik testler temel alınarak yeni potansiyel miRNA-çevresel faktör bağlantıları hesaplanmaktadır [104].

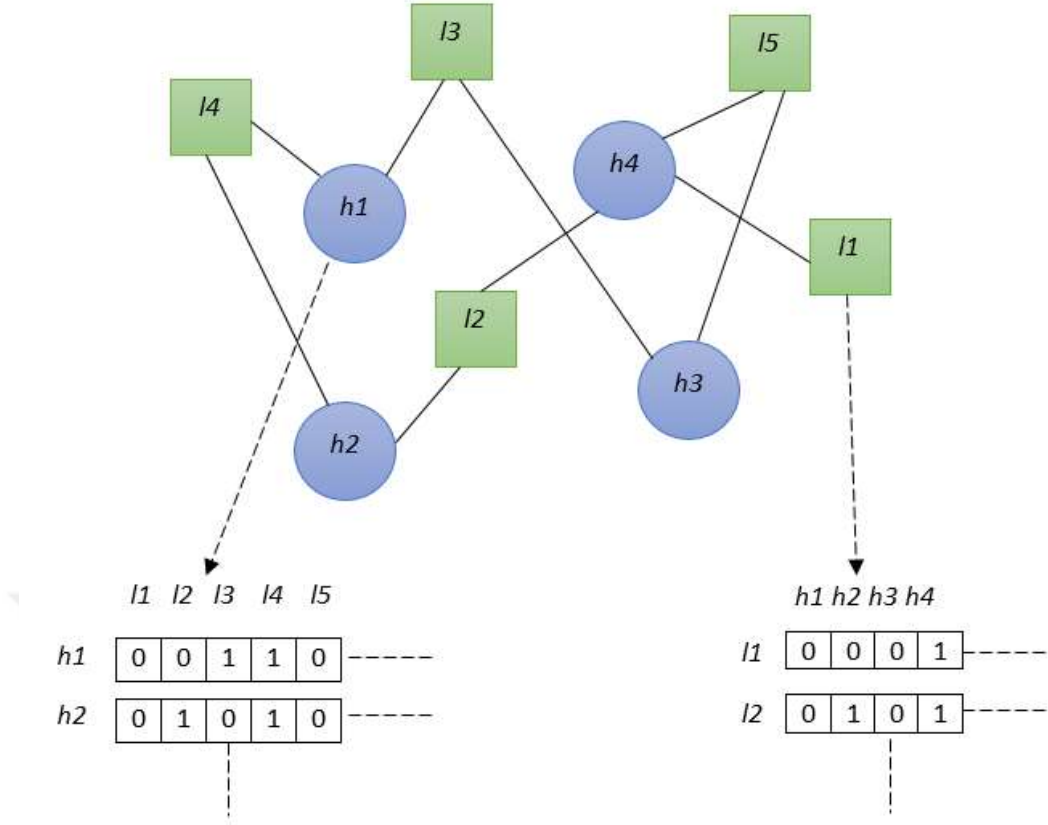
4.MATERYAL VE METOD

4.1. KATZ

KATZ graf-tabanlı bir modeldir ve daha önce sosyal ağlarda [105], ilaç-hedef etkileşimlerinde [106], hastalık-gen etkileşimlerinde [107] ve mikroRNA-hastalık etkileşimlerinde [108] yeni potansiyel bağlantıların tahmin edilmesi amacıyla kullanılmıştır. KATZ, sıralama algoritması temelinde çalışmaktadır. KATZ kullanılarak, heterojen bir ağda aralarında bağlantı olmayan iki düğümün benzerlik değeri, bu iki düğümü birleştiren farklı uzunluktaki yürüyüşler hesaplanarak elde edilmektedir. Elde edilen benzerlik değerleri büyükten küçüğe doğru sıralanmaktadır. En yüksek benzerlik değerleri bu iki düğüm arasında potansiyel bağlantı olup olamayacağının tahmininde kullanılmaktadır.

4.2. Gauss Etkileşim Profili Çekirdek Benzerliği

Gauss etkileşim profili çekirdek benzerliği ilk olarak ilaç-hedef etkileşimlerinin tahmin edilmesi çalışmasında kullanılmıştır [109]. Bu çalışmada aralarındaki bağlantıları bilinmeyen ilaç ve hedeflerin tahmin edilmesi amacıyla aralarındaki bağlantıları bilinen ilaç ve hedefler kullanılmıştır. İlaç-hedef bağlantı ağının topoloji bilgisinin analiz edilmesi amacıyla bir kernel oluşturulmuştur. Daha sonra, ilaç-hedef bağlantı ağında hedefler ile benzer şekilde bağlantılara sahip olan iki ilacın birbirleri ile benzer özelliklere sahip olduğu varsayımı temel alınarak bu iki ilacın diğer hedefler ile de benzer şekilde bağlantı kuracağı varsayılmıştır. Aynı varsayım ilaçlar ile benzer bağlantılara sahip olan iki hedefin de benzer özellikler göstereceğini temel almaktadır. Aynı hedefler ile benzer bağlantılara sahip iki ilacın benzer özellikler göstereceği varsayımı temel alınarak her ilaç, bir etkileşim profili olarak tanımlanmaktadır. Bu etkileşim profili, ikili bir vektör şeklinde tarif edilebilmektedir. Bir ilaç için ikili vektör şu şekilde oluşturulmaktadır; ilgili ilaç ile hedeflerin tümü arasında bağlantının olup olmaması bilgisi sırasıyla kontrol edilmekte ve bu bilgi vektördeki ilgili kısma yazılmaktadır. Bu işlem hedefler içinde aynı şekilde yapılmaktadır. Daha sonra elde edilen etkileşim profilleri ile Gauss etkileşim profil çekirdeği oluşturulmaktadır. Etkileşim profillerinin oluşturulma örnekleri aşağıdaki Şekil 4.1.'de gösterilmektedir.



Şekil 4.1. İlaç-hedef ağından hedefler için $h1, h2 \dots$ ve ilaçlar için $l1, l2 \dots$ etkileşim profillerinin oluşturulması.

Çalışmada etkileşim profillerinden bir çekirdek oluşturulmuştur. Bu çekirdek ilaç-hedef ağının topoloji bilgisini içermektedir. Kernel olarak Gauss çekirdeği seçilmiştir. Gauss çekirdeği d_i ve d_j ilacı için şu şekildedir.

$$KM(d(i), d(j)) = \exp(-\gamma_m \| IP(d(i)) - IP(d(j)) \|^2) \quad (4.1)$$

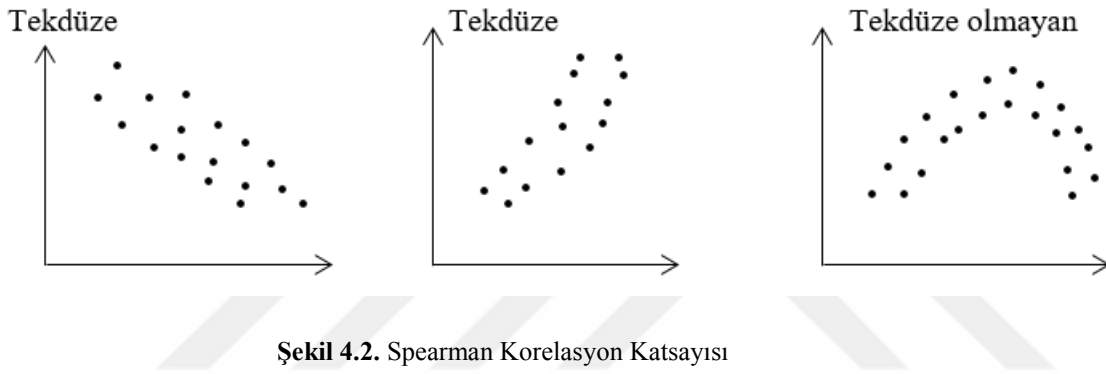
Formüldeki γ_m parametresi kernel bant genişliğini kontrol etmektedir. Bu parametre aşağıda gösterildiği şekilde hesaplanmaktadır.

$$\gamma_m = \gamma'_m / (\sum_{i=1}^{n_m} \| IP(d(i)) \|^2) \quad (4.2)$$

Yukarıdaki formül kullanılarak γ_m parametresi ilaç başına ortalama etkileşim sayısına bölünerek normalize edilmektedir. Bu seçim ile çekirdek değerlerinin veri setinin büyüklüğünden bağımsız olması sağlanmaktadır.

4.3. Spearman Korelasyon Katsayısı

Spearman Korelasyon katsayısı (Spearman Correlation Coefficient) biyoinformatik araştırmalarında bir çift arasındaki ekspresyon benzerliklerinin hesaplanmasında kullanılan genel bir yöntemdir [110].



Şekil 4.2. Spearman Korelasyon Katsayısı

Şekil 4.2.'de gösterilen tablolar da bulunan veriler üç farklı durumu göstermektedir. Birinci tabloda gösterilen Tekdüze durumda bir değişkenin değerinin artışına bağlı olarak diğer değişkenin değeri azalmaktadır. İkinci tabloda gösterilen Tekdüze durumda bir değişkenin değerinin artışına bağlı olarak diğer değişkenin değeri artmaktadır. Tekdüze olmayan durumda ise bir değişkenin artışına bağlı olarak diğer değişkenin artışı ve azalışı gösterilmektedir. Değişkenler arasındaki bu ilişkiyi saptamayı sağlayan Spearman Korelasyon Katsayısı algoritması Tekdüze durumlar için en optimum değer olan 1 veya 1'e yakın sonuçlar vermektedir. Tekdüze olmayan durumlarda ise 0 veya 0'a yakın sonuçlar vermektedir.

4.4. Hastalık Semantik Benzerliği

Hastalık Semantik Benzerliği (Disease Semantic Similarity) hesaplamalarında hastalıkların tıbbi konu başlıkları (MeSH, Medical Subject Headings) kullanılmaktadır. MeSH başlıkları Amerika Birleşik Devletleri Ulusal Tıp Kütüphanesi'nde [111] bulunan MeSH veritabanından

alınmaktadır. MeSH başlıkları 16 farklı kategoriye ayrılmaktadır: A kategorisi anatomik başlıkları, B kategorisi organizma başlıkları, C kategorisi hastalık başlıkları, D kategorisi ilaç ve kimyasal başlıkları içerecek şekilde 16 farklı kategori için başlık bilgileri bulunmaktadır.

4.4.1. Bir Hastalığın Semantik Değeri

Bir hastalığın semantik değeri hesaplanırken öncelikle bu hastalık için bir Yönlü Çevrimsiz Çizgesel (YÇÇ) oluşturulmaktadır. Örneğin bir A hastalığı için YÇÇ şu şekilde oluşturulmaktadır; $YÇÇ_A = (A, T_A, E_A)$. Bu çizgede T_A , A hastalığının kendisi ile birlikte tüm üst düğümlerini göstermekte, E_A ise bu düğümlere karşılık gelen bağlantıları göstermektedir. $YÇÇ_A$ çizgesindeki bir t hastalığının A hastalığının semantik değerine katkısı $H_A(t)$ şeklinde aşağıdaki denklemde gösterilmektedir:

$$\begin{cases} H_A(A) = 1 \\ H_A(t) = \text{maksimum} \{ \Delta * H_A(t') \mid t' \in t \text{ alt düğümleri} \} \text{ eğer } t \neq A \end{cases} \quad (4.3)$$

Yukarıdaki denklemde Δ , t hastalık düğümü ile bu düğümün alt düğümleri (t') arasındaki bağlantıların semantik katkı faktörünü göstermektedir. $YÇÇ_A$ çizgesinde, A hastalığı en yakın hastalık olduğundan semantik değerine katkısı 1 olarak kabul edilmektedir. A hastalığına yapılan semantik katkı üst düğümlere doğru azalmaktadır ve bunu sağlamak için Δ değeri 0 ile 1 aralığında seçilmektedir. Yukarıdaki denklem temel alınarak A hastalığının semantik değeri, $HS(A)$ olarak aşağıda şekilde hesaplanmaktadır:

$$HS(A) = \sum_{t \in T_A} H_A(t) \quad (4.4)$$

4.4.2. İki Hastalığın Semantik Benzerliği

İki hastalık arasındaki semantik benzerlik, bu hastalıkların MeSH veritabanındaki YÇÇ pozisyonlarının konumları kullanılarak hesaplanmaktadır. İki hastalığın semantik benzerliği aşağıdaki denklem de gösterilmektedir.

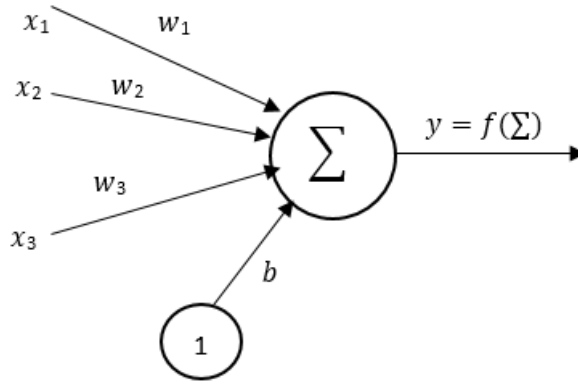
$$S(A,B) = \frac{\sum_{t \in T_{A \cap B}} (H_A(t) + H_B(t))}{HS(A) + HS(B)} \quad (4.5)$$

$H_A(t)$, t hastalığının A hastalığı ile olan semantik değerini ve $H_B(t)$ ise t hastalığının B hastalığı ile olan semantik değerini göstermektedir. Yukarıdaki formül ile iki hastalığın YÇÇ çizgilerindeki adresleri ve üst hastalık düğümleri ile olan semantik ilişkileri dikkate alınarak bu iki hastalığın semantik benzerlikleri hesaplanmaktadır.

4.5. Yapay Sinir Ağları

Yapay Sinir Ağları'nın (YSA) tarihi yarım yüzyıl öncesine kadar uzanmaktadır. Bu tarihten önce sinir ağları üzerinde çalışan araştırmacılar sadece beyindeki sinir sistemlerini araştırmaktaydılar. 1943 yılında McCulloch ve Pitts [112] bazı matematiksel kuralları temel alarak sinir ağları için bir hesaplamalı model geliştirmiştir. Bu sinir ağları ile ilgili bilinen ilk çalışma özelliği taşımaktadır ve daha sonra diğer araştırmacılar sinir ağlarını temel alarak farklı hesaplamalı modelleri geliştirmişlerdir. 1958 yılında Rosenblatt [113] beyindeki nöronları inceleyerek matematiksel olarak nöronu taklit edebilen algılayıcı (perceptron) isminde bir model geliştirmiştir. Bu ilk YSA'lerden biri olarak kabul edilmektedir. 1969 yılında yapılan bir çalışmada [114] YSA'ların eksik tarafları şu şekilde belirtilmiştir; çok fazla işlem gücü gerektirmesi, bilgisayarların gerekli hesaplamaları yapma konusunda yetersiz kalması ve ayrıca XOR (exclusive-or) işleminin uygulamasının YSA tarafından yapılamaması. Bu dezavantajlı durum nedeniyle YSA'lar ile ilgili yapılan çalışmaların sayıda azalış gözlemlenmiştir. 1970'li yıllar ile 1980'li yıllar arasında bir optimizasyon metodu olarak geriye yayılım (back-propagation) metodu geliştirilmiş ve YSA üzerinde uygulanmıştır. Bu sayede XOR problemi YSA ile çözülmüştür. Her ne kadar geriye yayılım metodu ile XOR problemine çözüm bulunmuş ise de bilgisayarlar daha hızlı işlem gücüne sahip olmadan önce çok katmanlı bir ağın hızlı bir sürede eğitilmesi mümkün olamamaktaydı. Bilgisayarların işlem gücünün artması ile birlikte YSA yeniden popülerlik kazanmaya başlamış ve YSA üzerine yapılan çalışmaların sayısında artış gözlemlenmiştir.

Nöronlar, bir yapay sinir ağındaki temel hesaplama birimleridir. Bir YSA'da tipik olarak çok sayıda nöron bulunmaktadır. Nöronlar kaynak veriden bilgi girişleri veya diğer nöronlardan aktivasyon değerleri alarak bağlantı ağırlıklarına göre tek bir çıkış verisi hesaplamaktadır. Bir nöron yapısı temel olarak aşağıdaki Şekil 4.3.'te gösterilmektedir.



Şekil 4.3. Bir nöronun temel yapısı.

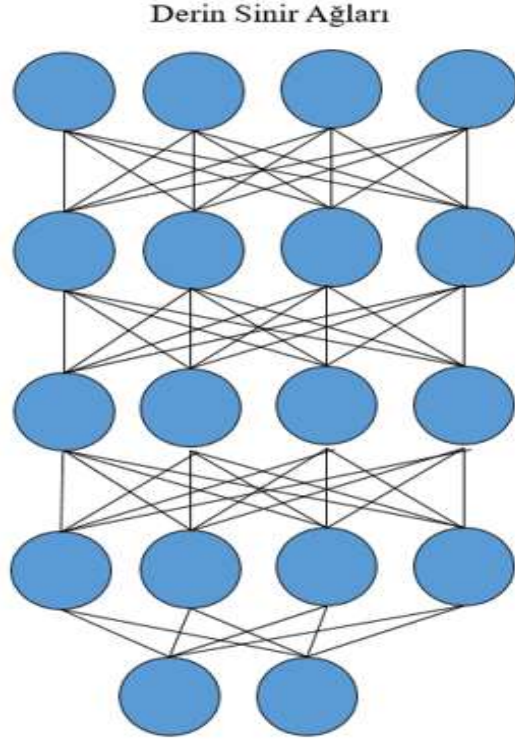
Yukarıdaki şekilde gösterildiği üzere sinyaller nörona farklı bağlantılar aracılığıyla bağlanır. Her bir bağlantının bir ağırlık değeri bulunmaktadır. Nöronun çıkışı ise; sinyaller ve bağlantıların ağırlık değerlerinin aşağıdaki fonksiyona göre hesaplanması ile elde edilmektedir.

$$\hat{y} = f(\sum_{i=1}^n w_i x_i + b) = f(w_1 x_1 + w_2 x_2 + w_i x_i \dots + w_n x_n + b) \quad (4.6)$$

Yukarıdaki formülde b nöronun eşik değerini (bias) göstermektedir. Eşik değeri genel olarak nörona gelen özel bir giriş sinyali olarak tanımlanmakta ve değeri 1 olarak kabul edilmektedir. Yukarıdaki formülde YSA'nın önemli bir ögesi olan aktivasyon fonksiyonu, f ile gösterilmektedir.

4.5.1. Derin Sinir Ağları

Derin Sinir Ağları, giriş ve çıkış katmanları arasında birden fazla gizli katman içeren ileri beslemeli (feed-forward) bir yapay sinir ağıdır. Her bir gizli birim, bir aktivasyon fonksiyon kullanarak önceki katmandan gelen toplam girişi skalar duruma çevirerek sonraki katmana aktarmaktadır. Aşağıdaki Şekil 4.4.'te birden fazla katmandan oluşan bir Derin Sinir Ağı örneği gösterilmiştir.



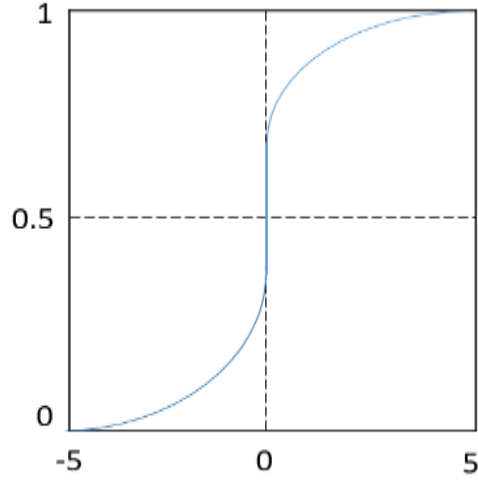
Şekil 4.4. Derin Sinir Ağları örneği

Derin sinir ağlarında kullanılan aktivasyon fonksiyonları, kayıp fonksiyonları, optimizasyon algoritmaları, gradyan inişi optimizasyon algoritmaları ve iletim sönümü hakkındaki bilgiler sırasıyla verilmiştir.

Sigmoid Fonksiyon: Sigmoid fonksiyon; giriş değerlerini alarak, (0,1) aralığında çıkış değerler üretilmesini sağlamaktadır. Aşağıdaki formülde gösterilmektedir.

$$f(x) = \frac{1}{1+e^{-x}} \quad (4.7)$$

Sigmoid fonksiyon aşağıdaki Şekil 4.5'te gösterilmektedir.

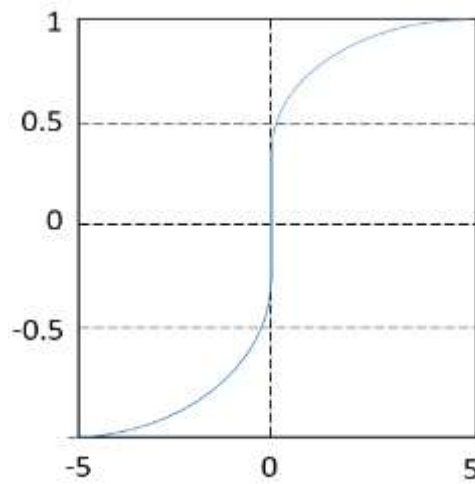


Şekil 4.5. Sigmoid fonksiyon

Hiperbolik Tanjant: Hiperbolik tanjant, tanh olarak adlandırılmaktadır ve giriş değerlerini $(-\infty, \infty)$ aralığından alarak $(-1, 1)$ aralığına çevirmektedir.

$$f(x) = \frac{2}{1+e^{-2x}} - 1 \quad (4.8)$$

Hiperbolik tanjant fonksiyonu aşağıdaki Şekil 4.6'ta gösterilmektedir.

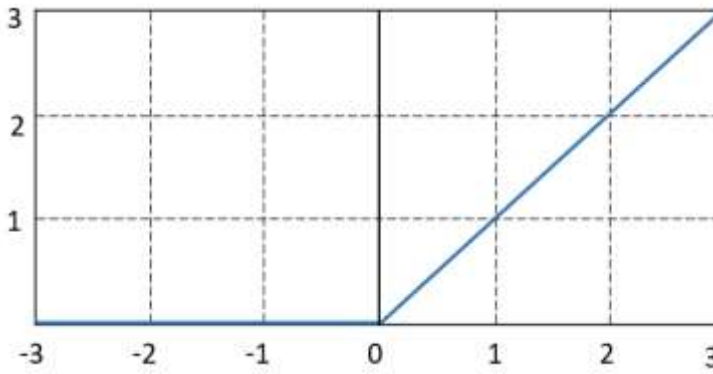


Şekil 4.6. Tanh fonksiyon

Düzeltilmiş Doğrusal Birim: Düzeltilmiş Doğrusal Birim (ReLU) aktivasyon fonksiyonu yapay sinir ağlarının eğitimini daha fazla geliştirmek için kullanılmaktadır [115]. Çalışma prensibi olarak, giriş değerlerini 0 eşik değerine göre hesaplamaktadır. Örneğin, giriş değerleri 0'dan küçük ise ($x < 0$) çıktı olarak 0 değeri üretilmekte ancak giriş değerleri 0'dan büyük veya eşit ise ($x \geq 0$) doğrusal bir fonksiyona göre çıktı üretilmektedir. Aşağıdaki formülde gösterilmektedir.

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (4.9)$$

ReLU fonksiyonu aşağıdaki Şekil 4.7'de gösterilmektedir.



Şekil 4.7. ReLU fonksiyon

4.5.1.1. Kayıp Fonksiyon

Derin sinir ağlarında bir veya daha fazla değeri gerçek bir sayıya eşleyen matematiksel bir fonksiyonu minimum veya maksimum yapmayı amaçlayan optimizasyon algoritmasına kayıp fonksiyon denilmektedir. Kayıp fonksiyonların türevi alınabilmektedir ve en çok kullanılan kayıp fonksiyonlar, Ortalama Karesel Hata (Mean Squared Error) ve Çapraz-Entropi (Cross-Entropy) fonksiyonlarıdır.

Ortalama Karesel Hata: Ortalama karesel hata fonksiyonu iki vektör arasındaki karesel uzaklığın ortalama değerini hesaplamaktadır. Regresyon hesaplama uygulamalarında yoğun

olarak kullanılmaktadır. Matematiksel olarak iki vektör $\vec{x} = [x_1, x_2, \dots, x_n]$ ve $\vec{y} = [y_1, y_2, \dots, y_n]$ ortalama karesel hata fonksiyonu ile aşağıdaki şekilde hesaplanmaktadır.

$$\text{Ortalama Karesel Hata } (\vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \quad (4.10)$$

Çapraz Entropi: YSA’da çapraz entropi hata fonksiyonu daha önceki çalışmada [116] şu şekilde tanımlanmıştır. N-elemanlı iki vektör için; hedef değerler $\vec{t} = [t_1, t_2, \dots, t_n]$ ve ağ çıkışı $\vec{y} = [y_1, y_2, \dots, y_n]$ olmak üzere bu iki vektör üzerinde çapraz entropi aşağıdaki şekilde hesaplanmaktadır.

$$E_n = -\frac{1}{n} [\sum_{k=1}^n (t_k \ln y_k + (1 - t_k) \ln(1 - y_k))] \quad (4.11)$$

Yukarıdaki formülde t_k hedef değeri ve y_k ise ağ çıkışını göstermektedir.

4.5.1.2. Optimizasyon Algoritmaları

Geriye yayılım (Backpropagation), genel olarak hatanın geriye doğru yayılım metodunu ifade etmektedir. Derin sinir ağında; eğitim setindeki kayıp fonksiyonu minimum yapmak için ağırlıkların güncellenmesi işlemi eğitim algoritmalarını oluşturmaktadır. Bu algoritmalar aşağıda açıklanmıştır.

Gradyan İnişi (Gradient Descent): Gradyan inişi, yapay sinir ağlarının optimizasyonunda kullanılan en yaygın metodlardan bir tanesidir. Gradyan inişi, yapay sinir ağlarında bir $J(\theta)$ kayıp fonksiyonunu θ parametrelerine bağlı olarak minimize etmek için kullanılmaktadır. Öğrenme oranı η , yerel bir minimum seviyeye ulaşmak için atılan adımların boyutunu belirlemektedir. Gradyan hesaplanmasında kullanılan veri miktarına bağlı olarak “Batch Gradient Descent”, “Mini-Batch Gradient Descent” ve “Stochastic Gradient Descent” olmak üzere üç farklı Gradyan İniş yöntemi bulunmaktadır.

Yığın Gradyan İnişi (Batch Gradient Descent): Yığın Gradyan İnişi tüm eğitim seti için θ parametrelerine bağlı olarak kayıp fonksiyonunun gradyanını hesaplamaktadır. Bir kez güncelleme işlemi yapmak için tüm veri kümesinin derecelerinin hesaplanması gerekmektedir. Aşağıdaki formülde bu işlem gösterilmektedir.

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta) \quad (4.12)$$

Yalnızca bir kez güncelleme yapılması için tüm veri kümesinin gradyanlarının hesaplanması gerektiğinden yığın gradyan inişi uzun sürede hesaplama yapabilmektedir.

Stokastik Gradyan İnişi (Stochastic Gradient Descent): Stokastik Gradyan İnişi yönteminde her bir eğitim seti örneği $x^{(i)}$ ve sınıf $y^{(i)}$ için bir parametre güncellemesi gerçekleştirilmektedir.

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}) \quad (4.13)$$

Stokastik Gradyan İnişi, bir seferde bir güncelleme işlemi gerçekleştirdiğinden hızlı hesaplama yapabilmektedir.

Minibatch Gradyan İnişi (Mini-Batch Gradient Descent): Minibatch gradyan inişi yönteminde n eğitim örneklerinin her minibatchı için bir parametre güncellemesi gerçekleştirilmektedir.

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta; x^{(i:i+n)}; y^{(i:i+n)}) \quad (4.14)$$

Minibatch gradyan inişi yöntemi parametre güncellemelerinin varyansını azaltmaktadır ve bu sayede daha kararlı yansıma sağlanabilmektedir.

4.5.1.3. Gradyan İnişi Optimizasyon Algoritmaları

Momentum: SGD yöntemi; bir boyutta bulunan eğrilerin başka bir boyutta sıklaşmasıyla birlikte minimum değeri bulma noktasında kararsız salınım sergilemektedir. Momentum algoritması ile SGD yöntemi minimum değeri bulma noktasında hızlandırılmakta ve salınım sayısı azaltılmaktadır [117].

$$\vartheta_t = \gamma \vartheta_{t-1} + \eta \cdot \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}) \quad (4.15)$$

$$\theta = \theta - \vartheta_t \quad (4.16)$$

Yukarıda formülde gösterildiği gibi Momentum algoritması geçmiş parametreleri güncellerken geçmiş gradyanları da hesaba katmaktadır.

Adagrad: Adagrad algoritması, gradyan inişi yöntemindeki öğrenme hızını parametrelere uyarlamaktadır [118]. Sık karşılaşılan özelliklerle ilişkili parametreler için daha küçük güncellemeler yapılırken nadir karşılaşılan özelliklerle ilişkili parametreler için daha büyük güncellemeler yapılmaktadır. Bu sayede sabit öğrenme katsayısı problemi giderilmektedir. t adım zamanındaki gradyan g_t ile gösterilmek üzere, t adım zamanındaki θ_i parametresine bağlı kayıp fonksiyonunun parçalı türevi ise $g_{t,i}$ ile gösterilmektedir.

$$g_{t,i} = \nabla_{\theta} J(\theta_{t,i}) \quad (4.17)$$

Stokastik gradyan inişi yöntemi; her t adım zamanında, her θ_i parametresinin güncellenmesi ile aşağıdaki forma dönüşmektedir:

$$\theta_{t+1,i} = \theta_{t,i} - \eta \cdot g_{t,i} \quad (4.18)$$

Adagrad; θ_i parametresi için hesaplanmış önceki gradyanları temel alarak her θ_i parametresi için her adım zamanı t 'de genel öğrenme oranı η' 'yi güncellemektedir:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,i} + \epsilon}} \cdot g_{t,i} \quad (4.19)$$

G_t , köşegen boyunca bütün θ paramterlerine bağlı olarak geçmiş gradyanların karelerinin toplamını içerdiğinden G_t ile g_t arasında matris-vektör çarpımı uygulanabilmektedir:

$$\theta_{t+1} = - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot g_{t,I} \quad (4.20)$$

Adadelta: Adadelta yöntemi [119], Adagrad yöntemindeki zamanla azalan öğrenme katsayısı problemini gidermeye yönelik olarak geliştirilmiştir. Adadelta yönteminde, önceki tüm kare gradyanların toplanması, sabit bir büyüklük w ile sınırlandırılmakta ve böylece t zaman adımındaki ortalama çalışma $E[g^2]_t$ sadece önceki ortalama ve güncel gradyan ile aşağıdaki şekilde hesaplanmaktadır:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma)g_t^2 \quad (4.21)$$

Yukarıdaki formülde γ momentum katsayısına benzer bir şekilde değer taşımaktadır. Bu sayede SGD formülü aşağıdaki formülde gösterilebilmektedir:

$$\Delta\theta_t = -\eta \cdot g_{t,i} \quad (4.22)$$

$$\theta_{t+1} = \theta_t + \Delta\theta_t \quad (4.23)$$

Adagrad yönteminde daha önce elde edilen güncelleme parametre vektörü aşağıdaki formüle dönüşmektedir:

$$\Delta\theta_t = -\frac{\eta}{\sqrt{G_t + \epsilon}} \odot g_t \quad (4.24)$$

Daha sonra köşegen matris G_t , $E[g^2]_t$ ile yer değiştirmektedir ve aşağıda gösterilmektedir.

$$\Delta\theta_t = -\frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \quad (4.25)$$

Payda, gradyanının ortalama karesel kök hata kriterini sağladığından $RMS[g]_t$ ile yer değiştirmektedir.

$$\Delta\theta_t = \theta_t - \frac{\eta}{RMS[g]_t} g_{t,i} \quad (4.26)$$

Güncellemenin aynı kuramsal birimlere sahip olması gerekliliğinden dolayı başka bir üssel azalan ortalama tanımlanmaktadır:

$$E[\Delta\theta^2]_t = \gamma E[\Delta\theta^2]_{t-1} + (1-\gamma) \Delta\theta_t^2 \quad (4.27)$$

Parametre güncellemelerinin ana ortalama karesel hatası aşağıdaki şekilde olmaktadır

$$RMS[\Delta\theta]_t = \sqrt{E[\theta^2]_t + \epsilon} \quad (4.28)$$

$RMS[\Delta\theta]_t$ bilinmediğinden dolayı önceki zaman adımına kadar olan parametre güncellemeleri RMS ile tahmin edilmeye çalışılmaktadır. Önceki güncelleme kuralındaki η öğrenme oranının $RMS[\Delta\theta]_t$ ile değiştirilmesinin ardından Adadelta güncelleme kuralı aşağıdaki formül ile gösterilmektedir

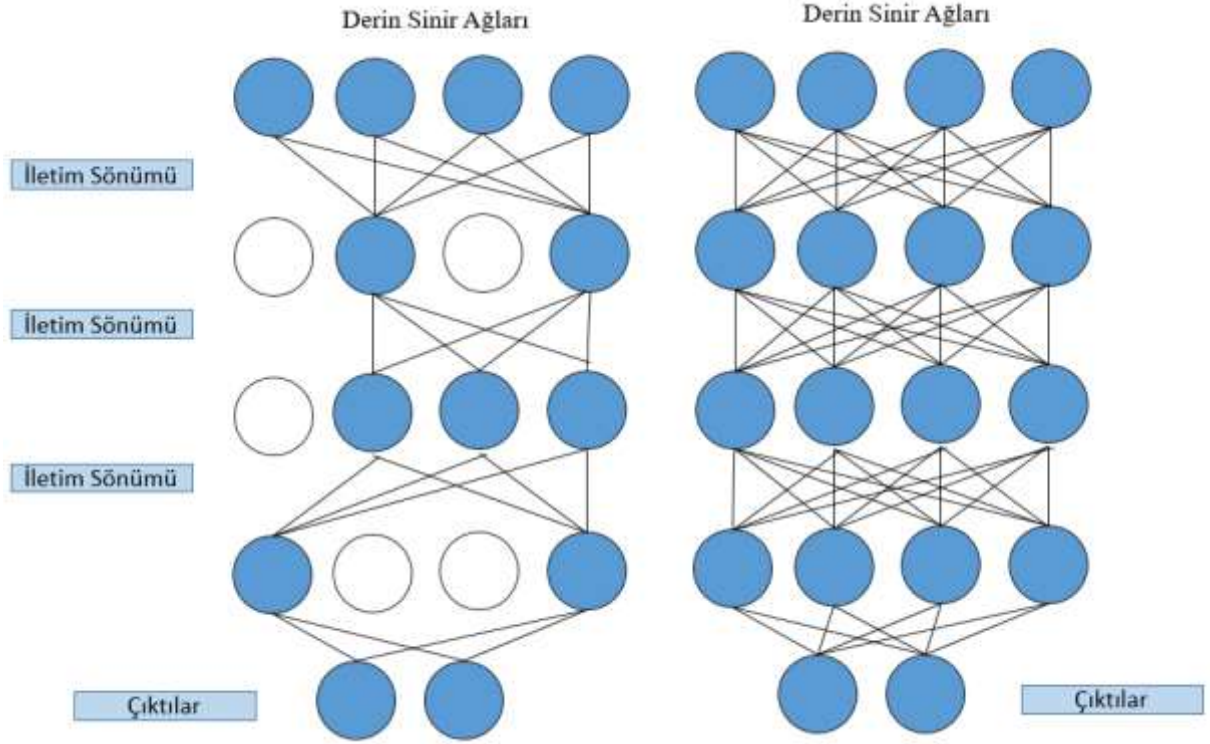
$$\Delta\theta_t = -\frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} g_t g_t \quad (4.29)$$

$$\theta_{t+1} = \theta_t + \Delta\theta_t \quad (4.30)$$

Adadelta'nın kullanılması ile varsayılan öğrenme oranının belirlenmesi ihtiyacı kalmamaktadır.

4.5.1.4. İletim Sönümü

Derin sinir ağlarının eğitimi sırasında, ağın aşırı öğrenmesinin (overfitting) azaltılması amacıyla iletim sönümü uygulanmaktadır [120]. Gizli katmandaki nöronlara iletim sönümü uygulanarak rastgele seçilen bazı nöronların işlevsiz olması sağlanmaktadır. Bu sayede bu nöronlar ağın sonraki aşamasında kullanılmamaktadır. Aşağıdaki Şekil 4.8.'de iletim sönümünün bir derin sinir ağında uygulaması gösterilmektedir.



Şekil 4.8. Derin sinir ağında iletim sönümü eklenmiş gösterim

4.6. İstif Edilmiş Otokodlayıcılar

Otokodlayıcılar (AutoEncoder) giriş verisine benzer bir çıkış verisi üretmeye çalışan belirli bir Derin Sinir Ağları mimarisidir [121]. Otokodlayıcılar bir denetimsiz öğrenme modelidir bu nedenle etiketlenmemiş verilere ihtiyaç duymaktadır. Üç ana kısımdan oluşmaktadır: Giriş katmanı, gizli katman ve çıkış katmanı. Otokodlayıcı modelinde giriş verisinin boyutu gizli katmanda azaltılıp çıkış katmanında orjinal boyutuna getirilmektedir. Gizli katmanda boyutun azaltılması sağlanılarak giriş verisi sıkıştırılmakta ve bu sayede sıkıştırılmış veriden önemli özelliklerin öğrenilmesi amaçlanmaktadır. Öğrenilen özellikler ile giriş verisi üzerinde; sınıflandırma, görselleştirme, iletişim vb. işlemlerinin kolaylaştırılması hedeflenmektedir. Örnek olarak, x girdi verisinden z çıktısının otokodlayıcı tarafından üretilmesi işlemi şu şekilde yapılmaktadır. Girdi verisinin kodlanma kısmı otokodlayıcı kullanılarak aşağıdaki formül kullanılarak hesaplanmaktadır:

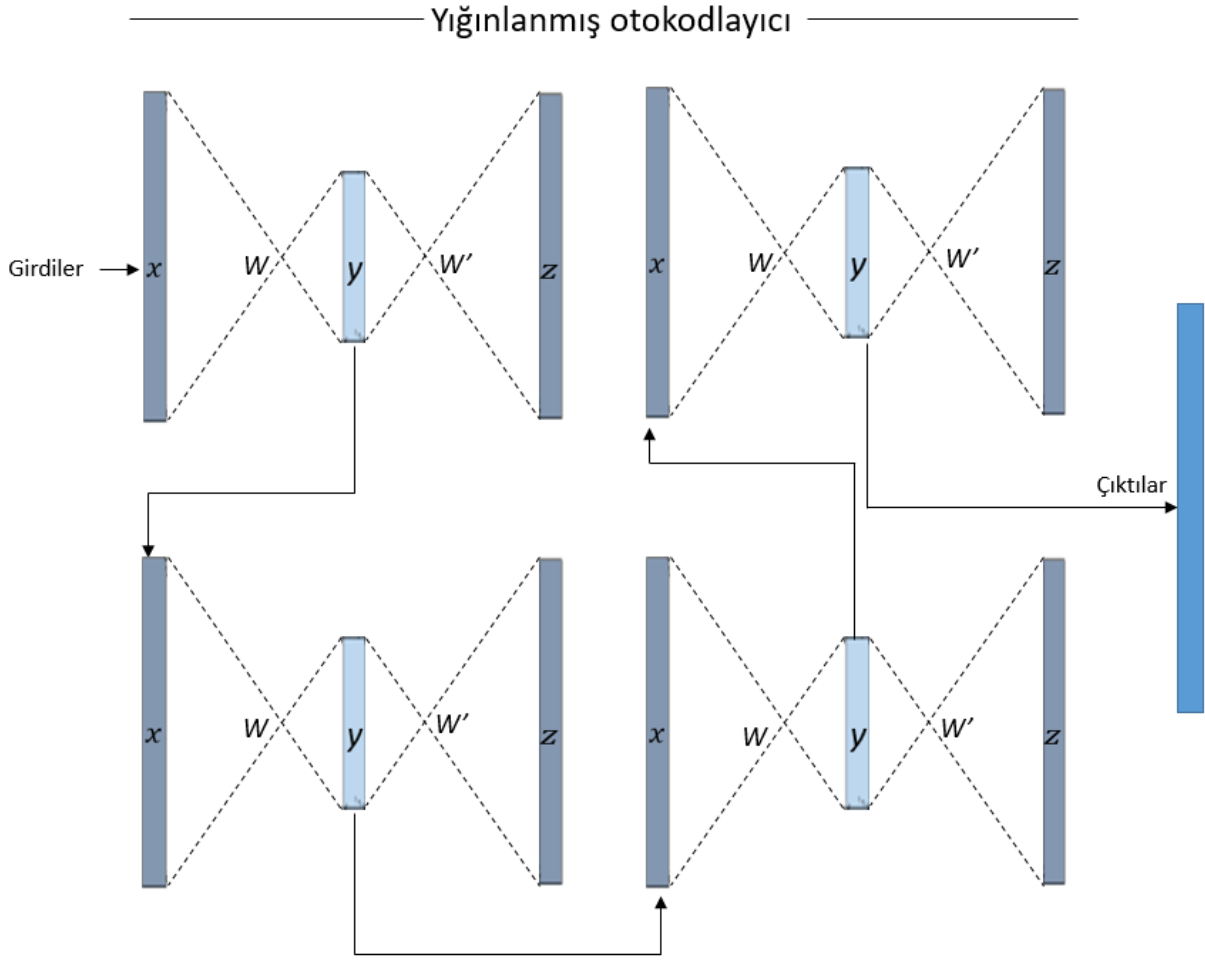
$$y = f(Wx+b) \quad (4.31)$$

Yukarıdaki formülde f lineer olmayan bir fonksiyondur ve x girdi verisinin lineer sonucunu, lineer olmayan bir uzaya dönüştürmektedir. Bu işlem sonucunda elde edilen y kodlanmış verisinden aşağıdaki formül kullanılarak z çıktı verisi elde edilmektedir. Elde edilen z çıktı verisi ile x girdi verisi aynı boyutlara sahiptirler.

$$z = g(W'y+b') \quad (4.32)$$

Yukarıdaki formülde W' yeniden yapılandırma ağırlıklandırma matrisidir, b' yeniden yapılandırma önyargısıdır (bias) ve g fonksiyonu f fonksiyonuna benzer şekilde lineer olmayan bir fonksiyondur.

Bir otokodlayıcı oluşturmak için üç öğeye ihtiyaç vardır; şifreleme fonksiyonu, şifre çözücü fonksiyon ve çıktı verisi ile istenilen hedef arasındaki farkı hesaplayan kayıp fonksiyon. Otokodlayıcı ile şifreleme fonksiyonu kullanılarak giriş verisinden şifrelenmiş veri ve şifrelenmiş veriden de şifre çözücü fonksiyon kullanılarak çıktı verisi üretilmektedir. Otokodlayıcıların art arda gelmesi ile istif edilmiş otokodlayıcı (stacked autoencoder) oluşturulmaktadır [122] ve aşağıdaki şekil 4.9'da gösterilmektedir.



Şekil 4.9. Birden fazla otokodlayıcıdan oluşan istif edilmiş otokodlayıcı örneği

Otokodlayıcılar esas olarak bir kaç önemli özelliği içeren boyut azaltma algoritmasıdır. Bu önemli özellikler şunlardır:

- **Veriye özgün:** Otokodlayıcılar yalnızca eğitildikleri veriler üzerinde anlamlı bir boyut azaltma işlemi yapabilmektedirler. Sadece verilen eğitim setine özgü özellikler öğrenebilmeleri nedeniyle standart sıkıştırma algoritmalarından farklıdır.
- **Kayıp:** Bir otokodlayıcının çıktı verisi ile girdi verisi birbirine benzer olmamakla birlikte tamamen aynı olmamalıdır.
- **Denetimsiz:** Otokodlayıcılar öğrenme işleminde açık etiketlere ihtiyaç duymadıkları için denetimsiz öğrenme tekniği olarak kabul edilmektedirler.

Bir otokodlayıcının öğrenme işlemine başlamadan önce 4 parametresinin kullanıcı tarafından belirlenmesi gerekmektedir:

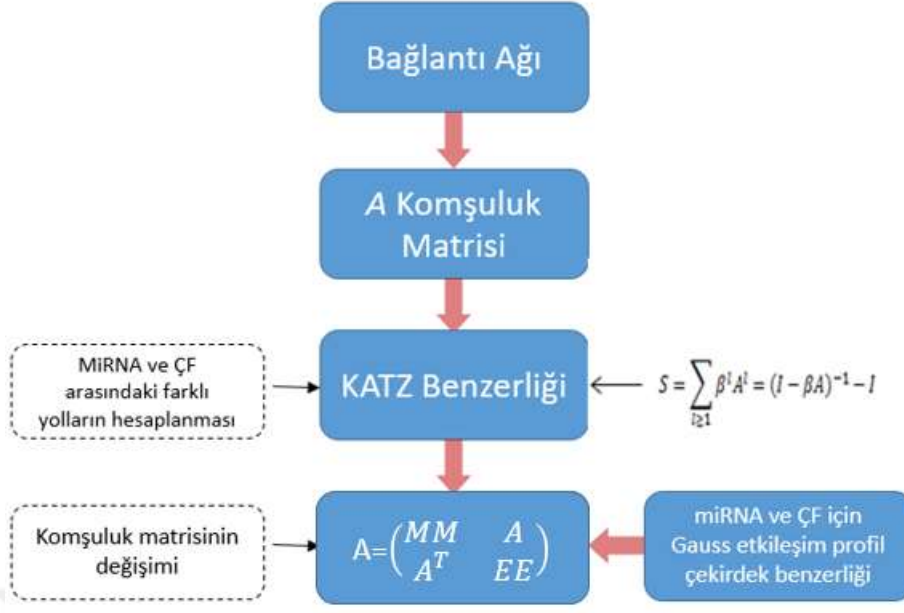
- Ara katman boyutu: Orta katmandaki nöronların sayısı. Sayı küçüldükçe boyut azaltma oranı artmaktadır.
- Ara katmanların sayısı: Otokodlayıcıda ara katman sayısı kullanıcı tarafından her hangi bir limit olmaksızın belirlenmektedir.
- Katmanlardaki nöron sayısı: Şifreleme fonksiyonun kullanıldığı her gizli katmanda düğüm sayısı orta katmana doğru giderek azalmakta ve şifre çözücünün kullanıldığı her gizli katmanda çıkış katmanına doğru giderek artmaktadır. Düğümlerin azalıp artış işlemi zorunlu olmamakla birlikte genel olarak uygulamalarda birbirinin simetrisi olacak şekilde kullanılmaktadır.
- Kayıp fonksiyon: Giriş verisi $[0,1]$ aralığında ise genel olarak çapraz entropi değilse ortalama karesel hata fonksiyonu kullanılmaktadır.



5. ÖNERİLEN MODELLER

5.1. MiRNA Çevresel Faktör Etkileşim Modeli

MiRNA ve çevresel faktör etkileşimlerinin insan sağlığı üzerindeki etkileri nedeniyle KATZ ölçütü temel alınarak miRNA'lar ve çevresel faktörler arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla yeni bir model geliştirilmiştir. Hastalıkların moleküler düzeyde tedavi edilmesi için moleküllerin çevresel faktörler ile olan etkileşimlerinin bilinmesi gerekmektedir. MiRNA'lar da bu moleküller arasındadır. Geliştirilen model ile daha önce aralarında ilişkileri bilinmeyen ancak potansiyel olarak ilişkili olma ihtimali yüksek miRNA-çevresel faktör çiftlerinin tahmin edilmesi amaçlanmıştır. Geliştirilen modelde, miRNA'lar arasında benzerlik değerleri elde edilmesi amacıyla bölüm 4.2'de gösterilen Gauss etkileşim profili çekirdek benzerliği metodu kullanılarak 665x665 boyutlarında miRNA-miRNA benzerlik matrisi oluşturulmuştur. Çevresel faktörler arasında benzerlik değerlerinin elde edilmesi amacıyla yine aynı şekilde Gauss etkileşim profili çekirdek benzerliği metodu kullanılarak 223x223 boyutlarında çevresel faktör-çevresel faktör benzerlik matrisi oluşturulmuştur. Elde edilen miRNA-miRNA benzerlik matrisi, çevresel faktör-çevresel faktör benzerlik matrisi ve miRNA-çevresel faktör komşuluk matrisi, KATZ ölçütü kullanılarak hesaplama yönteminde kullanılmaktadır. Modelimizin akış diyagramı başlangıç adımından bitiş adımına kadar aşağıdaki Şekil 5.1.'de gösterilmektedir. Çevresel faktörlerin ve miRNA'ların benzerliklerin hesaplanması ile elde edilen matrisler şekilde gösterildiği gibi komşuluk matrisine entegre edilmiştir.



Şekil 5.1. MiRNA-Çevresel Faktörler modelinin akış diyagramı

5.1.1. Veri Seti

MiRNA-çevresel faktör arasında kanıtlanmış olan bağlantı verileri miREnvironment veritabanından alınmıştır. miREnvironment veritabanı birden fazla canlı türü için toplamda 3723 kayıt içermektedir. Çalışmamızda bu veritabanından sadece insan miRNA'ları ve çevresel faktörlerin bağlantı bilgilerini içeren 2026 kayıt kullanılmıştır. Daha sonra, bu bağlantı bilgileri komşuluk matrisine aktarılmıştır. Komşuluk matrisinde miRNA ve çevresel faktör arasında bağlantı var ise karşılık gelen kısım 1, bağlantı yok ise 0 değerini almaktadır.

5.1.2. MikroRNA'lar için Gauss Etkileşim Profili Çekirdek Benzerliği

Fonksiyonel olarak benzer miRNA'lar çevresel faktörler ile benzer veya benzer olmayan etkileşim desenleri gösterme eğilimindedirler varsayımı temel alınarak [123], bütün miRNA çiftleri arasındaki benzerliklerin hesaplanması amacıyla gauss etkileşim profili çekirdek benzerliği metodu kullanılmıştır. İlk olarak her bir miRNA $m(i)$ için ikili vektör $IP(m(i))$ şeklinde bir etkileşim profili IP oluşturulmuştur ve çevresel faktörlerin her biri ile bağlantısının olup olmadığı kontrol edilerek bağlantı var ise karşılık gelen kısma 1 ve bağlantı

yok ise karşılık gelen kısma 0 değeri yazılmıştır. Gauss çekirdeği aşağıdaki formülde gösterilmiştir:

$$MM(m(i), m(j)) = \exp(-y_m \| IP(m(i)) - IP(m(j)) \|^2) \quad (5.1)$$

Bant genişliğini kontrol etmek için y_m parametresi kullanılmıştır. Bu parametre aşağıdaki formül ile hesaplanmaktadır:

$$y_m = y'_m / (\sum_{i=1}^{n_m} \| IP(m(i)) \|^2) \quad (5.2)$$

n_m miRNA'ların sayısını göstermektedir ve y'_m parametresi Laarhoven'in çalışması [109] temel alınarak 1 seçilmiştir.

5.1.3. Çevresel Faktörler için Gauss Etkileşim Profili Çekirdek Benzerliği

Çevresel faktörler arasındaki benzerlik değerlerin hesaplanması için yine gauss etkileşim profili çekirdek benzerliği kullanılmıştır. İlk olarak her bir çevresel $e(i)$ faktör için ikili vektör $IP(e(i))$ şeklinde bir etkileşim profili IP oluşturulmaktadır ve miRNA'ların her biri ile bağlantısının olup olmadığı kontrol edilerek eğer bağlantı var ise karşılık gelen kısma 1 ve bağlantı yok ise karşılık gelen kısma 0 değeri yazılmıştır. Gauss çekirdeği aşağıdaki formülde gösterilmiştir:

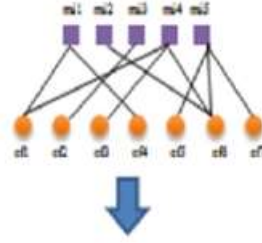
$$EE(e(i), e(j)) = \exp(-y_k \| IP(e(i)) - IP(e(j)) \|^2) \quad (5.3)$$

Bant genişliğini kontrol etmek için y_k parametresi kullanılmıştır. Bu parametre aşağıdaki formül ile hesaplanmaktadır:

$$y_k = y'_k / (\sum_{i=1}^{n_k} \| IP(e(i)) \|^2) \quad (5.4)$$

n_k çevresel faktörlerin sayısını göstermektedir. Gauss kullanılarak benzerlik matrislerinin elde edilmesi işleminin bir örneği aşağıdaki Şekil 5.2.'de gösterilmiştir.

miRNA Çevresel Faktörler bağlantı ağı



	mi1	mi2	mi3	mi4	mi5
ef1	1	0	0	1	0
ef2	0	0	1	0	0
ef3	0	0	0	1	0
ef4	1	0	0	0	0
ef5	0	0	0	0	1
ef6	0	1	0	1	1
ef7	0	0	0	0	1

A komşuluk matrisi

miRNA için Gauss Etkileşim
Profil Çekirdek Benzerliği

	mi1	mi2	mi3	mi4	mi5
mi1	1	0.2231	0.2231	0.0820	0.2231
mi2	0.2231	1	0.1353	0.1353	0.3678
mi3	0.2231	0.1353	1	0.1353	0.3678
mi4	0.0820	0.1353	0.1353	1	0.3678
mi5	0.2231	0.3678	0.3678	0.3678	1

ÇF için Gauss Etkileşim Profil
Çekirdek Benzerliği

	ef1	ef2	ef3	ef4	ef5	ef6	ef7
ef1	1	0.1224	0.4956	0.4956	0.1224	0.1224	0.1224
ef2	0.1224	1	0.2466	0.2466	0.2466	0.0608	0.2466
ef3	0.4956	0.2466	1	0.2466	0.2466	0.2466	0.2466
ef4	0.4956	0.2466	0.2466	1	0.2466	0.0608	0.2466
ef5	0.1224	0.2466	0.2466	0.2466	1	0.2466	1
ef6	0.1224	0.0608	0.2466	0.0608	0.2466	1	0.2466
ef7	0.1224	0.2466	0.2466	0.2466	1	0.2466	1

Şekil 5.2. Gauss kullanılarak komşuluk matrisinden benzerlik matrislerinin elde edilmesi

5.1.4. KATZ

KATZ ölçütünün modelimizde kullanılmasının nedeni aralarında bağlantı bulunmayan miRNA-çevresel faktör çiftleri arasında, birden fazla yürüyüş yolu kullanılarak bağlantı değerlerinin hesaplanabilmesidir. Bir ağda farklı uzunluklardaki yürüyüşler mevcut olduğundan farklı uzunluklardaki yürüyüşleri baskılamak için negatif olmayan bir β parametresi kullanılmaktadır. Yürüyüş uzunluğu arttırıldığında β parametresinin değeri azalmaktadır. β parametresi daha önceki yapılan çalışmaya göre $\beta < 1$ $\|A\|^2$ şeklinde seçilmiştir ve bu seçime göre yolların uzunluğu arttığında β parametresi 0 değerine

yaklaşmaktadır. Tüm miRNA-çevresel faktör çiftlerinin benzerlik değerlerinin hesaplanması amacıyla aşağıdaki formül kullanılmaktadır.

$$S = \sum_{l \geq 1} \beta^l A^l = (I - \beta A)^{-1} - I \quad (5.6)$$

Yukarıdaki formülde, A komşuluk matrisine miRNA benzerlik matrisi ve çevresel faktör benzerlik matrisinin eklenmesi amacıyla A komşuluk matrisi aşağıdaki forma dönüştürülmüştür.

$$A = \begin{bmatrix} MM & A \\ A^T & EE \end{bmatrix} \quad (5.7)$$

Yukarıdaki formda A^T , A komşuluk matrisinin transpoz matrisidir. Eğer k parametresi sırasıyla $k = 2$, $k = 3$ ve $k = 4$ seçilirse KATZ benzerlik ölçütü aşağıdaki şekilde hesaplanmaktadır.

$$S^{k=2}(A^*) = \beta \cdot A + \beta^2 \cdot (MM \cdot A + A \cdot EE) \quad (5.8)$$

$$S^{k=3}(A^*) = S^{k=2}(A^*) + \beta^3 \cdot (A \cdot A^T \cdot A + MM^2 \cdot EE + MM \cdot A \cdot EE + A \cdot EE^2) \quad (5.9)$$

$$S^{k=4}(A^*) = S^{k=3}(A^*) + \beta^4 \cdot (MM^3 \cdot A + A \cdot A^T \cdot MM \cdot A + MM \cdot A \cdot A^T \cdot A + A \cdot EE \cdot A^T \cdot A) + \beta^4 \cdot (A \cdot A^T \cdot A \cdot EE + MM^2 \cdot A \cdot EE + MM \cdot A \cdot EE^2 + A \cdot EE^3) \quad (5.10)$$

5.1.5. Değerlendirme

Geliştirilen miRNA-çevresel faktör modelinin performansının test edilmesi amacıyla tek-çıkışlı çapraz doğrulama, 2-katlı ve 5-katlı çapraz doğrulama yöntemi uygulanarak AUC değerleri hesaplanmıştır. AUC değerleri modelin yeni potansiyel bağlantılar yapma konusunda başarılı sonuçlar elde ettiğini göstermektedir. Sonuçlar $k=2$, $k=3$ ve $k=4$ bağlantı yollarının uzunluğuna bağlı olarak sırasıyla aşağıdaki Şekil 5.3., Şekil 5.4. ve Şekil 5.5.'te gösterilmektedir. k parametresinin değeri 2,3 ve 4 seçilerek çalışmada yer verilmiştir 4'ten büyük değerler düşük performans gösterdiği için yer verilmemiştir.

5.1.6. Global Tek Çıkarımlı Çapraz Doğrulama Yöntemi

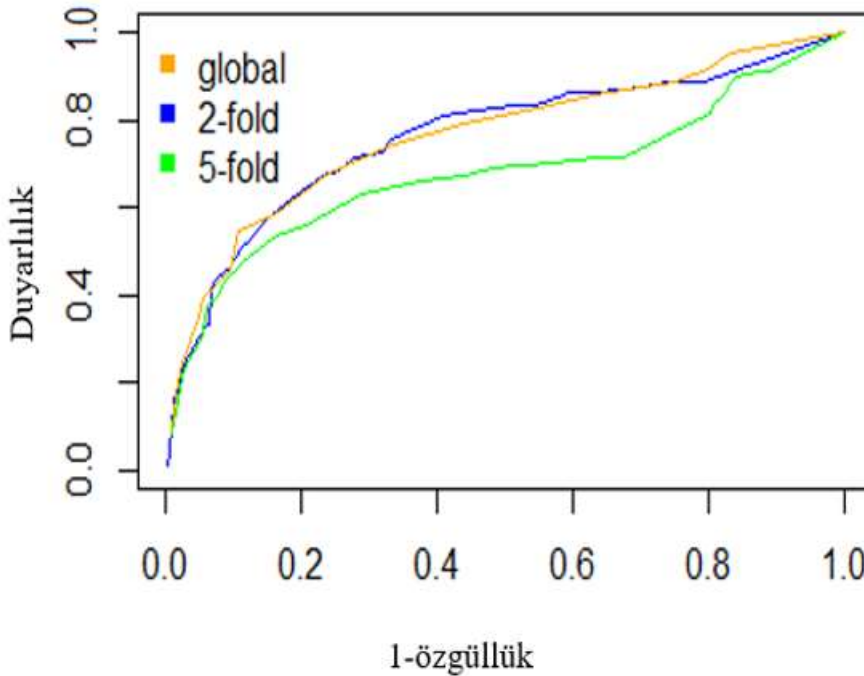
Global tek-çıkışlı çapraz doğrulama yönteminde değerleri 1 olan tüm bağlantılar yalnızca bir kez test örneği olarak kullanılmaktadır. Test örneği olarak kullanılan bir bağlantı eğitim setine dahil edilmemekte, değeri 1 olan diğer tüm bağlantılar ise eğitim setine dahil edilmektedir. Bağlantı değerleri 0 olan bağlantılar ise aday örnekler olarak değerlendirilmektedir. Çevresel faktörler ile miRNA'lar arasında bağlantı değeri 1 olan 1439 bağlantı olduğundan dolayı geliştirilen model her bir bağlantı için bir kez olmak üzere toplamda 1439 kez uygulanarak sonuçlar hesaplanmaktadır. Her bir uygulamada, test örneğinin sıralaması belirlenen eşik değerden (threshold) yüksek olduğunda bu test örneği başarılı olarak değerlendirilmektedir. Farklı eşik değerler ile doğru onaylanmış oranlar (duyarlılık) ve hatalı onaylanmış oranlar (1-özgüllük) hesaplanmaktadır. Eşik değerden yüksek çıkan test örneklerinin oranı ile duyarlılık, eşik değerden düşük çıkan aday örneklerin oranı ile özgüllük hesaplanmaktadır. Farklı eşik değerler oluşturularak, doğru onaylanmış oranlara karşı hatalı onaylanmış oranlar ile ROC (receiver operating characteristics) eğrisi çizilebilmektedir [124]. Geliştirilen modelin tahmin performansı ROC eğrisinin altında kalan ile hesaplanmakta ve bu alan AUC değerini göstermektedir. AUC değeri bu tarz modellerde tahmin performansını ölçmek için kullanılmaktadır [33]. AUC değeri şu şekilde değerlendirilmektedir; eğer AUC değeri 1 ise o zaman en yüksek performans sergilenmiştir eğer 0.5 ise ortalama bir performans sergilenmiştir ve eğer 0 ise en kötü performans sergilenmiştir. Belirtildiği üzere k parametresi yürüyüşlerin sayısını ve uzunluğunu göstermektedir ve bu sayede modelin tahmin performansını etkilemektedir. Modelin en yüksek AUC değeri 0.780 ile k parametresinin değeri 4 olduğunda elde edilmiştir bunun dışındaki değerlerde AUC değeri azalmaktadır. Örnek olarak farklı k parametreleri için elde edilen değerler $k=2$, $k=3$ ve $k=4$ için sırasıyla aşağıdaki Şekil 5.3., Şekil 5.4. ve Şekil 5.5.'te gösterilmektedir.

5.1.7. K-katlı Çapraz Doğrulama

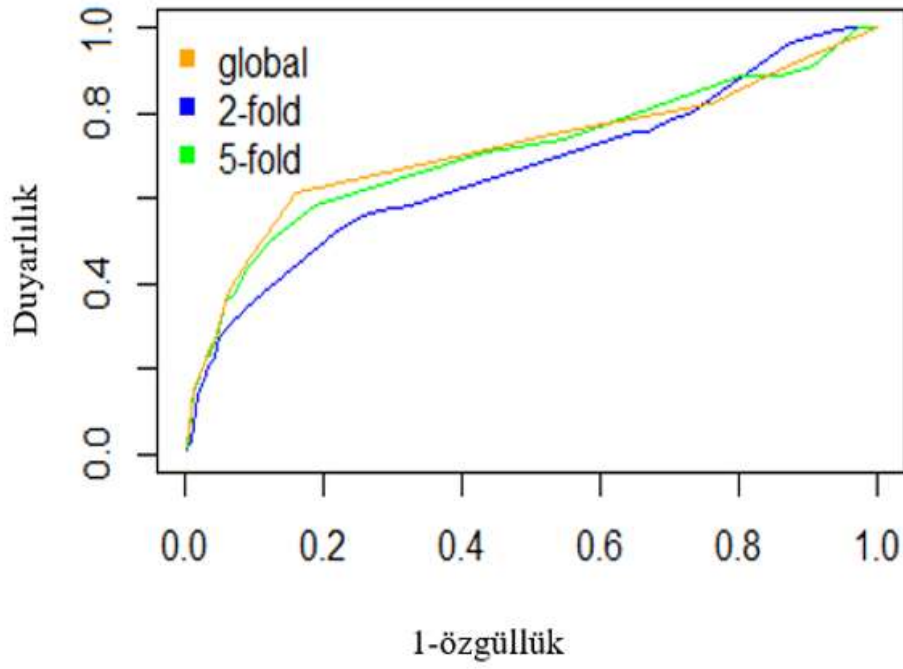
K-katlı çapraz doğrulama yönteminde ilk olarak bilinen tüm miRNA-çevresel faktör bağlantıları eşit olacak şekilde k parçaya bölünmektedir. Daha sonra bir parça test için geriye kalanlar ($k-1$) ise eğitim için kullanılmaktadır. Bu işlem k kez tekrarlanmaktadır ve her bir parça sırasıyla sadece bir kez test için seçilmekte diğerleri ise eğitim için seçilmektedir.

Global tek-çıkışlı çapraz doğrulama yöntemine benzer şekilde miRNA-çevresel faktörler arasındaki bilinmeyen bağlantılar örnek adaylar olarak kabul edilmektedir. Özel olarak, 2-katlı ve 5-katlı çapraz doğrulama yöntemleri uygulanmıştır. 2-katlı çapraz doğrulama yönteminde bilinen tüm bağlantılar iki parçaya bölünmektedir. İlk parça test için ikinci parça eğitim için seçilmekte ve daha sonra, test ve eğitim için seçilen parçalar değiştirilerek ilk parça eğitim ve ikinci parça test için seçilmektedir. Önyargıdan (bias) uzaklaşılması için bu işlem 100 kez tekrar edilmiştir. Bu sayede ROC eğrisi çizilmiş ve global tek-çıkışlı çapraz doğrulama yönteminde olduğu gibi AUC değeri elde edilmiştir. 2-katlı çapraz doğrulama yönteminde, modelin en yüksek AUC değeri 0.761 ile k parametresinin değeri 2 olduğunda elde edilmiştir bunun dışındaki değerlerde AUC değeri azalmaktadır. Farklı k parametreleri $k=2$, $k=3$ ve $k=4$ için elde edilen AUC değerleri sırasıyla 0.761, 0.663 ve 0.725'tir ve aşağıdaki Şekil 5.3., Şekil 5.4. ve Şekil 5.5'te sırasıyla gösterilmektedir.

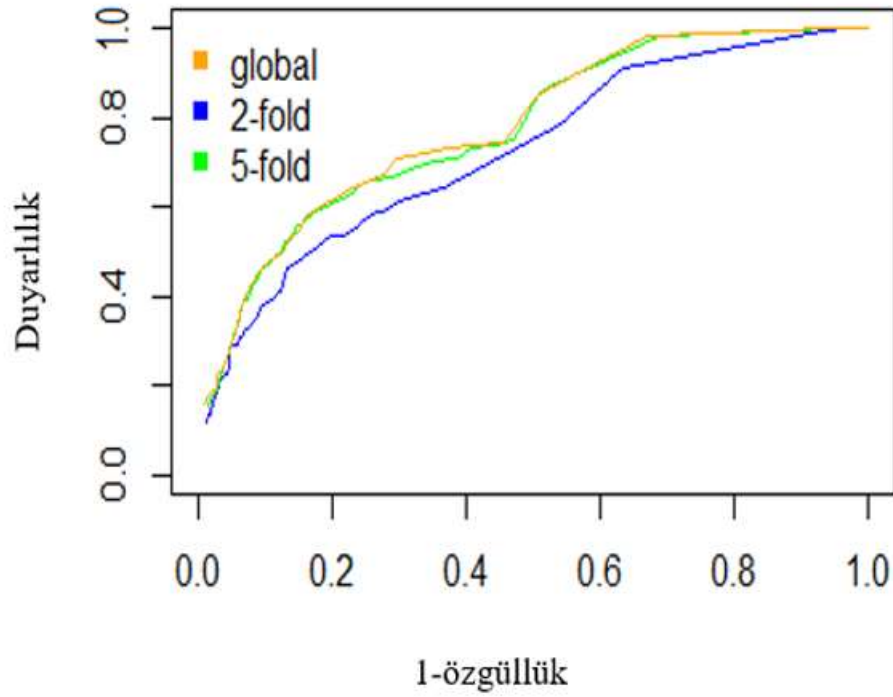
5-katlı çapraz doğrulama yönteminde, modelin en yüksek AUC değeri 0.773 ile k parametresinin değeri 4 olduğunda elde edilmektedir. Farklı k parametreleri $k=2$, $k=3$ ve $k=4$ için elde edilen AUC değerleri sırasıyla 0.676, 0.706 ve 0.773'tir ve aşağıdaki Şekil 5.3., Şekil 5.4. ve Şekil 5.5'te sırasıyla gösterilmektedir.



Şekil 5.3. k parametresinin değeri 2 seçildiğinde, AUC sonuçları tek-çıkışlı, 2-katlı ve 5-katlı çapraz doğrulama için sırasıyla 0.763, 0.761, 0.676'dir.



Şekil 5.4. k parametresinin değeri 3 seçildiğinde, AUC sonuçları tek-çıkartımlı, 2-katlı ve 5-katlı çapraz doğrulama için sırasıyla 0.713, 0.663, 0.706'dir.



Şekil 5.5. k parametresinin değeri 4 seçildiğinde, AUC sonuçları tek-çıkartımlı, 2-katlı ve 5-katlı çapraz doğrulama için sırasıyla 0.780, 0.725, 0.773'dir.

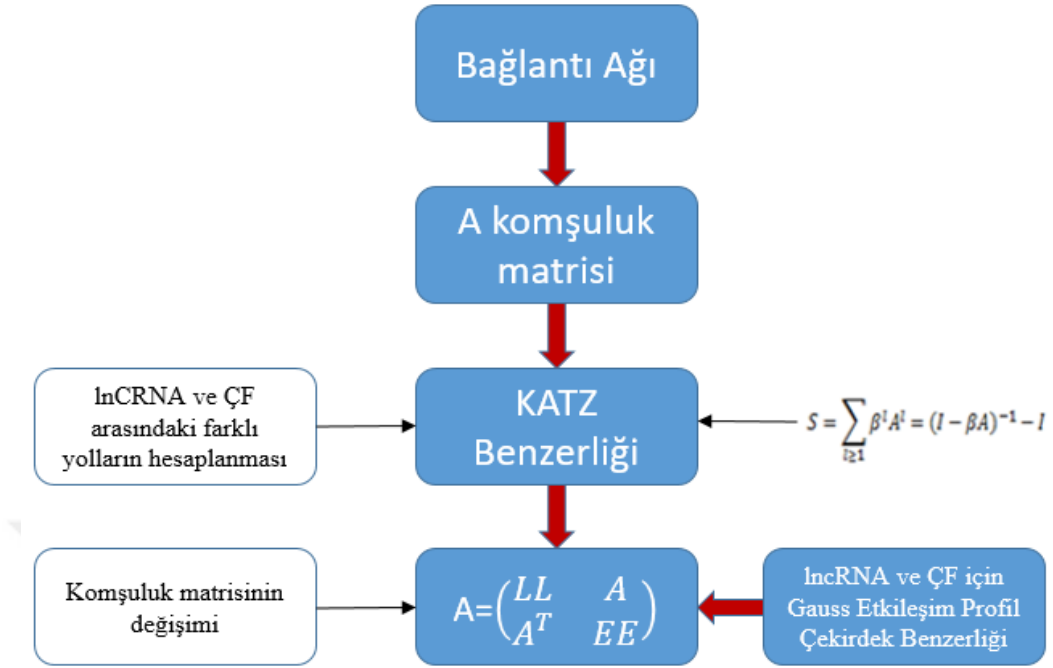
5.1.8. Sonular

MiRNA'lar ve evresel faktrlerin hastalıkların ortaya ıkmasındaki rollerinden dolayı miRNA'lar ve evresel faktrler arasındaki baėlantıların eksiksiz ortaya ıkarılmasının hastalıkların karmaşıık mekanizmalarının anlaşıılmasına katkı saėlayacaėı dşnlmektedir. Etkili hesaplama modelleri geliştiriilmesinin amacı laboratuvar yntemlerinde ihtiya duyulan zaman ve insan kaynaėı dezevantajını ortadan kaldırmaya yneliktir. Bu alıřmada, miRNA'lar ve evresel faktrler arasındaki yeni potansiyel iliřkileri tahmin etmek amacıyla KATZ ltne dayalı bir model geliştiriilmiřtir. MiRNA'lar ve evresel faktrlerin benzerlik matrisini hesaplanmasında Gauss etkileřim profili ekirdek benzerliėi kullanılmıřtır. Geliştirilen modelin tahmin performansının deėerlendirilmesinde global tek ıkarımlı ve k-katlı apraz doėrulama yntemleri uygulanmıřtır. Elde edilen AUC sonuları, modelin miRNA'lar ve evresel faktrler arasındaki yeni potansiyel iliřkileri tahmin etmek iin nemli bir gvenilirliėe sahip olduėunu gstermektedir. Aynı veri setini kullanan literatr alıřması eksikliėinden dolayı karřılařtırma yapılamamıřtır.

5.2. LncRNA evresel Faktr Etkileřim Modeli

alıřmamızda, lncRNA'lar ile evresel faktrler arasında yeni potansiyel baėlantıların bulunması amacıyla KATZ ltne dayalı yeni bir model geliştiriilmiřtir. Hastalıkların molekler dzeyde tedavi edilmesi iin molekllerin evresel faktrler ile olan etkileřimlerinin bilinmesi gerekmektedir. LncRNA'lar da bu molekller arasındadır. Modelimizin geliřtirilmesi amacı daha nce aralarında iliřkileri bilinmeyen ancak potansiyel olarak iliřkili olma ihtimali yksek lncRNA-evresel faktr iftlerinin hesaplanarak tahmin edilmesidir. Modelimizde lncRNA'lar arasında benzerlik deėerleri elde edilmesi amacıyla blm 4.2'de gsterilen Gauss etkileřim profili ekirdek benzerliėi metodu kullanılarak 406x406 boyutlarında lncRNA-lncRNA benzerlik matrisi oluřturulmuřtur. evresel faktrler arasında benzerlik deėerlerinin elde edilmesi amacıyla yine aynı řekilde Gauss etkileřim profili ekirdek benzerliėi metodu kullanılarak 124x124 boyutlarında evresel faktr-evresel faktr benzerlik matrisi oluřturulmuřtur. Elde edilen lncRNA-lncRNA benzerlik matrisi, evresel faktr-evresel faktr benzerlik matrisi ve lncRNA-evresel faktr komřuluk matrisi, KATZ lt ile deėerlendirilmiřtir. Ařaėıdaki řekil 5.6'da geliřtirilen modelin akıř

diyagramı gösterilmektedir.



Şekil 5.6. LncRNA-Çevresel Faktörler modelinin akış diyagramı

5.2.1. Veritabanı

Geliştirilen modelde, lncRNA-çevresel faktör bağlantıları DLREFD veritabanından alınmıştır. DLREFD veritabanında 406 lncRNA ve 124 çevresel faktör arasında 636 tane bağlantı bulunmaktadır. A matrisi, lncRNA-çevresel faktör bağlantılarını, komşuluk matrisi olarak göstermektedir ve boyutu 406 x 124'tür. Bu komşuluk matrisinde örneğin $l(i)$ lncRNA'sı ile $ef(j)$ çevresel faktörü arasında bağlantı var ise $A(i,j)$ 1 değerini almakta eğer yoksa 0 değerini almaktadır. LncRNA'lar ile çevresel faktörler arasında çok fazla bağlantı bulunmaması nedeniyle KATZ ölçütü kullanılarak yeni potansiyel bağlantıların tahmin edilmesi amaçlanmıştır.

5.2.2. LncRNA'lar için Gauss Etkileşim Profili Çekirdek Benzerliği

Geliştirilen modelde temel alınan varsayım, fonksiyonel olarak benzer lncRNA'ların çevresel faktörler ile benzer veya benzer olmayan etkileşim desenleri gösterme eğiliminde olmasıdır [125]. Bütün lncRNA çiftleri arasındaki benzerliklerin hesaplanması amacıyla

Gauss etkileşim profili çekirdek benzerliği metodu kullanılmıştır. İlk olarak her bir lncRNA $l(i)$ için ikili vektör $IP(l(i))$ şeklinde bir etkileşim profili IP oluşturulmuştur. Her bir lncRNA ile çevresel faktörlerin tümü karşılaştırılarak aralarında bağlantı olanlar için ikili vektörde karşılık gelen kısma 1, olmayanlar için karşılık gelen kısma 0 değeri yazılmaktadır. Gauss çekirdeği aşağıdaki formülde gösterilmiştir:

$$LL(l(i), l(j)) = \exp(-y_k \|IP(l(i)) - IP(l(j))\|^2) \quad (5.11)$$

Bant genişliğini kontrol etmek için y_k parametresi kullanılmıştır. Bu parametre aşağıdaki formül ile hesaplanmaktadır:

$$y_k = y'_k / (\sum_{i=1}^{n_k} \|IP(l(i))\|^2) \quad (5.12)$$

n_k lncRNA'ların sayısını göstermektedir.

5.2.3. Çevresel Faktörler için Gauss Etkileşim Profili Çekirdek Benzerliği

Çevresel faktörler arasındaki benzerlik değerlerin hesaplanması için lncRNA gauss etkileşim profil çekirdek benzerliği ile aynı şekilde çevresel faktörler içinde Gauss etkileşim profili çekirdek benzerliği oluşturulmuştur. İlk olarak her bir çevresel $ef(i)$ faktör için ikili vektör $IP(ef(i))$ şeklinde bir etkileşim profili IP oluşturulmuş ve lncRNA'ların her biri ile bağlantısının olup olmadığı kontrol edilerek eğer bağlantı var ise karşılık gelen kısma 1 ve bağlantı yok ise karşılık gelen kısma 0 değeri yazılmıştır. Gauss çekirdeği aşağıdaki formülde gösterilmiştir:

$$EE(ef(i), ef(j)) = \exp(-y_e \|IP(ef(i)) - IP(ef(j))\|^2) \quad (5.13)$$

Bant genişliğini kontrol etmek için y_e parametresi kullanılmıştır. Bu parametre aşağıdaki formül ile hesaplanmaktadır:

$$y_e = y'_e / (\sum_{e=1}^{n_e} \|IP(ef(e))\|^2) \quad (5.14)$$

n_e çevresel faktörlerin sayısını göstermektedir.

5.2.4. KATZ

KATZ ölçütünün avantajı aralarında bağlantı bulunmayan lncRNA-çevresel faktör çiftleri arasında birden fazla yürüyüş yolu hesaplanarak benzerlik değerlerinin hesaplanabilmesidir. Bir ağda farklı uzunluklardaki yürüyüşler mevcut olduğundan farklı uzunluklardaki yürüyüşleri baskılamak için negatif olmayan bir β parametresi kullanılmaktadır. Yürüyüş uzunluğu arttırıldığında β parametresinin değeri azalmaktadır. β parametresi daha önceki yapılan çalışmaya göre $\beta < 1 \|A\|^2$ şeklinde seçilmiştir ve bu seçime göre yolların uzunluğu arttığında β parametresi 0 değerine yaklaşmaktadır. Aşağıdaki KATZ formülü i lncRNA'sı ile j çevresel faktörü arasındaki bağlantının hesaplanmasında kullanılmaktadır:

$$S(l(i), ef(j)) = \sum_{l=1}^k \beta_l A^l(i, j) \quad (5.15)$$

Tüm lncRNA-çevresel faktör çiftlerinin benzerlik değerlerinin hesaplanması amacıyla aşağıdaki formül kullanılmaktadır.

$$S = \sum_{l \geq 1} \beta^l A^l = (I - \beta A)^{-1} - I \quad (5.16)$$

Yukarıdaki formülde, A komşuluk matrisine lncRNA benzerlik matrisi ve çevresel faktör benzerlik matrisinin eklenmesi amacıyla A komşuluk matrisi aşağıdaki forma dönüştürülmüştür.

$$A = \begin{bmatrix} LL & A \\ A^T & EE \end{bmatrix} \quad (5.17)$$

Yukarıdaki formda A^T , A komşuluk matrisinin transpoz matrisidir. Eğer k parametresi sırasıyla $k = 2$, $k = 3$ ve $k = 4$ seçilirse KATZ benzerlik ölçütü aşağıdaki şekilde hesaplanmaktadır.

$$S^{k=2}(A^*) = \beta \cdot A + \beta^2 \cdot (LL \cdot A + A \cdot EE) \quad (5.18)$$

$$S^{k=3}(A^*) = S^{k=2}(A^*) + \beta^3 \cdot (A \cdot A^T \cdot A + LL^2 \cdot EE + LL \cdot A \cdot EE + A \cdot EE^2) \quad (5.19)$$

$$S^{k=4}(A^*) = S^{k=3}(A^*) + \beta^4 \cdot (LL^3 \cdot A + A \cdot A^T \cdot LL \cdot A + LL \cdot A \cdot A^T \cdot A + A \cdot EE \cdot A^T \cdot A) + \beta^4 \cdot (A \cdot A^T \cdot A \cdot EE + LL^2 \cdot A \cdot EE + LL \cdot A \cdot EE^2 + A \cdot EE^3) \quad (5.20)$$

5.2.5. Değerlendirme

Geliştirilen modelin performansının test edilmesi amacıyla global tek-çıkarımlı çapraz doğrulama, 2-katlı ve 5-katlı çapraz doğrulama yöntemi uygulanarak AUC değerleri hesaplanmıştır ve elde edilen AUC değerleri modelin yeni potansiyel bağlantılar yapma konusunda başarılı sonuçlar ürettiğini göstermektedir. Sonuçlar $k=2$, $k=3$ ve $k=4$ bağlantı yollarının uzunluğuna bağlı olarak sırasıyla aşağıdaki Şekil 5.7., Şekil 5.8. ve Şekil 5.9.'da gösterilmiştir.

5.2.6. Global Tek Çıkarımlı Çapraz Doğrulama Yöntemi

Global tek-çıkarımlı çapraz doğrulama yönteminde değerleri 1 olan tüm bağlantılar yalnızca bir kez test örneği olarak kullanılmaktadır. Test örneği olarak kullanılan bir bağlantı eğitim setine dahil edilmemekte, değeri 1 olan diğer tüm bağlantılar ise eğitim setine dahil edilmektedir. Bağlantı değerleri 0 olan bağlantılar ise aday örnekler olarak değerlendirilmektedir. Çevresel faktörler ile lncRNA'lar arasında bağlantı değeri 1 olan 636 bağlantı olduğundan, geliştirilen model her bir bağlantı için bir kez olmak üzere toplamda 636 kez uygulanmaktadır. Her bir uygulamada, test örneğinin sıralaması belirlenen eşik değerden (threshold) yüksek çıktığında bu test örneği başarılı olarak değerlendirilmektedir. Farklı eşik değerler ile doğru onaylanmış oranlar (duyarlılık) ve hatalı onaylanmış oranlar (1-özgüllük) hesaplanmaktadır. Eşik değerden yüksek çıkan test örneklerinin oranı ile duyarlılık, eşik değerden düşük çıkan aday örneklerin oranı ile özgüllük hesaplanmaktadır. Farklı eşik değerler belirlenerek, doğru onaylanmış oranlara karşı hatalı onaylanmış oranlar ile ROC eğrisi çizilebilmektedir. Geliştirilen modelin tahmin performansı ROC eğrisinin altında kalan ile hesaplanmaktadır ve bu alan AUC değerini göstermektedir. Örnek olarak lncRNA-çevresel faktör bağlantılarının bir kısmı aşağıdaki Tablo 5.1.'de verilmektedir. Modelin belirtilen veriler üzerinde k parametresinin değeri 3 seçildiğinde elde edilen sonuçlar Tablo 5.2.'de gösterilmektedir. Modelin en yüksek AUC değeri 0.855 ile k parametresi 3 olarak seçildiğinde elde edilmiştir ve farklı seçimlerde AUC değeri azalmaktadır. Örnek olarak k parametresi $k=2$, $k=3$ ve $k=4$ olarak seçildiğinde elde edilen AUC değerleri sırasıyla 0.826, 0.855 ve 0.823'dir ve Şekil 5.7., Şekil 5.8. ve Şekil 5.9.'da gösterilmektedir.

Tablo 5.1. LncRNA-Çevresel Faktör bağlantı örnekleri

lncRNA-ÇF Komşuluk Matrisi	Cisplatin	Dihydrotest osterone (DHT)	Foetal Calf Serum (FCS)	Hepatitis B Virus
ABCC6P1	1	0	0	0
AC005682.8	0	0	0	0
AC006050.3	1	0	0	0
AC006369.2	0	1	0	0

Tablo 5.2. LncRNA-Çevresel Faktör bağlantı örnekleri üzerinde KATZ uygulama sonuçları

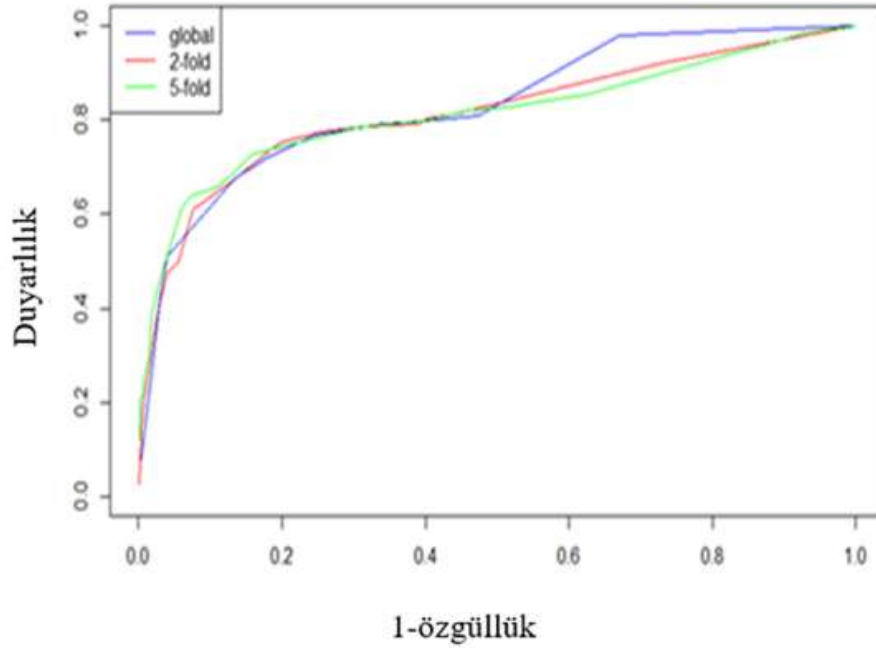
KATZ 3	Cisplatin	Dihydrotest osterone (DHT)	Foetal Calf Serum (FCS)	Hepatitis B Virus
ABCC6P1	26,57277	11,13599	4,865956	1,551295
AC005682.8	17,87066	12,58481	5,516111	1,748565
AC006050.3	26,57277	11,13599	4,865956	1,551295
AC006369.2	15,34849	16,33879	4,682368	1,457839

5.2.7. K-Katlı Çapraz Doğrulama

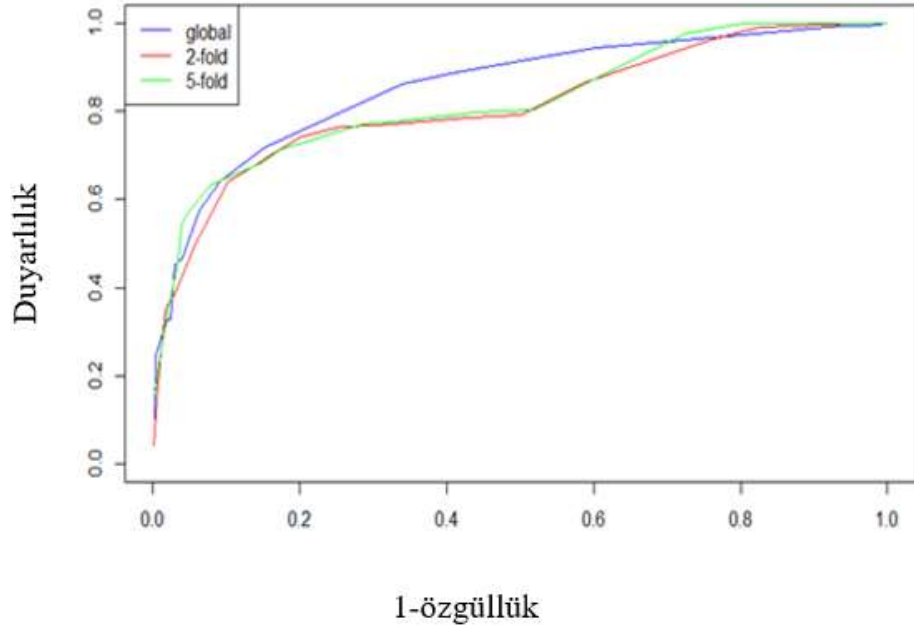
K-katlı çapraz doğrulama yönteminde ilk olarak bilinen tüm lncRNA-çevresel faktör bağlantıları eşit olacak şekilde k parçaya bölünmektedir. Daha sonra bir parça test için geriye kalanlar $(k-1)$ ise eğitim için kullanılmaktadır. Bu işlem k kez tekrarlanmaktadır ve her bir parça sadece bir kez test için seçilmekte diğerleri ise eğitim için seçilmektedir. Global tek-çıkışlı çapraz doğrulama yöntemine benzer şekilde lncRNA-çevresel faktörler arasındaki bilinmeyen bağlantılar örnek adaylar olarak kabul edilmektedir. Özel olarak, 2-katlı ve 5-katlı çapraz yöntemleri uygulanmıştır. 2-katlı çapraz doğrulama yönteminde bilinen tüm bağlantılar iki parçaya bölünmektedir. İlk parça test için ikinci parça eğitim için seçilmekte ve daha sonra test ve eğitim için seçilen parçalar değiştirilerek ilk parça eğitim ve ikinci parça test için seçilmektedir. Önyargıdan (bias) uzaklaşılması için bu işlem 100 kez tekrarlanmıştır.

Son olarak, ROC eğrisi çizilerek AUC değeri elde edilmiştir. 2-katlı çapraz doğrulama yönteminde, modelin en yüksek AUC değeri 0.827 ile k parametresinin değeri 4 olarak seçildiğinde elde edilmektedir bunun dışındaki değerlerde performans azalmaktadır. Farklı k parametreleri $k=2$, $k=3$ ve $k=4$ için elde edilen AUC değerleri sırasıyla 0.814, 0.813 ve 0.827'dir ve aşağıdaki Şekil 5.7., Şekil 5.8. ve Şekil 5.9.'da sırasıyla gösterilmektedir.

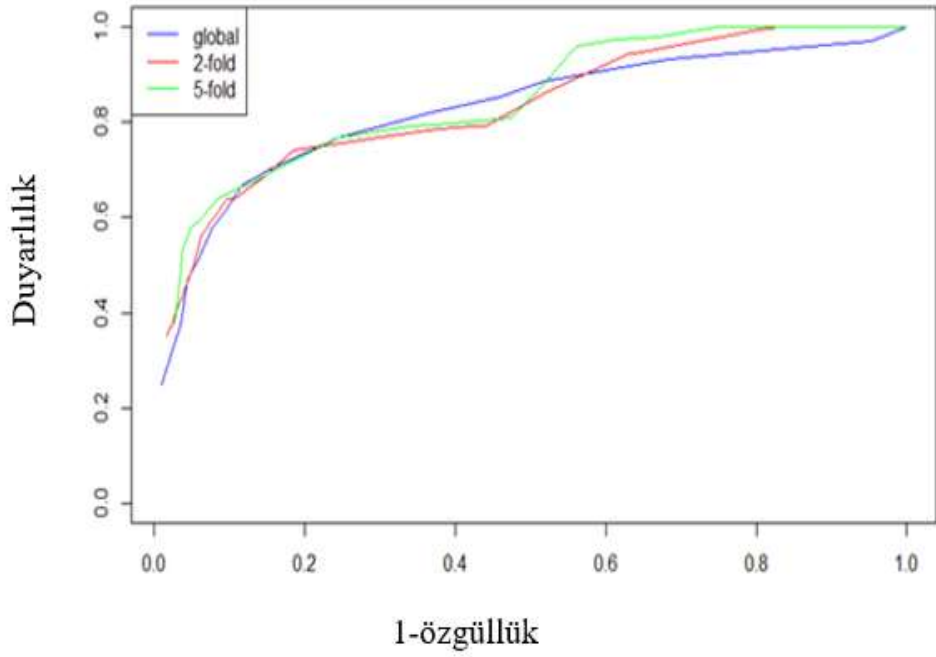
5-katlı çapraz doğrulama yönteminde, modelin en yüksek AUC değeri 0.838 ile k parametresinin değeri 4 olarak seçildiğinde elde edilmiştir. Farklı k parametreleri $k=2$, $k=3$ ve $k=4$ için elde edilen AUC değerleri sırasıyla 0.812, 0.823 ve 0.838'dir ve aşağıdaki Şekil 5.7., Şekil 5.8. ve Şekil 5.9.'da sırasıyla gösterilmektedir. k parametresinin farklı değerleri için 2-katlı ve 5-katlı çapraz doğrulama uygulamalarında elde edilen AUC değeri değişmektedir ve en yüksek AUC değeri k parametresi 4 olarak seçildiğinde elde edilmektedir. Bunun nedeni olarak, bağlantı sayılarının az olması değerlendirilmektedir.



Şekil 5.7. k parametresinin değeri 2 seçildiğinde, AUC sonuçları tek-çıkartımlı, 2-katlı ve 5-katlı çapraz doğrulama için sırasıyla 0.826, 0.814, 0.812'dir.



Şekil 5.8. k parametresinin değeri 3 seçildiğinde, AUC sonuçları tek-çıkarımlı, 2-katlı ve 5-katlı çapraz doğrulama için sırasıyla 0.855, 0.813 ve 0.823'dir.

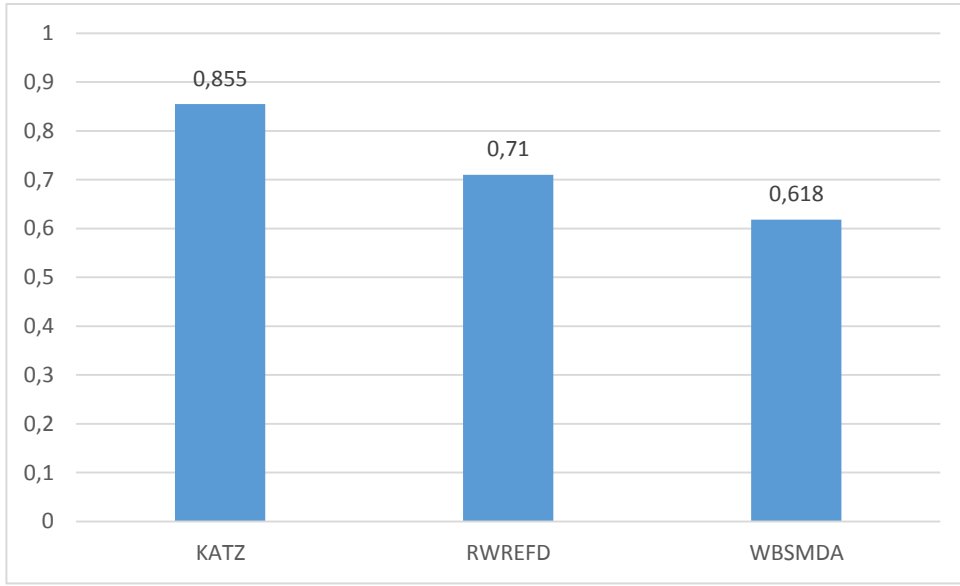


Şekil 5.9. k parametresinin değeri 4 seçildiğinde, AUC sonuçları tek-çıkarımlı, 2-katlı ve 5-katlı çapraz doğrulama için sırasıyla 0.823, 0.827 ve 0.838'dir.

5.2.8. Diğer Metodlar ile Karşılaştırma

LncRNA'lar ve çevresel faktörler arasındaki potansiyel yeni bağlantıların tahmin edilmesi amacıyla geliştirilen model daha önce geliştirilen modeller ile karşılaştırılmıştır. Bilindiği kadarıyla bu konuda yapılan ilk hesaplamalı model Zhou [126] tarafından geliştirilmiştir ve RWREFD olarak adlandırılmıştır. Bu çalışma; eğer bir lncRNA ve çevresel faktör ortak bir şekilde bir miRNA ile ilişkili ise o zaman bu lncRNA ile çevresel faktörün bağlantılı olma ihtimali yüksektir hipotezine dayanmaktadır. RWREFD modeli yeniden başlama rastgele yürüyüş algoritmasını (random walk with restart) temel almaktadır. Bu model ile tek-çıkışlı çapraz doğrulama yönteminde 0.71 AUC değeri elde edilmiştir.

Başka bir çalışma [127] iki parçalı ağlar üzerinde kaynak aktarma tekniği (resource transfer technique) kullanılarak geliştirilmiştir. Bu çalışmada, lncRNA'lar ve çevresel faktörler arasında yeni potansiyel bağlantılar tahmin edilerek elde edilen bağlantı tahminlerini içeren bir veritabanı oluşturulmuştur. Oluşturulan veritabanı ile lncRNA ve çevresel faktör bağlantılarını içeren başka bir veritabanı karşılaştırılmıştır. Zenginleştirme analiz tekniği (enrichment analysis technique) temel alınarak kendi modellerinin önceki modele göre daha başarılı sonuçlar gösterdiği çalışmanın yazarları tarafından belirtilmiştir ancak modelde sonuçların değerlendirilmesi için AUC değeri kullanılmadığından kendi modelimiz ile karşılaştırma yapılamamıştır. Modelimiz ile daha fazla model karşılaştırması yapmak amacıyla, miRNA'lar ile hastalıklar arasında potansiyel bağlantıların tahmin edilmesi amacıyla geliştirilmiş olan WBSMDA [128] modelini, kendi veri setimiz üzerinde uygulayarak, lncRNA ve çevresel faktörler arasında yeni potansiyel bağlantıların WBSMDA modeli ile tahmin edilmesi amaçlanmıştır. WBSMDA modeli kendi veri setimiz üzerinde uygulandığında 0.618 AUC değeri elde edilmiştir. Modelimiz, RWREFD ve WBSMDA modelleri ile karşılaştırıldığında aşağıdaki Şekil 5.10'da gösterildiği üzere diğer iki modelden daha yüksek AUC değeri elde ettiği görülmüştür. Sonuçlar, modelimizin lncRNA-çevresel faktör çiftleri arasındaki potansiyel tahminlerin hesaplanmasında daha yüksek başarı oranına sahip olduğunu göstermektedir.



Şekil 5.10. Modelimiz ile diğer yöntemlerin AUC değerlerinin karşılaştırılması.

5.2.9. Örnek Çalışmalar

Geliştirilen modelin tahmin performansını örnek çalışmalar ile değerlendirmek amacıyla iki önemli çevresel faktör olan sisplatin ve Hepatit B virüsü örnek adaylar olarak seçilmiştir. Sisplatin bir çok kanser hastalığının tedavisinde kullanılan önemli bir kemoterapi ilacıdır [129]. Sisplatinin etki şekli, DNA'daki pürin bazlarla çapraz bağlanma yeteneğinden kaynaklanmaktadır. Sisplatin ile bir çok lncRNA arasında kanıtlanmış ilişkiler bulunmaktadır. Sisplatin ile lncRNA'lar arasında örnek çalışma yapılırken mantıksal çemberden [130] kaçınmak için sisplatin ile lncRNA'lar arasındaki bütün bağlantılar eğitim setinden kaldırılmıştır. Modelimiz kullanılarak, sisplatin ile bağlantılı olabilecek potansiyel lncRNA'lar, bağlantı skorlarına göre büyükten küçüğe doğru sıralanmıştır. Literatür ve mevcut veritabanları ile karşılaştırılıp başarı oranı bulunmaya çalışılmıştır. Uygulama sonucunda elde edilen sıralamadaki ilk 10 lncRNA'dan 6 tanesinin DLREFD veritabanında, 2 tanesinin ise literatür çalışmalarında olduğu görülmüştür. Diğer bir çevresel faktör olarak Hepatit B virüsü seçilmiştir. Hepatit B virüsü insanlarda kronik enfeksiyonlara neden olan ve şu an dünya nüfusunun %3.5'ünü etkileyen önemli bir virüstür. Hepatit B virüsü ile çevresel faktörler arasında daha önce bulunmuş tüm bağlantılar mantıksal çemberden kaçınmak için kaldırılmıştır. Modelimiz kullanılarak, Hepatit B virüsü ile bağlantılı olabilecek potansiyel lncRNA'lar, bağlantı skorlarına göre büyükten küçüğe doğru sıralanmıştır. Literatür ve mevcut veritabanları ile karşılaştırılıp başarı oranı bulunmaya çalışılmıştır. Bağlantılar,

skorlarına göre büyükten küçüğe doğru sıralanmıştır. Sıralamada ki ilk 10 lncRNA'dan 1 tanesinin DLREFD veritabanında, 5 tanesinin ise literatür çalışmalarında olduğu görülmüştür.

5.2.10. Sonuçlar

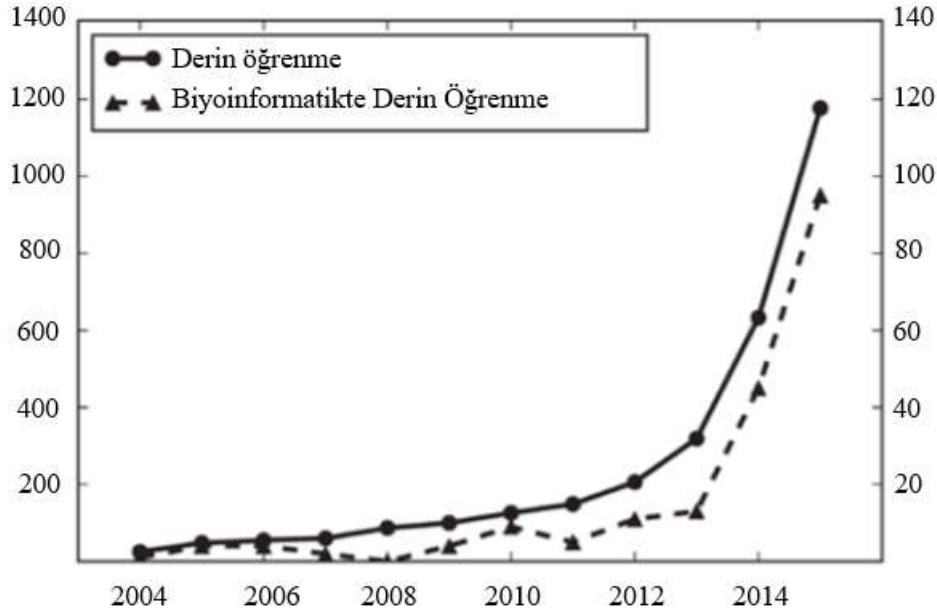
Genetik ve çevresel faktörler arasındaki ilişkilerin eksiksiz bir şekilde bilinmesi, birçok hastalığın karmaşık mekanizmalarının anlaşılmasına önemli bir katkı sağlayacaktır. LncRNA'lar ve ÇF'ler arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla daha önce geliştirilen birkaç model bulunmaktadır. Çalışmamızda geliştirilen lncRNA-ÇF modeli ile lncRNA'lar ve çevresel faktörler arasındaki yeni potansiyel bağlantıların daha yüksek başarı oranı ile tahmin edilmesi amaçlanmıştır. Veri seti olarak, lncRNA-ÇF'ler arasındaki kanıtlanmış bağlantıları içeren DLREFD veritabanı kullanılmıştır. Kullanılan DLREFD veritabanının sürümü Eylül 2017'de sunulmuştur. Veri seti temel alınarak, lncRNA-lncRNA ve ÇF-ÇF benzerlik matrisleri Gauss etkileşim profili çekirdek benzerliği kullanılarak hesaplanmıştır. Son olarak, KATZ ölçütü, yeni potansiyel lncRNA-ÇF bağlantılarının tahmin edilmesi amacıyla kullanılmıştır. Modelin başarısının test edilmesi amacıyla global LOOCV, 2 kat ve 5 kat çapraz doğrulama yöntemleri kullanılmıştır. Bu yöntemlerden elde edilen AUC değerleri geliştirilen modelin yüksek başarı oranına sahip olduğunu göstermektedir. Modelin tahmin performansının, örnekler üzerinde de test edilmesi amacıyla iki önemli çevresel faktör kullanılmıştır. Bu iki çevresel faktör üzerinde elde edilen başarı yüzdesi de önerilen modelin lncRNA'lar ve ÇF'ler arasındaki yeni potansiyel ilişkileri öngörmede yüksek performans değerlerine ve güvenilirliğe sahip olduğunu göstermektedir.

5.3. Otokodlayıcı ve Derin Öğrenme Tabanlı lncRNA-Hastalık Modeli

Günümüzde bir çok alanda ortaya çıkan veri miktarının artması sonucunda elde edilen büyük veriden önemli bilgilerin elde edilmesi giderek artan bir şekilde önem kazanmaktadır [131]. Biyoinformatik'te bu alanlardan bir tanesidir ve bu alanda sinyal verileri, görüntü ve omik verilerinden büyük veri elde edilmektedir. Elde edilen büyük veri, hem akademi hem de endüstri tarafından sağlık ve biyolojik çalışmalarında kullanılmaktadır. Örneğin, IBM onkoloji alanında hastaların tıbbi bilgilerini analiz eden ve klinik tedavi uzmanlarına tedavi seçenekleri konusunda destek sağlayan Watson platformunu geliştirmiştir [132]. Google, sağlık hizmetleri alanında önemli teknolojiler geliştirmek amacıyla yakın zamanda "DeepMind Health" uygulamasını geliştirmiştir [133]. Endüstri tarafından geliştirilen

biyoinformatik uygulamaların yanında akademik olarak biyoinformatik ile ilgili olarak yayınlanan makale sayısı giderek artış göstermektedir ve aşağıdaki Tablo 5.3.'te gösterilmektedir.

Tablo 5.3. Derin Öğrenme'nin toplam yayınlanan makale sayısı ile Derin Öğrenme'nin Biyoinformatik alanındaki makalele sayısı [134]



Çalışmamızda lncRNA'lar ile hastalıklar arasında yeni potansiyel bağlantıların bulunması amacıyla bir derin öğrenme algoritması olan istiflenmiş otokodlayıcılar ve derin sinir ağları kullanarak bir model geliştirilmiştir. Geliştirilen model iki kısımdan oluşan bir yapıdadır. İlk olarak lncRNA ve hastalık bilgileri vektörel formata dönüştürülüp istif edilmiş otokodlayıcılara verilmektedir. İstif edilmiş otokodlayıcılardan elde edilen öznitelik vektörleri birleştirilerek derin sinir ağlarına verilmektedir. Derin sinir ağları kullanılarak sınıflandırma işlemi yapılmaktadır. İstif edilmiş otokodlayıcılar ve derin sinir ağları biyoinformatik çalışmalarında daha önce kullanılmıştır ve aşağıdaki Tablo 5.4.'te gösterilmektedir.

Tablo 5.4. Derin Sinir Ağları ve İstif Edilmiş Otokodlayıcıların Biyoinformatik uygulamalarında kullanım alanları

Derin Sinir Ağları	Protein Yapısı [135]
	Gen ekspresyonu düzenleme [136]
	Protein sınıflandırması [137]
	Anomali sınıflandırması [138]
İstif Edilmiş Otokodlayıcılar	Gen ekspresyon tahmini [139]
	EEG tabanlı duygu sınıflandırması [140]
	Protein-protein etkileşim tahminleri [141]
	ncRNA-protein etkileşim tahminleri [142]

Geliştirilen model ilk olarak lncRNA-hastalık bağlantı bilgilerini veritabanından alarak bu bağlantı bilgilerinden lncRNA ve hastalık benzerlik matrisleri oluşturmaktadır. Daha fazla lncRNA ve hastalık benzerlik matrisi elde edilmesi amacıyla farklı veritabanlarından lncRNA ekspresyon bilgileri ve hastalık semantik bilgileri alınarak benzerlik matrisleri oluşturulmaktadır. Elde edilen lncRNA ve hastalık benzerlik matrisleri ayrı ayrı entegre edilerek, entegre benzerlik matrisleri oluşturulup istif edilmiş otokodlayıcılara aktarılmaktadır. İstif edilmiş otokodlayıcılardan elde edilen öznitelik bilgileri derin sinir ağlarına aktarılarak aralarında bağlantı bilinmeyen lncRNA-hastalık çiftleri arasından potansiyel bağlantı değerleri hesaplanmaktadır.

5.3.1. Veritabanı

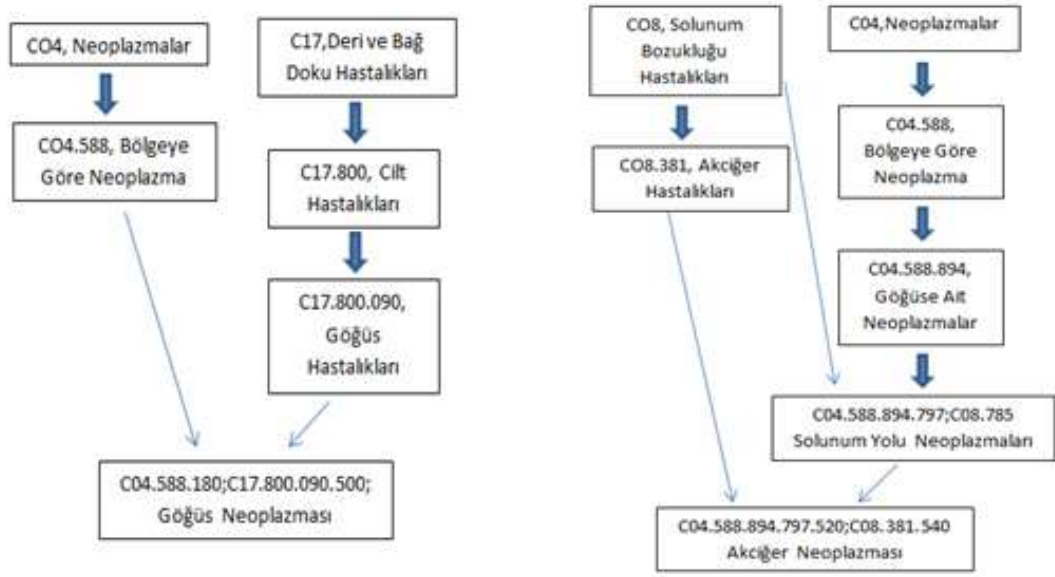
lncRNA-hastalık bağlantıları LncRNADisease veritabanından alınmıştır [87]. LncRNADisease veritabanında 246 hastalık ile 369 lncRNA arasında 687 tane bağlantı bulunmaktadır. Bu bağlantılar arasından MeSH veritabanında olmayan hastalıklar ve ArrayExpress [143] veritabanında olmayan lncRNA'lar çıkartıldığında geriye kalan 540 lncRNA-hastalık bağlantısı veriseti olarak kullanılmıştır.

5.3.2. LncRNA Ekspresyon Benzerliđi

LncRNA ekspresyon benzerlikleri ArrayExpress veritabanından alınmıřtır. ArrayExpress veritabanı 16 insan dokusu üzerinde 60000'den fazla lncRNA ekspresyon profili içermektedir. LncRNA ekspresyon benzerlikleri, her bir lncRNA çiftinin ekspresyon profiller arasında spearman korelasyonu katsayısı kullanılarak hesaplanmıřtır ve *EL* matrisinde gösterilmektedir. LncRNADisease veritabanında olup ArrayExpress veritabanında olmayan 254 lncRNA veri setimizden çıkartıldığında geriye 115 lncRNA kalmaktadır. LncRNA ekspresyon benzerlik matrisi 115x115 boyutundadır ve 115 lncRNA'nın birbirleri ile olan ekspresyon deđerlerini içermektedir.

5.3.3. Hastalık Semantik Benzerliđi

Geliřtirilen modelde, hastalık benzerlik matrisinin hesaplanmasında MeSH veritabanının C kategorisinde bulunan hastalık bařlıkları kullanılmıřtır. MeSH veritabanı yönlü çevrimsiz çizgesel (YÇÇ) řeklinde tarif edilebilmektedir; bu çizgede düđümler hastalıkları bađlantılar ise düđümler arasındaki iliřkileri temsil etmektedir. Her düđümün çizgedeki konumunu sayısal olarak tanımlayan bir veya daha fazla adresi vardır. Bir alt düđümün adresi bađlı bulunduđu üst düđüm veya düđümlerin adresine kendi özel adres kodunun eklenmesi ile oluřturulmaktadır. řekil 5.11'de bu iřlem gösterilmektedir. Çizgede alt düđümlere dođru hastalık bařlıkları daha spesifik yapılmakta iken üst düđümlerine dođru daha genel yapılmaktadır. Her hastalık YÇÇ řeklinde gösterilmektedir. Hastalıklar arasındaki benzerlikler, hastalıkların YÇÇ'leri temel alınarak hesaplanmaktadır.



Şekil 5.11. Hastalıkların yönlü çevrimsiz çizgesel olarak gösterilmesi

LncRNADisease veritabanında olup MeSH veritabanında olmayan 68 hastalık veritabanımızdan çıkartıldığında geriye 178 hastalık kalmaktadır. 178 hastalık için semantik benzerlik matrisi *SH* oluşturulmuştur ve bu matris 178x178 boyutundadır.

5.3.4. LncRNA'lar için Gauss Etkileşim Profili Çekirdek Benzerliği

Geliştirilen modelde; fonksiyonel olarak benzer lncRNA'lar, hastalıklar ile benzer veya benzer olmayan etkileşim desenleri gösterme eğilimindedir varsayımı [78] temel alınarak bütün lncRNA çiftleri arasındaki benzerliklerin hesaplanması amacıyla Gauss etkileşim profili çekirdek benzerliği metodu kullanılmıştır. İlk olarak her bir lncRNA $l(i)$ için ikili vektör $IP(l(i))$ şeklinde bir etkileşim profili IP oluşturulmuştur. Her bir lncRNA ile hastalıkların tümü karşılaştırılarak aralarında bağlantı olanlar için ikili vektörde karşılık gelen kısma 1, bağlantı olmayanlar için karşılık gelen kısma 0 değeri yazılmaktadır. Gauss çekirdeği aşağıdaki formülde gösterilmiştir;

$$GL(l(i), l(j)) = \exp(-\gamma_l \|IP(l(i)) - IP(l(j))\|^2) \quad (5.21)$$

Bant genişliğini kontrol etmek için y_l parametresi kullanılmıştır. Bu parametre aşağıdaki formül ile hesaplanmaktadır:

$$y_l = y'_l / (\sum_{k=1}^{n_k} ||IP(l(k))||^2) \quad (5.22)$$

n_k lncRNA'ların sayısını göstermektedir.

5.3.5. Hastalıklar için Gauss Etkileşim Profili Çekirdek Benzerliği

Hastalıklar arasındaki benzerlik değerlerin hesaplanması için lncRNA Gauss etkileşim profil çekirdek benzerliği ile benzer şekilde hastalıklar için Gauss etkileşim profili çekirdek benzerliği oluşturulmuştur. İlk olarak her bir hastalık $h(i)$ için ikili vektör $IP(h(i))$ şeklinde bir etkileşim profili IP oluşturulmuş ve lncRNA'ların her biri ile bağlantısının olup olmadığı kontrol edilerek eğer bağlantı var ise karşılık gelen kısma 1 ve bağlantı yok ise karşılık gelen kısma 0 değeri yazılmaktadır. Gauss çekirdeği aşağıdaki formülde gösterilmiştir:

$$GH(h(i), h(j)) = \exp(-y_e ||IP(h(i)) - IP(h(j))||^2) \quad (5.23)$$

Bant genişliğini kontrol etmek için y_e parametresi kullanılmıştır. Bu parametre aşağıdaki formül ile hesaplanmaktadır:

$$y_e = y'_e / (\sum_{e=1}^{n_e} ||IP(h(e))||^2) \quad (5.24)$$

n_e hastalıkların sayısını göstermektedir.

5.3.6. lncRNA'lar için Entegre Benzerlik Matrisi

Gauss lncRNA benzerlik matrisi ile lncRNA ekspresyon benzerlik matrisleri birleştirilerek lncRNA'lar için entegre benzerlik matrisi oluşturulmuştur. Gauss lncRNA benzerlik matrisinin boyutları ile lncRNA ekspresyon benzerlik matrisinin boyutları birbirine eşit

olduğundan dolayı bir çift $l(i), l(j)$ lncRNA için entegre benzerlik matrisi aşağıdaki şekilde hesaplanmaktadır:

$$LE(l(i), l(j)) = \left\{ \frac{GL(l(i), l(j)) + EL(l(i), l(j))}{2} \right\} \quad (5.25)$$

5.3.7. Hastalıklar için Entegre Benzerlik Matrisi

Gauss hastalık benzerlik matrisi ile hastalık semantik benzerlik matrisleri birleştirilerek hastalıklar için entegre benzerlik matrisi oluşturulmuştur. Gauss hastalık benzerlik matrisinin boyutları ile hastalık semantik benzerlik matrisinin boyutları birbirine eşit olduğundan dolayı bir çift $h(i), h(j)$ lncRNA için entegre benzerlik matrisi aşağıdaki şekilde hesaplanmaktadır:

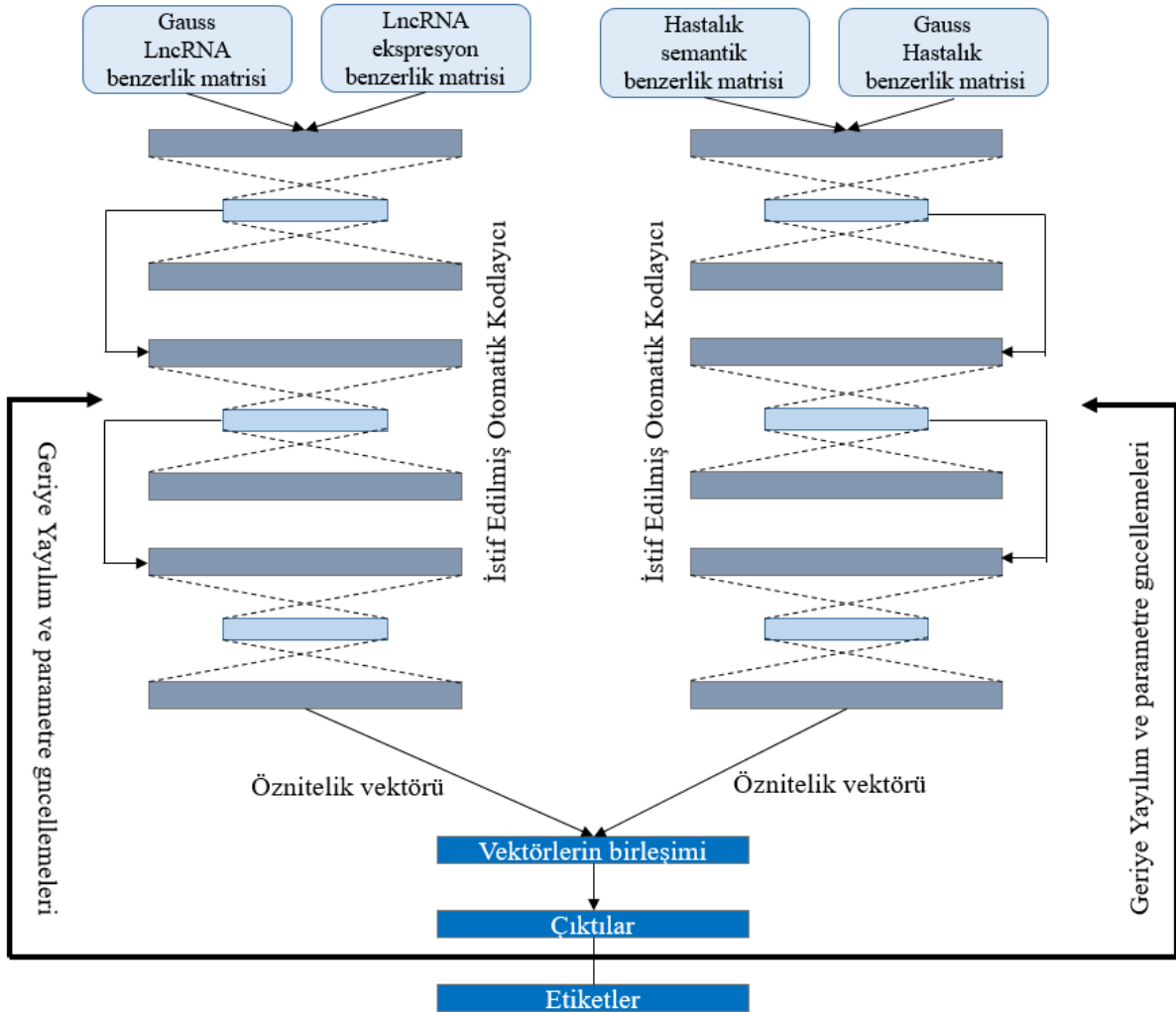
$$HE(h(i), h(j)) = \left\{ \frac{GH(h(i), h(j)) + SH(h(i), h(j))}{2} \right\} \quad (5.26)$$

5.3.8. Metod

Çalışmamızda lncRNA ile hastalıklar arasında yeni potansiyel bağlantıların tahmin edilmesi amacıyla bir derin öğrenme algoritması olan istif edilmiş otokodlayıcılar ve derin sinir ağları tabanlı model geliştirilmiştir. Geliştirilen model iki kısımdan oluşmaktadır; ilk kısımda istif edilmiş otokodlayıcılar ve sonraki kısımda derin sinir ağları bulunmaktadır.

lncRNA entegre benzerlik matrisi ve hastalık entegre benzerlik matrisinde bulunan veriler vektörel olarak ayrı ayrı istif edilmiş otokodlayıcılara aktarılmaktadır. Her bir istif edilmiş otokodlayıcı ile vektörlerden özniteliklerin çıkarılması amaçlanmaktadır. İstif edilmiş otokodlayıcılar üç otokodlayıcının birleştirilmesi ile oluşturulmuştur. İstiflenmiş otokodlayıcının vektörel çıktısının boyutu 32 olarak belirlenmiştir. Geliştirilen modelde, istiflenmiş otokodlayıcılardan özniteliklerin elde edilmesi amaçlanırken her bir otokodlayıcının katmanları üzerinde katmanların denetimsiz hırslı öneğitimi (greedy layer-wise unsupervised pretraining) uygulanmıştır [144]. Derin öğrenme algoritmalarında gizli katmanlara hırslı öneğitim uygulanmasının amacı gradyanın kaybolması (vanishing gradient)

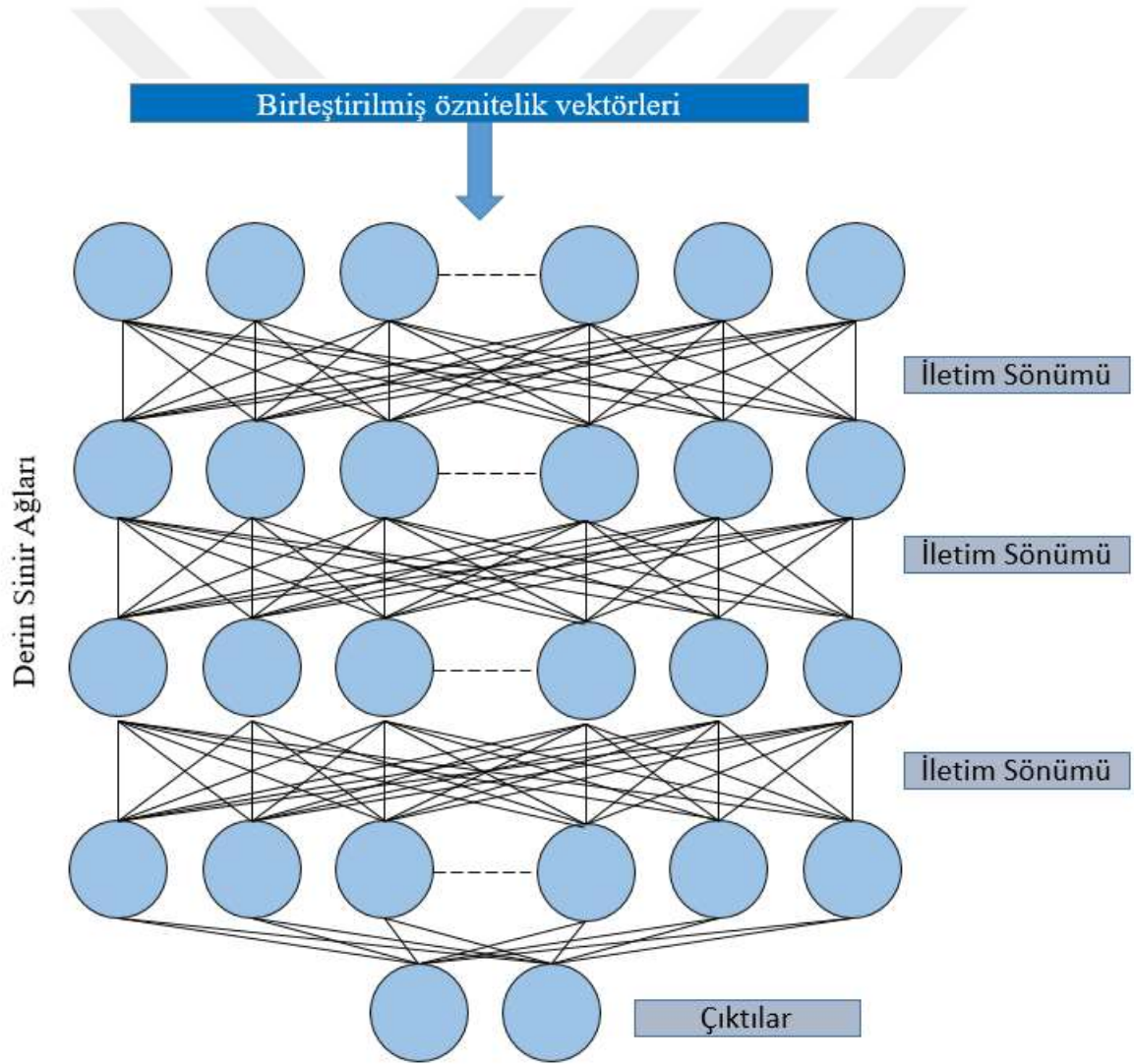
probleminin çözülmesidir [145]. Gradyanın kaybolması problemi, geri-yayılım (back-propagation) yapılırken ilk katmanlara doğru parametre güncellemelerinin yeterli düzeyde yapılamaması olarak tanımlanmaktadır. Her bir otokodlayıcıda aktivasyon fonksiyonu olarak ReLU kullanılmıştır. İstiflenmiş otokodlayıcılarda istenilen hedef ile çıktı arasındaki farkın ölçülmesinde kayıp fonksiyonlar kullanılmaktadır. Geliştirilen modelde kayıp fonksiyon olarak ortalama karesel hata kullanılmıştır.



Şekil 5.12. İstif edilmiş otokodlayıcılarda geriye-yayılım ve parametre güncellemeleri

Kayıp fonksiyon değerinin azaltılması için Şekil 5.12’de gösterildiği biçimde geri-yayılım temelinde hassas ayar uygulanmıştır. Geri-yayılımın amacı denetimli öğrenme kullanarak parametrelerin güncellenmesidir. Bunun sağlanması amacıyla son katmanda sigmoid

aktivasyonu kullanılarak softmax katmanı eklenmiştir ve eğitimin (parametre ve önyargı güncellemeleri) sağlanması amacıyla etiket bilgisinden yararlanılmıştır. Geri-yayılım uygulamasında kayıp fonksiyondan elde edilen değerin azaltılması amacıyla stokastik gradyan inişi kullanılmıştır. İstiflenmiş otokodlayıcılarda hassas ayar uygulanırken epok değeri 250, yığın büyüklüğü 1, öğrenme oranı 0.003 olarak belirlenmiş ve 0.9 değeri ile Momentum kullanılmıştır. İstif edilmiş otokodlayıcılarda parametrelerin başlangıç değerleri glorot düzgün başlatıcısı kullanılarak başlatılmıştır ve standart sapma 0.01 olarak belirlenmiştir. İstif edilmiş iki otokodlayıcı kullanılarak lncRNA ve hastalık vektörlerinden elde edilen özneliklerin oluşturduğu 32 boyutlu vektörler birleştirilerek 64 boyutlu entegre vektörler oluşturulmuştur. Oluşturulan entegre vektörler daha sonra üç katmanlı bir derin öğrenme ağına verilmektedir ve Şekil 5.13’de gösterilmektedir.

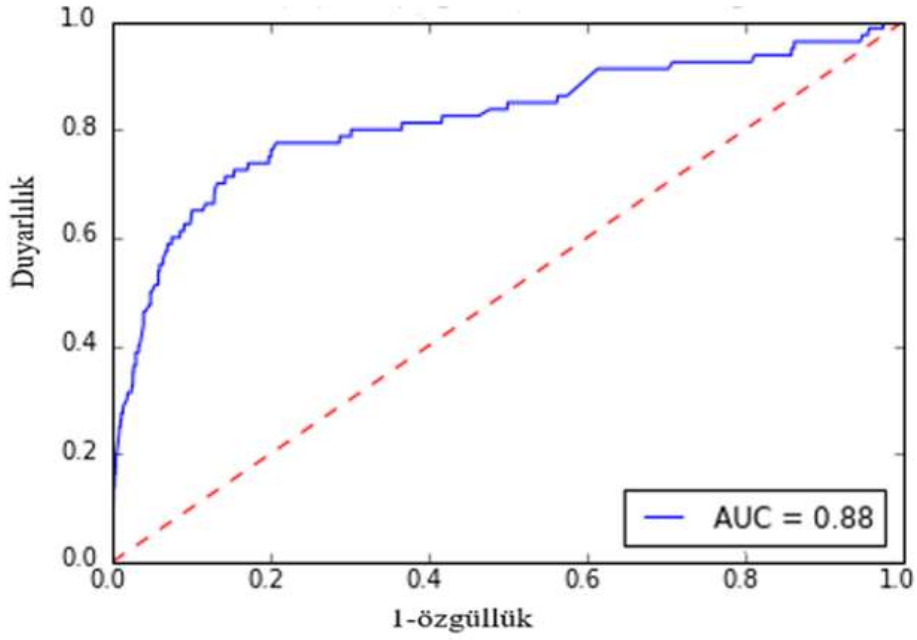


Şekil 5.13. İstif edilmiş otokodlayıcılardan elde edilen öznelik vektörlerinin derin sinir ağlarına aktarılması ve derin sinir ağlarından lncRNA-hastalık tahminlerinin elde edilmesi

Derin öğrenme katmanlarındaki nöron sayıları sırasıyla 64, 32 ve 16 olarak belirlenmiştir. Çıktı katmanı ise 2 nörondan oluşmaktadır ve lncRNA-hastalık çiftleri arasında potansiyel ilişkinin olup olmadığının belirlenmesi amacıyla bir eşik değeri belirlenmiştir. Derin öğrenme ağında aşırı eğitimin engellenmesi amacıyla iletim sönümü kullanılmıştır. Derin öğrenme ağında, her katmandan sonra ReLU aktivasyon fonksiyonu uygulanmıştır. Çıktı katmanında lncRNA-hastalık çiftlerinin tahmin edilmesi amacıyla bir sigmoid fonksiyonu kullanılmıştır. Derin sinir ağlarının çıktısı ile istenilen hedef arasındaki farkın hesaplanmasında ortalama karesel hata kayıp fonksiyonu kullanılmıştır. Bu farkın minimize edilmesi amacıyla bir öğrenme yapısı oluşturulmuştur. Öğrenme yapısında parametrelerin güncellenmesi amacıyla epok değeri 250 ve yığın büyüklüğü 1 olarak seçilmiştir. Derin sinir ağlarının eğitilmesinde stokastik gradyan inişi kullanılmıştır ve stokastik gradyan inişi algoritmasının optimizasyonunda Momentum değeri 0.7 seçilerek kullanılmıştır. Varsayılan öğrenme oranı 0.005 olarak seçilmiştir.

5.3.9. 5-katlı Çapraz Doğrulama Yöntemi

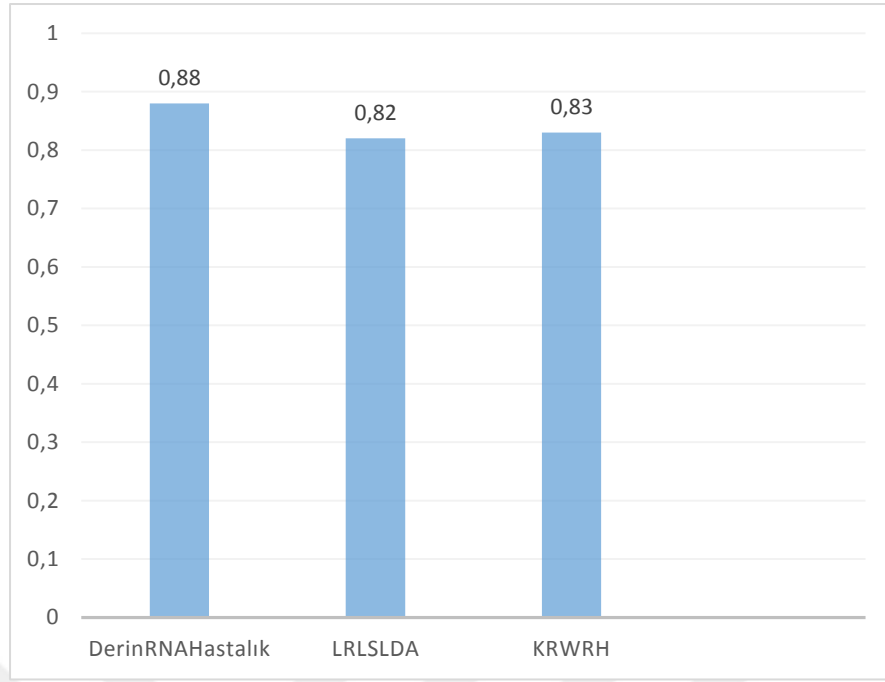
5-katlı çapraz doğrulama yönteminde bilinen tüm bağlantılar beş parçaya bölünmektedir. İlk parça test için diğerleri eğitim için seçilmektedir. Daha sonra, test ve eğitim için seçilen parçalar sırasıyla her biri için bir kez olmak üzere değiştirilmektedir ve bu sayede ROC eğrisi çizilerek AUC değeri elde edilmektedir. Geliştirilen modelin 5-katlı çapraz doğrulama yönteminde elde ettiği AUC değeri 0.886'dır ve aşağıdaki Şekil 5.14'te gösterilmektedir.



Şekil 5.14. Modelimizin 5-katlı çapraz doğrulama yönteminde elde ettiği AUC değeri

Geliştirilen modelin performansının ölçülmesinde daha önce geliştirilmiş olan LRLSLDA [95] ve KRWRH [146] modelleri ile karşılaştırma yapılmıştır. LRLSLDA modeli Laplas düzenlenmiş en küçük kareler yöntemi temel alınarak geliştirilmiş yarı öğrenimli bir öğrenme modelidir. KRWRH modeli ise lncRNA ekspresyon verileri, lncRNA-hastalık bağlantı bilgileri ve hastalık semantik bilgilerini heterojen bir ağda içeren ve bu ağ üzerinde yeniden başlama rastgele yürüyüş algoritmasını uygulayarak geliştirilen bir modeldir. Bahsedilen modellerin seçilmelerinin nedeni lncRNA-hastalık ağında aynı veri setini kullanarak en yüksek AUC sonuçlarını elde etmelerinden kaynaklanmaktadır.

LRLSLDA ve KRWRH modelleri, geliştirilen modelde kullanılan veri setine uygulandığında elde edilen AUC sonuçları aşağıdaki Şekil 5.15'te gösterilmektedir. Çalışmamızda geliştirilen model önceki modellere göre daha yüksek performans elde etmektedir.



Şekil 5.15. Modelimiz ile diğer yöntemlerin AUC değerlerinin karşılaştırılması

6.SONUÇLAR VE ÖNERİLER

Çalışmamızda insan vücudunda bulunan kodlama yapmayan uzun ve kısa boyutlu RNA'lar ile çevresel faktörler ve hastalıklar arasındaki bağlantıların tahmin edilmesi amacıyla modeller önerilmiştir. Çevresel faktörlerin, kısa boyutlu RNA çeşidi olan miRNA'ların ve uzun boyutlu RNA çeşidi olan lncRNA'ların ekspresyon düzeyini değiştirebildiği bilinmektedir. Ekspresyon düzeyi değişen miRNA ve lncRNA'ların ise hastalıklara neden olabildiği bilinmektedir. Bu nedenle, çevresel faktörler ile miRNA'lar ve lncRNA'lar arasındaki bağlantıların ve lncRNA'lar ile hastalıklar arasındaki bağlantıların bilinmesi, hastalıkların moleküler düzeyde teşhis ve tedavisi açısından önem arz etmektedir. Çok büyük sayıda miRNA, lncRNA, çevresel faktör ve hastalık olması nedeniyle yukarıda bahsedilen bağlantıların laboratuvar ortamında her biri için birer birer test edilerek bağlantının olup olmadığının belirlenmesi çok büyük insan kaynağı ve zaman gerektirmektedir. Geliştirilen lncRNA-çevresel faktör modeli, lncRNA'lar ve çevresel faktörler arasındaki potansiyel bağlantılı olabilecek lncRNA-çevresel faktör çiftlerini hesaplamaktadır. Geliştirilen model ile elde edilen potansiyel bağlantılı olabilecek lncRNA-çevresel faktör çiftlerinin laboratuvar ortamında test edilmesi ile lncRNA-çevresel faktör bağlantılarının hızlı bir şekilde belirlenmesi ve insan kaynağı ile zamandan tasarruf edilmesi amaçlanmıştır. Ek olarak, bu işlemin laboratuvar ortamında hızlanmasının moleküler düzeyde bağlantıların hızlı bir şekilde ortaya çıkarılmasına ve bu sayede hastalıkların teşhis ve tedavi yöntemlerinin hızlanmasına katkı sunulması amaçlanmıştır.

Geliştirilen ikinci model, miRNA ve çevresel faktörler arasında potansiyel bağlantılı olabilecek miRNA-çevresel faktör çiftlerini hesaplamaktadır. Tüm miRNA-çevresel faktör çiftlerinin laboratuvar ortamında birer birer test edilmesi büyük sayıda insan kaynağı ve zaman gerektirmektedir. Geliştirilen model ile elde edilen bağlantılı olabilecek miRNA-çevresel faktör çiftlerinin laboratuvar ortamında test edilmesi ile miRNA-çevresel faktör bağlantılarının hızlı bir şekilde belirlenmesi hedeflenmiştir ve bu sayede insan kaynağı ve zamandan tasarruf edilmesi amaçlanmıştır. MiRNA-çevresel faktörler arasındaki bağlantıların hızlı bir şekilde belirlenmesi ile hastalıkların moleküler düzeydeki teşhis ve tedavilerinin hızlanmasına katkı sunulması amaçlanmıştır.

LncRNA-çevresel faktörler ve miRNA-çevresel faktörler arasındaki potansiyel bağlantıların hesaplanması amacıyla geliştirilen modellere ek olarak lncRNA-hastalıklar arasındaki potansiyel bağlantıların hesaplanması amacıyla üçüncü bir model geliştirilmiştir.

Geliştirilen model ile lncRNA-hastalıklar arasındaki potansiyel bağlantıların hesaplanarak laboratuvar ortamında tüm lncRNA-hastalık çiftlerinin birer birer test edilmesi yerine modelden elde edilen potansiyel bağlantılı çiftlerin test edilerek lncRNA-hastalık bağlantılarının hızlı bir şekilde belirlenmesi ve bu sayede insan kaynağı ve zamandan tasarruf edilmesi amaçlanmıştır. LncRNA-hastalıklar arasındaki bağlantılarının belirlenmesi ile hastalıkların moleküler düzeyde teşhis ve tedavi yöntemlerinin hızlandırılmasına katkı sunulması amaçlanmıştır.

Çalışmamızda geliştirilen modellerden global tek çıkarımlı, 2-katlı ve 5-katlı çapraz geçерleme yöntemlerinde elde edilen AUC değeri daha önce geliştirilen modellerin AUC değeriinden daha yüksektir. Elde edilen sonuçlar geliştirilen modellerin başarısını göstermektedir. Bununla birlikte, modellerin yararlandığı veri tabanlarında bulunan verilerin artışı ile başarı oranının daha fazla artacağı düşünülmektedir. Benzerlik hesaplamalarında yeni metotların kullanılarak benzerlik matrislerinin artırılmasının da başarı oranını arttıracığı düşünülmektedir.

KAYNAKLAR

- [1] **Can T.**, 2014. Introduction to bioinformatics. *Methods Mol Biol.* 1107:51–71.
- [2] **Bayat A.**, 2002. Science, medicine, and the future: Bioinformatics. *BMJ.* 324:1018–1022.
- [3] **F. S. Collins, M. Morgan, and A. Patrinos**, 2003. The Human Genome Project: lessons from large-scale biology, *Science*, vol. 300, no. 5617, pp. 286–290.
- [4] **DiStefano JK.**, 2018. The emerging role of long noncoding RNAs in human disease. *Methods Mol Biol.* 1706:91–110.
- [5] **Karlsson O, Baccarelli AA.**, 2016. Environmental health and long non-coding RNAs. *Curr Environ Health Rep* 3:178–87
- [6] **X. Chen, M. X. Liu, Q. H. Cui, G. Y. Yan.**, 2012. Prediction of DiseaseRelated Interactions between MicroRNAs and Environmental Factors Based on a Semi-Supervised Classifier, *PloS one*, (7) e43425.
- [7] **L. Katz**, 1953. A new status index derived from sociometric analysis, *Psychometrika* 18 (1) 39–43.
- [8] **T. D Le, L. Liu, J. Zhang, B. Liu, J. Li**, 2015. From miRNA regulation to miRNA - TF co-regulation: computational approaches and challenges, *Briefings in Bioinformatics*, 16: 475-96.
- [9] **Wilusz JE, Sunwoo H, Spector DL.**, 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* 23: 1494–504.
- [10] **Guttman, M. et al.**, 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature.* 458, 223-7.
- [11] **Cabili, M.N. et al.**, 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25, 1915-27.
- [12] **Kung, J.T., Colognori, D. & Lee, J.T.**, 2013. Long noncoding RNAs: past, present, and future. *Genetics.* 193, 651-69.
- [13] **Mattick, J. S.**, 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* 2, 986–991.
- [14] **Taft RJ, Pheasant M, Mattick JS.**, 2007. The relationship between non- protein- coding DNA and eukaryotic complexity. *Bioessays* 29: 288–299.
- [15] **Mattick JS.**, 2001. Non- coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* 2: 986–991.
- [16] **Mattick JS.**, 2007. A new paradigm for developmental biology. *J Exp Biol* 210: 1526–1547.
- [17] **Tuncer HB and Bensan AE.**, 2014. Noncoding RNAs and cancer. *Turk. J. Biol.* 38 817–28.

- [18] **Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al.**, 2017. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*. 543:199–204.
- [19] **Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al.** 2015, The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*. 47:199–208.
- [20] **Long Y, Wang X, Youmans DT, Cech TR.**, 2017. How do lncRNAs regulate transcription? *Sci Adv*. 3:eaao2110.
- [21] **Fatica A, Bozzoni I.**, 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*. 15:7–21.
- [22] **Fortes P, Morris KV.**, 2016. Long noncoding RNAs in viral infections. *Virus Res*. 212:1–11.
- [23] **Heo JB, Sung S.**, 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*. 331:76–9.
- [24] **Lu Q, Ren S, Lu M., et al.**, 2013. Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14:651.
- [25] **Bussemakers MJ, van Bokhoven A, Verhaegh GW., et al.**, 1999. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res* 59:5975–9.
- [26] **Managadze D, Rogozin IB, Chernikova D., et al.**, 2011. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol* 3:1390–404.
- [27] **Bhartiya D, Kapoor S, Jalali S., et al.**, 2012. Conceptual approaches for lncRNA drug discovery and future strategies. *Expert Opin Drug Discov* 7:503–13.
- [28] **Zhao W, Luo J, Jiao S.**, 2014. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci Rep* 4:6591.
- [29] **Li J, Xuan Z, Liu C.**, 2013. Long non-coding RNAs and complex human diseases. *Int J Mol Sci* 14:18790–808.
- [30] **Schonrock N, Harvey RP, Mattick JS.** 2012. Long noncoding RNAs in cardiac development and pathophysiology. *Circ Res*, 111:1349–62.
- [31] **F. Wahid, A. Shehzad, T. Khan, Y.Y. Kim**, 2010. MicroRNAs: Synthesis, mechanism, function, and recent clinical trials, *Biochimica Biophysica Acta.*, 1803:1231–1243.
- [32] **Das J, Podder S, Ghosh TC.**, 2014. Insights into the miRNA regulations in human disease genes. *BMC Genomics.*, 15:1010.
- [33] **Giza DE, Vasilescu C, Calin GA.**, 2014. Key principles of miRNA involvement in human diseases. *Discoveries*. 2(4):e34.
- [34] **Bartel DP.**, 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 116(2):281–297.
- [35] **Volinia S, et al.** 2010. Reprogramming of miRNA networks in cancer and leukemia. *Genome Res*. 20(5):589–599.

- [36] **X. Chen**, 2016. miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method, *Molecular Biosyst.*, 624-33.
- [37] **Clayton D., McKeigue P.M.**, 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*. 358:1356–1360
- [38] **Kau AL, Ahern PP, Griffin NW, et al.** 2011. Human nutrition, the gut microbiome and the immune system. *Nature* 474:327e36.
- [39] **Epel ES, Lin J, Wilhelm FH, et al.** 2006. Cell aging in relation to stress arousal and cardiovascular disease risk factors. *Psychoneuroendocrinology* 31:277e87.
- [40] **Karagoz K, Sinha R, Arga KY.**, 2015. Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *Omics: A J Integr Biol* 19:115–30.
- [41] **Wang KC, Chang HY.**, 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* 43:904–14.
- [42] **Lottin S, Adriaenssens E, Dupressoir T., et al.**, 2002. Overexpression of an ectopic H19 gene enhances the tumorigenic properties of breast cancer cells. *Carcinogenesis* 23:1885–95
- [43] **Iacoangeli A, Lin Y, Morley EJ., et al.**, 2004. BC200 RNA in invasive and preinvasive breast cancer. *Carcinogenesis* 25:2125–33.
- [44] **Pasmant E, Laurendeau I, Héron D., et al.**, 2007. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res* 67:3963–9.
- [45] **Mourtada-Maarabouni M, Pickard M, Hedge V., et al.**, 2009. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* 28:195–208.
- [46] **Vincent-Salomon A, Ganem-Elbaz C, Manié E., et al.**, 2007. X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors. *Cancer Res* 67:5134–40.
- [47] **Wood SL, Pernemalm M, Crosbie PA., et al.**, 2015. Molecular histology of lung cancer: from targets to treatments. *Cancer Treat Rev* 41:361–75.
- [48] **Chen W, Böcker W, Brosius J., et al.**, 1997. Expression of neural BC200 RNA in human tumours. *J Pathol* 183:345–51.
- [49] **Barsyte-Lovejoy D, Lau SK, Boutros PC., et al.**, 2006. The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res* 66:5330–7.
- [50] **Shi X, Sun M, Liu H., et al.**, 2013. A critical role for the long non- coding RNA GAS5 in proliferation and apoptosis in non- small- cell lung cancer. *Mol Carcinog* 54:E1–12.
- [51] **Venook AP, Papandreou C, Furuse J., et al.**, 2010. The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective. *Oncologist* 15:5–13.

- [52] **Song P, Feng X, Zhang K., et al.**, 2013. Screening for and surveillance of high-risk patients with HBV-related chronic liver disease: promoting the early detection of hepatocellular carcinoma in China. *Biosci Trends* 7:1–6.
- [53] **Huang J-L, Zheng L, Hu Y-W.**, 2013. Characteristics of long noncoding RNA and its relation to hepatocellular carcinoma. *Carcinogenesis*, 35:507–14.
- [54] **He Y, Meng X-M, Huang C., et al.**, 2014. Long noncoding RNAs: novel insights into hepatocellular carcinoma. *Cancer Lett* 344:20–7.
- [55] **Panzitt K, Tschernatsch MM, Guelly C., et al.**, 2007. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology* 132:330–42.
- [56] **Yuan J-H, Yang F, Wang F., et al.**, 2014. A long noncoding RNA activated by TGF- β promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* 25:666–81.
- [57] **Huang J-L, Zheng L, Hu Y-W.**, 2013. Characteristics of long noncoding RNA and its relation to hepatocellular carcinoma. *Carcinogenesis* 35:507–14.
- [58] **Schonrock N, Götz J.**, 2012. Decoding the non-coding RNAs in Alzheimer's disease. *Cell Mol Life Sci* 69:3543–59.
- [59] **Palop JJ, Chin J, Mucke L.**, 2006. A network dysfunction perspective on neurodegenerative diseases. *Nature*, 443:768–73.
- [60] **Faghihi MA, Modarresi F, Khalil AM., et al.**, 2008. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β -secretase. *Nat Med*. 14:723–30.
- [61] **Ng S-Y, Lin L, Soh BS., et al.**, 2013. Long noncoding RNAs in development and disease of the central nervous system. *Trends Genet* 29:461–8.
- [62] **Kaya Z, Leib C, Katus HA.**, 2012. Autoantibodies in heart failure and cardiac dysfunction. *Circ Res* 110:145–58.
- [63] **Papait R, Kunderfranco P, Stirparo GG., et al.**, 2013. Long noncoding RNA: a new player of heart failure? *J Cardiovasc Transl Res* 6:876–83.
- [64] **Grote P, Wittler L, Hendrix D., et al.** 2013. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell*. 24:206–14.
- [65] **Schonrock N, Harvey RP, Mattick JS.**, 2012. Long noncoding RNAs in cardiac development and pathophysiology. *Circ Res* 111:1349–62.
- [66] **Kumarswamy R, Bauters C, Volkmann I., et al.** 2014. Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. *Circ Res* 114:1569–75.
- [67] **X. Chen, M. X. Liu, Q. H. Cui, G. Y. Yan**, 2012. Prediction of DiseaseRelated Interactions between MicroRNAs and Environmental Factors Based on a Semi-Supervised Classifier, *PloS one*, (7) e43425.

- [68] **H. Vural, B. Kaya, R. Alhajj, M. Kaya**, 2018. Prediction of New Potential MicroRNAs Environmental Factor Associations Based on KATZ Measure, *ASONAM*, 1;102:120-125.
- [69] **J. Wang, Q. H. Cui**, 2012. Specific roles of microRNAs in their interactions with environmental factors, *J Nucleic Acids*, 2012, 978384.
- [70] **R. Kulshreshtha, R.V. Davuluri, G.A. Calin, M. Ivan**, 2008. A microRNA component of the hypoxic response, *Cell Death&Differentiation*, 15:667-71.
- [71] **S. Xi , H. Xu, J. Shan**, 2013. Cigarette smoke mediates epigenetic repression of miR-487b during pulmonary carcinogenesis, *The Journal of Clinical Investigation*, 123 (3), 1241-1261.
- [72] **K.E. Cosselman, A. Navas-Acien, J.D.Kaufman**, 2015. Environmental factors in cardiovascular disease, *Nature Reviews Cardiology*, (12) 627-42.
- [73] **P. Boffetta, M. Hashibe**, 2006. Alcohol and cancer, *Lancet Oncol.*, (7) 149- 56.
- [74] **D. Mager**, 2006. Bacteria and cancer: cause, coincidence or cure? A review, *Journal of Translational Medicine*, 4;14.
- [75] **Kendler KS.**, 1986. Models for the joint effect of genotype and environment on liability to psychiatric illness. *Am J Psychiatry*, 143:279- 289.
- [76] **M.G. Andreassi**, 2009. Metabolic syndrome, diabetes and atherosclerosis: influence of gene-environment interaction, *Mutat. Res.* 667;35–43.
- [77] **E. Flowers, E.S. Froelicher, B.E. Aouizerat**, 2012. Gene-environment interactions in cardiovascular disease, *Eur. J. Cardiovasc. Nurs.* 11 (4); 472–478.
- [78] **Chen X., Yan G.Y.**, 2013. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinform.* 29;2617–2624.
- [79] **Flynn R.A., Chang H.Y.**, 2014. Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell.* 14;752–761.
- [80] **Karlsson O., Baccarelli A.A.**, 2016. Environmental health and long non-coding RNAs. *Curr. Environ. Health Rep.* 3;178–187.
- [81] **Szegedi K., Goblos A., Bacsá S. et al.** 2012. Expression and functional studies on the noncoding RNA, PRINS. *Int. J. Mol. Sci.* 14;205–225.
- [82] **Bhan A., Hussain I., Ansari K.I. et al.** 2014. Bisphenol-A and diethylstilbestrol exposure induces the expression of breast cancer associated long noncoding RNA HOTAIR in vitro and in vivo. *J. Steroid Biochem. Mol. Biol.* 141;160–170.
- [83] **Shamovsky I., Ivannikov M., Kandel E.S. et al.** 2006. RNA-mediated response to heat shock in mammalian cells. *Nature.* 440;556–560.
- [84] **Y. Zhao, H. Li, S. Fang, Y. Kang, W. Wu, Y. Hao, Z. Li, D. Bu, N. Sun, M. Q. Zhang, R. Chen**, 2016. NONCODE 2016: An informative and valuable data source of long non-coding RNAs, *Nucleic Acids Research* 44 (D1) D203–D208.
- [85] **Bhartiya D, Pal K, Ghosh S, et al.** 2013. lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database* 2013;2013:bat034.

- [86] **Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, et al.** 2015. lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* 43:D168–73.
- [87] **Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q.,** 2013. lncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41(Database issue):D983–D986.
- [88] **Ning S, Zhang J, Wang P., et al.** 2016. lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res* 44:D980–5.
- [89] **Yang Q, Qiu C, Yang J, et al.,** 2011. miREnvironment database: providing a bridge for microRNAs, environmental factors and phenotypes. *Bioinformatics* 27:3329–30.
- [90] **Li J, Wu Z, Cheng F, et al.,** 2014. Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology. *Sci Rep* 4:5576
- [91] **Sun YZ, Zhang DH, Ming Z, et al.,** 2017. DLREFD: a database providing associations of long non-coding RNAs, environmental factors and phenotypes. *Database* 2017: bax084.
- [92] **Li J, Ma W, Zeng P., et al.,**2015. lncTar: a tool for predicting the RNA targets of long noncoding RNAs. *Brief Bioinformatics.* 16:806–12.
- [93] **Ferrè F, Colantoni A, Helmer-Citterich M.,**2016. Revealing protein–lncRNA interaction. *Brief Bioinformatics.* 17:106–16.
- [94] **Yotsukura S, Hancock T, Natsume-Kitatani Y., et al.,** 2016. Computational recognition for long non-coding RNA (lncRNA): software and databases. *Brief Bioinformatics* 2016;bbv114.
- [95] **Chen X, Yan G-Y.,** 2013. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics.* 29:2617–24.
- [96] **Chen X, Yan CC, Luo C., et al.,**2015. Constructing lncRNA functional similarity network based on lncRNA–disease associations and disease semantic similarity. *Sci Rep* 5:11338.
- [97] **Zhao T, Xu J, Liu L., et al.,** 2015. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Mol Biosyst.* 11:126–36.
- [98] **Sun J, Shi H, Wang Z., et al.,** 2014. Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. *Mol Biosyst.* 10:2074–81.
- [99] **Liu Y, Zhang R, Qiu F., et al.** Construction of a lncRNA–PCG bipartite network and identification of cancer-related lncRNAs: a case study in prostate cancer.
- [100] **Liu M-X, Chen X, Chen G., et al.,** 2014. A computational framework to infer human disease-associated long noncoding RNAs. *PLoS One* 9:e84408.
- [101] **M. Zhou, L. Han, J. Zhang, D. Hao, Y. Cai, Z. Wang, H. Zhou, J. Sun,** 2014. A computational frame and resource for understanding the lncRNA–environmental factor

associations and prediction of environmental factors implicated in diseases, *Mol. Biosyst.* 10; 3264–3271.

[102] **J. Zhou, Y.Y. Shi.**, 2018. A bipartite network and resource transfer-based approach to infer lncRNA-environmental factor associations, *IEEE ACM Trans. Comput. Biol. Bioinf* 15 (3) 753–759.

[103] **Chen X.**, 2016. miREFRWR: a novel disease-related microRNA–environmental factor interactions prediction method. *Mol Biosyst* 12:624–33.

[104] **Jiang W., Chen X., Liao M., Li W., Lian B., Wang L., Meng F., Liu X., Chen X., Jin Y., et al.**, 2012. Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Sci. Rep.* 2:282.

[105] **L. Katz**, 1953. A new status index derived from sociometric analysis, *Psychometrika* 18 (1) 39–43.

[106] **Y. Lu, Y. Guo, A. Korhonen**, 2017. Link prediction in drug-target interactions network using similarity indices, *BMC Bioinf.* 18 (1) 39.

[107] **U.M. Singh-Blom, N. Natarajan, A. Tewari, J.O. Woods, I.S. Dhillon, E.M. Marcotte**, 2013. Prediction and validation of gene-disease associations using methods inspired by social network analyses, *PLoS One* 8:e58977.

[108] **Y. Qu, H. Zhang, C. Liang, X. Dong**, 2018. KATZMDA: Prediction of miRNA-disease Associations Based on KATZ Model, *IEEE Access*, pp. 3943–3950.

[109] **van Laarhoven, T., Nabuurs, S. B. & Marchiori, E.**, 2011. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043.

[110] **Caruso, J.C., & Cliff, N.**, 1997. Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, 57, 637–654.

[111] **Bhattacharya S., Ha-Thuc V., Srinivasan P.**, 2011. MeSH: a window into full text for document summarization. *Bioinformatics*. 2011; 27:i120–i128.

[112] **Warren S. McCulloch, W. Pitts.**, 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, vol. 5 (4), p115–133.

[113] **Frank Rosenblatt**, 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, vol. 65(6), p386–408.

[114] **Marvin Minsky, S. Papert**, 1969. Perceptrons: An Introduction to Computational Geometry, *MIT Press*.

[115] **Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung**. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 6789 ;947.

[116] **George. E. Nasr, E. Badr, C. Joun**, 2002. Cross Entropy Error Function in Neural Networks: Forecasting Gasoline Demand. *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, p381–384.

- [117] **Sutton, R. S.**, 1986. Two problems with backpropagation and other steepest-descent learning procedures for networks. *Proc. 8th Annual Conf. Cognitive Science Society*.
- [118] **Duchi, J., Hazan, E., & Singer, Y.**, 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12, 2121–2159.
- [119] **Zeiler, M. D.**, 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701*.
- [120] **N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov**, *Dropout: a simple way to prevent neural networks from overfitting*, *Journal of Machine Learning Research*, p1929-1958, 2014.
- [121] **Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol PA.**, 2010. *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion*, *J Mach Learn Res*, 11:3371–408.
- [121] **R. Hinton, G. and Salakhutdinov**, 2006. *Reducing the dimensionality of data with neural networks*, *Science*, vol. 313, no. 5786, pp. 504–507.
- [122] **Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol PA.**, 2010. *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion*, *J Mach Learn Res*, 11:3371–408.
- [123] **X. Chen, M.X. Liu, Q.H. Cui, G.Y. Yan**, 2012. Prediction of disease-related interactions between micromas and environmental factors based on a semi-supervised classifier, *PloS one*, vol. 7, no. 8, p. e43425.
- [124] **T. Fawcett**, 2006. An introduction to ROC analysis, *Pattern Recogn. Lett.* 27 861–874.
- [125] **Zhou M, Han L, Zhang J.**, 2014. A computational frame and resource for understanding the lncRNA-environmental factor associations and prediction of environmental factors implicated in diseases. *Mol Biosyst* 10:3264–71.
- [126] **M. Zhou, L. Han, J. Zhang, D. Hao, Y. Cai, Z. Wang, H. Zhou, J. Sun**, 2014. A computational frame and resource for understanding the lncRNA-environmental factor associations and prediction of environmental factors implicated in diseases, *Mol. Biosyst.* 10 3264–3271.
- [127] **J. Zhou, Y.Y. Shi**, 2018. A bipartite network and resource transfer-based approach to infer lncRNA-environmental factor associations, *IEEE ACM Trans. Comput. Biol. Bioinf* 15 (3) 753–759.
- [128] **X. Chen, C.C. Yan, X. Zhang, et al.**, 2016. WBSMDA: within and between score for MiRNA–disease association prediction, *Sci. Rep.* 6;21106.
- [129] **A. Florea, D. Büsselberg**, 2011. Cisplatin as an anti-tumor drug: cellular mechanisms of activity, drug resistance and induced side effects, *Cancers* 3 (1) 1351–1371.
- [130] **P. Jiajie, H. Weiwei, L. Qianqian, C. Bolin, W. Zhongyu, S. Xuequn**, 2018. A learning-based framework for miRNA-disease association prediction using neural networks, *bioRxiv*.

- [131] **Manyika J, Chui M, Brown B et al.**, 2011. Big data: The next frontier for innovation, competition, and productivity.
- [132] **Ferrucci D., Brown E.**, 2010. Chu-Carroll J et al. Building Watson: An overview of the DeepQA project. *AI magazine* 31(3):59-79.
- [133] DeepMind Health. 2016. Google DeepMind. <https://www.deepmind.com/health>.
- [134] **S. Min, B. Lee, S. Yoon**, 2017. Deep learning in bioinformatics. *Brief Bioinform.* 18:851-69.
- [135] **Spencer M, Eickholt J, Cheng J.**, 2015. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on.* 12(1):103-12.
- [136] **Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C.**, 2016. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* 44(4):e32.
- [137] **Asgari E, Mofrad MR.** 2015. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PloS one.* 10(11):e0141287.
- [138] **Fakoor R, Ladhak F, Nazi A et al.** 2013. Using deep learning to enhance cancer diagnosis and classification. *Proceedings of the International Conference on Machine Learning.*
- [139] **Xie, R., Wen, J., Quitadamo, A., Cheng, J., Shi, X.**, 2017. A deep auto-encoder model for gene expression prediction. *BMC Genomics.* 18 (Suppl. 9), 845.
- [140] **Jirayucharoensak S, Pan-Ngum S, Israsena P.**, 2014. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal.* 2014:627892-1–627892-10.
- [141] **Wang, Y.B., You, Z.H., Li, X. Jiang, T.H. Chen, X., Zhou, X., Wang, L.**, 2017. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. Biosyst.* 13, 1336–1344.
- [142] **Zhao Z., Li J., Yong Z., Li L., Hai Y.**, 2019. BGFE: A Deep Learning Model for ncRNA-Protein Interaction Predictions Based on Improved Sequence Information. *Int J Mol Sci* 20(4): 978.
- [143] **Parkinson H.**, 2007. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic acids research.* 2007;35:D747–D750.
- [144] **Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.**, 2006. Greedy layer-wise training of deep networks. In *Proc. Advances in Neural Information Processing Systems*, 19 153–160.
- [145] **Hochreiter S.**, 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

[146] **Ganegoda, G. U., Li, M., Wang, W. & Feng, Q.**, 2015. Heterogeneous network model to infer human disease-long intergenic non-coding RNA associations. *IEEE transactions on nanobioscience* 14, 175–183.



ÖZGEÇMİŞ

Hüseyin VURAL

Eğitim

Lise

Adıyaman Anadolu Öğretmen Lisesi (2002)

Lisans

Bilkent Üniversitesi/Bilgisayar Mühendisliği (2008)

Yüksek Lisans

Trakya Üniversitesi/Bilgisayar Mühendisliği/Bilgisayar Donanımı A.B.D

Simetrik Şifreleme Tekniklerinde Anahtar Planlama (2014)

Danışman: Doç. Dr. Muharrem Tolga SAKALLI

Doktora

Fırat Üniversitesi/Bilgisayar Mühendisliği/ Kuramsal Temeller A.B.D.

Kodlama Yapmayan Uzun Boyutlu RNA'lar ile Hastalıklar Arasında Yeni Bağlantılar Tahmini

Danışman: Prof. Dr. Mehmet KAYA

İş Deneyimi

- Türk Ekonomi Bankası (TEB)/Genel Müdürlük /Bilgi Teknolojileri Grubu/
Uygulama Geliştirme Bölümü/Bilgisayar Mühendisi
İstanbul (2008-2011)
- Bilgisayar Teknolojileri Bölümü, Kahta MYO,
Adıyaman Üniversitesi (2011-..)