TOPLULUK ÖĞRENME YÖNTEMLERİNE DAYALI BİLGİSAYAR DESTEKLİ TANI SİSTEMİNİN GELİŞTİRİLMESİ: GENOMİK TEKNOLOJİLERİ ÜZERİNE UYGULAMASI

Zeynep KÜÇÜKAKÇALI

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

Tez Danışmanı
Doç. Dr. Harika Gözde GÖZÜKARA BAĞ

Doktora Tezi- 2023

T.C.
İNÖNÜ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TOPLULUK ÖĞRENME YÖNTEMLERİNE DAYALI BİLGİSAYAR DESTEKLİ TANI SİSTEMİNİN GELİŞTİRİLMESİ: GENOMİK TEKNOLOJİLERİ ÜZERİNE UYGULAMASI

Zeynep KÜÇÜKAKÇALI

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı
Doktora Tezi

Tez Danışmanı
Doç. Dr. Harika Gözde GÖZÜKARA BAĞ

MALATYA
2023

T.C. İNÖNÜ ÜNİVERSİTESİ

Sağlık Bilimleri Enstitüsü Müdürlüğüne

ETİK BEYANI

İnönü Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak "Doç. Dr. Harika Gözde GÖZÜKARA BAĞ" danışmanlığında hazırlayıp sunduğum "Topluluk Öğrenme Yöntemlerine Dayalı Bilgisayar Destekli Tanı Sisteminin Geliştirilmesi: Genomik Teknolojileri Üzerine Uygulaması" başlıklı Doktora tezim içinde elde ettiğim verileri, bilgileri, belgeleri akademik ve etik kurallar çerçevesinde elde ettiğimi, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tezimde yararlandığım eserlere bilimsel kurallara uygun atıfta bulunarak kaynak gösterdiğimi, tezimin özgün olduğunu, tezimin çalışma ve yazımında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. 12/07/2023

Zeynep KÜÇÜKAKÇALI

İÇİNDEKİLER

TEŞEKKÜR

Bu çalışmanın planlanması, yürütülmesi sürecinde benden desteğini esirgemeyen, lisansüstü eğitimim boyunca, bilgi ve hoşgörülerinden yararlandığım danışman hocam Sayın Doç. Dr. Harika Gözde GÖZÜKARA BAĞ'a, lisansüstü eğitimim boyunca desteğini hiç esirgemeyen ve her zaman deneyimlerini paylaşarak bana ışık tutan değerli hocam Sayın Prof. Dr. Cemil ÇOLAK' a, tez çalışma süresince her daim desteğini hissettiğim bölüm kurucumuz Prof. Dr. Saim YOLOĞLU'na ve anabilim dalımız diğer kıymetli hocaları Doç. Dr. Emek GÜLDOĞAN ve Dr. Öğr. Üyesi Ahmet Kadir ARSLAN'a, bu tez çalışmasının izleme raporlarında değerli görüşleriyle çalışmama katkı sağlayan Sayın Prof. Dr. Hakan PARLAKPINAR'a, akademik çalışmalarımda yanımda olan, bilgi ve birikimlerini esirgemeden benimle paylaşan Prof. Dr. Ahmet Sami AKBULUT hocama,

17 yıldır her anımda yanımda olup destekleyen canım arkadaşım, kardeşim ve aynı zamanda aynı bölümde görev yaptığım kıymetli asistan arkadaşım Arş. Gör. İpek BALIKÇI ÇİÇEK'e, aynı anabilim dalında görev yaptığım bir diğer asistan arkadaşım ama sadece mesai ile kalmayıp her anımda yanımda olan, aslında en çok beni seven can dostum, arkadaşım Arş. Gör. Şeyma YAŞAR'a ve son olarak destekleri için asistan arkadaşım Fatma Hilal YAĞIN'a teşekkür ederim.

Bu tez çalışmamda beni hiçbir zaman yalnız bırakmayan sevgili kızım Okyanus Balın ve sevgili oğlum Rüzgar Diren'e sonsuz teşekkür ederim. Son olarak her zaman yanımda olan sevgili kardeşime ve beni yetiştirip bugünlere gelmemi sağlayan sevgili anne ve babama sonsuz teşekkürlerimi sunarım.

Arş. Gör. Zeynep KÜÇÜKAKÇALI

ÖZET

Topluluk Öğrenme Yöntemlerine Dayalı Bilgisayar Destekli Tanı Sisteminin Geliştirilmesi: Genomik Teknolojileri Üzerine Uygulaması

Amaç: Bu çalışmanın amacı sisplatin ile hepatotoksisite oluşturulmuş sıçanlar ile patoloji oluşturulmamış sıçanlardan alınan karaciğer doku örneklerinin genomik analizlerinin yapılması ile elde edilen büyük verilerin biyoinformatik analizlerinin yapılması amacıyla bir bir web-tabanlı yazılımın geliştirilmesidir. Ayrıca verilerin topluluk öğrenmesi metodları ile modellenmesi sonucunda hepatotoksisite ile ilişkili tanıya/erken tanıya yönelik olası biyobelirteçlerin tespit edilmesi amaçlanmıştır.

Materyal ve Metot: Çalışmada 20 adet dişi Sprague-Dawley sıçanın alınmasıyla oluşturulan bir deney düzeneğinden elde edilen genomik veriler kullanılmıştır. Biyoinformatik analizlerin yapılması için geliştirilen web tabanlı yazılım R programlama dili temelinde interaktif web tabanlı uygulamaların tasarlanmasına izin veren Shiny kütüphanesi ve ggplot2, ggrepel, DT, shinyWidgets, shinyLP, shinydashboard ve limma paketleri kullanılmıştır. Modellemelerde ise topluluk öğrenme yöntemlerinden bagging, boosting, stacking ve XGBoost modelleri kullanılmıştır.

Bulgular: Çalışmada kullanılan genomik veri seti 16.386 ifade içermektedir. Biyoinformatik analiz sonuçlarına göre hepatotoksisite ve kontrol grupları için 589 adet lncRNA gruplarda farklı ekspresyon göstermiştir. Bunlardan 450 tanesi yukarı ekspresyon göstermişken 139 tanesi aşağı ekspresyon göstermiştir. Değişken seçimi ile seçilen lncRNA lar ile yapılan modellemeler sonucunda XGBoost modeli en yüksek performans metriklerine sahip bulunmuştur.

Sonuç: Bu çalışma ile biyoinformatik analiz yapılmasına olanak sağlayan bir yazılım geliştirilmiştir. Ayrıca yapılan biyoinformatik analiz ve modellemeler sonucunda hepatotoksisitesi olan sıçanlar ile kontrol grubunda yer alan sıçanların lncRNA ekspresyon verileri kullanılarak hepatotoksisite için potansiyel genomik biyobelirteçleri belirlenmiştir.

Anahtar Kelimeler: Bilgisayar Destekli Tanı Sistemi, Genomik, Makine Öğrenmesi, Sınıflandırma, Topluluk Öğrenme, Yapay Zekâ

ABSTRACT

Development of Computer-Aided Diagnosis System Based on Ensemble Learning Methods: Application on Genomic Technologies

Aim: The aim of this study is to develop web-based software for bioinformatic analysis of large data obtained by performing genomic analysis of liver tissue samples from rats with cisplatin hepatotoxicity and rats without pathology. In addition, as a result of modeling the data with community learning methods, it was aimed to determine possible biomarkers for diagnosis/early diagnosis related to hepatotoxicity.

Material and Method: Genomic data obtained from an experimental setup created by taking 20 female Sprague-Dawley rats were used in the study. The web-based software developed for bioinformatics analysis is designed using the Shiny library, which allows interactive web-based applications to be designed based on the R programming language, and ggplot2, ggrepel, DT, shinyWidgets, shinyLP, shinydashboard, and limma the packages. In the models, bagging, boosting, stacking, and XGBoost models from ensemble learning methods were used.

Results: The genomic dataset used in the study contains 16,386 expressions. According to the results of the bioinformatics analysis, 589 lncRNAs for hepatotoxicity and control groups showed different expressions in the groups. Of these, 450 showed up-expression, while 139 showed down-expression. As a result of modeling with lncRNAs selected by the variable selection, the XGBoost model had the highest performance metrics.

Conclusion: With this study, a software was developed that allows bioinformatic analysis. In addition, as a result of bioinformatic analysis and modeling, potential genomic biomarkers for hepatotoxicity were determined using lncRNA expression data of rats with and without hepatotoxicity.

Keywords: Computer Aided Diagnosis System, Genomics, Machine Learning, Classification, Ensemble Learning, Artificial Intelligence

SİMGELER VE KISALTMALAR DİZİNİ

cDNA :Hedef DNA veya complementary DNA

DGGE :Temel Bileşenler Regresyonu

DVM :Destek vektör makinesi

lncRNA :Long non-coding RNA

log2FC :log2-kat değişim

ncRNA :non-coding RNA

NGS :Yeni nesil DNA dizileme yöntemleri

RF :Random Forest

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

TABLOLAR DİZİNİ

1. GİRİŞ

Organizmada bulunan bazı sistemlerin iç etmenlerden dolayı etkinliğinin azalması ile durma aşamasına gelmesi ve ayrıca maruz kalınan radyasyon veya günümüzde artan teknolojik kaynakların fazlalığından dolayı teknolojik aletlerin olduğundan aşırı kullanımı gibi çeşitli dış etmenlerden dolayı oluşan kanser hastalığının önemi gün geçtikçe artmakta ve daha çok ilgi çekmektedir. Kanser gerek sebebiyet verdiği mortalite ve morbidite, gerekse de tanı ve tedavisinde oluşturduğu oldukça büyük ekonomik yük sebebiyle önemi artan bir halk sağlığı sıkıntısı olmaktadır (1). Kemoterapi kanser hastalıklarının tedavisi amacıyla günümüzde başlıca tercih edilen yöntemler arasında olup radyoterapi ve cerrahi ile bir arada kullanımı tercih edilen etkin bir tedavi yöntemidir. Kemoterapide asıl amaç kemoterapötik ajanlar yardımıyla kanser hücrelerini öldürmektir. Sitotoksik anti-neoplastik ajanlar kemoterapiyle tedavide başrol oynamaktadırlar (2). Kanser hastalığı tedavisinde kullanılan ve güvenilirlikleri oldukça yüksek boyutta olan antineoplastik ilaçlar, hedef olan kanser hücrelerini imha etmeye çalışırken hiç beklenilmeyen etkilere neden olup sağlıklı hücreleri de etkileyip onlara da zarar verebilmekte ve öldürmektedir. Antineoplastik ilaçlar tedavide DNA üzerine direkt etki mekanizması kurarak veya hücre bölünmesini durdurarak kanser ile savaşmaktadır (3). Bir antineoplastik ilaç olan sisplatin (cis-diaminodikloroplatinum (II)) organik bir platin türevi olup, kanser tedavisinde en çok kullanımı tercih edilen ve en çok bilinen kemoterapötik ajanlardan biridir. Akciğer, mesane, yumurtalık, baş, boyun ve testis kanserleri dâhil olmak üzere birçok kanser çeşidinin tedavisinde tercih edilerek kullanılmaktadır. Kullanım dozu ile ilgili olarak sisplatin kullanılmasını sınırlandıran yan etkilerin en başında, hepatotoksisite, nefrotoksisite, nörotoksisite, testiküler toksisite ve gastrointestinal bozukluklar görülmektedir (4). Sisplatin'in yüksek dozlarda kullanılması karaciğerde oldukça ciddi yan etkilere sebep olmakla birlikte bireyi de hepatotoksisite açısından oldukça olumsuz şekilde etkileyebilmektedir (5). Birçok kimyasal ajan ve ilacın metabolizmasında neden olduğu fonksiyonlar nedeniyle karaciğer ilaç toksisitelerinden bizzat ve en çok etkilenen organdır. Karaciğer hasarının en büyük nedenlerinden biri olarak ilaçlar nedeniyle gerçekleşen karaciğer toksisiteleri görülmektedir (6). İlaca bağlı oluşan hepatotoksisiteler sonucunda siroz, akut karaciğer yetmezliği ve karaciğer kanseri gibi oldukça çeşitli klinik tablolar meydana gelmektedir (7). Sisplatin nedeniyle ortaya çıkan karaciğer hasarında oldukça önemli mekanizmalardan biri olarak görülen oksidatif

stres sonucunda açığa çıkan serbest oksijen radikalleri (SOR) tarafından oluşan lipit peroksidasyonu; kalp hastalıkları, toksik hücre hasarı ve kanserler gibi birçok hastalığın ortaya çıkış aşamasında ve patogenezinde rol oynamaktadır (8).

Hastalıkların araştırılmasında önemli veri kaynaklarından biri genomik verilerdir. Bu veriler sayesinde hastalıklar gen bazında incelenebilmekte olup bu çalışmalar bireye ait genetik profillerin ortaya çıkarılması, hastaların tanısının, tedavi seçeneklerinin, ilaç dozlarının, tedavilerin yan etkilerinin kişi bazlı ortaya konulması yönünde çalışmalar yapılmasına olanak sağlanmaktadır (9). Biyoinformatik ise biyolojik bilimler ile bilgisayar bilimi, matematik ve istatistik gibi bilimlerin arasında bir köprü oluşturmakta ve biyolojik süreçler sonucu elde edilen verilerin belirli programlar ve yazılımlar kullanılarak analiz edilip sonuçlarının görselleştirilerek sunulmasını sağlar (10). Son yıllarda, sisplatin kaynaklı hepatotoksisite altında yatan mekanizmayı anlamak için sıklıkla genom çalışmaları yapılmıştır (11, 12). Fakat ilaç kaynaklı hepatotoksisite için yapılan genomik çalışmalar sınırlı sayıda olup long non-coding RNA'lar (lncRNA) ile ilişkili çalışmalara pek rastlanamamıştır. Dolayısıyla, bu alanda yer alan eksiklikleri giderebilmek adına genomik çalışmalara ihtiyaç vardır.

Yapay zekanın bir alt dalı olan makine öğreniminin amacı, veriye dayalı öğrenme yöntemini kullanarak yeni verilere maruz kaldıkça önceki çıkarımlarından yeni tahminlerde bulunmaktır. Araştırmacıların genel odak noktası, bilgisayarlara karmaşık yapıdaki kalıpları tespit etme ve veriye dayalı olarak kararlar verme yeteneği kazandırmaktır (13). Hastalıkların teşhis edilmesinde ve klinik karar destek sistemlerinde son yıllarda gittikçe yaygın bir şekilde kullanılmaya başlanan teknolojilerin yaygın olanlarından biri de makine öğrenmesi yöntemleridir. Bu tür yaklaşımlar çok geniş bir uygulama alanına sahip olup son zamanlarda giderek daha popüler hale gelip kullanım alanları artmıştır. Makine öğrenimi teknikleri, genellikle sağlık alanında hastalık tahmin sürecinde kullanılır. Tahminlemede ise genellikle sınıflandırma sürecini yürütür. Sağlık alanında oldukça etkin bir uygulama alanına sahip olan makine öğrenmesi; kanser hastalıklarının ve kronik hastalıkların erken teşhisi ve bunların ortaya çıkmasına neden olan risk faktörlerinin belirlenmesi, genetik hastalıkların belirlenmesi ve altta yatan sebeplerin genetik alt yapısının ortaya konulması, ve tıbbi görüntülemede örüntülerin belirlenmesi gibi uygulamaların temel altyapısını oluşturmaktadır. Son on yılda artan hesaplama gücü ile makine öğrenmesi yöntemleri sağlık alanında çok yüksek performanslar ve başarılar ortaya koymuş ve tercih sebebi olmuştur (14, 15). Makine

öğreniminin önemli bir alt alanı olan topluluk öğrenmesinin mantığı, tek bir temel sınıflandırıcı kullanılarak elde edilen doğru tahmin oranını arttırmak için birçok sınıflandırıcıyı birleştirme işlemi yapılabileceği düşüncesine dayanmaktadır. Diğer bir deyişle, topluluk öğrenmesi yöntemi bir temel sınıflandırıcının (modelin) elde edebileceği sınıflandırma başarısı ile karşılaştırıldığında daha doğru ve güvenilir bir model (meta sınıflandırıcı) elde etmek için birçok temel sınıflandırıcıyı birleştirme düşüncesine dayanmaktadır (16). Topluluk öğrenme metodlarının diğer makine öğrenmesi yöntemlerine göre daha çok tercih edilmesinin nedeni, zayıf olarak tanımlanabilen sınıflandırıcıların (weak classifiers), kesin çıkarımlar ortaya koyabilen güçlü sınıflandırıcılara (strong classifiers) dönüştürülmesini sağlamasıdır. Böylece topluluk öğrenme yöntemleri ile zayıf sınıflandırıcıların tahminleme yeteneği arttırılmış olur. Ayrıca kullanılan sınıflandırma metodunun eğitim verisinde ezbere başvurmasını engellemek ve sınıflandırma performansının daha yüksek elde edilmesi beklendiğinde topluluk öğrenme yöntemlerine başvurulur (17).

Bu çalışma kapsamında, sisplatin ile hepatotoksisite oluşturulmuş sıçanlar ile patoloji oluşturulmamış sıçanlardan alınan karaciğer doku örneklerinin genomik analizlerinin yapılması ile elde edilen büyük verilerin biyoinformatik analizlerinin yapılması amacıyla bir web-tabanlı yazılımın geliştirilmesi ve bu yazılım sonucunda elde edilen bulgulardan yola çıkarak verilerin makine öğrenmesi yöntemlerinden topluluk öğrenmesi metodları ile modellenmesi sonucunda hepatotoksisite ile ilişkili tanıya/erken tanıya yönelik olası biyobelirteçlerin tespit edilmesi amaçlanmıştır.

2. GENEL BİLGİLER

2.1. Karaciğer

Vücudumuzda bulunan en büyük bezi olan karaciğer, ikinci en büyük organ olarak kabul edilmektedir. Sindirim kanalı boyunca emilen besinleri işlemesi ve vücudun diğer kısımlarında kullanılabilmesi için depo yapan hem ekzokrin hem de endokrin fonksiyonları olan bir organdır (18). Yaşam için oldukça önemli olan metabolik fonksiyonların gerçekleştirilmesine katkıda bulunan karaciğer, abdominal boşlukta üst kısımda diyafram kasının altında, vücudumuzun ise sağ tarafında olup tam olarak 7-11. kaburgaların arkasında yerleşmektedir. Cinsiyete, yaşa ve vücut boyutuna bağlı olarak karaciğer ebatı değişir. Yetişkin bir bireyde karaciğerin yaklaşık ağırlığı 1.2-1.5 kg arasına değişmektedir (19).

Fibrinojen, lipoproteinler, globülinler, albümin, protrombin gibi çeşitli protein türlerini doğrudan kana vermesi sebebiyle endokrin bir bez özelliği gösterirken, safra kanalları yoluyla duodenuma safrayı boşaltması sebebiyle ise ekzokrin bir bez özelliği göstermektedir (20). Sağ ve sol olmak üzere iki lobdan oluşan karaciğerde sağ lob sol loba göre daha geniş olup sağ lobun büyük bir kısmı göğüs kafesi ile çevrelenip korunmaktadır. Karaciğeri en dışta seröz zar periton sarmakta olup onun altında ise elastif lifler ve kollajen yapıdan oluşan Glisson kapsülü adı verilen yapılar bulunmaktadır. Karaciğerin iç kısımlarına doğru uzanan Glisson kapsülleri karaciğeri lob, segment ve lobüllere ayrıştırmaktadır (18). Karaciğerin yapısında hepatositler başta olmak üzere stellat hücreleri, sinuzoidal endotel hücreleri, Kupffer hücreleri ile safra kanalı epitel hücreleri gibi farklı türde hücreler yer almaktadır. Karaciğer hacminin oldukça büyük kısmını oluşturan hepatositler yağ asitleri, fosfolipidler ve trigliseridlerin metabolizmasında yer alan hücreler olarak bilinirler (21). Lobülden santrale doğru yönelimleri ile tek sıra dizilen hepatositler yan yana gelip birleşerek safra kanallarını oluşturmaktadırlar. Hepatositlerin çok yüksek derecede rejenerasyon özelliğine sahiptirler ve yaklaşık 150 günlük bir yaşam süreleri vardır. Bu özellikleri sayesinde çeşitli travmatik olaylar ve toksik madde sebebiyle ortaya çıkan karaciğer hasarlarından sonra oldukça kısa sürelerde kaybedilen doku yerine konularak karaciğerin yenilenmesi sağlanabilmektedir (21).

2.2. Toksisite

Toksik madde canlı organizmalar için zarar teşkil eden mineraller, bitkisel, hayvansal veya sentetik maddeler olarak tanımlanırken bu tür maddelere maruz kalan organizmanın geçici veya sürekli olarak bozulması durumuna ise toksisite (zehirlenme, intoksikasyon) denilmektedir. Kısacası toksisite, kimyasal maddelerin organizmada olumsuz etkiler oluşturmasıdır.

2.2.1. Hepatotoksisite

Birçok kimyasal reaksiyonun gerçekleştirilmesine sağladığı olanakla yaşamımızda oldukça önemli bir rolde olan karaciğer, sindirim kanalı yoluyla emilen besinlerin işlendiği ve vücudumuzdaki diğer kısımların yararlanmasını sağlamak adına bazılarının depolandığı, bazılarının ise direk dolaşıma aktarıldığı bir organdır. Karaciğerin önemli görevleri arasında safra üretimi, karbonhidratların, lipitlerin ve proteinlerin metabolizmaları, ve detoksifikasyon yer almaktadır (22). Karaciğer bahsedilen bu fizyolojik ve biyokimyasal görevleri sebebiyle birçok ilaç ve toksik maddeye maruz kalmaktadır. Pek çok maddenin metabolizmasından iç veya dış nedenlerden dolayı sorumlu olması karaciğeri ilaç kaynaklı toksisitelerde önemli bir yere oturtmaktadır (23). Hepatotoksisiteye neden olan ajanları doğal toksik ajanlar, kimyasal ajanlar, ilaçlar ve vitaminler olarak sıralayabiliriz. Hepatotoksisite akut ve kronik hasar, siroz oluşumu ve tümör gözlenmesi gibi birçok klinik durumla ortaya çıkabilmektedir. İlaca bağlı hepatotoksisite, hepatositlerde ve safra kanalı hücrelerinde kolestaza yol açar. Kolestaz, sırayla, toksik safra asitleri ve boşaltım ürünlerinin intrahepatik birikimine neden olur ve bu da karaciğer hasarını arttırır. Karaciğerde büyük rejeneratif kapasite vardır. Ama nekroz ve apoptoz hücre ölümü sebebiyle kaybolan hepatositlerin rejenerasyonu ilaca bağlı hasarın belirlenebilmesini engelleyebilmektedir. Böylece hepatositlerin aktif çoğalma tepkisi karaciğeri karsinojenler için hedef organ yapar. Lökositler, makrofajlar ve nötrofiller ile Kupffer ve sinüzoidal endotelyal hücreleri, stellat hücreleri karaciğer hasarının patogenezinde oldukça önemli bir rol oynamaktadır (24).

2.3. Sisplatin

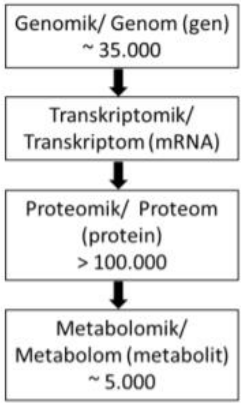
Kanser tedavisinde kullanılan antineoplastik ilaçlar DNA ile doğrudan etkileşime girerek ya da hücre bölünmesi aşamasını inhibe ederek tümör hücrelerini yok etmelerinin yanı sıra sağlıklı hücrelere de oldukça büyük zararlar verebilmektedirler (25).

Antineoplastik ilaçlardan biri olan sisplatin organik bir platin türevi olup, bilinirlik açısından oldukça popüler ve kanser tedavisinde en çok kullanıma sahip kemoterapötik ajanlardan birisidir. Cis-diaminodikloroplatinum (II) olarak da bilinen sisplatin, ilk kez 1844 yılında M. Peyrone tarafınca sentezlenerek literatüre kazandırılmış olup kimyasal yapısı ise ilk olarak 1893'te Alfred Werner tarafından ortaya konmuştur. 1960'lı yıllara gelinene kadar bu bileşikle ilgili olarak bilimsel çalışma ve araştırmalar yapılmamıştır. 1960 yılında ise Rosenberg ve arkadaşlarının ilk çalışmalarıyla elde edilen gözlemleri sonucuyla, sisplatinin yapısında bulunan platin ağ elektrotlarında yer alan bazı elektroliz ürünlerinin Escherichia coli'de hücre bölünmesine engel olabildiği ve bu ürünlerin de kanser kemoterapisinde kullanımının oldukça büyük ilgi yaratacağını belirlemişlerdir. DNA üzerine etki mekanizması, DNA üzerinde yer alan pürin bazlarıyla çapraz bağlanma, DNA onarım mekanizmalarına müdahalede bulunma, DNA hasarına sebep olma ve en son kanser hücrelerinde apoptozu indükleme yeteneği ile bağlantılıdır (4). Sisplatin kullanımını sınırlayan en önemli yan etkilerin başında kullanım dozu ile alakalı olarak hepatotoksisite, nefrotoksisite, nörotoksisite, testiküler toksisite, gastrointestinal bozuklukların yer aldığı görülmektedir (26). Sisplatinin oldukça yüksek dozlarda kullanımının olması karaciğerde oldukça ciddi yan etkilere sebebiyet verip bireyi hepatotoksisite yönünden oldukça olumsuz etkileyebilmektedir. Sisplatinin neden olduğu karaciğer hasarı için önemli mekanizmalardan biri olarak bilinen oksidatif stresin sonucunda açığa çıkan serbest oksijen radikallerinin (SOR) oluşumunu sağladığı lipit peroksidasyonu; kalp hastalıkları, toksik hücre hasarı, ve kanser gibi bir çok sayıda hastalığın ortaya çıkış evresinde ve patogenezinde rol oynamakta olup ortaya çıkışlarına oldukça büyük katkı sağlamaktadır (27).

2.4. Genomik

İnsan genomuna ait dizinin tümüyle çıkarılması işlemi ile sistem biyolojisi denilen alanda omik teknolojileri adı verilen yeni bir çalışma alanı ortaya çıkmıştır. Omik yaklaşımlar biyolojik sistemlerin oldukça kapsamlı olan bir alanı olarak ifade edilebilir. Ortaya çıkan yeni omik yaklaşımlarla birlikte biyoinformatik araçlar, yapılan araştırmalarda araştırıcılara oldukça büyük bir potansiyel sunmakta olup çalışmaları farklı odaklara çekmektedir. Son yıllarda oldukça çeşitli omik alt alanları ortaya atılmış olup, her birinin kendine ait bir takım çözümleyici teknikleri, araçları, analizler için yazılımları ve kendilerine özgü belirteçleri bulunmaktadır. Omik araçları ve teknolojilerinin ortaya çıkmasıyla, oldukça karmaşık olduğu bilinen biyolojik sistemlerin

tanımlanabildiği deneysel yaklaşımlar çarpıcı bir şekilde değişme eğilimi göstermiştir (28). Son yıllarda oldukça çok sayıda çalışma alanı için omik yaklaşımlar gündeme gelmekte ve bilinirlik açısından en çok ilgi gören omik yaklaşımları arasında gen düzeyindeki çalışmalar için genomik, mRNA düzeyindeki çalışmalar için transkriptomik, protein düzeyindeki araştırmalar için proteomik ve metabolitler düzeyindeki araştırmalarda ise metabolomik teknolojileri yer almaktadır (29).



Şekil 2.1. Omik teknolojileri ve aralarındaki ilişki (29)

Son yıllarda üzerinde en çok çalışılan ve tercih edilen omik yaklaşımlardan birisi olan genomik, tüm hücreli yapıdaki canlıların ve bazı tür virüslerin biyolojik gelişimleri açısından gerekli olacak genetik materyali taşıyan DNA'nın yapısal özelliklerini ve fonksiyonel yapısını detaylı bir şekilde inceleyerek ortaya koyan bir bilim dalı olarak tanımlanmaktadır (30). Genomik teknolojisi ise bir organizmanın/türün genomunda bulunan tüm genetik elementlerin belirlenebilmesi, dizi analizlerinin yapılarak haritasının ortaya konulması gibi DNA yapı ve işlevlerinin kapsamlı bir şeklide incelenmesi çalışmaları olarak açıklanabilmektedir. Genetik, tek bir genin yapısını ve yapısal faaliyetlerini, genomik ise genomda yer alan tüm genlerin yani kısacası DNA'nın tam olarak analiz edilmesi yoluyla etki ve işleyiş mekanizmalarını inceleyen bilim alanıdır (28).

Genomik yaklaşımlar yapısal genomik ve fonksiyonel genomik olarak iki alt gruba ayrılmaktadır. Yapısal genomik yaklaşımlarının başlıca çalışma alanları arasında kromozomlar üzerinde genlerin konumlandığı bölgelerin gösterilmesine yardımcı olan genom haritalaması ve DNA nükleotid dizi saptanmasını gerçekleştiren DNA dizi analizleri (sekanslama) yöntemleri bulunmaktadır. Yapısal genomik yaklaşımları bu

çalışma alanları aracılığıyla organizmaların genetik bilgi ve yapılarının açıklanmasına aracılık eder (31). Bu yaklaşımlar organizmaların genetik bilgi ve yapılarının ortaya çıkartılmasını ve açıklanmasını sağlasa bile bu elde edilen bilgi fotografiktir ve genlerin fonksiyonları ve yapısal işleyişleri ile ilgili bilgileri içermez. Önemli olan durum ise, bu dizilerin fonksiyonel yapı ve işleyişlerini belirlemektir. Bu tür yapısal genomik yaklaşımları içerdikleri yöntem sınırlılıklardan dolayı günümüzde pek tercih edilmemektedir ve yerlerini fonksiyonel genomik yaklaşımlara bırakmışlardır. Fonksiyonel genomik yaklaşımlar genomik bilgiyi biyolojik olarak anlamlandırmayı hedefleyen, genlerin fonksiyonel yapılarına, gen ekspresyon paternlerine ve genlerin birbirleriyle olan ilişkilerine ışık tutabilecek genom çapında sistematik olan tüm yaklaşımları içeren bir çalışma alanıdır. Bu tür yapısal genomik yaklaşımlar genlerin fonksiyonlarının ortaya konulmasının yanı sıra, organizma yapısı açısından önemlerinin ortaya konulması ve anlaşılmasına da aracılık eder (32). Bu çalışmalar genel olarak DNA'da ifade edilen (ekspresyon) bölgelerin ürünleri olan mRNA'ların analizlerini kapsamaktadır. Bunun sebebi ise DNA'nın fonksiyonel yapısını mRNA'lar aracılığıyla ortaya çıkarmasıdır (33).

Çıktılarını bilişim teknolojileriyle harmanlayarak ortaya koyan ve anlamlandırarak depolayan genomik, otomasyon ve biyoinformatik alanlarında gerçekleşen ilerleyişin gelişmesine katkıda bulunan bir bilim dalıdır (34). Genomik hemen hemen her konuda ve alanda değerlendirme yapmak için araştırma yapmaya uygun bir disiplindir. İyi kurgulanmış bir senaryo ve doğru seçilmiş bir karşılaştırma tekniği ile tıbbın tüm alanlarında (Mikrobiyoloji, Onkoloji, Farmakoloji, Biyokimya ve İmmünoloji, vb) araştırmalar yapmaya elverişlidir. Kurgulanmış karşılaştırmalı çalışmalar yolu ile kanserleşme ve prognoz tahmini, ilaca yanıt tahmini ve bireye özgü ilaç ve tedavi gelişimi, immün yanıtın mekanizması ve transplantasyon sonuçlarının önceden tahmin edilmesi gibi oldukça çeşitli araştırmalar yapılabilir (35). Günümüzde modern tıbbın çalışma alanları, bireysel genetik profillerin hazırlanması, hastaların tanılarının bireyselleştirilmesi, bireye özgü tedavi seçeneklerinin oluşturulması, bireysel ilaç dozlarının belirlenmesi, ilaçların neden olduğu yan etkilerin bireysel olarak ortaya konulması yönünde gelişmektedir. Gen ve gen, gen ve çevre birlikteliklerinin belirlenebildiği bilgisayar yazılım sistemleri ile moleküle ait DNA, RNA ve proteinlerin özelliklerini ve bunların ait olduğu metabolik yolakların fonksiyonel bir örüntü içerisine yerleştirilmesini sağlayan sistem biyolojisi oldukça hızlı gelişmektedir. Genomik tıbbın

asıl amacı hastalıktan korunma ve erken tanı, tanılamanın ve tedavinin bireye göre düzenlenebilmesi ve bu yolla insan sağlığının korunmasıdır (9).

Genomiği kısaca özetleyecek olursak;

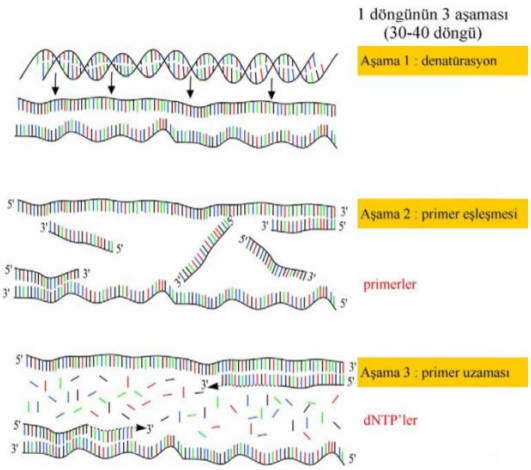
- Hastalığa neden olan genleri tanımlar ve fizyolojik süreçlerle ilişkili genleri ortaya koyar.
- Genlerin birbirleriyle ilişkilerinden yola çıkarak aralarındaki yapısal ve fonksiyonel ilişkileri ortaya çıkarır.
- Gelişim boyunca genlerin rollerini belirler ve genlerin ekspresyon profillerini ortaya koyar.
- Farklı organizmaların genetik zeminde karşılaştırılmalarına olanak tanır (35).

Genomik yaklaşımlarda analizler geleneksel ve yüksek verimli yöntemler olarak bilinen yöntemler aracılığıyla gerçekleştirilir. Geleneksel yöntemler arasında polimeraz zincir reaksiyonu temeline dayanan yöntemler, elektrofaz temeline dayanan yöntemler, blotlama temeline dayanan yöntemler ve DNA dizileme analizleri yer alırken, yüksek verimli yöntemlerin temeli mikroarray yöntemine dayanmaktadır (36).

2.4.1. Polimeraz zincir reaksiyonu temelli yöntemler

Yüzyılın en önemli buluşları arasında yer alan Polimeraz Zincir Reaksiyonu (Polymerase Chain Reaction, PCR) yöntemi ilk olarak 1985 yılında kullanılmıştır. Çok yaygın olarak kullanılan bu yöntem, kısaca tanımlanacak olursa nükleik asitlerin uygun ve elverişli koşullar altında çoğaltılması olayıdır. Temel olarak bu yöntem DNA'da yer alan iki zincirinin oldukça yüksek ısı yardımıyla birbirlerinden ayrılması (denatürasyon) evresi, sonra ise primer adı verilen sentetik oligonükleotidlerin hedef DNA'ya bağlanmasıyla sonuçlanan yapışma evresi ve son olarak zincirin uzamasını kapsayan döngünün belirli bir sayı kadar tekrarlanmasına dayanmaktadır. Bu tekrarlanan döngüler sonucunda DNA'ya ait milyarlarca kopya üretilebilmektedir. Bu teknik sayesinde laboratuvar çalışmalarında oldukça büyük bir hız ve kesinlik elde edilmiştir. Çoğu moleküler laboratuvarda kullanılabilen polimeraz zincir reaksiyonu yöntemi moleküler biyoloji alanı ve biyomedikal araştırmaların tasarlanmasından, adli tıp, genetik teşhis gerektiren oldukça farklı alanlara yayılan geniş bir kullanım alanına hizmet etmektedir (37, 38).

Şekil 2.2 polimeraz zincir reaksiyonuna ait bir görüntüyü göstermektedir.



Şekil 2.2. Polimeraz zincir reaksiyonun basamakları (39)

2.4.2. Denatüran Gradiyent Jel Elektroforezi

Denatüran Gradiyent Jel Elektroforezi (DGGE) yöntemi aynı boyutta olan fakat nükleotid dizilimleri farklı çift zincirli DNA moleküllerinin ayrılmasını sağlayan elektroforetik bir tekniktir. DGGE ile öncelikle PCR yöntemiyle çoğaltılarak elde edilen belirteç olan gen ürünleri (16S rRNA veya 18S r RNA genleri) denatüre edici ajanlardan olan üre ve formamid gibi ajanlar eşliğinde ve derecesi gitgide artan bir konsantrasyon yapısına sahip poliakrilamid jel yapı içerisinde elektriksel bir alana maruz bırakılarak ayrıştırılırlar (40). DGGE'nin temel basamakları şunlardır (41).

i) Biyolojik verilerden nükleik asit (DNA veya RNA) izolasyonunun gerçekleştirilmesi,

ii) Belirlenen gen bölgelerinin PCR aracılığıyla çoğaltılması,

iii) PCR ile elde edilen ürünlerin DGGE yöntemiyle ile ayrıştırılması ve sonrasında görüntülenmesi,

iv) Dizi analizlerinin yapılması ve ortaya konan dizilerin gen bankalarında yer verilen diziler ile karşılaştırılması işlemlerinin yapılması.

2.4.3. Blotlama

Bir genin kopyasının oluşturulması işlemidir. DNA taşıyıcı birimi olarak bilinen vektörler ile istenilen gen bir konak hücreye yerleştirilir ve burada çoğaltma işlemi yapılır. Bu işlem ile genler ayrıntılı olacak şekilde incelenebilir. Blotlamanın aşamaları şu şekildedir (42).

i) DNA'nın izolasyonu sağlanarak PCR yöntemi ile çoğaltılması,

i) DNA'nın uygun taşıyıcı birim olan vektörlere eklenmesi,

iii) Belirlenen konak hücreler içerisinde istenilen DNA'nın klonlanması,

iv) Klon kütüphanelerinin taramasının yapılması,

v) Biyoinformatik analizlerin yapılması.

2.4.4. DNA Dizileme Analizi

Nükleik asit moleküllerinin yapısında yer alan nükleotid dizisi dizileme (sekanslama) yöntemleri aracılığıyla belirlenmektedir. Dizileme işlemi bir PCR ürününün kimliğinin belirlenmesinde kullanılan hassasiyeti yüksek bir tekniktir. Yaygın olarak en çok kullanılan ve bilinen, zincir sonlandırma yöntemi Sanger yöntemidir (43, 44). Bu teknikte DNA polimeraz enzimi kullanılması yardımıyla tek zincirli DNA yapısının bir kopyasının oluşturulması sonucu dizi belirlenmiş olur. Elde edilen DNA parçaları farklı uzunluklara sahip olup bu parçalar birbirlerinden elektroforez yöntemi ile ayrılmışlardır. Sonlandırma ürünlerine denk gelen bantların belirlenmesi ise radyografi veya floresan uygulamaları aracılığı ile gerçekleşir. Fakat Sanger dizi yönteminde bulunan bir kısıtlamadan dolayı az sayıda nükleotid okunabilmesinden dolayı uzun bir DNA'nın dizilemesine ihtiyaç duyulduğunda bu yöntem kullanım açısından elverişliliğini yitirmektedir. Bu nedenden yola çıkarak bilgisayar destekli dizileme yöntemlerinin geliştirilmesi sağlanmıştır. Bu işlemlerde floresan boyalar yardımıyla işlemler gerçekleştirilir. DNA parçaları jel içerisinde ilerlemesi sırasında bir lazer ışını yardımıyla taranmaktadır ve her bir parça farklı floresanla boyanmış olan ddNTP'leri uyarır. Oluşan ışık toplandıktan sonra elde edilen bilgi bilgisayara aktarılma yoluyla DNA dizisi belirlenmiş olur (45, 46). Günümüzde yukarıda bahsedilen yöntemlerden daha verimli, oldukça daha ucuz ve hızlı olan birçok yöntem mevcuttur. Yeni nesil DNA dizileme yöntemleri (NGS), genomik çalışmalarda çığır açmış, filogenetiği ve işlevsel çeşitliliğini

araştırmamıza imkan veren oldukça gelişmiş floresan görüntüleme yöntemlerinden oluşan yeni teknolojilerdir (47). NGS dizileme yöntemlerinin en büyük özelliği çok yüksek doğruluk ve hız ile dizileme yapabilmesidir. Bu yöntem ile milyonlarca kısa dizilemenin aynı anda yapılabilmesi yolu açılmıştır. Oldukça çeşitli NGS platformu mevcuttur. Bu platformlardan bazıları Roche 454, Illumina, ion torrent dir. Her bir platformda kullanılan yöntem birbirinden farklıdır (48). Roche 454 platformunda çoğaltılması istenilen DNA boncuklara bağlanarak PCR tekniği ile çoğaltılarak ardından kuyulara konur ve bu işlem sonunda pirosekanslama yapılır. Pirosekanslama işlemi, DNA sentezi esnasında serbest bırakılan bir pirofosfat molekülünün (PPi) tespit edilmesine dayalı olarak geliştirilen yeni nesil DNA dizileme yöntemlerinden biridir (49, 50).

2.4.5. Mikrodizi (Microarray)

Mikrodizi ileri derece olarak bilinen genotip ve gen ekspresyon analizlerinin yapımı amacıyla geliştirilen bir yöntemdir. Mikrodizi yöntemi küçük örnek hacimlerinin kullanımına olanak tanımasıyla tek deneyde, tek nükleotid polimorfizmi (SNP) veya hastalıklı ve normal fizyolojik durumlardaki değiştirilmiş gen ekspresyonunun (mRNA'daki artış ve azalışlar) oldukça hızlı çalışılmasına olanak sağlamaktadır. Mikrodizi yöntemi yaklaşık olarak 1.8*1.8 cm ebatlarında olan bir cam yapı üzerinde birden çok DNA bölgesinin (hemen hemen tüm bir genomunu) bir seferde oldukça yüksek hassasiyet ile incelemesine olanak sağlar (51). Tipik bir mikrodizi çalışması beş temel basamaktan oluşmaktadır (31, 52).

Hedef DNA veya complementary DNA (cDNA) dizileri için tamamlayıcı dizilerden oluşan probların immobilizasyonunun yapılacağı mikroarray platformunun çalışmaya hazır hale getirilmesi, hazırlanması veya temini.

İncelenecek olan numuneden floresan yardımıyla işaretlenmiş cDNA, DNA ve complementary RNA (cRNA) 'nın oluşturulması.

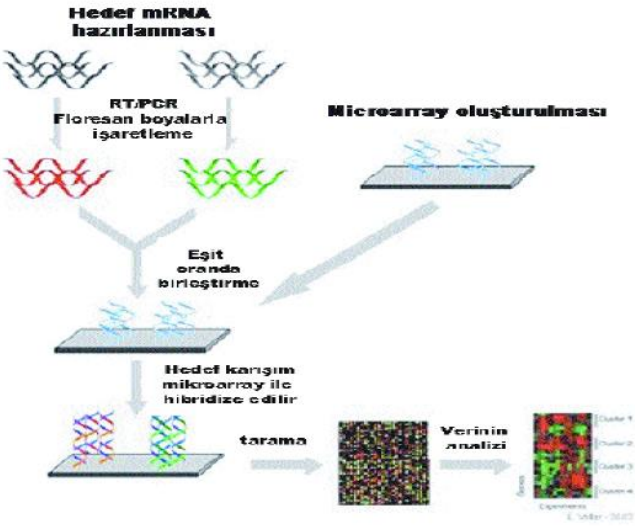
İşaretlenmiş cDNA/DNA/cRNA ile mikroarray platformunun hibridizasyon solüsyonu içerisinde karşılaştırılması.

Yıkama sonrasında hibridizasyonun varlığının okuyucu yardımıyla analiz edilmesi ve görüntünün bir bilgisayarda depolanması.

Bir yazılım yardımıyla depolanan görüntünün değerlendirilmesi ile mikroçip platformu üzerindeki hibridizasyonun hangi noktalarda olup olmadığının ve niceliksel

olarak düzeyinin (koyu mavi-yeşil-sarıkırmızı renk yelpazesinde) değerlendirilmesi ve yorumlanması.

Mikroçip yönteminin ayrıntıları şekil 2.3 de gösterilmiştir.



Şekil 2.3. Mikroçip yönteminin şematik gösterimi (31)

2.5. RNA-Dizi Analizi (RNA-Seq)

Son zamanlarda öne çıkan yöntemlerden biri olan RNA dizi analizi teknolojileri hücreler tarafından üretilen tüm RNA transkriptlerinin karakterizasyonunu ortaya koymak adına üstün sonuçlar ortaya koymuştur. Transkriptomik için devrim niteliğinde olan RNA-seq, organizmanın transkriptomunu keşfedebilmek adına oldukça yüksek verimde ve kantitatif bir şekilde tüm transkriptomların analizinin gerçekleştirilmesine izin veren ilk dizileme-temelli yöntemdir. Hibridizasyon temelli yöntemlerden farklı RNA-seq yöntemi sadece bilinen transkriptlerin belirlenmesine odaklanmayıp, bilinen transkriptleri araştırıp ortaya koyarken diğer yandan yeni transkripleri keşfetmeyi ve ilişkilendirmeyi de amaçlar. RNA-seq yönteminin bir başka önemli avantajı ise ekspresyon seviyelerini göreceli değerlerle ölçmek yerine oldukça geniş ve dinamik bir aralıkta ölçülmesine olanak tanır. Elde edilen verilerin doğruluk ve güvenilirliğini yüksek seviyelere çıkaran düşük arka plan kirliliğinin var olması ve ekzonlar ile intronların sınırlarının kesin bir şekilde tanımlanması ile transkriptler içinde yer alan SNP ve diğer varyantların tanımlanabilmesi ve belirlenebilmesi önemli RNA-seq yöntemlerinin önemli avantajları arasında yer alır. Tüm bu avantajlarından dolayı RNA-seq yöntemi buluşa

dayanan çoğu araştırma için ideal bir yöntemdir. Referans genomu eksik olan bir organizma var olsa bile novo transkriptomları belirleyebilecek veya düşük miktarlarda var olan transkriptleri bile tanımlayıp ortaya koyabilecek, kritik düzeydeki biyolojik izoformları ayırt edebilme gücüne sahip, genetik varyantları bulup onları tanımlayabilecek ve farklı düzeyde ekspresyonların analiz edilebilmesinde kullanılabilecek üstünlükte bir yöntemdir (53).

2.5.1. Transkriptomiks

Transkriptom, belli bir zaman diliminde bir hücre, doku veya organizmaya ait genom tarafından üretimi gerçekleştirilen RNA moleküllerinin tamamı (mRNA, tRNA, rRNA ve kodlanmayan RNA'lar) olarak tanımlanabilir. Gen ekspresyonunda oluşan farklılıklardan dolayı transkriptomlar oldukça dinamik bir yapıya sahip olup sürekli değişim içerisindedir. Bir hücre için kabaca sabit olan yani değişimi olmayan genomun tam tersine, transkriptom çevrede var olan pH değişikliklerinden, besinin çeşitliliğinden, ısı kaynaklı değişikliklerden ve diğer hücrelerden gelen sinyallerle olan etkileşim gibi çevresel faktörlerle değişebilmektedir. Hücre içinde farklı işlevler gerçekleştirilirken bu işlevlerle alakalı olan genlere ait transkripsiyonlar artıp azalabilme gösterebileceğinden ve transkriptom hücre içinde yer alan tüm mRNA'ları içerdiğinden, belli bir zamanda ve durum karşısında aktif olarak bulunan genleri yansıtmaktadır. Çevresel faktörlerden dolayı genlerin ifadelerindeki değişiminin ortaya konulması çevre-sistem etkileşimlerinin önemini ortaya koymuştur (53, 54).

Transkriptomik ise hücre genomundan transkripsiyonla oluşumu gerçekleşen tüm mRNA transkriptlerini eş zamanlı olarak inceleyip onların ekspresyon profillerini ortaya koyan bir çalışma alanıdır. Bahsedilen ekspresyon profillerini ortaya çıkarırken yaygın olarak kullanılan transkriptomik uygulamalar olan mikroarray ve yeni nesil dizileme gibi yüksek çıktı sağlayan teknolojiler transkriptomda belirli sebeplerle ve belirli zamanda oluşan değişikliklerin incelenmesini temel almaktadır. Bu transkriptomik uygulamaları sonucu ortaya çıkarılan ekspresyon profillerine de transkriptom adı verilmektedir (54).

Son zamanlarda transkriptomik alanındaki gerçekleştirilen çalışmalar, özellikle genetik varyantların ekspresyonunda oluşan değişikliklerin (artış-azalış) kanser gibi bir çok kompleks hastalığın oluşumuna ne şekilde katkıda bulunduğunun anlaşılmasına ve etkilerinin ilişkilendirilerek ortaya konulmasına ışık tutması ile oldukça önem kazanmış ve bu tür araştırmalar oldukça artmıştır. Ayrıca, bilim insanları ve araştırıcılar bu yapılan

çalışmalar sayesinde hücrelere ait yaşam döngülerini ve dolayısı ile de hastalık seyrini düzenleyip ona yön veren biyokimyasal yolaklar ile moleküler mekanizmalar hakkında daha fazla bilgiye ulaşabilip hastalık ile ilgili önemli bilgilere erişebilmektedir (54, 55).

Bilinen başlıca mRNA analiz yöntemleri; Northern blot (tek gen RNA ekspresyon analizi), Ters Transkripsiyon Polimeraz Zincir Reaksiyonu (RT-PCR) (tek veya çoklu gen analizi) ve mikrodizi (Microarray)'dir. Günümüzde bahsedilen bu yöntemlerin yerini yeni nesil dizileme teknolojilerine almaya başlamıştır. Dizileme yöntemlerinde mevcut olan yüksek verimlilik ve hız limitasyonlarını aşabilmek adına geliştirilmiş ve ortaya konmuş olan yeni nesil dizileme teknolojileri, tek bir örnekten alınmış olan ve milyonlarca parçaya bölünmüş olan bir DNA molekülüne ait her bir parçayı aynı anda ve bir uyum dahilinde paralel bir şekilde işlenmesini temel almaktadır. Yeni nesil dizileme teknolojisi, sahip olduğu bu avantajları sebebiyle birçok alanda gerçekleştirilen biyolojik araştırmaların vazgeçilmez bir parçası olmuştur (53, 55).

Son yıllarda gerçekleşen genetik kodun keşfi çalışmaları ile birlikte bilim insanları tüm insan transkriptini analiz etmeye ve değerlendirmeye başlamışlardır. Bu alanda yapılan gerçek ilerleme 1980'lerde otomatik DNA dizi analizinin ortaya atılmasıyla gerçekleşti. 1990'lı yıllarda eksprese dizi işaretleri (EST) analizinin ortaya atılmasıyla birçok insan dokusunda eksprese edilen genler oldukça hızlı bir şekilde tanımlanmaya ve keşfedilmeye başlandı. EST dizi analizleri oldukça yüksek-verimliliğe sahip bir teknik olmasına rağmen yüksek maliyeti ve teknik sınırlamalarından dolayı tüm transkript profillerinin elde edilmesini ve ortaya çıkarılmasına engel olmuştur. Bu nedenle klasik EST yaklaşımının sınırlılıklarını ortadan kaldıracak ve buna alternatif olabilecek ve EST'yi tamamlayacak çeşitli teknolojiler geliştirilmiş ve ortaya atılmıştır. Gen ekspresyonunun seri analizi (SAGE), gen ekspresyonu kap analizi (CAGE) ve kitlesel paralel imza dizilemesi (MPSS) bu yöntemler arasında yer almaktadır. Ayrıca EST yaklaşımı dışında işaretleme metotları da gen-ekspresyon seviyelerini ortaya koyabilmektedir. Ancak bu yöntemler özgül izoformları ayırt edemezler. Çoğunun çalışma prensibi klasik Sanger dizileme yöntemine dayanmakta olup oldukça yüksek maliyetlere sahip tekniklerdir. Hibridizasyona dayalı mikroçip yöntemler ise geniş ölçüde transkript belirleyebilen ve ölçebilen, ve işaretleme yöntemlerine kıyasla daha ucuz tekniklerdir. Bu yöntemler kullanılan prob yoğunluğuna bağlı bir şekilde transkripsiyon bölgelerinin 5 bç'den 50 baz çiftine kadar yüksek çözünürlüklerde haritalanmalarına olanak sağlar (54, 55).

2.5.2. RNA ve non-coding RNA (ncRNA)

Yüksek organizmaya sahip canlılar için genetik materyal DNA olup DNA da yer alan bilginin RNA'ya transkripe edilmesiyle RNA'nın işlevi bu aktarılan bilgi ve şifreyi proteine çevirmektir. RNA; protein olarak çevrilebilen mRNA ve protein olarak çevrilemeyen yani kodlanamayan RNA (ncRNA) olarak ifade edilebilen 2 kısımdan meydana gelmektedir (56).

1990'lı yıllarda insan genom projesinin başlamasıyla genom dizilemesinin daha başında beklentilerin çok fazla üstünde transkriptin var olduğu bildirilmiştir. Çalışmalar sonucunda genomun %80'lik kısmının aktif bir şekilde transkribe edilebildiği fakat bu işlemi gerçekleştiren kodlayıcı genlerin ise genomun %2 den daha az bir kısmı olduğu ve geriye kalan bu büyük kısmının ise kodlayıcı olmayan (non-coding) RNA (ncRNA) genler tarafından oluştuğu ortaya konmuştur. Uzun yıllar boyunca sayıları çokça fazla olan bu kodlamayan olarak adlandırılan transkriptler "junk/çöp" DNA ya da "transkripsiyonel gürültü" olarak isimlendirilip işlevsellik olarak düşünülerek dikkate alınmamıştır. Fakat daha sonraları yapılan çalışmalar ile bu kodlamayan transkriptlerin oldukça önemli bir kısmının aslında işlevsel olarak aktif RNA molekülleri olduğu belirlenmiştr. Kodlamayan bu genlerin sayıları türlere bağlı olarak değişkenlik göstermekle birlikte bu genlerin sayıca fazlalığı ise organizmanın komplekslik derecesi ile ilişkilendirilmekte olup bu durum ncRNA ların oldukça önemli olduğunu göstermektedir (57, 58).

2.5.3. Uzun kodlamayan (long non-coding) RNA lar (lncRNA)

lncRNA'lar 200 nükleotidden daha uzun olan ve açık okuma çerçeveleri bulunmayan buna ek olarak DNA, mRNA, protein ve miRNA yapıları ile etkileşime geçerek çeşitli mekanizma yapıları ile epigenetik, transkripsiyonel, post transkripsiyonel ve translasyonel işlev düzeylerinde gen ekspresyonunu düzenleyerek gerçekleşmesine katkıda bulunan moleküller olarak bilinirler (59).

lncRNA lar protein kodlayabilen genlere ait intronik ya da intergenik olarak bilinen bölgelerinde bulunabilmektedirler. lncRNA lar genellikle transkripsiyonda etkili olan RNA polimeraz II olarak bilinen bir multiprotein kompleksi aracılığıyla sentezlenmektedir. Bazı lncRNA'ların mRNA'ların yapılarında olduğu gibi benzer olarak alternatip kırpılma, poliadenilayon ve 5' şapka formasyonu gösterdikleri bilinmektedir. lncRNA'lar düzenleyici rolleri göz önüne alınarak iskele, tuzak, sinyal ve rehber olarak

fonksiyonlarda yer almaktadır. lncRNA'ların transkripsiyon ve daha sonra gerçekleşen işlevsel seviyelerde gen ekspresyonu gerçekleşirken birçok düzenleyici mekanizmada bulunarak, oldukça farklı biyolojik süreç için ön hazırlık fonksiyonlarını gerçekleştirdiği bildirilmiştir (60).

1980'li yılarda bilim insanları doku spesifik genleri ve ekspresyon paternlerini ortaya çıkarabilmek amacıyla cDNA (complementary DNA) kütüphanelerinin görüntülenmesinde bilinenlerden farklı hibridizasyon yöntemleri kullanma fikrini ortaya atmıştır. Bu çalışmaların amacı kodlayıcı genlerin ortaya konması iken sonraki zamanlarda RNA'nın kodlayıcı olarak yerine getirdiği görev potansiyeli göz ardı edilerek bir yaklaşım benimsenmiş ve bu yaklaşım sonucunda o zamanlarda mRNA olarak sınıflandırılmış olan ve ilk ökaryotik lncRNA olarak bilinen "H19" 1989 yılında keşfedilerek literatürde yer almıştır. Yapılan çalışmalar ile bu "H19" ekspresyonunun transgenik fareler için prenatal dönemde ölümcül olduğu ortaya konmuştur. Böylece "H19" un embriyonik dönemde ne kadar önemli olduğu ortaya konmuştur (61).

X kromozom inaktivasyonu (XKİ) olayı iki X kromozomuna sahip olan dişi bireylerde bir X kromozomunun, gelişim evresinin erken dönemlerinde X inaktivasyon merkezi (Xic) olarak bilinen bir lokus tarafından inaktive edilmesi durumudur. 1990'lı yılların başında yapılan çalışmalar ile bu bahsedilen lokusun "Xist" (X inactivation specific transcript) ismi verilen bir lncRNA eksprese ettiği ortaya konulmuştur. Rastgele XKİ'ni başlatabilmek için inaktive olacak X (Xi) kromozomundan mono allelik gösteren bir yol ile Xist ekspresyonunun başlayarak aktive olduğu ve Xist'in ise Xi üzerine yayılmak suretiyle cis pozisyonda gen susturmayı başlattığı belirlenmiştir. "H19" ile "Xist" ile ilgili bu oldukça önemli ve lncRNA'lar açısından öncü olan bu çalışmalar, lncRNA'ların biyolojik anlamlarını ortaya koymak için protein kodlamayan genlere ait işleyiş mekanizmalarını ve onların fonksiyonları hakkında bilim insanlarının bakış açılarının kökten değişmesine sebep oluştur. Bu nedenle önemli oldukları sonucuna varılan lncRNA'ları tanımlamak ve işleyiş mekanizmalarını karakterize etmek amacıyla 2000'li yıllarda küresel çapta bir çalışma başlatılmıştır (61).

Oldukça çeşitli epigenetik mekanizmalar sayesinde gen ekspresyonları gerçekleşmekte ve kontrol edilmektedir. DNA metilasyonu ve histon modifikasyonları gibi bilinen kovalent modifikasyonların tam tersine lncRNA'lar, mRNA degredasyonu ve kromatin yapının yeniden düzenlenmesi işlevlerinde rol aldıkları için nükleozom

yeniden düzenlenmesi işleminin en önemli kovalent olmayan epigenetik modifikasyonları olarak bilinirler (62).

Bilinen ncRNA'ların bir çoğu hücresel yapılardaki farklılaşma, dozaj kompanzasyonu, genomik damgalama (impriting) gibi biyolojik birçok süreçte oldukça önemli bir yere sahip olmakla birlikte kanserler, Alzheimer hastalığı ve kardiyovasküler hastalıklar olmak üzere oldukça çok sayıda kompleks hastalığın etiyolojisi ve başlangıcıyla da güçlü bir şekilde ilişki göstermektedir. Hastalıklarla ilişkili olan tek nükleotid polimorfizmlerinin (SNP) %90'dan daha fazlasının genomda yer alan kodlamayan bölgelerde veya kodlamayan RNA genlerinde yer aldığı bildirilmiştir. Bu sebeplerle hastalıkların altında yatan nedenleri anlayabilmek ve etiyolojilerini ortaya koyabilmek, daha iyi sağlık sonuçları elde edip hastalara uygun ve daha etkili tedaviler geliştirmek ve sunabilmek için ncRNA türlerini bilmek ve yapılarını belirlemek önemlidir (63, 64). Kodlamayan RNA'lar arasında lncRNA'lar en yaygın olan grup olup fonksiyonel olarak da diğer ncRA'lardan farklı bir sınıfta yer almaktadır. lncRNA'lar virüsler, maya, bitkiler, hayvanlar gibi oldukça çok çeşitli türde gözlenebilmekte olup türler arasında oldukça zayıf bir şekilde korunmaktadır.

lncRNA'ların DNA, RNA ve bunların dışındaki diğer transkripsiyonel moleküller ile etkileşime girerek birçok biyolojik süreçte ve hastalığın ortaya çıkış aşamasında çok yönlü olarak işlev gösterdikleri bildirilmiştir. Gen ekspresyonunun gerçekleşmesi, gen susturma, ısı şok cevabı, damgalama (imprinting) ve embriyogenez oldukça farklı ve çok sayıdaki biyolojik olayın gerçekleşmesinde oldukça önemli ve fonksiyonel oldukları gösterilmiştir. Ayrıca lncRNA'larda gerçekleşen mutasyonların virüs enfeksiyonları, kanser, nörodejeneratif hastalıklar da dahil olmak üzere oldukça çok sayıdaki hastalık ile ilişkisi de ortaya konmuştur. lncRNA'larda meydana gelen herhangi bir düzensizliğin hücre anjiyogenezi durumunu tetikleme, proliferasyon, apoptoz direnci, tümör baskılayıcılarından kaçma durumu gibi birçok normal hücresel süreci etkilediği de bildirilmiştir. Fakat şimdiye kadar lncRNA'lar ile ilişkili olan oldukça az sayıda çalışma yapılmıştır (65).

Gen regülasyonunda meydana gelen bozuklukların birçok hastalıkla ilişkisinin olduğunun ortaya çıkması lncRNA ile ilgili çalışmaların önemini gündeme getirmiştir. lncRNA'lar çeşitli seviyelerdeki gen regülasyonu olayında düzenleyici rolde olup bu düzenleyici rolünde herhangi bir bozulma olması durumu kompleks çeşitli hastalığın

ortaya çıkmasına, gelişip ilerlemesine katkı sağlamaktadır. lncRNA'lar birçok fizyolojik ve patolojik süreçte işlevsel olduğu için özellikle dizileme ve üç boyutlu yapılarında oluşabilecek anormallikler ile DNA-protein arasındaki bilinen etkileşim ve ekspresyon seviyelerinde gerçekleşen anormalliklerin, çoğu hastalığın oluşum aşamasıyla yakından ilişkili olduğu bildirilmiştir. Bu sebeplerle lncRNA'ların çoğu önemli hastalık için terapötik hedef ve biyobelirteç olarak gösterilmesi önerilmektedir. Şimdiye kadar yapılan çalışmalarda toplam 270.044 lncRNA geni sınıflandırılmasına rağmen işlevsellik ve hastalık gelişimi üzerindeki fonksiyonları açısından oldukça az sayıda transkriptin çalışıldığı ve ilişkilerinin belirlendiği bildirilmiştir. miRNA'lar ile ilgili literatürde oldukça kapsamlı çalışmalar yapılmasına rağmen lncRNA'ya dayalı çalışmaların oldukça sınırlı kaldığı gözlenmiştir (66).

Genomik mutasyonlar, tümörlerin oluşumu, DNA'da meydana gelen hasar ve immun kaçış çeşitli biyolojik süreç metabolik olarak gelişebilmektedir. lncRNAlar bu bahsedilen süreçlerde translasyonel düzenleyiciler olarak işlev yapmaktadır. Ayrıca lncRNA'ların proliferasyon, apoptoz gelişim mekanizması, anjiogenez ve metastaz durumu gibi süreçleri trankripsiyon ve daha sonraki seviyelerde düzenlediği ortaya konmuştur (67). Kanser gelişim sürecinde lncRNA'lar oldukça önemli rol oynamakta olup tümör baskılayıcı veya onkogenik fonksiyonlarda yer alabilmektedirler. Örneğin "MEG3" normal beyin dokusunda yer alan ve yüksek oranda eksprese edilen ayrıca gliomalarda ise aşağı regüle edilen tumor baskılayıcı bir lncRNA'dır. "PCAT-1 (prostate cancer-associated ncRNA transkript 1) metastatik olan prostat kanserinde aşırı ekspresyon gösteren ve BRCA2 gen ekspresyonunu ise aşağı regüle ederek hücre proliferasyonun indüklenmesini sağlayan bir onkogenik lncRNA'dır (68).

Kompleks hastalıklar olarak bilinen hastalıklar birçok çevresel faktörler ile birlikte düşük penetransa sahip oldukça çok sayıdaki genetik varyantın kombinasyonu ile oluşabilen ve Mendeliyen kalıtım modellerine uyum göstermeyen multifaktöriyel hastalıklardır. Nörolojik hastalıklar, KAH, otoimmun hastalıklar, kanserler ve çoğu hastalık bu sınıflandırmada yer alan hastalıklardandır. lncRNA'ların bu çeşitli kompleks hastalıklar durumunda düzensizleşerek hastalık başlaması ve ilerlemesi ile ilişkisi ortaya konmuştur (69).

lncRNA'nın spesifik işlevleri ve rolleri net olarak açıklığa kavuşturulmamış olsa da karaciğer hastalıklarının lncRNA'ların anormal ekspresyonu ile ilişkili olduğu

kanıtlanmıştır (70, 71). Birkaç araştırmacı ise, anormal lncRNA ekspresyonunun, karaciğer fibrozu ve kanser dahil olmak üzere çeşitli bozuklukların gelişimi ile ilişkili olduğunu bildirmiştir (72, 73).

2.6. Biyoinformatik

Biyoinformatik genel olarak biyoloji, tıp, davranışsal veya sağlık bilimlerindeki bir alanda hem teoriye hem de pratiğe dayalı olarak elde edilen verilerin toplanabilmesi, saklanabilmesi, düzenlenip arşivlenmesi, analiz edilip sonuçların görselleştirilerek sunulmasını içerir. Ek olarak çalışmalar sonucu elde edilen verilerin kullanımını ve işlenmesini genişletmeye yönelik olacak hesaplama araçları ve yaklaşımların araştırılarak geliştirilmesi veya bilinen yöntemlerin uygulanmasıyla da ilgilenir. Biyoinformatik analizler bir bakıma biyolojik bilimler ile bilgisayar bilimi, matematik ve istatistik gibi kantitatif bilimlerin arasında bir köprü oluşturur. Temelinde oldukça gelişmiş algoritmaların yer aldığı ve kullanıldığı bu analizler, yazılım programları ya da çevrimiçi ağ tabanlı hizmetler kullanılması aracılığıyla gerçekleştirilir. Biyoinformatik analizlerde; incelenecek biyolojik soru, molekül ya da yapıya uygun olacak şekilde bir veri tabanı ve biyoinformatik analiz gerçekleştirmeye olanak tanıyan program seçilerek analizler yapılmaktadır. Gerçekleştirilen analizler sonucu elde edilen veriler ve sonuçlar literatürde konuyla alakalı önceden bilinen bilgiler ışığında harmanlanıp, analitik olarak yorumlanır (10).

2.7. Veri Madenciliği, Makine Öğrenmesi ve Topluluk Öğrenmesi

Veri madenciliği veri tabanlarındaki gizli kalıpları ortaya çıkarmak için kullanılan yöntemler bütünüdür. Veri madenciliği yöntemleri oldukça büyük miktardaki veri örüntüleri içerisinden gelecekle ilgili tahmin yapabilmemize imkan tanıyacak olan bağıntılar ve kurallar dizilerinin bilgisayar programları yardımıyla bulunmasıdır. Veri madenciliği tanımından yola çıkarak çok fazla miktardaki verinin veri ambarında tutulabilmesi ve bu verilerin işlenerek anlamlı bilgiler elde edilebilir hale getirilmesi asıl amaçtır (74). Veri madenciliği; veri tabanı bilgi teknolojisi, istatistik bilimleri, makine öğrenimi, örüntü tanımlama, sinir ağları, verilerin görselleştirilebilmesi ve uzaysal veri analizi gibi oldukça farklı çalışma alanlarında yer alan yöntemlerin bir birleşiminden oluşmaktadır (75). Bu tekniklerden biri olan makine öğrenimi veriye dayalı bir öğrenim yöntemi gerçekleştirerek yeni gelen verilere maruz bırakıldıklarında bu veri için tahminler ortaya koyabilmeyi amaçlayan yapay zekânın bir alt çalışma alanıdır.

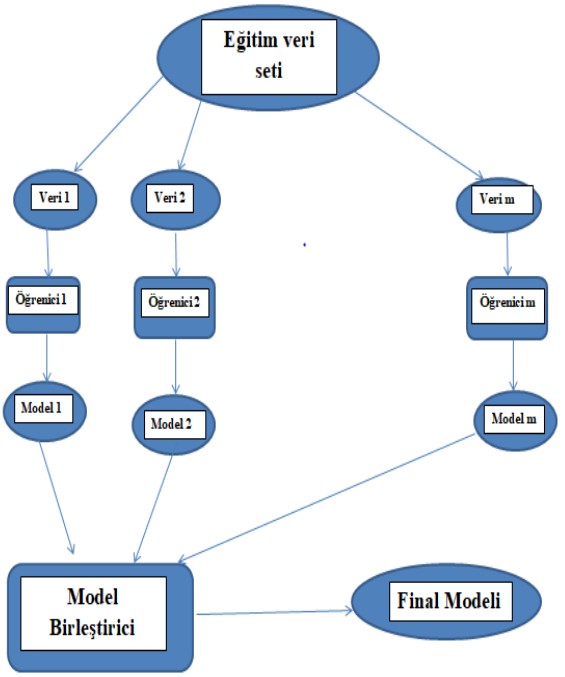
Araştırmacıların çalışmalarında odaklandıkları konu bilgisayarları karmaşık örüntülere karşı bilgilendirmek, onları algılamalarını sağlamak ve veriye dayalı olarak kararlar geliştirebilme yeteneği kazandırmaktır. Makine öğrenmesi, verilere dayalı öğrenimi gerçekleştirmeyi amaçlayan algoritmaların tasarım ve geliştirme süreçlerini içerir. Makine öğrenmesi sistemleri, insan sezgisine ihtiyaç gerektiren durumları tümüyle ortadan kaldırmayı veya insan-makine arasındaki işbirliği ile karar verebilme becerisini elde edebilmeyi amaçlar (13). Çalışmalarda eldeki veri seti ile model oluşturulurken amaç kullanılacak algoritma yardımıyla çalışma için en yüksek performansı elde etmek olmalıdır. Bu nedenle farklı durumlarda ve şartlarda kullanılmak üzere bir çok makine öğrenmesi yöntemi geliştirilerek kullanıma sunulmuştur. Bunlardan en çok bilinenleri ve popüler olanları destek vektör makinaları, sinir ağları, k-en yakın komşuluk algoritması, Naive Bayes sınıflandırıcı, karar ağaçlarıdır. Bu yaklaşımlar genellikle tahmin ve kestirim ve sınıflandırma yapabilme yeteneklerine sahiptir (76). Makine öğreniminin oldukça önemli bir alt çalışma alanı olan topluluk öğrenmesinin mantığı, bilinen tek bir temel sınıflandırıcı modeli kullanılması yoluyla elde edilen doğru tahmin değerini arttırabilmek amacıyla birçok sınıflandırıcı modelini birleştirip modellemenin bu yeni sınıflandırıcılarla yapılabileceği düşüncesine dayanmaktadır. Başka bir deyişle, topluluk öğrenmesi yöntemi tek bir temel sınıflandırıcının (modelin) modelleme ile elde edebileceği sınıflandırma başarısı ile karşılaştırıldığında daha doğru ve güvenilir bir model (meta sınıflandırıcı) elde etmek için birçok temel olarak kabul edilen sınıflandırıcıyı birleştirerek yeni bir sınıflandırma modeli elde edilmesi düşüncesine dayanmaktadır (16). Yapay zekâ, makine öğrenmesi ve topluluk öğrenme disiplinlerine ilişkin hiyerarşi Şekil 2.4'de verilmiştir.

Şekil 2.4. Yapay zekâ, makine öğrenmesi ve topluluk öğrenme disiplinlerine ilişkin hiyerarşi

Topluluk öğrenme yöntemlerinin diğer yöntemlere oranla daha çok tercih edilmesinin sebebi, rastgele tahminlemeler oluşturabilen zayıf sınıflandırıcıların (weak classifiers),

kesin tahminlemeler ortaya koyabilen güçlü sınıflandırıcılara (strong classifiers) göre daha da güçlendirebilir olmalarıdır. Bu güçlendirebilme yeteneğinden dolayı topluluk öğrenme yöntemleri zayıf sınıflandırıcıların geliştirilerek tahminleme güçlerinin artışını sağlamaktadır. Diğer yandan, topluluk öğrenme yöntemleri, kullanılan tek bir sınıflandırma yönteminin eğitim veri seti için ezbere başvurma olayının önlenmesini sağlar ve sınıflandırma başarısının yüksek oranlarda elde edilmesinin beklendiği durumlarda tercih edilir (17).

Topluluk öğrenme mimarisine ilişkin genel yapı Şekil 2.5'de açıklanmaktadır.



Şekil 2.5. Topluluk öğrenme mimarisine ilişkin genel yapı

3. MATERYAL VE METOT

3.1. Veri Seti

İlaca bağlı hepatotoksisiteye ilişkin olası biyobelirteçlerin saptanabilmesi ve bu modelle oluşturulan bilgisayar destekli bir sistemin tasarlanarak hepatotoksisite tanısına yönelik bir yazılım geliştirilmesi ve klinik düzeyde hepatotoksisite durumunun sınıflandırılmasının yapılabilmesi için açık erişimli bir veri seti kullanılmıştır. Tez çalışması kapsamında kullanılan veri seti lncRNA lardan oluşan bir veri seti olup bu veriler aşağıdaki koşullarda gerçekleşen bir sıçan deneyinden elde edilmiştir.

20 adet dişi Sprague-Dawley sıçanın (yaş: 3 aylık; ağırlık: 250±20 g) alınmasıyla deney düzeneği oluşturulmuştur. Deney düzeneğinde

Kontrol grubu: Deneyin ilk günü sisplatinin taşıt çözücüsünün intraperitoneal verildiği grup.

Hepatotoksisite grubu: Deneyin ilk günü 7 mg/kg sisplatinin intraperitoneal verildiği grup.

şeklindedir.

Yapılan deneyde 4. gün yüksek doz anestezi ile ketamin (225 mg/kg i.p.) ve ksilazin (24 mg/kg i.p.) uygulanan sıçanlardan karaciğer doku örnekleri alınarak genomik analizler yapılmış ve veriler elde edilmiştir (77). Tez çalışması için İnönü Üniversitesi Sağlık Bilimleri Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu Başkanlığı'ndan 09.05.20231 tarih ve 2023/4634 sayılı etik kurul onayı alındı.

3.2. Cross Industry Standard Processing – Data Mining (CRISP-DM) (Veri madenciliği için CROSS Endüstriler Arası Standart İşleme Süreci)

Veri bilimi günümüzde oldukça hızlı gelişen ve gün geçtikçe yeni sektörlerde gündeme gelen ve birçok yeni probleme odaklanabilen oldukça önemli çalışma alanlarından birisi olarak görülmektedir. Veri biliminin yıllar içindeki gelişimi ve veri bilimi yoluyla oldukça çok ve farklı alanda yapılan proje ve çalışmaların gereksinimlerinden doğan ihtiyaçlar bu alan için belirli standartlaşmaların gündeme gelmesine sebep olmuştur. Yıllar içerisinde doğan bu gereksinimler için farklı

yaklaşımlar geliştirilmiş olsa da varılan son noktada, gerçekleştirilecek bir veri bilimi projesine nereden ve nasıl başlanacağı, hangi adımların izlenerek projenin yürütüleceği, projenin tüm aşamalarına ait çıktıları ve proje boyunca ölçülebilir olan adımları Cross Industry Standard Processing – Data Mining olarak bilinen CRISP-DM olarak kısaltılan yöntemle yönetilebilmektedir.

Veri madenciliği oldukça farklı aşamaların entegrasyonu ile oluşur. Hemen hemen tüm veri madenciliği çalışmalarında verilerin kaynaklardan toplanması, toplanan verilerin işlenebilecek hale gelebilmesi için temizlenmesi, çalışmada kullanılacak modelin oluşturulması, modelin veri üzerinde denenmesi ve elde edilen sonuçların sunulabilmesi için hazırlık adımları karşımıza çıkmaktadır. CRISP-DM endüstri ve kullanılan yazılımdan bağımsız olarak ele alınan bir veri madenciliği süreç modelidir (77).

6 adımlı bir süreçten oluşan CRISP-DM aşamaları şu şekildedir.

İş Süreçlerinin anlaşılması aşaması: Bu aşamada üzerinde çalışılacak problemin tanımlaması yapılır. Yani bir problemin neden meydana geldiği, problemin çözülmesiyle gerçekleşmesi beklenilenler, problemin etkilediği veri kaynakları ve veri akışlarının tespiti sağlanır. Son olarak problemin çözümü ile hangi çıktıların elde edileceği tanımlanır.

Verinin anlaşılması: Bu adımda, problem tanımına uygun olacak şekilde veri toplanır veya var olan mevcut veri incelenir. Bu adıma geçilmeden önce problemin tanımının doğru ve kesin bir şekilde yapılması oldukça gerekli ve önemlidir. Bir önceki aşamada yapılan tanımlama hatasından dolayı eldeki veri tamamının gereksiz yere işlenebilir ve amaca hizmet etmeyebilir. Ayrıca bu aşamada kullanılacak veri ile ilgili olan problemler de incelenmelidir. Örneğin verinin yapısal tespiti bu aşamada gerçekleştirilir. Ayrıca verinin gürültülü veya kirli olması, eksik gözlem içermesi gibi sorunların tespiti gibi işlemler de bu aşamada kontrol edilir. Yapılan incelemeler sonucunda uygun görülürse ek veri toplama veya eldeki veri üzerinden ne tür işlemler yardımıyla verinin çalışmaya uygun hale getirileceğine bu aşamada karar verilir.

Veri Ön İşleme aşaması: Bu aşamada veri üzerinde gerçekleştirilmesi planlanan işlemlerin hangi yöntemler yardımıyla gerçekleştirileceğine karar verilir. Bir önceki süreçte, veride eksik veri olduğu tespit edildiyse, eksik olan veri kısmının çalışmaya dahil edilmemesi sisteme hiç dahil edilmemesi veya veride eksik olan kısımların tamamlanması

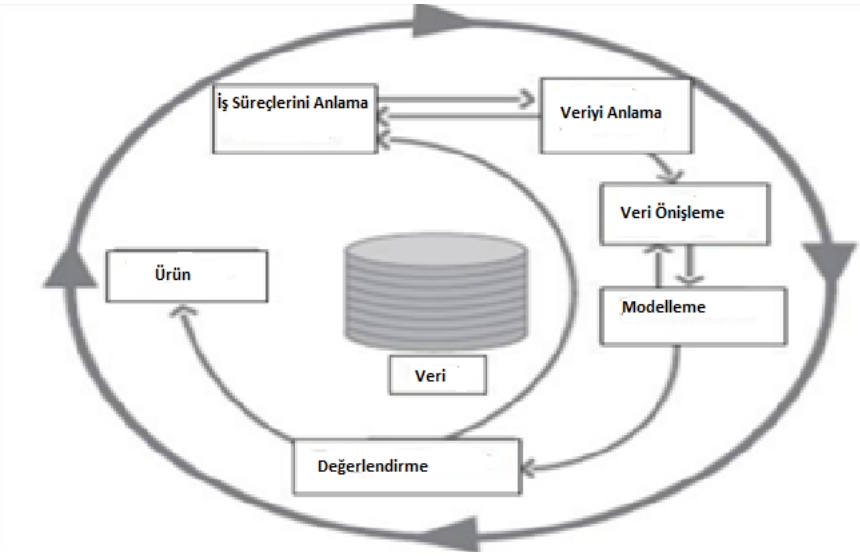
aşamalarına karar verilir. Eğer eksik veriler tamamlanacak ise, bu tamamlamanın nasıl yapılacağı ve hangi yöntem ile giderileceği kararına bu aşamada karar verilmektedir.

Model aşaması: Bu aşama veri analitiği projelerinin, veri madenciliği bilimine doğrudan dokunduğu aşama olarak bilinir. Bu aşamada ilk adımda tanımı gerçekleştirilen probleme yönelik veri kaynağı üzerinde bir makine öğrenmesi modeli ya da istatistiksel bir model geliştirilir. Geliştirilen bu model, tanımlanan problemin çözümüne uygun olacak şekilde çeşitli yöntemler ile iyileştirilebilir (optimization). Bu aşamada ayrıca birçok çalışma için verinin modele uygun olarak düzenlenmesi gerekebilir. Örneğin bazı modellerin çalışabilmesi için dengeli veri durumu gerekmektedir. Bazı modeller için ise modelin çalışması aykırı verinin ayıklanması durumuna bağlı olup bazı modeller için de veri dönüşümleri gerekebilmektedir.

Değerlendirme aşaması: Bu aşamada, önceki adımların hepsi için bir değerlendirme yapılır ve ilk adımda tanımlanan problem çözümü için konulan başarı kriterlerinin ne ölçüde gerçekleştirildiği ve çözümün sağladığı test edilir.

Ürün aşaması: Bu son aşamada daha önceki aşamalar ile elde edilen çıktılar çalışma ortamına uygun hale getirilmek amacıyla işlenme süreci başlar. Örneğin elde edilen çıktılar kullanılan programlama dilleri yardımıyla veri akışına uygun hale getirilerek elde edilen sistemin çalışan bir uygulaması ortaya konur.

CRISP-DM adımları, akışı ve bu süreçler arasındaki geçişler Şekil 3'te gösterilmiştir.



Şekil 3.1. CRISP-DM adımları, akışı ve bu süreçler arasındaki geçişler (77)

3.3. İş süreçlerinin anlaşılması

Bu aşamada hedef değişken olan ilaca bağlı hepatotoksisite modeli oluşturulan sıçanlar ile patoloji oluşturulmamış sağlıklı sıçanlardan alınan karaciğer doku örneklerinin genomik analizlerinden elde edilen büyük verilerin öznitelikleri açıklayıcı/tahminleyici/girdi değişkeni, hepatotoksisite'nin varlığı/yokluğu ise sonuç/çıktı/hedef değişkenleri olarak alınmıştır.

3.4. Verinin anlaşılması

Biyolojik örneklerden genomik analizler sonucunda elde edilen büyük veriler üzerinde analizlere başlamadan önce veride gürültülü/aşırı/aykırı, kayıp/eksik, hatalı veya sapan gözlemlerin olup olmadığı tespit edilecektir. İlgili verilerdeki saptanan sorunların ve eksiklerin giderilebilmesi için hangi işlemlerin yapılabileceği ve verinin hangi yolla analizlere uygun hale getirilebileceğine bu aşamada karar verilecektir.

3.5. Veri önişleme aşaması

Bu aşamada, kullanılacak büyük verilerde kayıp değer varsa, kayıp değer sayılarının değişken bazındaki durumuna göre, ilgili değişkenin veriden çıkarılması veya kayıp değerlerin uygun algoritmalar kullanılarak tamamlanması gerçekleştirilecektir. Kayıp değer içeren gözlemlerin veri setinden çıkarılması durumu, örneklem hacminin küçülmesine neden olmakta ve bu durum ise yapılan analizlerin istatistiksel gücünün azalması sonucunu gündeme getirmektedir (78). Bu nedenle bu çalışmada kayıp veriler ile karşılaşılması durumunda ilgili değişken(ler)in yapısına uygun olabilecek random forest, lojistik regresyon, vb. modeller kullanılarak tamamlanacaktır. Ayrıca bu aşamada kullanılacak verilerden nicel olanlar gerek duyulduğu takdirde standardizasyon, normalizasyon gibi veri dönüşümü yöntemleri ile uygun formlara getirilecektir. Böylelikle kullanılacak modellerin sonuçlarının daha geçerli ve güvenilir olması sağlanabilecektir. Kullanılacak modellerin eğitilmesinde karşılaşılabilecek olası aşırı/eksik uyum (over/under fitting) problemleri ve modellerin performansını artırmak gerektiğinde değişken seçim tekniklerinden Lasso, Elastik net ve Boruta, algoritmalarından uygun olan(lar)ı kullanılacaktır.

3.5.1. Değişken seçimi

Değişken seçimi, herhangi bir öngörücü modelleme projesinde yer alan en önemli adımlardan biridir. Genellikle 'özellik seçimi' olarak adlandırılan değişken seçimi

istatistiksel bir model geliştirmenin oldukça önemli bir adımı olup, hangi verilerin/değişkenlerin çalışmaya dahil edileceğine karar verme sürecidir. Oldukça büyük veri kümelerinde ve yüksek hesaplama maliyetleri olan modellerle çalışmadan önce bir veri setini tanımlayabilecek en kullanışlı özellikleri belirleyerek yüksek verimlilik ve sonuçlar elde edilebilir. Özellik seçimi işlemi bağımlı değişken üzerinde etkisi olan bir veri kümesi üstünde özellik tanımlama işlemidir. Bağımlı değişken üzerinde etkisi olan açıklayıcı değişkenlerin yüksek boyutsallığa sahip olması hem uzun hesaplama sürelerine hem de verilerin aşırı öğrenme ile sonuç türetmesine neden olabilir. Ayrıca, birçok özellik ile oluşturulan modellerin yorumlanması ve sonuçlarının ortaya konması süreci oldukça zordur. İstatistiksel modelleme yapmadan önce önemli özellikler yani modeli en çok açıklama özelliğine sahip değişkenler seçilmelidir (79). Birçok makine öğrenimi modeli ve veri madenciliği yöntemi yüksek boyutlu veriler ile çalışırken zorlanabilir veya etkisiz kalabilir. Sonuç olarak, verideki boyutsallık azaltıldığında bu yöntemler daha etkili ve doğru sonuçlar üretebilir (80).

Gen ekspresyonu ile elde edilen veri kümeleri oldukça büyük hacimlere sahiptir. Gen ekspresyonu veri setlerinin sahip olduğu bu büyük boyutu nedeniyle, kullanılan veri modelleme analizleri oldukça uzun zamanlar alır ve bu veri kümeleri analizlerde hesaplama verimsizliğine neden olabilir. Belkide modelin performansı, yüksek boyutsallık sorununun bir sonucu olarak yetersiz kalabilir. Gen ekspresyonu ile elde edilen veri setlerindeki oldukça çok sayıda olan genler, sınıflandırma algoritmasında eğitim örneklerinin aşırı öğrenilmesine ve yeni örnekleri genellenememesine neden olabilir.

3.5.2. Lasso

Lasso değişken seçimi yöntemi ilk olarak 1996 yılında Robert Tibshirani tarafından ortaya atılmıştır. LASSO yöntemi, model parametrelerinin mutlak değerlerinin toplamına bir kısıt koyar ve bu toplamın bir sabit değerden (üst sınır) daha az olması gerekmektedir. Bu durumu gerçekleştirmek için yöntem, regresyon değişkenlerinin katsayılarını cezalandıran ve bazılarının sıfıra düşmesine neden olan bir kısma (daraltma-düzenlileştirme) işlemi kullanır. Bu yöntem özellikle az sayıda gözlem ve çok sayıda değişken içeren veri setlerinde kullanışlıdır. Ayrıca LASSO yöntemi, yanıt/hedef değişkeni ile ilişkili olmayan alakasız değişkenlerin ortadan kaldırılmasını sağlayarak

modelin yorumlanabilirliğini artırmaya yardımcı olur ve bu şekilde aşırı öğrenme sorunu da giderilebilmektedir (81).

3.5.3. Elastik net

Elastik net yöntemi Ridge ve Lasso değişken seçimi yöntemlerinin özelliklerinden yararlanmak için aynı anda bu iki yöntemin cezalarını kullanır. Elastic-net tahmin edicisine ait denklem

$$\beta_{e-net} = min\|y - X\beta\|^2_2 + \lambda[(1-\alpha)\|\beta\|^2_2 + \alpha\|\beta\|_1]$$

şeklindedir. Burada yer verilen "λ" büzülme (shrinkage) parametresini belirtmektedir. α ise bir cezalandırma parametresi olarak bilinir. α=1 alındığı durumlarda denklem LASSO tahmin edicisi, α=0 alındığında ise Ridge tahmin edicisi olarak elde edilir. Denklemde α değerinin 0 ile 1 arasında (0 ve 1'den farklı) bir değer alması ile Elastic-net tahmin edicisi elde edilir. Bu özellik Elastic net algoritmasının, çoklu bağlantı problemi karşısında etkili olduğunu ve aynı zamanda değişken seçimi görevini de yerine getirdiğinin göstergesidir (82).

Bu değişken seçim yöntemi genom çapında ilişki çalışmalarında (GWAS) yaygın olan şiddetli çoklu doğrusallığı daha etkili bir şekilde ele alma yeteneği sunmaktadır (83, 84).

3.5.4. Boruta

İlk kez 2010 yılında Kursa ve Rudnicki ikilisi tarafından ortaya atılan Boruta değişken seçim yöntemi Random Forest yönteminin temel çalışma algoritmasına bağlı bir değişken seçim yöntemidir. Boruta algoritması, randomForest R paketinde uygulanan rasgele orman sınıflandırma algoritması üzerine kurulmuş bir modeldir. Kısacası rastgele bir orman sınıflandırıcısı üzerine inşa edilmiş bir sarmal yaklaşım kullanır (85).

Boruta yöntemi çalışma prensibi aşağıdaki adımları içerir:

İlk olarak orijinal veri setine ait kopyalar oluşturulur. Bu oluşturulan kopya veri setlerinde yer alan değişken değerleri (gölge değişkenler olarak adlandırılır) karıştırılır. Daha sonra orijinal ve karıştırılmış veriler, değişken önemliliğini ölçümü yapan bir RF modelini eğitmek adına için birleştirilir.

Sonra her bir değişken için Z skoru hesaplanır. Bu hesaplanan Z skoru, Random Forest modelinden elde edilen değişken önemlilik değerlerinin standardize edilmesiyle elde edilen değerdir.

Gölge değişkenler arasında en yüksek Z skoru olan değer belirlenir (Buna maxZF adını verelim).

Elde edilen Z skorlarından değeri maxZF'den büyük olan orijinal değişkenler önemli olarak belirlenirken, Z skoru maxZF'den küçük olan değişkenler ise önemsiz olarak belirlenir.

Tüm değişkenler için belirleme işlemi tamamlanana kadar yukarıdaki işlemler aynı şekilde tekrarlanır.

3.6. Model aşaması

Bu aşamada tanımı yapılan problem ve probleme uygun olarak elde edilen veri seti üzerinde bir makine öğrenmesi yöntemi veya bir istatistiksel model uygulanacaktır. Kullanılacak model/yöntem, tanımı yapılan problem çözümüne yönelik olarak elverişli optimizasyon yöntemleri yardımıyla iyileştirilir (86).

Bu çerçevede, çalışmada CRISP-DM metodolojisinin model aşamasındaki işlemler için kullanılabilecek modeller ve algoritmalara ilişkin detaylar aşağıdadır:

3.6.1. Sınıflandırma yöntemleri

Sınıflandırma, önceden belirlenen bir sınıf etiketi baz alınarak veri öğelerini sınıflandırmaya yarayan kontrollü bir öğrenme tekniğidir. Sınıflandırma işleminin amacı yeni bir veri öğesinin niteliklerini inceleyerek bu ögeyi önceden tanımlanmış bir sınıfa atamaktır. Kısaca sınıflandırmada sınıflandırılmamış verilere sınıflandırılması için uygulanabilecek bir tür model oluşturmaktır. Sınıflandırma işleminin yapılabilmesi için birçok istatistik ve makine öğrenimi temeline dayalı olan yöntem geliştirilmiştir (87). Bunlardan Bayes sınıflandırma, doğrusal regresyon analizi, lojistik regresyon analizi, ayırma analizi istatistik temelleri olan sınıflandırma yöntemlerindendir. Karar ağaçları, en yakın komşu yöntemi, yapay sinir ağları, destek vektör makineleri gibi yöntemler ise başlıca makine öğrenimi temelli sınıflandırma yöntemleri olarak bilinir.

3.6.2. Lojistik Regresyon Analizi

Lojistik regresyon, bağımlı/hedef/yanıt değişkeninin ikili veya çoklu kategorik olarak gözlenebildiği durumlarda bağımsız değişkenlerle olan sebep-sonuç ilişkisini belirleyebilmek adına yararlanılan bir yöntemdir. Bağımsız değişkenlerin aldığı değerlere göre bağımlı değişkenin beklenen değerine ait olasılığı belirleme yöntemi olarak bilinen lojistik regresyon, aynı zamanda bağımsız değişkenlerin etkileri göz önüne alınarak

verilerin sınıflandırılması amacıyla da kullanılabilmektedir. Bu yöntemde bağımlı/hedef/yanıt değişken üzerinde bağımsız değişkenlerin yarattığı etkiler olasılık olarak elde edilmesi yolu ile aslında risk faktörlerinin olasılık olarak ortaya konması sağlanır. Lojistik regresyon yönteminin tıp/sağlık alanında yer alan uygulamalarında bağımsız değişkenler, risk değişkenleri/faktörleri ya da bir hastalığın ortaya çıkıp çıkmamasını belirleyen değişkenler/faktörler olarak adlandırılır. Kısacası lojistik regresyon, bağımsız değişkenlere göre bağımlı/hedef/yanıt değişkenin beklenen değerinin olasılık olarak ifade edilebildiği atama ve sınıflama işlemi yapmaya yardımcı bir yöntemidir (88).

3.6.3. Destek Vektör Makinesi (DVM)

DVM yöntemi son yılların oldukça popüler algoritmaları arasında yerini almıştır. Vapnik - Chervonenkis tarafından ortaya atılan DVM sınıflandırmanın yanı sıra regresyon için de kullanılan bir makine öğrenmesi yöntemidir. DVM yöntemi verileri dönüştürebilmek adına çekirdek çözümü (kernel trick) adı verilen bir teknikten faydalanmaktadır. Çekirdek çözümü yöntemlerinin alt yapısı veri dönüşümü modellerine dayanmakta olup olası sonuçların arasından en uygun olabilecek sınırı belirler. Kısacası çekirdek çözümü yöntemleri ilk olarak verilere karmaşık veri dönüşümleri uygulayarak verilerin tanımı yapılmış olan etiketler ya da sonuçlara bağlı olarak ne şekilde ayrılabileceğini belirler (89, 90). DVM'nin asıl amacı hedef/yanıt değişkenine ait sınıfları birbirlerinden en uygun olabilecek şekilde ayırsama yapabilecek bir hiperdüzlemin elde edilmesi olayıdır. DVM yöntemi için karşılaşılabilecek iki olası durum bulunmaktadır. Bunlar verilerin doğrusal olarak ayrılabilir bir yapıda olması veya doğrusal olarak ayrılamayan bir yapıdan oluşması durumlarıdır. DVM yöntemi ile sınıflandırma işlemi de regresyon modeli de yapılacak olsa doğrusal olmayan durumlardan oluşan yapılar için çözüm için çekirdek fonksiyonlarından faydalanılır (91, 92). Doğrusal olmayan DVM yöntemi değişik yapılardan oluşan çekirdek fonksiyonlarını maksimum marja sahip hiperdüzleme uygulayarak doğrusal olmayan sınıflandırıcılar elde etmektedir. Elde edilen yeni algoritma doğrusal olan DVM ile bazı benzer özellikler taşır. Fakat yeni algoritmada her iç çarpım yerine doğrusal olmayan bir çekirdek fonksiyonu kullanılmaktadır. Özetle doğrusal olmayan DVM yöntemi çekirdek yardımıyla bütün değerlerin tekrar tekrar çarpım değerlerinin hesaplanıp bulunması yerine direk çekirdek fonksiyonunda değerlerin yerlerine koyulmasıyla nitelik uzayındaki değerinin elde edilmesi sağlanmaktadır. Böylece elde edilen algoritma, maksimum aralığa sahip hiperdüzlemin

dönüştürülmüş olan örnek uzaya yerleştirilmesine izin vermiş Böylelikle oldukça yüksek boyuta sahip bir nitelik uzayı ile uğraşma durumu ortadan kalkmaktadır (93).

3.6.4. Random Forest (RF)

RF; Leo Breiman ile Adele Cutler tarafından geliştirilmiş olan oylama metoduna dayalı bir sınıflama yöntemidir. Birden çok karar ağacının birleşmesiyle oluşmaktadır ve bireysel ağaçların oylanması ile kazanan sınıf belirlenmektedir. Ormanda olan karar ağaçlarının her biri birbirlerinden bağımsız olup veri setinden bootstrap yöntemi ile çekilmiş örnekler ile oluşturulur (94). RF yöntemi oldukça çok sayıda karar ağacının birleşmesiyle oluşturulmuş orman sınıflayıcısı olup bu yöntem yardımıyla sınıflama veya regresyon ağaçları kurulabilmektedir. Veri setindeki "sınıf değişkeni" kategorik ise sınıflama, sürekli ise regresyon ağaçları oluşturulabilmektedir (14). RF yönteminde en önemli husus dallanma kriterlerinin belirlenerek uygun olabilecek bir budama yönteminin belirlenmesiyle modelin oluşturulmasıdır. RF sınıflandırıcının dallanma kriterlerinin belirleyebilmek amacıyla Gini indeksi yöntemi kullanılmaktadır. Gini indeksi yöntemi sınıf özniteliklerinin zayıflık derecelerini belirlemeye yarar (95). RF yönteminde de birçok sınıflandırma yönteminde olduğu gibi kullanıcı tarafından önceden belirlenmesi gerekmekte olan parametreler bulunmaktadır. Bahsedilen bu parametreler, ağaç yapısının oluşturulması adımı için gerekli olan ve her bir düğüm için kullanılacak örneklerin sayısı ile oluşturulacak ağaçların sayısıdır (96).

3.6.5. XGBoost

Gradient Boosting yöntemi, Friedman tarafından 2001 yılında tanıtımı yapılan oldukça güçlü bir teknik olup bu model; bir tahmin modeli için genel olarak karar ağaçları gibi zayıf bilinen tahmin modellerinin topluluk formlarını üreterek regresyon ve sınıflandırma problemlerinin çözümüne uygun bir makine öğrenme yöntemi olarak tanımlanmaktadır. Gradient Boosting'in, çalışma prensibi boosting tekniklerine dayanmakta olup çok sayıda olan zayıf öğreniciyi belirlemek ve onları karmaşık olan bir modelde birleştirmek üzerine kuruludur. Bu yöntem de birçok model gibi regresyon ve sınıflandırma problemlerinde kullanılabilmektedir (97).

Extreme Gradient Boosting'in kısaltılmış hali olan XGBoost, temel yapısı itibariyle gradient boosting ile karar ağacı yöntemlerine dayalı olan bir makine öğrenme yöntemidir. XGBoost yönteminin bilinen orijinal formu Friedman tarafından 2002 yılında ortaya konmuştur (98). Daha sonra ise Washington Üniversitesi'nde bulunan iki

araştırmacı Tianqi Chen ve Carlos Guestrin tarafından Bilgi İşlem Makinalarının Bilgi Keşfi ve Veri Madenciliği Özel İlgi Grubu Derneği 2016 konferansında makale olarak sunumu sonrasında makine öğrenmesi yöntemleri arasında oldukça popüler hale gelmiştir (99).

XGBoost, oldukça popüler bir algoritma olup sağlık, enerji, finans vb. alanlarda kendine uygulama alanı bulmuştur. Diğer algoritmalara kıyasla hızlı olması ve performansı düşünülecek olursa oldukça avantajlı olduğu görülür. Ayrıca XGBoost yöntemi oldukça yüksek tahminleme gücüne sahip olup bir yöntem olup, diğer yöntemlerden 10 kat daha hızlı olup modelin genel performansı iyileştirebilen ve aşırı uyum ile aşırı öğrenmeyi azaltmaya yarayan bir dizi regularization içermektedir. Gradient boosting yöntemi güçlü bir sınıflandırıcı oluşturabilmek adına çok sayıda zayıf sınıflandırıcıyı boosting algoritması yardımıyla birleştiren topluluk yöntemidir. Gradient boosting ile XGBoost yöntemleri aynı çalışma prensibini izlemektedir. Aralarındaki temel farklılıklar ise uygulama detaylarında bulunmaktadır. XGBoost yöntemi farklı regularization teknikleri kullanması yardımıyla ağaçların karmaşık durumlarını kontrol ederek daha iyi model performansı elde etmeyi başarmaktadır (100).

3.6.6. Meta Learning (Meta Öğrenme) ve Topluluk (Ensemble) Öğrenme Yöntemleri

Meta öğrenme otomatik öğrenme algoritmalarının makine öğrenme deneyleri hakkındaki meta verilere uygulandığı makine öğreniminin bir alt alanıdır. Meta-öğrenme en yaygın olarak, çoklu öğrenme bölümlerinde bir öğrenme algoritmasını geliştirme sürecini ifade eden öğrenmeyi öğrenme süreci olarak düşünülür. Meta öğrenmede bir faktör, gelecekteki yeni görevlerin öğrenilmesini kolaylaştırmayı amaçlayan, gözlemlenen görevlerden bilgi çıkarır. Gelecekteki görevlerin önceki görevlerle ilgili olduğu varsayımı altında, biriken bilgi öğrenilen görevler arasında ortak yapıyı yakalayacak şekilde öğrenilmeli ve öğrenicinin yeni görevlerin yeni yönlerine uyum sağlayabilmesi için yeterli esnekliğe izin verilmelidir. Öğrenme, gözlemlenen görevlere dayalı hipotezler üzerinde dağılımın inşa edilmesi ve yeni bir görev öğrenmek için kullanılması yoluyla gerçekleşir. Bu nedenle, önceki bilgi, yeni görevler için deneyime bağlı bir ön ayar yapılarak dahil edilir. Meta öğrenme, 'nasıl öğrenileceğini' öğrenerek performansı artırmayı amaçlar. Bu yöntemde amaç genellikle görevler arasında genelleme yapabilen ve her yeni görevin öncekinden daha iyi öğrenilmesini sağlayan genel amaçlı bir öğrenme algoritması öğrenmektir (101).

Makine öğrenme metodları, son zamanlarda yapay zekanın kullanıldığı çalışma alanlarında ve uygulamalı bilimlerde oldukça geniş bir kullanım alanına ulaşmıştır. Makine öğrenmesi barındırdığı güçlü algoritmalar yardımıyla birçok karmaşık veri seti üzerinde yüksek tahmin performansı gösterirken bazı veri setleri üzerinde ise yüksek varyans ve düşük doğruluk değerlerine sahip tahminler gerçekleştirmektedir. Sınıflandırma ve regresyon problemleri için oluşan bu performans eksikliğini giderebilmek adına farklı yöntemler ortaya atılmıştır. Bu yöntemlerden biri ise topluluk öğrenme yöntemleridir. Topluluk öğrenme yöntemleri ilk ortaya çıktığındaki amacı makine öğrenme yöntemlerinde bazı modellemeler sonucunda oluşan yüksek varyansı azaltmak ve bu yolla elde edilen doğruluk oranını yükseltmekti (102). Topluluk öğrenme yönteminin temel çalışma mantığı elde edilen birçok kararın, tek bir olası karardan oldukça daha sağlıklı ve daha doğru sonuçlar içereceği kabulüne dayanmaktadır. Bir modelin yapacağı tahminin hatalı olma olasılığı, birden fazla modelin vereceği ortak tahmin sonucunun hatalı olma olasılığından daha yüksektir (102). Topluluk öğrenme yöntemleri genel olarak, bagging (torbalama), boosting (artırma) ve stacking (istifleme) olmak üzere üç kısımda incelenir.

3.6.7. Bagging (Bootstrap aggregating; Torbalama)

Bagging yöntemi, bilinen bir eğitim setinden yerine koyma yolu ile rastgele seçim yaparak yeni eğitim setleri oluşturup temel öğrenicinin yeniden eğitilmesini amaçlayan bir yöntemdir (103). Özetle, Bagging yönteminin temel amacı, eğitim verisini kullanarak rastgele olacak şekilde yeni veri setleri elde edip farklılıklar oluşturarak sınıflandırma başarısını artırmaktır. Bagging yönteminde ilk olarak, veri seti eğitim ve test verisi olarak bölünür. N adet örnek içeren eğitim setinden yerine koyarak rastgele seçim yöntemiyle yine n örnekten oluşan bir ya da birden fazla yeni eğitim seti elde edilir. Sonuçta eğitim setinde yer alan örneklerin bazıları elde edilen yeni eğitim setlerinde bulunmazken bazıları ise birden fazla kez yer almaktadır. Bagging yöntemiyle elde edilen toplulukta bulunan her bir temel sınıflandırıcı bu şekilde elde edilmiş farklı örnekler içeren eğitim setleri ile eğitilmektedir. Son olarak her bir temel sınıflandırıcıdan elde edilen sonuç çoğunluk oylaması ile birleştirilir (104).

3.6.8. Boosting (Yükseltme/Artırma)

Boosting yöntemi Schapire tarafından 1990 yılında ortaya atılan ve 2000'li yıllara kadar gelişmekte olan bir topluluk öğrenme yöntemidir (105-107). "Artırma" terimi, zayıf

öğrenme yöntemlerini güçlü öğrenme tekniklerine dönüştüren bir algoritma ailesini ifade eder. Boosting, herhangi bir öğrenme algoritmasının model tahminlerini geliştirmek için bir topluluk yöntemi olup Bagging yönteminden farklı olarak bu yöntemde tahmin ediciler birbirlerinden bağımsız olmamayla birlikte sıralı olacak şekilde oluşturulmaktadır. Bu yöntemde amaç zayıf tahminleyicileri birleştirme yoluyla güçlü tahminleyici(ler) ortaya koymaktır. Modeller gözlemlere ağırlıkların ataması yapılmasıyla oluşturulur. Boosting yönteminde de bagging yönteminin çalışma prensibinde olduğu gibi N adet eğitim seti oluşturulmaktadır. Var olan eğitim setinde yer alan gözlemlere başlangıç durumunda eşit ağırlıklar atanmaktadır. İlk olarak eldeki eğitim seti yardımıyla model oluşturulmaktadır ve bu modelden sonuçlar elde edilmektedir. Ortaya çıkan sonuçlara göre yanlış olarak sınıflandırılması yapılan gözlemlerin ağırlıkları arttırılırken, doğru olarak sınıflandırılması yapılan gözlemlerin ağırlık değerleri ise azaltılır. Yenilenen bu ağırlıklar yardımıyla model tekrar kurulur ve aynı şekilde bir önceki adımda olduğu gibi sınıflandırma sonuçları baz alınarak ağırlıklar tekrar düzenlenir. Bu işlem N kez tekrarlanmaktadır. Bu yöntemde hem bagging yönteminin var olması hem de gözlemlere ağırlık atanması yolu ile varyansı ve yanlılığı oldukça düşük modeller elde edilmektedir (107).

3.6.9. Stacking (İstifleme)

Stacking yöntemi, iki veya daha fazla temel olarak kabul edilen çoklu sınıflandırma modellerinin birleşimi ile bir meta sınıflandırıcı oluşumunu sağlayan bir topluluk öğrenmesi tekniğidir. Kullanılan sınıflandırma modellerinin tahmin sonuçlarının birleştirilmesi yoluyla eğitimin gerçekleştirildiği bir topluluk öğrenme modeli olarak bilinir. Temel sınıflandırıcı yöntemler tarafından oluşturulan modellerden yapılan tahminlemeler, modelde bulunan sıralı katmanların her biri için girdi olarak yer almakta ve yeni bir tahmin seti oluşturabilmek için birleştirilmektedir. Stacking yönteminin çalışma prensibinde, temel olarak bilinen sınıflandırma modelleri orijinal eğitim veri seti üzerinde eğitilmekte ve daha sonra meta-sınıflandırıcı ise topluluk modelde yer verilen temel sınıflandırma yöntemlerinin çıktılarına, (sonuçlarına, tahminlerine) dayanarak elde edilmektedir. Elde edilen Meta-sınıflandırıcı ise tahminleme ile elde edilen sınıf etiketleri üzerinde eğitim işlemi gerçekleştirerek sınıflandırma işlemi yapmaktadır (108).

3.7. Veri analizi

Yapılan biyoinformatik ve biyoistatistik analizler ile modellemeler aşağıdadır.

3.7.1. Biyoinformatik analizler

lncRNA ekspresyon profilleri incelenen hepatotoksisite oluşturulan sıçanlar ile kontrol grubu sıçanların örnekleri için, R programlama dilindeki limma paketi kullanılarak diferansiyel ekspresyon analizleri yapıldı (109). Diferansiyel ifade analizi, hasta-kontrol veya tedavi kolları arasındaki ifade aktivitelerinde nicel farklılıkları bulmak için normalleştirilmiş okuma sayısı verilerinin istatistiksel analizidir. R yazılım ortamı aracılığıyla ilgili analizler için bir ardışık düzen (pipeline) tasarlanmıştır. Elde edilen sonuçlar, önem sırasına göre bir gen tablosundan ve farklı şekilde ifade edilen genleri görselleştirmek için bir grafikten (volkan grafiği) oluşmaktadır. Sonuç tablosu, ayarlanmış P ve log2-kat değişim (log2FC) değerlerini içermekte ve en küçük p değerlerine sahip genler en güvenilir genler olacaktır. Yukarı regüle edilmiş genleri tanımlamak için Log2FC>1 ve aşağı regüle edilmiş genleri tanımlamak için log2FC<-1 kullanıldı (110). Farklı ifade edilen lncRNAları görsel olarak gözlemleyebilmek için bir yanardağ grafiği çizilmiştir. Grafikte yukarı ekspresyon gösteren lncRNAlar kırmızı, aşağı ekspresyon gösteren lncRNAlar mavi ve 2 grup arasında ifadelerinde değişiklik olmayan lncRNAlar siyah renk ile gösterilmiştir.

3.7.2. Biyoistatistiksel analizler ve Modelleme

Değişkenlerin normal dağılım gösterip göstermediği Shapiro Wilk testi ile incelendi. Veriler ortanca (minimum-maksimum) veya ortalama± standart sapma olarak sunuldu. Normal dağılmayan verileri karşılaştırmak için Mann-Whitney U testi ve normal dağılım gösteren verileri karşılaştırmak için bağımsız örneklem t testi kullanıldı. Her bir genin olasılık oranını (etki büyüklüğünün bir ölçüsü) tahmin etmek için lojistik regresyon analizi yapıldı. Lojistik regresyon için Hosmer ve Lemeshow'un uyum iyiliği testi ve omnibus model katsayıları testi hesaplanmıştır. p değeri <0.05 istatistiksel olarak anlamlı kabul edildi. Analizde IBM SPSS Statistics 26.0 programı kullanılmıştır.

Proje kapsamında kullanılması planlanan bireysel ve topluluk öğrenme modellerinin uygulanmasında R programlama dili tabanlı RStudio kullanılmıştır (111). Model geçerliliği için 10 katlı cross validasyon yöntemi kullanıldı. n-katlı çapraz doğrulama yönteminde veri önce n parçaya bölünür ve kullanılan model n parçaya uygulanır. n parçadan biri test için kullanılırken, diğer n-1 parça modeli eğitmek için kullanılır. Elde edilen değerlerin ortalaması, çapraz doğrulama yöntemi için değerlendirilir. Bu çalışmada, Bagging ve Boosting topluluk öğrenme yöntemleri için

sınıflandırıcı olarak karar ağaçları yöntemi kullanılmıştır. Stacking topluluk öğrenme yöntemi için ise sınıflandırıcı olarak lojistik regresyon, yapay sinir ağları, Random Forest ve destek vektör makineleri yöntemi kullanılmıştır. Meta sınıflandırıcı olarak Random Forest yöntemi kullanılmıştır.

3.8. Değerlendirme aşaması

Bu projede, doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skoru gibi metrikler %95 güven aralığı ile analizlerin performans değerlendirmesinde kullanılmıştır.

3.9. Ürün aşaması

Bu tez kapsamında biyoinformatik analizleri gerçekleştirmek için R programlama dilinde yazılan ardışık kod dizilimi R programlama dili tabanlı ve bir RStudio projesi olan Shiny kütüphanesi ile web tabanlı bir yazılım olarak sunulmuştur (112).

3.9.1. Geliştirilen Web Tabanlı Yazılım

Geliştirilen web tabanlı yazılım, 2 gruplu verilerde diferansiyel olarak ifade edilen genleri tanımlamak için R programlama dili temelinde interaktif web tabanlı uygulamaların tasarlanmasına izin veren Shiny kütüphanesi kullanılarak tasarlanmıştır (112). Ayrıca ek olarak ara yüzün geliştirilmesinde ggplot2 (113), ggrepel (114), DT (115), shinyWidgets (116), shinyLP(117), shinydashboard (118) ve limma (109) paketleri kullanılmıştır. Yazılıma ait ana ve alt menüler aşağıda açıklanmıştır. Geliştirilen yazılım "Giriş", "Veri İşlemleri" ve "Analiz" olmak üzere 3 ana bölümden oluşmaktadır. Ayrıca geliştirilen yazılım Türkçe ve İngilizce dil seçeneklerini içermektedir.

3.9.2. Bioconductor, Limma

Bioconductor, yüksek verimli genomik verileri analiz etmek için araçlar sağlayan ve R programlama dilini temel alan açık kaynaklı bir yazılım projesidir. Biyolojik verileri yorumlamak için yeni hesaplamalı yaklaşımlar oluşturulurken, Bioconductor projesi, R programlama dilinde yazılmış çeşitli istatistiksel araçları barındıran açık kaynaklı bir yazılım deposudur. Birçok Bioconductor paketi, R'nin kapsamlı istatistiksel ve grafiksel özelliklerini kullanarak çeşitli veri analizi taleplerini karşılamak için oluşturulmuştur.
Limma (Linear Models for Microarray Analysis), özellikle tasarlanmış deneyleri analiz etmek ve diferansiyel ekspresyonun değerlendirilmesi için lineer modellerin kullanımı

olmak üzere, gen ekspresyonu mikrodizi verilerinin analizi için kullanılan bir Bioconductor paketidir. Ampirik Bayes yöntemleri, dizi sayısı az olduğunda bile kararlı sonuçlar sağlamak için kullanılır. Doğrusal model ve diferansiyel ekspresyon fonksiyonları, mikrodiziler, RNA-seq ve kantitatif PCR dahil olmak üzere tüm gen ekspresyon teknolojileri için geçerlidir (109).

3.9.3. "Giriş" Sekmesi

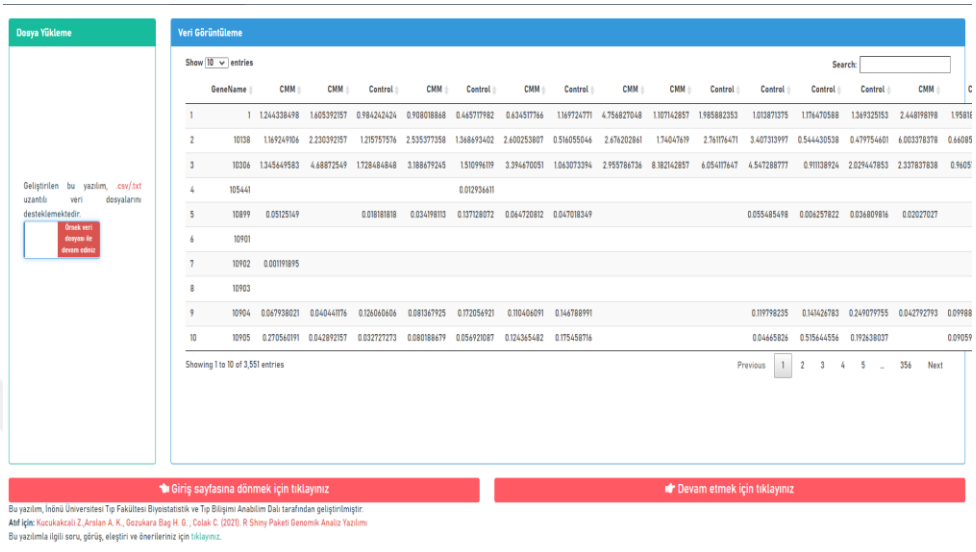
Bu sekmede yazılım ile ilgili genel bilgilerin yer aldığı ve geliştirilen yazılımın anlatıldığı bir bölüm ve yazılımda kullanılacak veri setini anlatan bilgilerin yer aldığı bir bilgilendirme bölümü yer almaktadır. Bu bölümünde omik teknolojisi ve genomik hakkında kısaca bilgi verilmiş ve yazılım hakkında kısa bilgiler sunulmuştur. Geliştirilen yazılım, ilk sütunda gen adlarını, diğer sütunlarda ise gruplardaki gen dizi değerlerini içeren .csv/.txt uzantılı veri setleri ile çalışmaktadır. Yazılımdaki sonuçlar, önem sırasına göre sıralanmış bir gen tablosundan ve diferansiyel olarak ifade edilen genleri görselleştirmek için bir grafikten oluşur. Sayfada yer alan "Devam etmek için tıklayın" sekmesi ile "Veri İşlemleri" menüsüne geçilir. Giriş bölümünün görünümü Şekil 3.2'de verilmiştir.

Şekil 3.2. "Giriş" Sekmesi

3.9.4. "Veri İşlemleri" Sekmesi

"Veri işlemleri" sekmesinde yazılıma uygun veri dosyası yüklenir. Bu bölümde giriş menüsünden veri işlemleri menüsüne geçiş yapan kullanıcılara kolaylık sağlamak

ve yazılım hakkında bilgi vermek amacıyla örnek bir veri seti ile kullanıcılar karşılanmaktadır. Böylece kullanıcılar, yazılımın hangi koşullar altında kullanılabileceği hakkında fikir sahibi olurlar. Bu aşamadan sonra kullanıcılar kendi veri dosyalarını yazılıma yükleyerek analize devam edebilirler. Veri işlemleri bölümünün görünümü Şekil 3.3'de verilmiştir.



Şekil 3.3. "Veri İşlemleri" Sekmesi

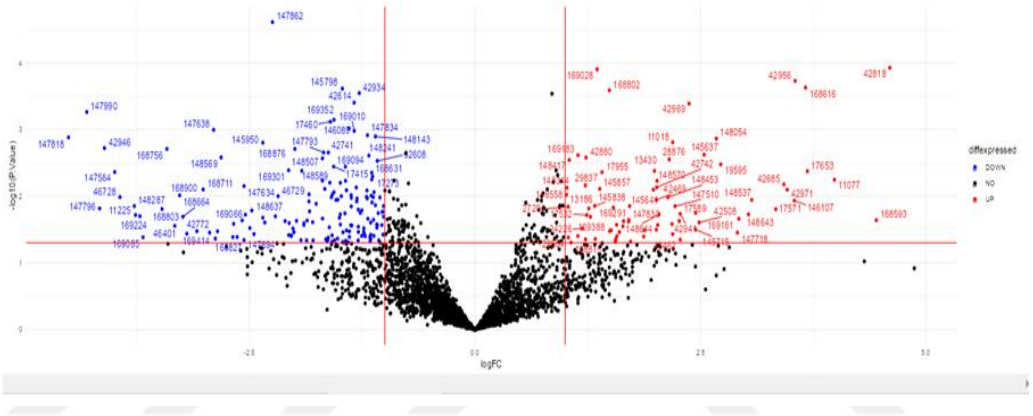
3.9.5. "Analiz" Sekmesi

"Analiz" sekmesinde, önem sırasına göre gruplarda farklı ifade edilen genleri görebileceğimiz bir tablo ve genleri görselleştirmeye yardımcı olan bir volkan grafiği bulunmaktadır. Sonuç tablosunda gen adları ve bunların LogFC değerleri, t istatistiksel değeri, p-değeri, düzeltilmiş p değerleri ve B istatistikleri bulunmaktadır. Bu değerlerden Log FC, iki deneysel koşul arasındaki log2 kat değişimi, B istatistiği ise genin farklı şekilde ifade edildiği log oranıdır. Tablodaki en küçük düzeltilmiş p-değerine sahip gen, en önemli gendir. Sonuçlara ait ekrana ilişkin görüntü şekil 3.4 ile verilmiştir.

	Gene	logFC	AveExpr	t	P.Value	adj.P.Val	B	diffexpressed
829	147862	-2.24507359065164	-4.66485091769294	-5.44766371525307	0.00002447964662735 09	0.0731451841225244	2.61395169557543	DOWN
2091	169028	1.35476083358266	1.60548765709344	4.64366429495029	0.00012432968622916 1	0.0961914699719285	1.20121207218414	UP
303	145798	-1.47048714923671	2.49309758248047	-4.369860296099	0.00024286628910830 6	0.0961914699719285	0.604062201485615	DOWN
1866	168802	1.49116912077059	-1.45058132158937	4.34464846450631	0.00025832379901809 5	0.0961914699719285	0.548984490326794	UP
3254	42934	-1.28299236827946	-2.63544795928595	-4.30576834532122	0.00028411217515966 2	0.0961914699719285	0.464034629123877	DOWN
728	147722	0.85415501468273	0.676260743249418	4.29776304640563	0.00028973334328894 1	0.0961914699719285	0.446542447661153	NO
3162	42818	4.59944736890558	-3.99738515527021	6.41618361153918	0.00011865743900911 3	0.0961914699719285	0.366909573066285	UP
1680	168616	3.66553518289413	-4.25831482303477	5.33324415634003	0.00023462150180329 1	0.0961914699719285	0.249297494710972	UP
2998	42614	-1.33989527193533	-2.1603605965879	-4.17134561792024	0.00039475999332129 3	0.110973953518119	0.170322292378993	DOWN
3270	42956	3.55073079850098	-7.51970519325288	6.041065421544 34	0.00018644862900923 1	0.0961914699719285	0.120963784845101	UP

Showing 1 to 10 of 3,551 entries Previous 1 2 3 4 5 … 356 Next

Not: Yukarıdaki tabloda, adj.Pvalue değeri en küçük olan gen en önemli gendir. B istatistiği genin farklı şekilde ifade edildiği log-orandır. Log FC iki deneysel koşul arasında log2 kat değişikliğidir.



Şekil 3.4. Sonuçlara ait görüntü

Geliştirilen interaktif web tabanlı yazılıma http://biostatapps.inonu.edu.tr/GAS/ adresinden ücretsiz olarak erişilebilir.

4. BULGULAR

4.1. Biyoinformatik Analiz Sonuçları

Çalışmada kullanılan genomik veri seti 16,386 ifade içermektedir. Biyoinformatik analiz sonuçlarına göre C (hapatotoksisite grubu) ve CK (kontrol grubu) grupları için 589 adet lncRNA gruplarda farklı ekspresyon göstermiştir. Bunlardan 450 tanesi yukarı ekspresyon (log2FC > 1) göstermişken 139 tanesi aşağı ekspresyon (log2FC < - 1) göstermiştir. Biyoinformatik analizine göre, minimum ayarlanmış-p değerleri ile ilgili ilk yirmi sonuç Tablo 4.1'de özetlenmiştir.

Tablo 4.1. Biyoinformatik Analiz Sonuçları

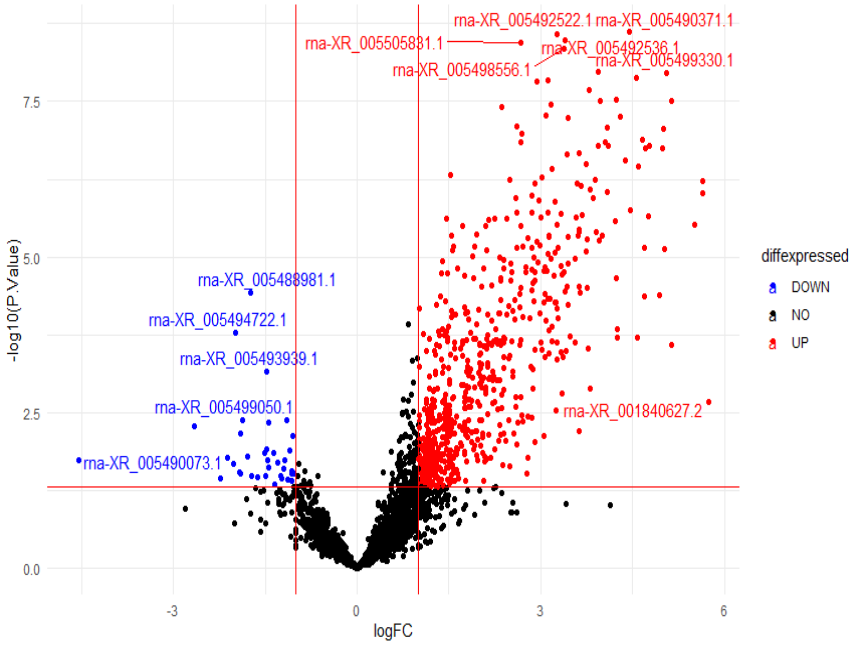
ID	Gen Adı	Log2FC	P değeri	Adj P Value	B	Diff-expressed
rna-XR_005490371.1	LOC120095150	4.45	2.38E-09	2.23981E-06	11.65227095	UP
rna-XR_005492522.1	LOC120096269	3.27	2.56E-09	2.23981E-06	11.58642019	UP
rna-XR_005492536.1	LOC120096276	3.40	3.34E-09	2.23981E-06	11.32455052	UP
rna-XR_005505831.1	LOC102557053	2.67	3.60E-09	2.23981E-06	11.25279171	UP
rna-XR_005498556.1	LOC120099381	3.38	4.57E-09	2.27068E-06	11.02479159	UP
rna-XR_005499033.1	LOC120099800	3.93	1.06E-08	3.72068E-06	10.20744289	UP
rna-XR_005499330.1	LOC120099889	5.06	1.11E-08	3.72068E-06	10.16394634	UP
rna-XR_005498350.1	LOC120099280	4.56	1.29E-08	3.72068E-06	10.02162517	UP
rna-XR_005497566.1	LOC120098880	3.12	1.44E-08	3.72068E-06	9.915694938	UP

rna-XR_005500842.1	LOC120100788	2.93	1.50E-08	3.72068E-06	9.87722245	UP
rna-XR_005503535.1	LOC120102261	3.79	2.11E-08	4.75824E-06	9.54588763	UP
rna-XR_005499333.1	LOC120099890	5.14	3.05E-08	5.44326E-06	9.21053834	UP
rna-XR_005502684.1	LOC102554065	4.24	3.01E-08	5.44326E-06	9.19910796	UP
rna-XR_005495694.1	LOC102556228	3.97	3.07E-08	5.44326E-06	9.18052193	UP
rna-XR_005498555.1	LOC120099380	3.16	3.58E-08	5.92918E-06	9.03006776	UP
rna-XR_005495063.1	LOC120097545	2.36	3.83E-08	5.9574E-06	8.96256015	UP
rna-XR_005497830.1	LOC120098996	3.09	5.22E-08	7.43714E-06	8.68939470	UP
rna-XR_005496283.1	LOC108348808	4.30	5.63E-08	7.43714E-06	8.61692319	UP
rna-XR_005501312.1	LOC120100109	3.45	5.68E-08	7.43714E-06	8.57857614	UP
rna-XR_005496522.1	LOC120098335	2.60	8.05E-08	9.79694E-06	8.23841920	UP

Tablo 4.1’e göre verilen lncRNA ifadelerinden hepsi Log2FC değerleri birden büyük olduğu için yukarı ekspresyon göstermiştir. Bu ifadelerden en çok ekspresyon gösterenlerden ilk 10 tanesi sırası ile 5.14, 5.06, 4.56, 4.45, 4.30, 4.24, 3.97, 3.93, 3.79 ve 3.45 Log2FC değerleri ile rna-XR_005499333.1 (LOC120099890), rna-XR_005499330.1 (LOC120099889), rna-XR_005498350.1 (LOC120099280), rna-XR_005490371.1 (LOC120095150), rna-XR_005496283.1 (LOC108348808), rna-XR_005502684.1 (LOC102554065), rna-XR_005495694.1 (LOC102556228), rna-

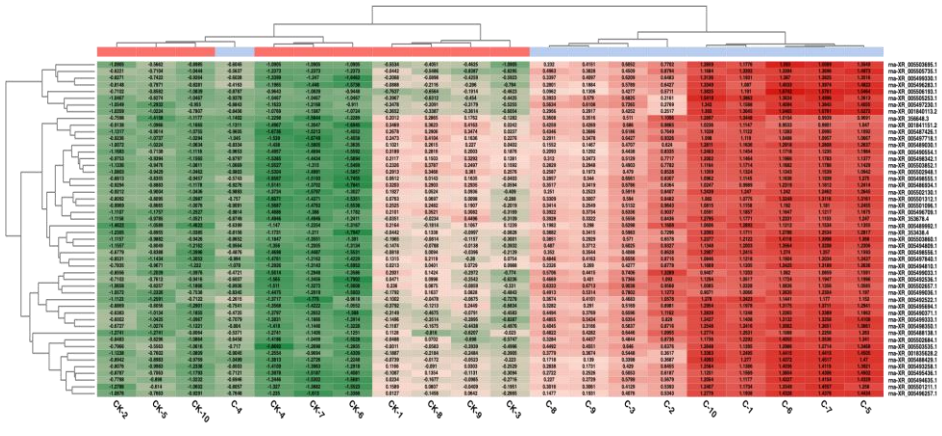
XR_005499033.1 (LOC120099800), rna-XR_005503535.1 (LOC120102261) ve rna-XR_005501312.1 (LOC108350109) id numaralı lncRNA lar olarak elde edilmiştir.

Diferansiyel olarak ifade edilen genleri görselleştirmek için kullanılan yanardağ grafiği Şekil 4.1'de verilmiştir. Volkan grafiği y ve x eksenlerinde diferansiyel olarak ifade edilen genleri hızlı bir şekilde gözlemlemek için log2 bazındaki kat değişimine karşı önemi çizer.



Şekil 4.1. Farklı ekspresyon gösteren lncRNA' lar için volkan grafiği

İfade seviyeleri kıyaslamasında en fazla değişiklik gösteren 50 lncRNA ifadesi için heatmap gösterimi aşağıdaki gibi şekil 4.2'de gösterilmiştir.



Şekil 4.2. En Fazla Değişiklik Gösteren 50 lncRNA İfadesi İçin Heatmap

C ve CK kıyaslamasında en fazla değişiklik gösteren 50 lncRNA için fazla ekspres olanlar kırmızı ve baskılanan ifade seviyesi için ise yeşil renkli gösterimde uygulama numunelerinin kontrole gore açık bir şekilde farklı ifade profili sergilediği görülmektedir. C-4 numunesi uygulama numunelerinden farklı bir profil göstermiştir.

4.2. Biyoistatistik Analiz ve Modelleme Sonuçları

Veri setinde mevcut olan 16.386 lncRNA transkriptinden farklı regülasyonlar gösteren (aşağı-yukarı) 589 lncRNA dan ElastikNet değişken seçimi yöntemi ile 17 adet lncRNA seçilmiştir. Seçilen bu ifadeler ile yapılan istatistik analiz sonuçları tablo 4.2'de verilmiştir. Tablo 4.2 seçilen ifadeler ile veri setinin açıklamaları, incelenen hedef değişken için seçilen ifadelere ait tanımlayıcıları, istatistik anlamlılığı ve hedef değişken için gen başına olasılık oranını (OR) içermektedir.

Tablo 4.2. İstatistik Analiz Sonuçları

Gen Adı	ID	Grup				OR	*p*
		C		CK			
		Ortalama± SD	Medyan (Min-Maks)	Ortalama ±SD	Medyan (Min-Maks)		
LOC120094596	**rna_XR_005488981.1**	12.9±9.291	11.5(0-32)	43.6±15.665	39(20-74)	0.80	<0.001*
LOC120097437	**rna_XR_005494824.1**	3.8±1.033	4(2-6)	0.6±0.843	0(0-2)	161095026.78	<0.001**
LOC120096352	**rna_XR_005492760.1**	12.5±5.126	13.5(4-18)	1.8±2.044	1(0-5)	4.22	<0.001**
LOC120102815	**rna_XR_005504579.1**	8±3.712	8(2-14)	1.5±1.269	1.5(0-3)	4.33	<0.001*
LOC120101756	**rna_XR_005502657.1**	402.7±151.646	424(43-548)	31.8±31.091	18.5(7-108)	1.02	<0.001**
LOC120099881	**rna_XR_005499304.1**	49.8±18.66	55.5(7-75)	4.7±4.572	4(0-14)	1.21	<0.001*
LOC102557053	**rna_XR_005505831.1**	351.9±115.959	364.5(107-484)	54.3±19.402	50(28-89)	4.88	<0.001*
LOC103693406	**rna_XR_005490393.1**	57.8±17.536	63(16-77)	8.9±7.534	7(1-23)	0.92	<0.001**
LOC120098296	**rna_XR_005496472.1**	13.4±5.358	13(7-23)	3.9±4.332	3(0-14)	0.02	0.001**
LOC120094640	**rna_XR_005489054.1**	28.6±8.708	29(9-38)	5±5.85	2.5(0-17)	1.33	<0.001**
LOC102552566	**rna_XR_001836079.2**	9.6±3.565	10(3-16)	1.2±0.919	1(0-3)	11233861.03	0.001*
LOC120096990	**rna_XR_005493970.1**	20.6±8.669	21(3-31)	2.1±1.37	2.5(0-4)	2.31	<0.001*
LOC120096276	**rna_XR_005492536.1**	2101.1±720.824	2139(586-2927)	213.2±123.706	194.5(64-482)	1.31	<0.001*
LOC120099800	**rna_XR_005499033.1**	168.9±67.258	162.5(49-310)	14.5±15.58	8(4-54)	1.13	<0.001**
LOC102553540	**rna_XR_360532.3**	17.7±7.273	19(5-30)	1.4±1.713	1(0-5)	2521.45	<0.001**
LOC120096269	**rna_XR_005492522.1**	50.7±11.47	55(30-64)	5.4±3.627	5(0-11)	6.31	<0.001*

LOC120094133	**rna_XR_005488138.1**	56.7±21.386	60.5(11-81)	2.7±3.889	1(0-12)	1.72	<0.001**

*: Bağımsız örneklerde t testi; **: Mann Whitney U testi; OR. Odds ratio; SD: Standard sapma

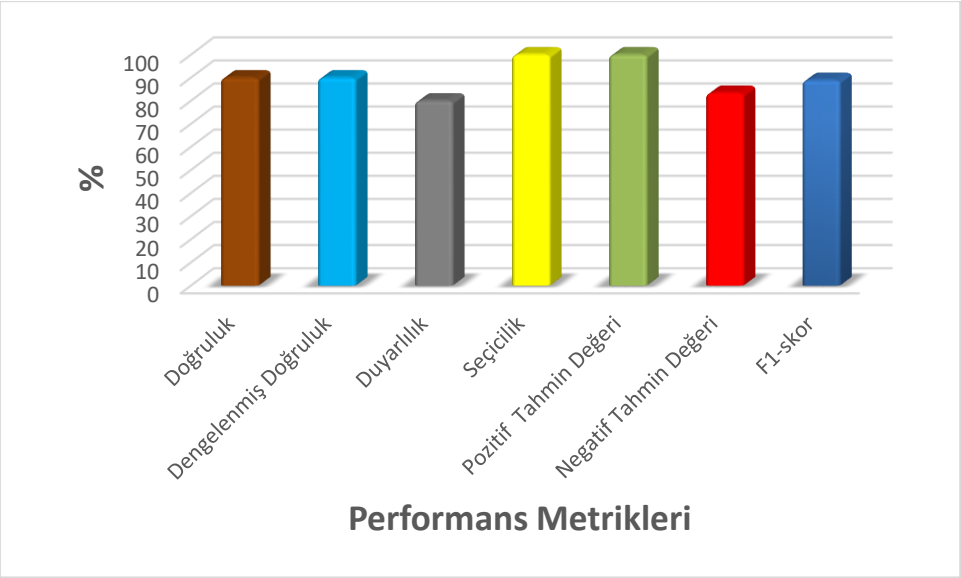
Tablo 4.2'deki istatistik analiz sonuçlarına göre tüm lncRNA ifadelerinde gruplar arasında istatistiksel olarak önemli farklılıklar tespit edildi ($p<0.05$). Tüm ifadeler için yapılan lojistik regresyon modelleri anlamlı olup OR oranları belirlenmiştir.

XGboost modelinden elde edilen performans metriklerinin bulguları Tablo 4.3'de verilmiştir.

Tablo 4.3. XGboost modelinden elde edilen performans metrikleri

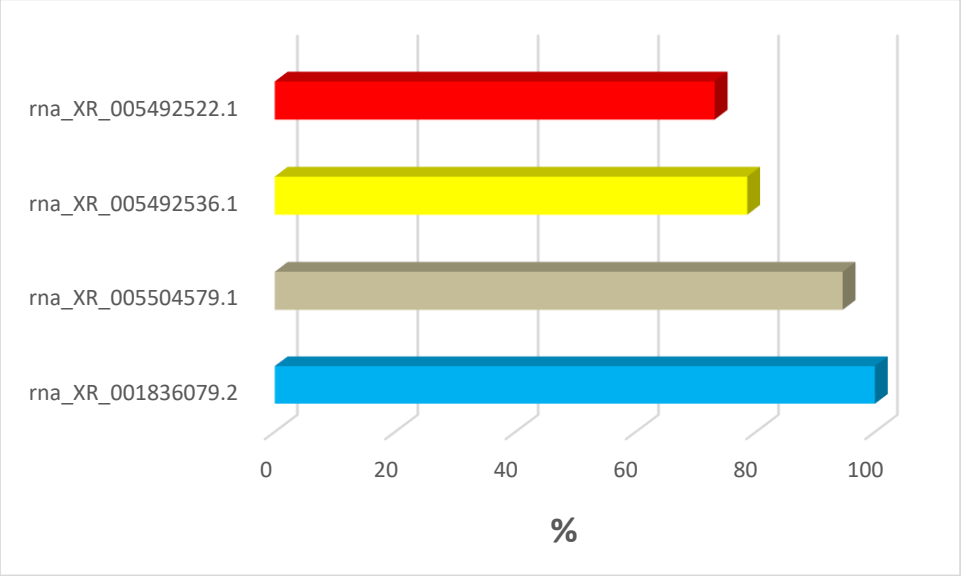
Metrik	Değer (%) (95% CI)
Doğruluk	95 (85.4-100)
Dengelenmiş Doğruluk	95 (85.4-100)
Duyarlılık	100 (69.15-100)
Seçicilik	90 (55.5-99.75)
Pozitif Tahmin Değeri	90.91 (60.9-98.47)
Negatif Tahmin Değeri	100 (87.5-100)
F1-skor	98.6 (0.954-1)

XGboost modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skoru sırasıyla %95, %95, %100, %90, %90.91, %100 ve %98.6 olarak elde edilmiştir. Şekil 4.3'de XGboost modeli için performans metriklerine ait grafik verilmektedir.



Şekil 4.3. XGboost modeli için performans metriklerine ait grafik

Şekil 4.4, çıktı değişkenini açıklamada seçilen genler için ifadelerin değişken önemlilik düzeylerini göstermektedir.



Şekil 4.4. XGBoost modeline ilişkin Değişken önemliliği grafiği

rna_XR_001836079.2 (LOC102552566) id li lncRNA % 100 ile en yüksek değişken önemliliğine sahip olup rna_XR_005504579.1 (LOC120102815) id li lncRNA % 94.572, rna_XR_005492536.1 (LOC120096276) id li lncRNA % 78.525 ve rna_XR_005492522.1 (LOC120096269) id li lncRNA % 73.1 ile sırasıyla en yüksek değişken önemliliklerine sahiptir.

Topluluk öğrenme modellerinden Bagging, Boosting ve Stacking modelleri sonucu elde edilen performans metriklerinin bulguları Tablo 4.4'de verilmiştir.

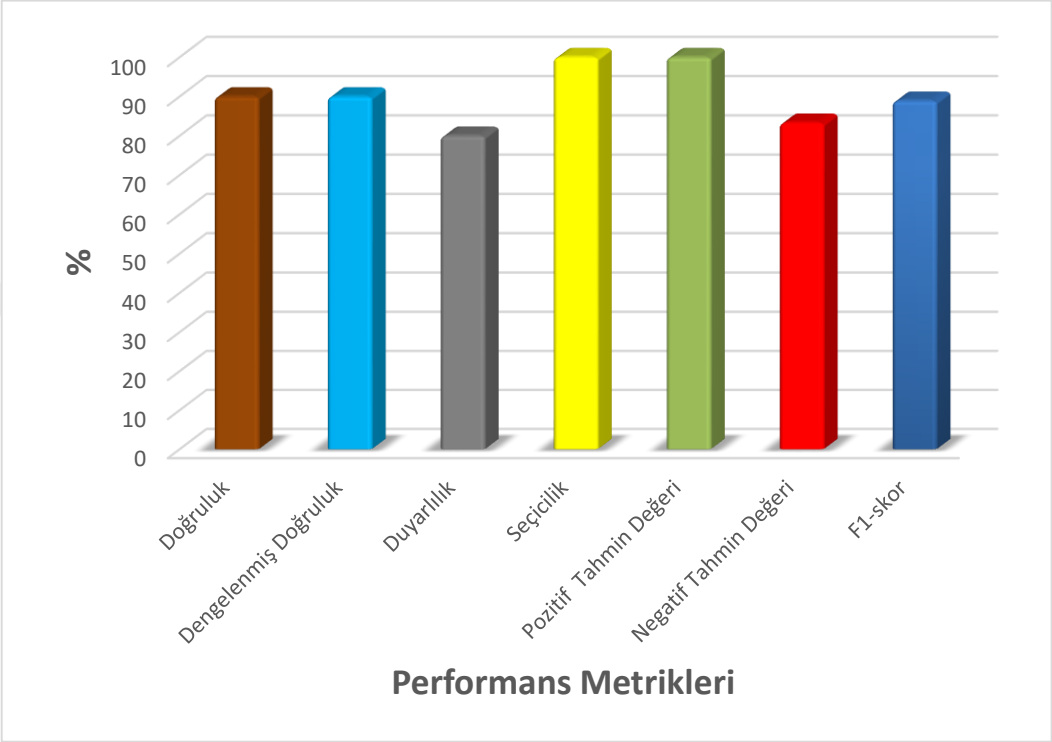
Tablo 4.4. Bagging, Boosting ve Stacking modellerinin sınıflandırma performansına ait metriklere ilişkin değerler

Modeller	Metric	**Value (%) (95% CI)**
Bagging	Doğruluk	85 (69.4-100)
	Dengelenmiş Doğruluk	85 (69.4-100)
	Duyarlılık	70 (34.75-93.33)
	Seçicilik	100 (69.15-100)
	Pozitif Tahmnin Değeri	100 (59-100)
	Negatif Tahmin Değeri	76.92 (56.39-89.57)
	F1-skor	82.4 (65.6-99.1)
Boosting	Doğruluk	85 (69.4-100)
	Dengelenmiş Doğruluk	85 (69.4-100)
	Duyarlılık	70 (34.75-93.33)
	Seçicilik	100 (69.15-100)
	Pozitif Tahmin Değeri	100 (59-100)
	Negatif Tahmin Değeri	76.92 (56.39-89.57)
	F1-skor	82.4 (65.6-99.1)
Stacking	Doğruluk	90 (68.3-98.77)
	Dengelenmiş Doğruluk	90 (68.3-98.77)
	Duyarlılık	80 (44.39-97.48)
	Seçicilik	100 (69.15-100)
	Pozitif Tahmin Değeri	100 (87.5-100)
	Negatif Tahmin Değeri	83.3 (59.14-94.53)
	F1-skor	88.9 (75.1-100)

Bagging, Boosting ve Stacking modellerinin sınıflandırma performansına göre Bagging modelinden doğruluk %85, dengelenmiş doğruluk %85, duyarlılık %70, seçicilik %100, pozitif tahmin değer %100, negatif tahmin değer %76.92 ve F1-skor %82.4 olarak elde edilmiştir. Boosting modelinden doğruluk %85, dengelenmiş doğruluk %85, duyarlılık %70, seçicilik %100, pozitif tahmin değer %100, negatif tahmin değer

%76.92 ve F1-skor 82.4 olarak elde edilmiştir. Stacking modelinden doğruluk %90, dengelenmiş doğruluk %90, duyarlılık %80, seçicilik %100, pozitif tahmin değer %100, negatif tahmin değer %83.3 ve F1-skor %88.9 olarak olarak hesaplanmıştır.

Kullanılan ensemble modellerden en iyi sonucu veren Stacking modeli için performans metriklerine ait grafik şekil 4.5'de verilmektedir.



Şekil 4.5. Stacking Modeli için performans metriklerine ait grafik

5. TARTIŞMA

Oluşumu ilaçlara dayanan karaciğer toksisiteleri, karaciğer hasarı için önde gelen nedenlerden biri olarak bilinmektedir (6). İlaca bağlı olarak ortaya çıkan hepatotoksisiteler, non-spesifik değişiklikler olan akut karaciğer yetmezliği, siroz ve karaciğer kanseri gibi çeşitli klinik tablolara yol açmaktadır (119). Karaciğer, birçok ilacın ve kimyasal olarak bilinen ajanın metabolizmasında yer alan işlevi nedeniyle ilaç toksisitelerine maruz kalıp bizzat etkilenen organdır. Tedavi amacıyla kullanımı gerçekleştirilen ilaçların yanında alkol, zehirli türde olabilecek mantar türleri, endüstriyel olarak kullanımı sağlanan kimyasal ilaçlar ve özellikle son yıllarda kullanılanımı oldukça artan ve alternatif tıp adı altında kullanımları kabul gören bitkilerin bazılarının da karaciğer toksisitesine neden oluşturabileceği belirtilmiştir (120, 121). Kullanılan ilaç dozu miktarı sebebiyle gelişen hepatoksisite özellikle Amerika Birleşik Devlatleri (ABD), İngiltere ve bazi batı ülkelerinde ortaya çıkan tüm akut karaciğer yetmezliği durumunun hemen hemen % 50'sine yakınından sorumlu tutulmaktadır (121). İlaçla oluşan hepatotoksisite olgularının % 39'unun parasetamol kullanımına, % 13'ünün ise diğer ilaçların kullanımına bağlı olarak geliştiği bildirilmiştir. Yapılan çalışmaların sonuçlarına göre ülkemizde yıl bazlı 150.000 e yakın toksisite vakasının görüldüğü düşünülmektedir. Ortaya çıkan karaciğer toksisitelerinin etki mekanizmalarının açıklanması ve buna bağlı olarak tedavi yöntemlerinin geliştirilmesi hastaların mortalite riski açısından oldukça önemlidir (122). Çalışmada kullanılan veri seti elde edilirken cisplatin ile oluşturulmuş hepatotoksisitesi olan sıçanlar ile kontrol grubundan oluşan sıçanlardan karaciğer örnekleri alınmış ve genomik analizler sonucu ifade verisi elde edilmiştir.

Araştırma ortamlarında insan genetik çeşitliliğini güvenilir ve ucuz bir şekilde ölçme ve hastalık gen ilişkilerini ortaya çıkarma yeteneği, sağlık hizmetlerinde kişiselleştirilmiş tıbba doğru hareketi ateşledi ve şekillendirdi. Kişiselleştirilmiş tıbbın birçok tanımı olmasına rağmen, çoğu, herhangi bir hastanın sağlığının en iyi şekilde, önleyici tedbirlerin ve tedavinin kişisel tercihlere ve ayrıca hastanın genomik dahil olmak üzere hastanın özel çevresel ve biyolojik özelliklerine göre uyarlanmasıyla yönetildiği şeklindeki temel fikri paylaşmaktadır.

Genomik bilgilerin getirdiği bu yeni tıp tanımları ve kavramları ile genomik verileri işleyen programlar ve yazılımlar çok önemli hale geldi. Bu nedenle genomik bilgilere kolay erişim sağlayan birçok yazılım, program ve veri tabanı sistemi geliştirilmiş ve araştırmacıların hizmetine sunulmuştur.

Çalışmada kullanılan veri setinin biyoinformatik analizlerinin gerçekleştirilmesi için bu tez kapsamında biyoinformatik analiz yapılmasına olanak sağlayan bir web arayüzü geliştirilmiştir. Bilindiği gibi NCBI Gene Expression Omnibus (GEO), mikrodizi verilerinin halka açık bir deposudur ve çoğu araştırmacı buradan elde edebilecekleri verilerle çalışmalarına yön vermektedir. GEO Serisinden alınan verileri kullanarak karşılaştırmalar yapmak için geliştirilmiş etkileşimli bir yazılım olan GEO2R, diferansiyel olarak ifade edilen genleri tanımlamak için geliştirilmiştir. Bu yazılımda bulunan veriler, NCBI GEO veri tabanındaki GEO serisinden verilerdir. Ancak geliştirilen yazılımda kullanılacak verilerin GEO serisinden olması gerekmez, herhangi bir analiz şirketinden alınan çıktı formatı olan .txt uzantılı olması gerekir. Bu, araştırmacıların kendi verilerini işlemesini kolaylaştırır. Geliştirilen bu web arayüzünde 2 gruplu verilerde diferansiyel olarak ifade edilen genleri tanımlamak mümkün olup ayrıca elde edilen sonuçlar görsel olarak da kullanıcıya sunulmaktadır. Geliştirilen yazılım sonuçları bir tablo olarak vermekte ve görselleştirmek için volkan grafiğini kullanmaktadır. Tabloda farklı ifade edilen genomik yapılar ilk sıralarda verilmiş ve diğerleri tablo boyunca listelenmiştir. Bu da kullanıcılar için yorumlama kolaylığı sağlar. R programlama dili ile geliştirilen yazılım kullanıcı dostu olup bir web ara yüzüne sahiptir. Geliştirilen yazılım kullanılarak biyoinformatik analizler gerçekleştirilmiştir.

Bu çalışmada hepatotoksitesi olan sıçanlar ile kontrol grubu sıçanların karaciğer dokularından elde edilen genomik veriler kullanılmıştır. Elde edilen genomik veri setinde 16.386 adet lncRNA ya ait ifade bulunmaktadır. Yapılan biyoinformatik analizlerin bulgularına göre (Tablo 3'de ayrıntılı olarak verilen) iki grup arasındaki ekspresyon kat değişikliklerini belirlemek için kullanılan Log2FC değerlerine göre, rna-XR_005499333.1 (LOC120099890) id li lncRNA, hepatotoksisitesi olan grupta kontrol grubuna göre 35.26 kat daha yüksek gen ekspresyonuna sahiptir. Benzer şekilde Log2FC değerleri en yüksek olan rna-XR_005499330.1 (LOC120099889), rna-XR_005498350.1 (LOC120099280), rna-XR_005490371.1 (LOC120095150), rna-XR_005496283.1 (LOC108348808), rna-XR_005502684.1 (LOC102554065), rna-XR_005495694.1 (LOC102556228), rna-XR_005499033.1 (LOC120099800), rna-XR_005503535.1

(LOC120102261) ve rna-XR_005501312.1 (LOC108350109) id numaralı lncRNA lar sırası ile 33.35 kat , 23.58 kat, 21.85 kat, 19.69 kat, 18.89 kat, 15.67 kat, 15.24 kat, 13.83 kat ve 10.92 kat daha yüksek gen ekspresyonuna sahiptir.

Çalışmada kullanılan ElastikNet değişken seçimi yöntemi ile seçilen 17 lncRNA ile hedef değişkeni alınarak yapılan topluluk öğrenme modellerinden Bagging, Boosting ve Stacking modelleri ile yapılan modellemeler sonucunda elde edilen performans ölçüleri şu şekildedir: Bagging modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skoru sırasıyla %85, %85, %70, %100, %100, %76.92, %82.4; Boosting modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skoru sırasıyla %85, %85, %70, %100, % 100, %76.92 ve %82.4 olarak elde edilmiştir. Stacking modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skoru sırasıyla %90, % 90, %80, %100, % 100, % 83.3 ve % 88.9 olarak olarak hesaplanmıştır. Değişken seçimi yöntemi ile seçilen lncRNA lar ile yapılan XGboost modellemesi sonucunda ise elde edilen performans ölçüleri şu şekildedir: doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skoru sırasıyla %95, %95, %100, %90, %90.91, %100 ve %98.6 dir. Performans metrikleri sonuçları, önerilen modellerden XGboost yönteminin iki grubu elde edilen yüksek performans metrikleri değerleri göz ününde bulundurularak doğru bir şekilde sınıflandırabileceğini göstermektedir. XGboost modellemesi sonucunda elde edilen değişken önemliliklerine göre, rna_XR_001836079.2 (LOC102552566) id li lncRNA, rna_XR_005504579.1 (LOC120102815) id li lncRNA, rna_XR_005492536.1 (LOC120096276) id li lncRNA ve rna_XR_005492522.1 (LOC120096269)id li lncRNA en yüksek değişken önemliliklerine sahip olup hepatotoksisite için öngörücü biyobelirteç adayları olarak kullanılabilir.

Yapılan istatistik analiz sonuçlarına göre, değişken seçimi ile elde edilen 17 gen, iki grup için istatistiksel olarak anlamlı farklılıklar gösterdi.

Odds değerleri hesaplanan lncRNA lardan, rna_XR_005494824.1 ve rna_XR_001836079.2 aşırı derecede büyük odds oranları ile hepatotoksisitesi olan grupta önemli ölçüde daha yüksek katlarda ekspresyon göstermiştir. rna_XR_005488981.1 id li lncRNA hariç seçilen tüm lncRNA ların OR oranları 1 den büyük olup bu lncRNA ların

hasta grubunda yüksek katlarda yukari ekspresyonu gözlenmiştir. rna_XR_005488981.1 id li lncRNA ise aşağı regülasyon göstermektedir. Çalışmada hesaplanan OR değerleri ile Log2FC değerleri birbirini desteklemektedir. Ayrıca değişken önemliliğine göre elde edilen lncRNA larda yüksek OR oranlarını desteklemektedir.

6. SONUÇ VE ÖNERİLER

- Sonuç olarak, bu çalışma ile biyoinformatik analiz yapılmasına olanak sağlayan bir yazılım geliştirilmiştir.
- Bu çalışma, hepatotoksisitesi olan sıçanlar ile kontrol grubunda yer alan sıçanların lncRNA ekspresyon verilerini kullanarak hepatoksisite için potansiyel genomik biyobelirteçleri tanımladı.
- Gelecekteki daha kapsamlı analizlerle keşfedilen genlerin güvenilirliği değerlendirilebilir, bu genlere dayalı olarak terapi teknikleri tasarlanabilir ve klinik kullanımları detaylandırılabilir.

KAYNAKLAR

1. Eren OÖ. Cancer screening and prevention. *Klinik Tıp Aile Hekimliği Dergisi* 2017, 9: 7-14.
2. Baykara O. Current modalities in treatment of cancer. *Balıkesir Sağlık Bil Derg* 2016, 5: 154-65.
3. Gansler T, Ganz PA, Grant M, Greene FL, Johnstone P, Mahoney M, Newman LA, Oh WK, Thomas CR Jr, Thun MJ, Vickers AJ, Wender RC, Brawley OW. Sixty years of CA: a cancer journal for clinicians. *CA Cancer J Clin* 2010, 60: 345-50.
4. Dasari S, Tchounwou PB. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur J Pharmacol* 2014, 740: 364-78.
5. Aksoy A, Karaoglu A, Akpolat N, Naziroglu M, Ozturk T, Karagoz ZK. Protective role of selenium and high dose vitamin e against cisplatin - induced nephrotoxicty in rats. *Asian Pac J Cancer Prev* 2015, 16: 6877-82.
6. Bjornsson HK, Bjornsson ES. Drug-induced liver injury: Pathogenesis, epidemiology, clinical features, and practical management. *Eur J Intern Med* 2022, 97: 26-31.
7. Kaplowitz N. Drug-induced liver disorders: implications for drug development and regulation. *Drug Saf* 2001, 24: 483-90.
8. Halliwell B, Chirico S. Lipid peroxidation: its mechanism, measurement, and significance. *Am J Clin Nutr* 1993, 57: 715-4.
9. Agırbaslı D, Ulman YI. Coronary artery disease from a perspective of genomic risk score, ethical approaches and suggestions. *Anadolu Kardiyol Derg* 2012, 12: 171-7.
10. Atalay RC. Neden Biyoinformatik. *Avrasya Dosyası, Moleküler Biyoloji ve Gen Teknolojileri Özel,* 2002, 8: 129-41.
11. Bai H, Wu T, Jiao L, Wu Q, Zhao Z, Song J, Liu T, Lv Y, Lu X, Ying B. Association of ABCC Gene Polymorphism With Susceptibility to Antituberculosis Drug-Induced Hepatotoxicity in Western Han Patients With Tuberculosis. *J Clin Pharmacol* 2020, 60: 361-8.
12. Aubrecht J, Schomaker SJ, Amacher DE. Emerging hepatotoxicity biomarkers and their potential to improve understanding and management of drug-induced liver injury. *Genome Med* 2013, 5: 85.

13. Polikar R. Ensemble learning. In: Zhang C, Ma Y (eds). *Ensemble machine learning*, 1st ed. Newyork, Springer, 2012: 1-34.
14. Akman M, GenC Y, Ankarali H. Random Forests Methods and an Application in Health Science. *Turkiye Klinikleri J Biostat* 2011, 3: 36-48.
15. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record* 2002, 31: 76-7.
16. Rokach L, Maimon O. Clustering methods. In: Rokach L, Maimon O (eds). *Data mining and knowledge discovery handbook*. 2nd ed. Newyork, Springer, 2005: 321-52.
17. Younis EM, Zaki SM, Kanjo E, Houssein EH. Evaluating Ensemble Learning Methods for Multi-Modal Emotion Recognition Using Sensor Data Fusion. *Sensors* 2022, 22(15): 5611.
18. Abdel Misih SR, Bloomston M. Liver anatomy. *Surg Clin North Am* 2010, 90: 643-53.
19. Srinivasan S, Friedman LS. *Sitaraman and Friedman's Essentials of Gastroenterology*, 2nd ed. USA, Wiley-Blackwell Publishing, 2018: 165-95.
20. Junqueira LCU, Carneiro J. *Junqueira's Basic Histology: Text and Atlas*, 16th ed. New York, McGraw Hill Medical Book, 2021: 45-68.
21. Jakab SS. Clinical hepatology : principles and practice of hepatobiliary diseases. *J Clin Gastroenterol* 2010, 44: 662.
22. Macpherson AJ, Heikenwalder M, Ganal Vonarburg SC. The Liver at the Nexus of Host-Microbial Interactions. *Cell Host Microbe* 2016, 20: 561-71.
23. Zimmerman HJ. Drug-induced liver disease. *Clin Liver Dis* 2000, 4: 73-96.
24. Jaeschke H, Gores GJ, Cederbaum AI, Hinson JA, Pessayre D, Lemasters JJ. Mechanisms of hepatotoxicity. *Toxicol Sci* 2002, 65: 166-76.
25. Kintzel PE. Anticancer drug-induced kidney disorders. *Drug Saf* 2001, 24: 19-38.
26. Atessahin A, Karahan I, Turk G, Gur S, Yilmaz S, Ceribasi AO. Protective role of lycopene on cisplatin-induced changes in sperm characteristics, testicular damage and oxidative stress in rats. *Reprod Toxicol* 2006, 21: 42-7.
27. Naziroglu M, Karaoglu A, Aksoy AO. Selenium and high dose vitamin E administration protects cisplatin-induced oxidative damage to renal, liver and lens tissues in rats. *Toxicology* 2004, 195: 221-30.
28. Budak SÖ, Dönmez S. Novel omics technologies in food science. *Journal of Food* 2012, 37: 173-9.

29. Horgan RP, Kenny LC. 'Omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet Gynecol* 2011, 13: 189-95.
30. Ordovas JM, Corella D. Nutritional genomics. *Annu Rev Genomics Hum Genet* 2004, 5: 71-118.
31. Başaran E, Sümer A, Cansaran Duman D. General outlook and applications of genomics, proteomics and metabolomics. *Turk Hij Den Biyol Derg* 2010, 67: 85-96.
32. Lotze MT, Thomson AW. *Measuring immunity: basic biology and clinical assessment*, 1st ed. Netherlands, Elsevier Academic Press, 2005: 187-92.
33. Marco ML, Wells-Bennik MHJ. Impact of bacterial genomics on determining quality and safety in the dairy production chain. *Int Dairy J* 2008, 18: 486-95.
34. Martin DB, Nelson PS. From genomics to proteomics: techniques and applications in cancer research. *Trends Cell Biol* 2001, 11: 60-5.
35. Del Boccio P, Urbani A. Homo sapiens proteomics: clinical perspectives. *Ann Ist Super Sanita* 2005, 41: 479-82.
36. Bal SH, Budak F General Overwiev to the Concepts of Genomics, Proteomics and Application Areas. *Uludağ Üniversitesi Tıp Fakültesi Dergisi* 2013, 39: 65-9.
37. Weining S, Langridge P. Identification and mapping of polymorphisms in cereals based on the polymerase chain reaction. *Theor Appl Genet* 1991, 82: 209-16.
38. Yildirim A, Kandemir N, Sonmezoğlu OA, Gülec T. Polymerase chain reaction in wheat and optimization of the possible problems. *Anadolu Tarım Bilim* Derg 2011, 26: 36-9.
39. Tekin K, Aygar İs, Hoşbul T. Polimeraz Zincir Reaksiyonu Teknolojisinin Temel Prensipleri. *J Mol Virol Immunol* 2020, 1(1): 57-66.
40. Muyzer G, de Waal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* 1993, 59: 695-700.
41. Muyzer G, Smalla K. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek* 1998, 73: 127-41.
42. Mocali S, Benedetti A. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Res Microbiol* 2010, 161: 497-505.

43. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 1975, 94: 441-8.
44. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 1977, 74: 560-4.
45. Mecler I, Nawrot U. [Molecular techniques used in microbiological diagnostics]. *Mikol Lek* 2007, 14: 280-4.
46. Raszka A, Ziembinska A, Wiechetek A. Molecular techniques used in microbiological diagnostics. *Srodowisko* 2009, 2: 101-14.
47. Ahmad I, Ahmad F, Pichtel J. *Microbes and microbial technology: agricultural and environmental applications,* 1st ed Newyork, Springer, 2011.
48. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010, 11: 31-46.
49. Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I. Characteristics of 454 pyrosequencing data-enabling realistic simulation with flowsim. *Bioinform* 2010, 26: 420-5.
50. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. *Genome Res* 2001, 11: 3-11.
51. Mewes HW, Albermann K, Bähr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A. Overview of the yeast genome. *Nature* 1997, 387: 7-65.
52. Erkan Y, Deveci O. Nanomedicine, microarrays and their applications in clinical microbiology. *Dicle Tıp Dergisi* 2010, 37: 422-8.
53. Pertea M. The human transcriptome: an unfinished story. *Genes (Basel)* 2012, 3: 344- 60.
54. Lee KJ, Yin W, Arafat D, Tang Y, Uppal K, Tran V, Cabrera-Mora M, Lapp S, Moreno A, Meyer E, DeBarry JD, Pakala S, Nayak V, Kissinger JC, Jones DP, Galinski M, Styczynski MP, Gibson G. Comparative transcriptomics and metabolomics in a rhesus macaque drug administration study. *Front Cell Dev Biol 2*014, 2: 54.
55. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, 5: 621-8.
56. Güzelgül F, Aksoy K. Kodlanmayan RNA'ların işlevi ve tıpda kullanım alanları. *Arşiv Kaynak Tarama Dergisi* 2009, 18: 141-55.

57. Uchida S, Dimmeler S. Long noncoding RNAs in cardiovascular diseases. *Cir Res* 2015, 116: 737-50.
58. Li J, Liu C. Coding or Noncoding, the Converging Concepts of RNAs. *Front Genet* 2019, 10: 496.
59. Zhang X, Wang W, Zhu W, Dong J, Cheng Y, Yin Z, Shen F. Mechanisms and functions of long non-coding rnas at multiple regulatory levels. *Int J Mol Sci* 2019, 20: 5573.
60. Dahariya S, Paddibhatla I, Kumar S, Raghuwanshi S, Pallepati A, Gutti RK. Long non-coding RNA: classification, biogenesis and functions in blood cells. *Mol Immunol* 2019, 112: 82-92.
61. Jarroux J, Morillon A, Pinskaya M. History, discovery, and classification of lncRNAs. *Adv Exp Med Biol* 2017, 1008: 1-46.
62. Mehler MF. Epigenetic principles and mechanisms underlying nervous system functions in health and disease. *Prog Neurobiol* 2008, 86: 305-41.
63. Amin NR, McGrath A, Chen YPP. Evaluation of deep learning in non-coding RNA classification. *Nat Mach Intell* 2019, 1: 246-56.
64. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim Biophys Acta* 2014, 1842: 1910-22.
65. Dhanoa JK, Sethi RS, Verma R, Arora JS, Mukhopadhyay CS. Long non-coding RNA: its evolutionary relics and biological implications in mammals: a review. *J Anim Sci Technol* 2018, 60: 25.
66. Zhang X, Hong R, Chen W, Xu M, Wang L. The role of long noncoding RNA in major human disease. *Bioorg Chem* 2019, 92: 103214.
67. Jiang MC, Ni JJ, Cui WY, Wang BY, Zhuo W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res* 2019, 9: 1354-66.
68. Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. *Cancer Res* 2017, 77: 3965-81.
69. Li J, Xuan Z, Liu C. Long non-coding RNAs and complex human diseases. *Int J Mol Sci* 2013, 14: 18790-808.
70. Yang L, Peng X, Li Y, Zhang X, Ma Y, Wu C, Fan Q, Wei S, Li H, Liu J. Long non-coding RNA HOTAIR promotes exosome secretion by regulating RAB35 and SNAP23 in hepatocellular carcinoma. *Mol Cancer* 2019, 18(1): 78.

71. Merrick BA, Chang JS, Phadke DP, Bostrom MA, Shah RR, Wang X, Gordon O, Wright GM. HAfTs are novel lncRNA transcripts from aflatoxin exposure. *PloS One* 2018, 13: 0190992.
72. Huang Y, Xiang B, Liu Y, Wang Y, Kan H. LncRNA CDKN2B-AS1 promotes tumor growth and metastasis of human hepatocellular carcinoma by targeting let-7c-5p/NAP1L1 axis. *Cancer Lett* 2018, 437: 56-66.
73. Wu JC, Luo SZ, Liu T, Lu LG, Xu MY. linc-SCRG1 accelerates liver fibrosis by decreasing RNA-binding protein tristetraprolin. *FASEB J* 2019, 33: 2105-15.
74. Akpınar H. Veri tabanlarında bilgi keşfi ve veri madenciliği. *Istanbul Business Research* 2000, 29: 1-22.
75. Hand DJ. Principles of data mining. *Drug Saf* 2007, 30: 621-2.
76. Atalay M, Çelik E. Artificial intelligence and machine learning applications in big data analysis. *MAKU SOBED* 2017, 9: 155-72.
77. Kucukakcali Z, Colak C, Gozukara Bag HG, Balikci Cicek I, Ozhan O, Yildiz A, Danis N, Koc A, Parlakpinar H, Akbulut S. Modeling Based on Ensemble Learning Methods for Detection of Diagnostic Biomarkers from LncRNA Data in Rats Treated with Cis-Platinum-Induced Hepatotoxicity. Diagnostics. 2023, 13:1583.
78. Zia A, Aziz M, Popa I, Khan SA, Hamedani AF, Asif AR. Artificial Intelligence-Based Medical Data Mining. *J Pers Med* 2022, 12(9): 1359.
79. Brown ML, Kros JF. Data mining and the impact of missing data. *Ind Manag Data Syst* 2003, 103: 611-21.
80. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinform* 2007, 23: 2507-17.
81. Espadoto M, Martins RM, Kerren A, Hirata NST, Telea AC. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Trans Vis Comput Graph* 2021, 27: 2153-73.
82. Kang C, Huo Y, Xin L, Tian B, Yu B. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *J Theor Biol* 2019, 463: 77-91.
83. Bulut H. R *Uygulamaları ile Çok Değişkenli İstatistiksel Yöntemler*, 1. Baskı. Ankara, Nobel Akademik Yayıncılık, 2018: 263-75.
84. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006, 101: 1418-29.

85. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005, 67: 301-20.
86. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010, 36: 1-13
87. Saltz JS, Krasteva I. Current approaches for executing big data science projects-a systematic literature review. *PeerJ Comput Sci* 2022, 8: 862.
88. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst* 2012, 36: 2431-48.
89. Bircan H. Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *KOSBED* 2004, 8: 185- 208.
90. Vapnik V. *The Nature of Statistical Learning Theory*, 2nd ed. New York, Springer, 2013: 138-67.
91. Güldogan E, Arslan AK, Yagmur J. The Performance Exploration of Support Vector Machines Models Constructed with Various Kernel Functions: A Clinical Application. *Fırat Tıp Dergisi* 2017, 22: 136-42.
92. Yakut E, Elmas B, Yavuz S. Predicting stock-exchange index using methods of neural networks and support vector machines. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi* 2014, 19: 139-57.
93. Tunc Z, Colak C, Ozdemir R. Classification of Hydrocephalus Disease and Determination of Related Factors by Machine Learning Method. *Ann Health Sci Res* 2018, 7: 14-20.
94. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006; 24(12): 1565-7.
95. Breiman L. Random forests. *Machine learning* 2001, 45: 5-32.
96. Biro TS, Neda Z. Gintropy: Gini Index Based Generalization of Entropy. *Entropy (Basel)* 2020, 22: 879.
97. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens* 2005, 26: 217-22.
98. Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. *Appl Sci* 2018, 8: 689.
99. Hu L, Li L. Using Tree-Based Machine Learning for Health Studies: Literature Review and Case Series. *Int J Environ Res Public Health* 2022, 19: 16080.

100. Song X, Zhu J, Tan X, Yu W, Wang Q, Shen D, Chen W. XGBoost-Based Feature Learning Method for Mining COVID-19 Novel Diagnostic Markers. *Front Public Health* 2022, 10: 926069.
101. Ramraj S, Uzir N, Sunil R, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. *Int J Control Theory Appl* 2016, 9(40): 651-62.
102. Langdon A, Botvinick M, Nakahara H, Tanaka K, Matsumoto M, Kanai R. Meta-learning, social cognition and consciousness in brains and machines. *Neural Netw* 2022, 145: 80-9.
103. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. *Front Comput Sci* 2020, 14: 241-58.
104. Breiman L. Bagging predictors. *Machine learning* 1996, 24: 123-40.
105. Dahiya S, Handa S, Singh N. A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. *Expert Syst* 2017, 34(6): 12217.
106. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001, 29: 1189-232.
107. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Statist* 2000, 28: 337-407.
108. Tuğ Karoğlu T. Lisans Yerleştirme Sınavında Yerleşme Başarısının Karar Ağaçlarına Göre 'Bagging' Ve 'Boosting' Yöntemleriyle Sınıflandırılması. Fen Bilimleri Enstitüsü, Zootekni Anabilim Dalı. Doktora tezi, Van: Yüzüncü Yil Üniversitesi, 2018.
109. Divina F, Gilson A, Gomez-Vela F, Garcia Torres M, Torres JF. Stacking ensemble learning for short-term electricity consumption forecasting. *Energies* 2018, 11: 949.
110. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S (eds). *Bioinformatics and computational biology solutions using R and Bioconductor*: New York, Springer, 2005: 397-420.
111. Yan H, Zheng G, Qu J, Liu Y, Huang X, Zhang E, Cai Z. Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. *J Cell Physiol* 2019, 234: 23785-97.
112. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *J Appl Econ* 2012, 27: 167-72.

113. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: web application framework for R. *R package version* 2017, 1(5): 2017.
114. Wickham H. ggplot2. *Wiley Interdiscip Rev Comput Stat* 2011, (2): 180-5.
115. Wickham H, Chang W, Wickham MH. Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics Version. *R topics documented* 2016, 2(1): 1-189.
116. Xie Y, Cheng J, Tan X. DT: a wrapper of the JavaScript library 'DataTables'. *R package version* 2018, 4: 1-21.
117. Perrier V, Meyer F, Granjon D. shinyWidgets: Custom inputs widgets for Shiny. *R package version* 2019, 4: 1-223.
118. Dumas J. shinyLP: Bootstrap Landing Home Pages for Shiny Applications. *R package version* 2019, 1: 2.
119. Chang W, Ribeiro BB, Studio A, Chang MW. Package 'shinydashboard'. *R topics documented* 2022, 7: 1-27.
120. Bioulac Sage P, Balabaud C. Toxic and Drug-Induced Disorders of the Liver. In: Odze RD, Goldblum JR (eds). S*urgical Pathology of the GI Tract, Liver, Biliary Tract, and Pancreas.* 2nd ed. Philadelphia, Saunders, USA, 2009: 1059-86.
121. Teoh NC, Farrell GC. Liver Disease Caused by Drugs. In: Feldman (ed). *Sleisenger and Fordtran's Gastrointestinal and Liver Disease.* 8th ed. Philadelphia, Saunders, USA, 2006: 1807- 43.
122. Lee WM. Acute liver failure. *Semin Respir Crit Care Med*. 2012, 33: 36-45.
123. Holt MP, Ju C. Mechanisms of drug-induced liver injury. *AAPS J* 2006, 8: 48-54.

EKLER

EK-1. Özgeçmiş

EK-2. Etik Kurul Onayı