BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GEN İFADE TAHMİNİ İÇİN VERİ BÜTÜNLEŞTİRME

TUNCAY BAYRAK

DOKTORA TEZİ
2019

GEN İFADE TAHMİNİ İÇİN VERİ BÜTÜNLEŞTİRME

DATA INTEGRATION FOR PREDICTING GENE EXPRESSION

TUNCAY BAYRAK

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
DOKTORA TEZİ
olarak hazırlanmıştır.

2019

“Gen İfade Tahmini için Veri Bütünleştirme” başlıklı bu çalışma, jürimiz tarafından, ../…../……tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda DOKTORA TEZİ** olarak kabul edilmiştir.

Başkan	:….....
	(Prof. Dr. Mustafa KOCAKULAK)
Üye (Danışman)	:….....
	(Prof. Dr. Hasan OĞUL)
Üye	:….....
	(Prof. Dr. Hamit ERDEM)
Üye	:….....
	(Doç. Dr. Ahmet Burak CAN)
Üye	:….....
	(Dr. Öğr. Üyesi Mustafa SERT)

ONAY

..../..../.....

Prof. Dr. Ömer Faruk ELALDI

Fen Bilimleri Enstitüsü Müdürü

BAŞKENT ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 19.08.2019

Öğrencinin Adı, Soyadı: Tuncay BAYRAK
Öğrencinin Numarası: 21220041
Anabilim Dalı: Bilgisayar Mühendisliği
Programı: Doktora
Danışmanın Unvanı/Adı, Soyadı: Prof. Dr. Hasan OĞUL

Tez Başlığı: Gen İfade Tahmini için Veri Bütünleştirme
Yukarıda başlığı belirtilen Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 106 sayfalık kısmına ilişkin, 19/08/2019 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %3'tür.

Uygulanan filtrelemeler:
1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.
Öğrenci İmzası:

Onay
19/08/2019

Prof. Dr. Hasan OĞUL

TEŞEKKÜR

Yazar, bu çalışmanın gerçekleşmesinde katkılarından dolayı, aşağıda adı geçen kişi ve kuruluşlara içtenlikle teşekkür eder.

Sayın Prof. Dr. Hasan OĞUL'a (tez danışmanı), çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için…

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) 'na 2211-A BİDEB Yurtiçi Lisansüstü Burs Programı kapsamında bu tez çalışmasına destek verdiği için…

Çalışmakta olduğum Türkiye İlaç ve Tıbbi Cihaz Kurumu'na doktora sürecimde her türlü imkânı ve desteği sağladıkları için…

Kıymetli eşim Melike'ye ve oğlum Kerem Ali'ye yanımda oldukları için…

ÖZ

GEN İFADE TAHMİNİ İÇİN VERİ BÜTÜNLEŞTİRME

Tuncay BAYRAK

Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Canlı formunun sürdürülebilirliğinin temelinde protein sentezi yer almaktadır. Protein sentezinde, insan genomundaki kodlayıcı genleri düzenleyen küçük nükleotid dizilerinin (mikro RNA) ve diğer yönetici genlerin (Transkripsiyon Faktör, TF) önemli görevleri vardır. Bu çalışmanın amacı, mikro RNA ve TF'lerin düzenleme bilgisinin protein kodlayıcı genlerin ifade tam değerlerinin kestirim performansına etkisini araştırmaktır. Gen ifade tam değerini tahmin etmek için regresyon tabanlı modelleri içeren sistematik yaklaşımlar ortaya konulmuştur.

Öncelikle, gen ifade ölçümlerinde yaygın olarak karşılaşılan kayıp veri (missing data) problemini çözmek için doğrusal, k-NN ve İlişkisel Vektör Makinesi (RVM) regresyon modelleri uygulanmıştır. Regresyon modelinin eğitiminde genellikle aynı genin farklı deneylere ait ifade değerlerinden oluşan vektörler kullanılmaktadır. Daha sonra, bu ifade vektörlerine aynı deneye ait farklı gen ifade değerlerinin dâhil edilmesinin gen ifade tahminine etkisi araştırılmıştır. Bunun için İki Yönlü İşbirlikçi Filtreleme (Two-way collaborative filtering) yöntemi kullanılarak gen ifade değerlerinden oluşan tek yönlü veri matrisi iki yönlü veri matrisine dönüştürülmüş ve regresyon modeli bu yeni veri matrisi ile oluşturulmuştur. Gen ifade tahmini için ilk defa kullanılan bu yeni öznitelik

sunum tekniği ile kestirim performansının artırıldığı görülmüştür. Ayrıca farklı kanser türlerine ait gen ifade verilerinin bütünleştirilmesinin gen ifade tahminine etkisi de araştırılmıştır. Burada, prostat kanserine ait gen ifade değerlerinin tahmin edilmesinde kolon kanseri verisinin model öğrenmede kullanılmasının kestirim performansını artırdığı görülmüştür. Literatürde gen ifade değerleri kullanılarak gen düzenleyici moleküller ile genler arasındaki ilişkinin tespit edilmesine yönelik çok sayıda çalışma bulunmaktadır. Ancak hücrede meydana gelen bu etkileşimler kullanılarak gen ifade tam değerinin tespitine yönelik çalışmalar oldukça kısıtlıdır. Son olarak, farklı veri yapısındaki miRNA-gen ve TF-gen regülasyon bilgileri ile gen ifade değerleri bütünleştirilmiş olup doğrusal ve RVM regresyon modelleri kullanılarak kestirim performansına etkisi araştırılmıştır. Veri bütünleştirme yaklaşımlarında Öklid, Affine Dönüşüm ve Bhattacharya uzaklık ölçütleri kullanılmıştır. Gen ifade matrisleri; Gene Expression Omnibus veritabanından, TF-gen regülasyon bilgisi TRANSFAC veritabanından ve miRNA-gen regülasyon bilgisi ise mirDB, mirTarbase ve mirConnX veri tabanlarından alınmıştır. Kestirim performansının değerlendirilmesinde Spearman benzerlik katsayısı, Pearson benzerlik katsayısı ve Hata Kareleri Ortalamasının Karekökü (RMSE) ölçütleri kullanılmıştır. miRNA-gen regülasyon bilgisinin bütünleştirilmesi ile gen ifade tahmini performansının artırıldığı görülmüştür.

ANAHTAR SÖZCÜKLER: İlişkisel vektör makineleri, regresyon, veri bütünleştirme, mikro-RNA, gen ifadesi tahmini, iki yönlü işbirlikçi filtreleme, transkripsiyon faktör.

Danışman: Prof. Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

DATA INTEGRATION FOR PREDICTING GENE EXPRESSION

Tuncay BAYRAK

Baskent University Institution of Science and Engineering
Department of Computer Engineering

Protein synthesis is the basis of the sustainability of the living form. Small nucleotide sequences (micro-RNA) and other executive genes (Transcription Factor, TF) that regulate coding genes play an important role in the protein synthesis. The aim of this study was to investigate the effect of regulation information of micro-RNA and TFs on the performance of predicting the exact value of expressions of protein coding genes. In order to predict the exact value of gene expression, systematic approaches that includes regression-based models are introduced.

First, linear, k-NN and Relational Vector Machine (RVM) regression models were applied to solve the common problem of missing data in gene expression measurements. The expression vectors used in the training phase of the regression model are generally composed of the expression values of the same gene that belongs to different experiments. After that, the effect of the inclusion of different gene expression values of the same experiment on these expression vectors was investigated. For this, the one-way data matrix, consisting of gene expression values, was transformed into a two-way data matrix using Two-way Collaborative Filtering method and the regression model was built with this new data matrix. It is observed

that this new feature representation technique that is first used in this study for gene expression predicting increases the performance of predicting. In addition, the effect of integrating gene expression values of different cancer types on gene expression predicting is also investigated. Here, it is observed that the use of colon cancer data in model learning to predict the gene expression of prostate cancer increases prediction performance. There are many studies in the literature to determine the relationship between regulating molecules and genes using gene expression values. However, there are very limited studies based on predicting the exact value of gene expression by using these relations in the cell. Finally, miRNA-gene and TF-gene interaction information and gene expression values were integrated and the prediction performance outcomes obtained by using linear and RVM regression models were discussed. Euclidean, Affine Transformation and Bhattacharya distance measures were used in data integration approaches. Gene expression matrices from Gene Expression Omnibus; TF-gene regulation information from TRANSFAC; miRNA-gene regulation information from mirDB, mirTarbase and mirConnX were used. Spearman similarity coefficient, Pearson similarity coefficient and Root Mean Squared Error (RMSE) were used to evaluate the performance of predicting. It is observed that the performance of predicting gene expression is increased by integrating of miRNA-gene regulation information.

KEYWORDS: Relevance vector machines, regression, data integration, micro RNA, gene expression prediction, two-way collaborative filtering.

Advisor: Prof. Dr. Hasan OGUL, Baskent University, Department of Computer Engineering

İÇİNDEKİLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

μ	ortalama
σ	çekirdek fonksiyonu sigma parametresi
γ	çekirdek fonksiyonu gama parametresi
ε_i	gürültü

cDNA	Tamamlayıcı DNA
DNA	Deoksiribonükleik asit
DR	Doğrusal Regresyon
EBI	European Bioinformatic Institute
GEO	Gene Expression Omnibus
KK	Korelasyon katsayısı
mRNA	Mesajcı RNA
NIH	The National Institute of Health
NLM	National Library of Medicine
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
RNA	Ribonükleik asit
rRNA	Ribozomal RNA
RT	Ters transkriptaz (Reverse transkriptaz)
RVM	Relevance Vector Machines
SVM	Support Vector Machines
tRNA	Taşıyıcı RNA

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

1. GİRİŞ

Yaşamın yerkürenin oluşumundan 750 milyon yıl sonra günümüzden 3.8 milyar yıl önce ilk olarak ortaya çıktığı düşünülmektedir. Hücreler bir araya gelerek dokuları, dokular organları, organlar sistemleri ve sistemler de organizmayı oluşturmaktadır. Bu hiyerarşide en alt seviyede hücreler yer almaktadır. Prokaryot ve ökaryot olmak üzere iki çeşit hücre tipi bulunmaktadır. Çok daha gelişmiş, büyük ve genetik materyalin nükleus olarak bilinen hücre çekirdeği içinde saklandığı ökaryotik hücre tipi gelişmiş canlıların da temel bileşenidir. Tüm canlılardaki ortak özellik ise her hücrenin içinde aynı genetik materyalin bulunmasıdır. İnsan vücudunda epitel, bağ, kan, kas ve sinir dokusu ile bu beş dokunun bileşenleri olarak iki yüzden fazla hücre çeşidinin bulunduğu bilinmektedir. Bu hücre çeşitliliği hücrede bulunan genetik materyalin bir ürünüdür. Hücrede genetik bilgi kromozom adı verilen yapılarda yer almaktadır. İnsanda toplam 23 kromozom çifti bulunmakta olup bunların 22'si otozomal kromozom çiftidir ve kalan 1 kromozom çifti ise cinsiyeti belirlemektedir. Bu kromozomlar; bir azotlu organik bazların bir araya gelerek oluşturduğu Deoksiribonükleik asit (DNA) paketleri olarak bilinmektedir. İnsanda yaklaşık 3 milyar baz çifti bulunmaktadır. Bu baz çiftlerinden oluşan nükleotid zincirinin protein sentezi sürecinde rol alan anlamlı parçalarına gen adı verilmektedir. İnsan genomunda yaklaşık 20 bin protein kodlayıcı gen bulunmaktadır. Protein kodlayan genlerin sayısı önceleri 100 bin civarında olarak bilinirken yeni gen dizilim teknolojilerinin geliştirilmesi ile aslında çok daha az sayıda genin protein sentezi için ifade verdiği bilgisine ulaşılmıştır. İnsan fizyolojisinde meydana gelen faaliyetlerin veya patolojik durumların tümünde gen aktiviteleri rol oynamaktadır. Bir gen birden fazla proteinin sentezinde rol aldığı gibi bazı genlerin tüm proteinlerin sentezinde rol aldığı bilinmektedir. Protein kodlayıcı dizilerin insan genomunun %2'si bile olmadığı bilinmektedir. İnsanda protein-kodlayıcı gen eşleşmesinin daha az gelişmiş diğer canlı türlerine göre oldukça az olduğu bilinmekte olup protein sentezi sürecinde kodlama dışında düzenleyici görevleri bulunan nükleotid dizilerinin de bulunduğu görülmüştür. İşlevsellik açısından bu anlamlı dizilerin tüm genomun yaklaşık %97'sini oluşturduğu düşünülmektedir. Protein sentezinde kodlayıcı genler dışında bu genlerin çalışmasını etkileyen diğer düzenleyici dizilerin çalışma şekli süreci daha da karmaşık hale getirmektedir. Protein sentezinde mesajcı Ribonükleik Asit (mRNA), taşıyıcı RNA (tRNA), ribozom gibi birçok molekül ve organellerin görev

aldığı moleküller arası etkileşimler söz konusudur. Bununla birlikte yaklaşık son 10 yıldır üzerinde çalışılan bir diğer nükleotid zinciri mikro-RNA (miRNA) molekülleridir. Küçük zincirlerden oluşan bu moleküllerin bazı genlerin çalışmasını hızlandırırken diğerlerini baskıladığı ve böylece protein sentezi sürecinde önemli rol aldığı bilinmektedir. Bu düzenleyici işlevlerin insan fenotipine etkisi de kaçınılmazdır. miRNA nükleotid dizilerinin yanında bazı genlerin yönetici gen olarak protein sentezi sürecinde rol aldığı da görülmüştür. Bu yönetici genlerin diğer genleri düzenlemesinin yanında protein kodlama ve miRNA'lar ile etkileşim halindedirler. Hücrede bu şekilde farklı moleküllerin protein sentezi sürecinde görev alması ve etkileşim halinde olması hesaplamalı biyoloji çalışmalarının temelini oluşturmaktadır. Ayrıca organizmanın genetik yapısı olarak bilinen genotip ile fiziksel görünüşü veya fonksiyonu olarak bilinen fenotip arasındaki ilişkinin moleküler düzeyde kurulması, bu değişkenlerin doğru analizine bağlıdır. Bu nedenle uzun yıllardır gen ifade profilleri ile hastalık veya fenotip farklılıkları arasındaki ilişkinin daha iyi anlaşılması için çalışmalar devam etmektedir.

Gen ifadelerinin hücrelerde, dokularda ve bunların işlevlerinde meydana getirdiği farklılaşmaların fenotipe yansımaları detaylı ve çok yönlü gen analizleri ile daha iyi açıklanabilir. Gen ifadelerinin fenotipe olan etkisi geçmişte çokça işlenmiş olup özellikle protein sentezi sürecindeki faktörlerin hastalıklarla ilişkilendirmesine yönelik literatürde çok sayıda çalışma bulunmaktadır. Ayrıca düzenleyici genler ve miRNA'lar ile gen ifade değerlerine bakılarak hastalıkların oluşma olasılığına, metastazına veya prognozuna ilişkin tahmin yürütülmesi konusunda literatürde çeşitli çalışmalar mevcuttur. Buna karşılık protein sentezi sürecinde düzenleyici görevleri bulunan miRNA ve bazı yönetici genlerin regülasyon bilgisinin protein kodlayıcı genlerin ifade tahmininde kullanılmasına yönelik oldukça kısıtlı sayıda çalışma mevcuttur. Hesaplamalı biyoloji çalışmalarında genler, proteinler ve hastalıklar arasındaki moleküler düzeydeki bağlantılara ilişkin yapılan araştırmaların temelinde gen ifade profillerinin analizi yer almaktadır. Gen ifade profilleri binlerce genin aktivitesini kantitatif olarak ortaya koyan veri matrislerinden oluşmaktadır. Bu veri matrislerindeki protein kodlayıcı gen ifade değerleri üzerinden yapılan analizler ile genotip-fenotip ilişkisi daha doğru kurulabilir.

Bu tezin amacı, protein sentezi sürecinde; düzenleyici görevleri bulunan yönetici genler ve miRNA'lar kullanılarak protein kodlayıcı genlerin ifade değerlerinin tahmin edilmesi için regresyon tabanlı yeni yaklaşımlar ortaya koymaktır. Bu kapsamda sistematik bir çalışma yapılmış olup üç bölümde sunulmuştur. Birinci bölümde; Kayıp Veri Atama (missing data imputation) probleminin çözümüne ilişkin farklı regresyon modelleri ile İlişkisel Vektör Makinesi (Relevance Vector Machine-RVM) regresyon modeli karşılaştırılmıştır. İnsandan tükürük, kan veya doku parçası gibi DNA içeren numunelerden deneysel çalışmalar ile elde edilen gen ifade profillerinde çevresel nedenlerden ve deneysel hatalardan kaynaklı bazı genlerin ifade değerleri eksik olabilir. Bu deneylerin tekrarlanması maliyet-etkin olmadığından dolayı bu kayıp verilerin tahmin edilmesi için literatürde makine öğrenme temelli çok sayıda çalışma mevcuttur. Tezin bu bölümünde bu amaçla ilk defa RVM regresyon modeli kullanılmıştır ve sonuçlar tartışılmıştır.

İkinci bölümde; farklı bir problemin çözümünde kullanılan İki Yönlü İşbirlikçi Filtreleme (Two-way Collaborative Filtering) yöntemi ilk defa bu tezin amacına uygun olarak adapte edilmiş ve gen ifade tam değerinin tespitinde kullanılmıştır. Ayrıca bu bölümde gen ifade değerlerinin kestirim performansını artırmak için farklı kanser dokularından ölçülen gen ifade değerleri bütünleştirilmiş ve sonuçlar tartışılmıştır.

Literatürde miRNA gibi düzenleyici moleküller ile genlerin ifade değerleri birlikte analiz edilerek moleküller arası etkileşimlerin tespit edilmesine yönelik çok sayıda çalışma vardır. Ancak birbirinden farklı veri tiplerine sahip olan regülasyon bilgisi ve gen ifade değerlerinin aynı modelde kullanılmak üzere bütünleştirilerek gen ifade değerlerinin tahmin edilmesine yönelik bir çalışmaya rastlanılmamıştır. Bu çalışmanın üçüncü bölümünde ise protein kodlayıcı genlerin ifade tahmininde miRNA'ların ve düzenleyici genlerin (transkripsiyon faktör) regülasyon bilgilerinin kullanıldığı farklı yaklaşımlar regresyon tabanlı modeller kullanılarak test edilmiştir. Farklı kanser türlerine ait gen ifade verileri, gen regülasyon bilgisinin elde edildiği birden fazla veritabanı ve farklı regresyon modelleri kullanılarak karşılaştırmalı bir çalışma sunulmuştur.

Hastalıkların genlerle veya diğer moleküllerle ilişkisinin tamamen anlaşılmasının teşhis ve tedavi süreçlerine katkı sağlayacağı düşünülmektedir. Bilişim sistemlerinin ve yeni yöntemlerin gelişmesiyle beraber gün geçtikçe moleküler biyoloji ve genetik alanında yapılan çalışmaların niteliği artmakta ve kapsamı genişlemektedir. Bilgisayar bilimlerinin bu alanda uygulanması ile özellikle hastalık kökenlerinin ve prognozunun bağlı olduğu diğer hücresel boyutta meydana gelen olaylar daha kolay tespit edilmektedir.

2. TEMEL BİLGİLER

2.1. Gen Regülasyonu

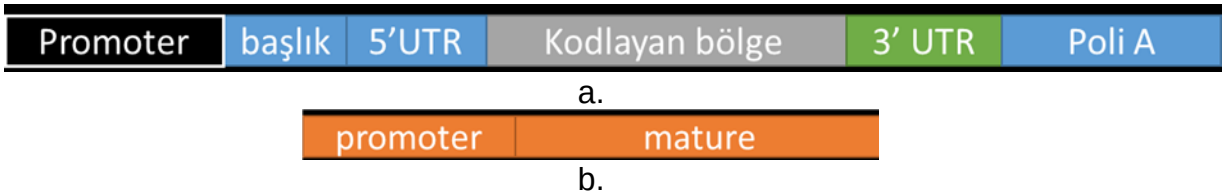
Bir organizmanın tüm hücrelerinde aynı nükleotid dizilimine sahip ve içinde genetik bilginin yer aldığı moleküller Doksiribonükleik asit (DNA) olarak bilinir. Hücrenin farklı işlevler göstermesini sağlayan anlamlı DNA dizisi parçalarına gen adı verilmektedir. Genomun fonksiyonel kısımları olarak tanımlanan genler tüm hücrelerde aynı şekilde bulunmasına rağmen, her hücrede aktif ve pasif genler değişiklik göstermektedir. Genlerin aktif veya pasif olması hücrelerin farklı fonksiyonlar göstermesi ile sonuçlanmaktadır. Aktif genler ifade vererek o gene spesifik proteinlerin sentezinde rol alırlar. Bu sayede aynı DNA'ya sahip ve farklı fonksiyonları olan hücreler oluşmaktadır. Bu farklı fonksiyonlara sahip hücreler ise canlıdaki doku ve organ farklılıklarını ortaya çıkarmaktadır.

Protein sentezi, nükleaz enzimi ile denatüre olan DNA'nın kendisini eşlemesi ile başlar ve RNA moleküllerinin aracılığıyla ribozom üzerinde gerçekleşir. Bu süreç, DNA zincirinin denatüre olması ve bu zincire karşılık gelen mRNA zincirinin üretimi ile başlar (transkripsiyon-yazılım). mRNA zincirinin taşıdığı şifreli mesaj ribozomlarda okunarak proteine dönüştürülür. Bu işleme çevirim (translasyon) denir. Bu işlemin gerçekleşmesinde, amino asitleri ribozoma taşıyan taşıyıcı-RNA (tRNA) molekülleri büyük rol oynar. Bir proteini oluşturan polipeptit zincirindeki aminoasitlerin sırası mRNA'da bulunan 3 nükleotidden oluşan kodon adı verilen yapıların sırayla okunması ile belirlenir. Transkripsiyon ve translasyon arasındaki sürece post-transkripsiyon adı verilmektedir. Her mRNA bir protein bilgisi kodlar. mRNA'daki her 3 baz proteindeki bir amino aside karşılık gelmektedir. Bu süreçte oluşturulan amino asit zinciri daha sonra katlanarak (protein folding) üç boyutlu protein halini almaktadır.

Protein sentezini baskılayan veya artıran moleküler düzeyde mekanizmaların bulunduğu bilinmektedir. Son zamanlarda protein sentezi sürecinde kodlama yapmayan (noncoding) fakat bu süreçte etkin rol alan küçük RNA molekülleri üzerindeki çalışmalar yoğunlaşmıştır. Bu RNA molekülleri, yapısal ve düzenleyici (regulatory) olarak ikiye ayrılır. Düzenleyici RNA'lar, uzun kodlama yapmayan (long

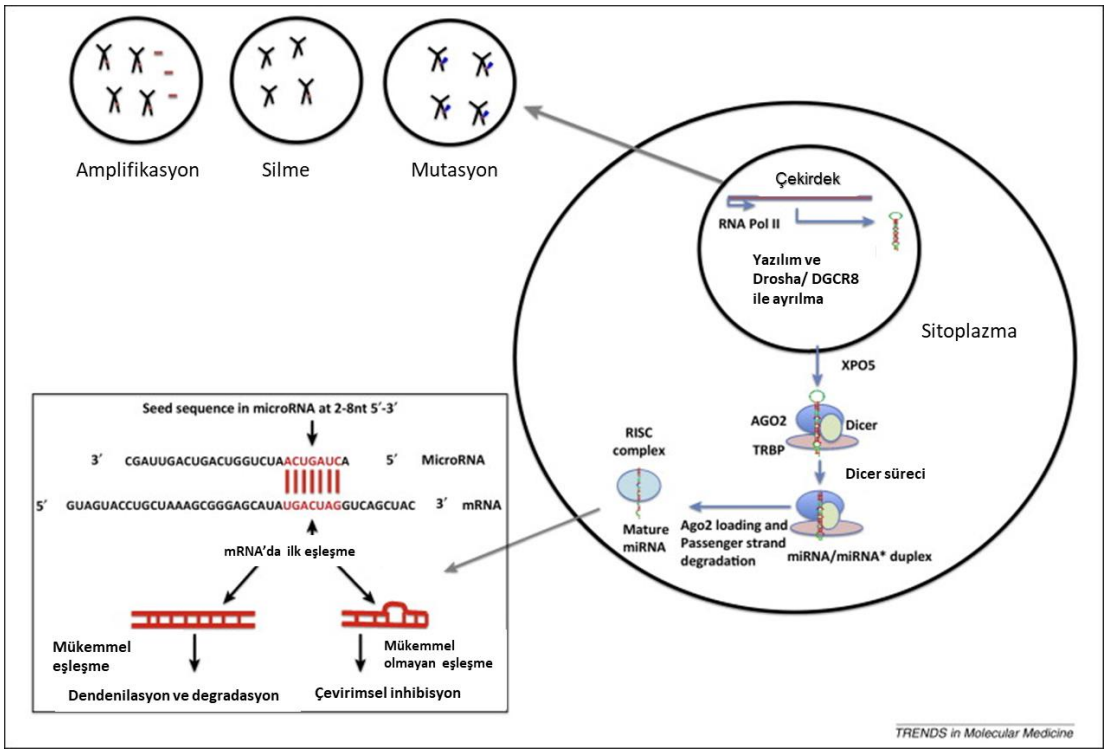
non-coding RNA-lncRNA) ve küçük kodlama yapmayan RNA (small non-coding RNA)'lar olarak ikiye ayrılır. Burada küçük kodlama yapmayan RNA'lar üçe ayrılır ve bunlar küçük bozucu (interferring, siRNA), piwi bağlantılı (piRNA) olanlar ve micro RNA'lardır [1].

Protein sentezi sürecinde miRNA molekülleri aktif rol oynamaktadır ve ilk olarak hayvanlarda, bitkilerde ve virüslerde keşfedilmiştir. Deneysel ve bilgisayar temelli yaklaşımlar ile 2008 sonu itibariyle yaklaşık 700 insan miRNA'sı keşfedilmiştir. Bunların 180 tanesi protein kodlama bölgelerinde, 381'i intronic bölgelerde ve geri kalanı intergenic (genler arası) bölgelerde bulunmaktadır. Genlerin ifade verme sürecinde, intergenic bölgede bulunan miRNA'lar, mRNA üzerindeki kendilerine tanımlı başlatıcı (promoter) dizilerini kullanırken, intronic bölgede bulunanlar ise içinde bulunduğu mRNA'lar ile koordineli bir şekilde hareket ederler. miRNA'lar; RNA polimeraz II ve polimeraz III tarafından çekirdekte yazılır, bu ilk moleküle pri-miRNA adı verilir. pri-miRNA'lar yapı itibariyle mRNA'lara çok benzerler ve bir ucunda başlık (cap) diğerinde poli-A yapısı vardır (Şekil 2.1). Bu molekül olgun bir miRNA'dan daha uzundur. Bunlar Drosha ve Dicer adı verilen iki adet RNase III enzimi ile ardışık bölünmelere bağlı olarak olgun miRNA'lara dönüşürler. Bu süreçte, öncelikle Drosha, pri-miRNA'yı işleyerek 70 nükleotid dizisine sahip pre-miRNA'ya dönüştürür. Daha sonra oluşan pre-miRNA molekülleri, XPO5 geni tarafından üretilen ve görevi çekirdek ile sitoplazma arasında taşımacılık yapmak olan exportin-5 proteini tarafından sitoplazmaya taşınır. Burada Dicer devreye girer ve bunları 22 nükleotid dizili miRNA çiftine ayırır. Bu çiftten sadece biri ribonükleoprotein kompleksine (RISC) bağlanabilir. Burada ayırt edici özellik; RISC içindeki argonaute proteinin 5' ucu en kararlı olan miRNA zincirini seçmesidir. RISC, translasyon baskılamada aktif rol oynayan bir multiproteindir ve böylece miRNA'lar tarafından düzenlenmiş olurlar [2]. Şekil 2.2'de bu sürecin şematik gösterimi yer almaktadır [7].



Şekil 2.1 a) mRNA ve b) miRNA yapılarının şematik gösterimi

miRNA molekülü 20-24 nükleotidden oluşan küçük RNA zincirleridir[3-5]. İnsanda üretilen 1000 civarındaki miRNA molekülünün genlerin %30'unu düzenlediği bilinmektedir. miRNA başlatıcı ve olgun (mature) adı verilen kısımlardan oluşmaktadır ve olgun zincir parçası özel olarak mRNA'nın 3-çevrilmemiş bölgesine (untranslated regions-UTR) bağlanarak, mRNA'nın protein sentezleme işlemini inhibe eder. Burada bağlanma biçimi adenin-urasil, guanin-sitozin şeklinde bazların komplementeri şeklinde gerçekleşir[6].

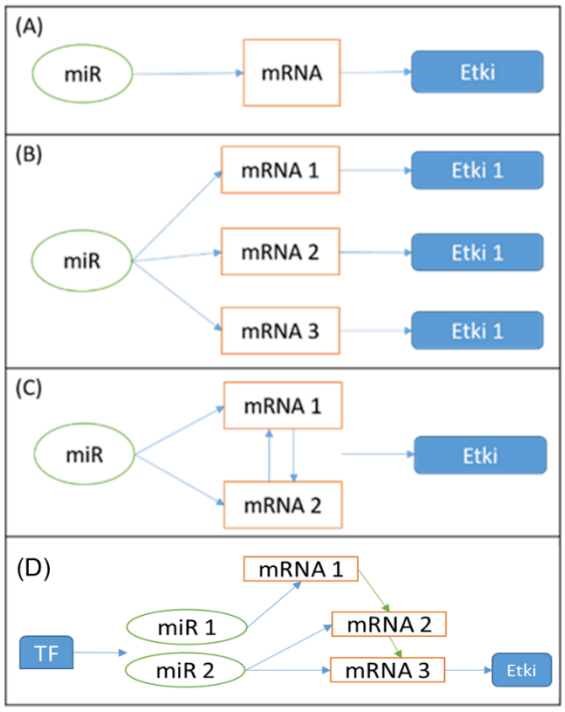


Şekil 2.2 miRNA üretimi ve translasyon baskılama

Birçok çalışmada, miRNA moleküllerinin gen düzenleyici mekanizmasındaki bozuklukların kolon, troid, özafagus gibi insan vücudunun farklı bölgelerinde meydana gelen kanser türlerinin ortaya çıkmasında, ilerlemesinde ve tümör farklılıklarının oluşmasında önemli rol oynadığı belirtilmiştir. Hastalıkların ortaya çıkmasının yanında, tedavi sonrası hastalığın farklı hedef bölgelerde nüksetmesinde de (metastaz) miRNA'ların etkin olduğu bildirilmiştir [8-14]. Aslında bu durum kanser çeşitlerinin altında yatan patolojik süreçlerin benzer olması ile ilgili olduğu düşünülebilir. Örneğin Kras proteini KRAS geni tarafından üretilmektedir ve bu protein üretiminin baskılanması kolon kanserinin prognozunda oldukça etkilidir. KRAS geninin mutasyonu ile karakterize olan kolon kanserinde let-7 ailesi, miR-

133b, miR-34 ailesi, miR-126, miR-143 ve miR-145 miRNA'larının aktif rol oynadığı bilinmektedir [15].

miRNA etkileşimleri Şekil 2.3'teki gibi dört farklı yaklaşım ile özetlenebilir [7]. Bu etkileşimler içinde en yaygın olanı miRNA'nın doğrudan mRNA ile olan etkileşimidir (A). Bir miRNA molekülünün birden fazla hedef mRNA'ya bağlandığı diğer etkileşim tipi geniş yaklaşım (broad approach) olarak bilinmektedir (B). Üçüncü etkileşim türü biraz daha karmaşık ve daha fonksiyoneldir. Burada miRNA, birbiri ile etkileşim içinde olan birden fazla mRNA'yı hedef olarak belirlemektedir (C). Son etkileşim türünde ise transkripsiyon faktörleri (TF) birden fazla miRNA ile etkileşimde bulunarak birden fazla mRNA'yı etkileyebilmektedir (D). Burada oluşan yolak (pathway) ve düğümler süreç sonundaki etkiyi belirlemektedir (pathway based approach). Sadece miRNA ve mRNA etkileşim bilgilerinin bulunduğu veri tabanları da mevcuttur. Bu çalışmada, bu veri tabanları da kullanılmıştır.



Şekil 2.3 miRNA etkileşimleri

Büyüme faktörü uyarımı (growth factor signalling) bazal durumda hücrelerin yaşamını sürdürebilmesi için gerekli molekül ve ATP sentezinde kullanılacak yeterli

minerallere erişmeleri için gerekli olan biyolojik bir izin mekanizmasıdır. Bunun için normal hücreler, aralarında bir yolak geliştirip sürece devam ederken, kanserli hücreler normal fizyolojik kısıtların dışına çıkarak kontrolsüz iletim yolları oluştururlar. Aslında genetik birçok faktörün sonucu olarak değerlendirilip birçok kanser türünün bazı özel genlerle ilişkili olduğu bilinmektedir [16]. Keşfedilen ilk tümör baskılayıcı miRNA'lar; miR-15a ve miR-16-1'dir. Bazı miRNA'lar bazı hücre tiplerinde onkojen gibi davranırken bazılarında ise tümör baskılamaktadır. Örneğin miR-221, hepatosellüler kanserde PTEN geninin ifadesini aşağı yönde düzenlerken, eritroblastik lösemide KIT onkojenini baskılayarak bir tümör baskılayıcı gibi davranmaktadır[17]. Bu bize miRNA'nın olduğu süreçlerin tekdüze gerçekleşmediğini, yeni bir klinik ve prognostik araç olarak kullanılabileceğini göstermektedir. Örneğin miR-135a ve miR-135b mikroRNA moleküllerinin adenomatous polyposis coli (APC) genini tetikleyerek kolon kanseri başlangıcında etkili olduğu bilinmektedir [18]. Bunu yanında literatürde miRNA ifade profilleri kullanılarak kanser sınıflandırılması yapılmış çalışmalar da yer almaktadır. Örneğin 334 kanser dokusuna ait mRNA ve 218 miRNA ifade profilleri analiz edilerek kolon, karaciğer, pankreas ve mide kanseri dokuları sınıflandırılabilmiştir [19]. Bu tip çalışmalar özetle; kanserde miRNA–mRNA etkileşimlerinin biyoişaretleyici (biomarker) olarak değerlendirilebildiği sonucunu ortaya koymaktadır. Bu nedenle bu etkileşimler üzerine yapılan çalışmalar kanserin erken teşhis ve tedavisinde büyük önem taşımaktadır. Ayrıca hastalıkların ortaya çıkmasında rol oynayan miRNA'ların inhibe edilmesine yönelik çalışmalar da mevcuttur [20].

Bir başka çalışmada, miRNA-mRNA etkileşimi üzerine internet tabanlı bir yazılım aracı geliştirilmiştir. Ön işlem olarak miRNA ve mRNA ifadelerinden düşük olanlar ayıklanmıştır. TF ve miRNA arasındaki bağlantılar istatistiksel olarak ölçülmüş ve hastalık veya diğer deneysel durumlarla bağlantılı bir ilişki ağı (*assocation network)* kurulmuştur. Bu ağ yönsüz bir çizge özelliğindedir ve daha sonra bağlanan TF motiflerinden, miRNA hedef tahmininden ve literatürdeki diğer bilgilerden elde edilen türe özel öncelikli ağ yapısı (species prior network) ile birleştirilmiştir. Türe özel öncelikli ağ yapısı, bağlantıları ağırlıklandırılmış yönlü bir çizgedir. Bu iki çizge bir entegrasyon fonksiyonu ile birleştirilerek yeni bir çizge elde edilir. mirConnX, bu sürecin görselleştirilmesi ve uygulanması için geliştirilmiş bir araçtır. İçerdiği bilgilere

göre hastalığa özel genetik varyasyon çerçevesinde bir düzenleyici ağ yapısı kurar [21]. Bir başka çalışmada TF'leri kontrol eden aktif miRNA'ları tanımlayan bir yöntem geliştirilmiştir. Burada miRNA-TF, miRNA-kinase-TF ve TF-TF arasındaki ağ etkileşimleri istatistiksel testler ile karşılaştırma yapılarak kullanılmıştır. 17'si kanser ile ilişkili miRNA içeren 43 adet transfeksiyon deneyinde yapılan testlerde yöntemin doğruluğu kanıtlanmıştır [22]. Bir diğer çalışmada ise birden fazla düzenleyici süreçte rol alan moleküllerin arasındaki etkileşimi çizge tabanlı bir yaklaşım ile ortaya koyan MIR@TN@N adı verilen bir araç ve geniş bir veritabanı geliştirilmiştir. İnternet ortamında ücretsiz olarak sunulan bu sistemde kullanıcı bazı filtreleme seçeneklerini kullanarak veri elde edebilir. Girilen kantitatif gen ifade profilleri ve TF/miRNA/mRNA listesi ile protein sentezi sürecinde anahtar rol oynayan faktörleri, etkileşim ağ yapısını ve alt ağ yapılarını tahmin eden bir sistemdir [23].

Kanser tedavisinde başvurulan yöntemler; cerrahi müdahale, kemoterapi, biyolojik terapi veya radyoterapidir. Ancak bu yöntemlerin kanserli dokunun yanında sağlıklı dokuları da yok etmesi, ilaçlara direnç göstermesi ve metastaza neden olması gibi etkileri vardır. Aynı zamanda bu tedavi prosedürleri hastanın yaşam koşullarını olumsuz etkilemektedir[24]. Bu nedenle, kişiye özel gen terapi yöntemlerinin yaygınlaşarak daha etkin ve moleküler düzeyde imkan sunulan tedavi prosedürlerinin oluşturulması üzerine çalışmalar önemli görülmektedir.

2.2. Gen İfade Analizi

İnsan genomunda yaklaşık 3 milyar nükleotid çifti ve 25 bine yakın gen vardır. Genomun sadece yaklaşık %2'sinin protein kodlama özelliği olduğu düşünüldüğünde protein sentezi sürecinde gen aktivitelerinin oldukça önemli olduğu görülmektedir. Gen ifade miktarının ölçümü ilk defa 1977 yılında Northern Blot adı verilen bir yöntemle yapılmıştır. Bu yöntem ile bir veya birkaç genin ifade miktarı ölçülebilmektedir. Ancak gelişen teknoloji ile hücrede meydana gelen gen aktivitelerine ve protein sentezi süreçlerine olan bakış açısı da değişmiştir.

Her bir genin ifade miktarı yerine genler arasındaki etkileşim ve genler ile diğer moleküller arasındaki etkileşimin araştırılmasının sistem biyolojisinin temelini

oluşturduğu görülmektedir. Hücrede, genler ve moleküller arası etkileşimin kantitatif veya kalitatif ölçümleri sayesinde; hastalıkların birbirleri ile ilişkisi veya farklı metabolik ve çevresel durumların protein sentezine olan etkisi gibi moleküler düzeyde çeşitli araştırmaların yapılması mümkündür. Gen ifade miktarı ölçüm yöntemlerinin gelişmesi ve bu yöntemlerden daha nitelikli verilerin sağlanması; biyoinformatik alanında ve daha etkin tedavi yöntemlerinin geliştirilmesinde önemli faydalar sağlamaktadır. Ölçüm yöntemlerindeki maliyetler; tamamen bu yöntemlerin gerektirdiği teknolojilere ve çalışmalarda belirlenen hedeflere göre değerlendirilmelidir. Araştırmalarda hangi düzeyde bilgiye ihtiyaç varsa ona göre bir ölçüm yönteminin belirlenmesi daha uygundur.

2.2.1. Gen İfade Tahmini

DNA, daha önce ifade edildiği gibi canlının genetik bilgisini taşıyan ve nesilden nesile aktarılmasını sağlayan hücre çekirdeğinde çift sarmallı yapıdan oluşan bir moleküldür. Protein sentezi sürecinde DNA'nın anlamlı parçaları olan genler tarafından kodlama yapılır. Sentezlenen proteinler ise hücre fonksiyonlarını oluşturur. İnsan genomunda 25 bin civarında gen bulunduğu bilinmektedir [25].

Bir genin ifade vermesi; içerdiği nükleotid dizilerinin kullanılması ile protein gibi işlevsel bir ürünün üretilmesi anlamına gelmektedir. Gen ürünleri çoğunlukla hücrenin fonksiyonunun yerine getirilmesi için gerekli olan enzim, hormon ve reseptör gibi proteinlerdir. Bir hücre içindeki binlerce genin aktif veya pasif olması o hücrenin ne iş yapacağı ve diğer hücrelerle ilişkisini belirlemektedir [26]. Hücrelerin farklı işlevlere sahip olması; farklı doku, organ ve sistemlerin oluşmasını sağlamaktadır.

Gen ifade tahminine yönelik kısıtlı sayıda çalışma mevcuttur. Bu tez çalışmasında önerilen regresyon tabanlı model ile diğer benzer çalışmalar Çizelge 2.1'de karşılaştırılmıştır. Gen ifade tahminine yönelik ilk çalışma 2004 yılında Beer ve Tavazoie tarafından yapılmıştır. Bu çalışmada, Hartigan tarafından 1975 yılında önerilen k-means kümeleme yöntemi ile mRNA başlatıcı dizileri kullanılarak gen ifadeleri tahmin edilmiştir. Burada gen ifadesi belirleyicisi, DNA sekans örüntülerine

bağlanan transkripsiyon faktörlerdir ve motif olarak adlandırılmışlardır. mRNA ifade tahminleri düzenleyici sekanslar kullanılarak gerçekleştirilmiştir. Verinin gürültü içermesinden dolayı analiz işlemlerinde bazı kısıtlar meydana getirdiği ifade edilmiştir. Performans değerlendirmesi için veri setinin %80'i eğitim ve %20'si test amaçlı kullanılarak bu işlem 5 kez tekrarlanmıştır (5-fold cross validation) [27]. Beer ve Tavazoie'nin çalışması, 2007 yılında Naive Bayes temelli daha basit bir sınıflandırma ile tekrarlanmış ve kestirim doğruluğunun %10 arttığı görülmüştür. Burada kullanılan özniteliklere, ki-kare testi ile belli bir eşik değere göre belirlenen ikili (binary) değer atanmıştır. Bu şekilde bir yaklaşım işlem yükü açısından çok kolaylık sağlasa da transkripsiyon faktörlerinin DNA'ya bağlanma sürecini temsil etmede yetersiz kalmaktadır. Genlerdeki her bir motif ile büyük oranda eşleşen bağlanma bölgeleri için yön ve pozisyon bilgileri kullanılmıştır [28].

Çizelge 2.1 Önceki benzer yöntemlerin karşılaştırması

Çalışma	**Tahmini yapılan Molekül**	**Tahmini yapılan parametre**	**Girdi parametresi**	**Yöntem**	**Kullanılan öznitelik**	**Kullanılan teknoloji**
Beer ve Tavazoies 2004	mRNA	Binary (ifade var mı yok mu)	mRNA başlatıcı dizisi	Sınıflandırma	TFBS	Mikrodizi
Yuan ve diğerleri, 2007	mRNA	Binary (ifade var mı yok mu)	mRNA başlatıcı dizisi	Naive Bayes Sınıflandırma	TFBS	Mikrodizi
Ogul ve Tuncer, 2015	miRNA	İfade tam değeri	miRNA başlatıcı dizisi	İlişkisel Vektör Makineleri Regresyonu	TFBS	Mikrodizi

*TFBS: Transkripsiyon faktör bağlanma bölgesi

Yukarıdaki çizelgede gösterilen ilk iki çalışmada, mRNA başlatıcı dizisi kullanılarak mRNA ifadesinin tahmin edilmesi amaçlanmıştır. Bunlara benzer olarak bir başka çalışmada ise miRNA gen ifadesi tahmin edilmeye çalışılmıştır. Burada ise miRNA başlatıcı dizisi ve regresyon temelli bir model kullanılmıştır. Yalnız bu çalışmada farklı olarak miRNA ifade tam değerleri tahmin edilmeye çalışılmıştır. Bir genin veya miRNA'nın ifade verip vermediği sınıflandırma problemi olarak tanımlanırken, ifade

tam değerinin tahmin edilmesi regresyon problemi olarak ele alınmaktadır. Çalışmada farklı kanser tipinde 255 birey için 217 insan miRNA'sı ve bunların mRNA üzerinde tanımlı başlatıcı dizilerine ait elde edilmiş mikrodizi verileri kullanılmıştır. Kullanılan bu veri kümesinde 163 sağlıklı ve 92 hastalıklı dokulardan elde edilen miRNA ifade verileri bulunmaktadır. En iyi ifade tahmini performansına akciğer kanseri dokularından elde edilmiş örneklerde ulaşılmıştır (%80). k-NN ve doğrusal regresyon modelleri kötü performans gösterirken, RVM regresyonu ile daha iyi kestirim performansı elde edilmiştir. Tüm sağlıklı dokular için ortalama Spearman benzerlik katsayısı 0.68 olarak hesaplanmıştır [29].

2.2.2. Ölçüm yöntemleri

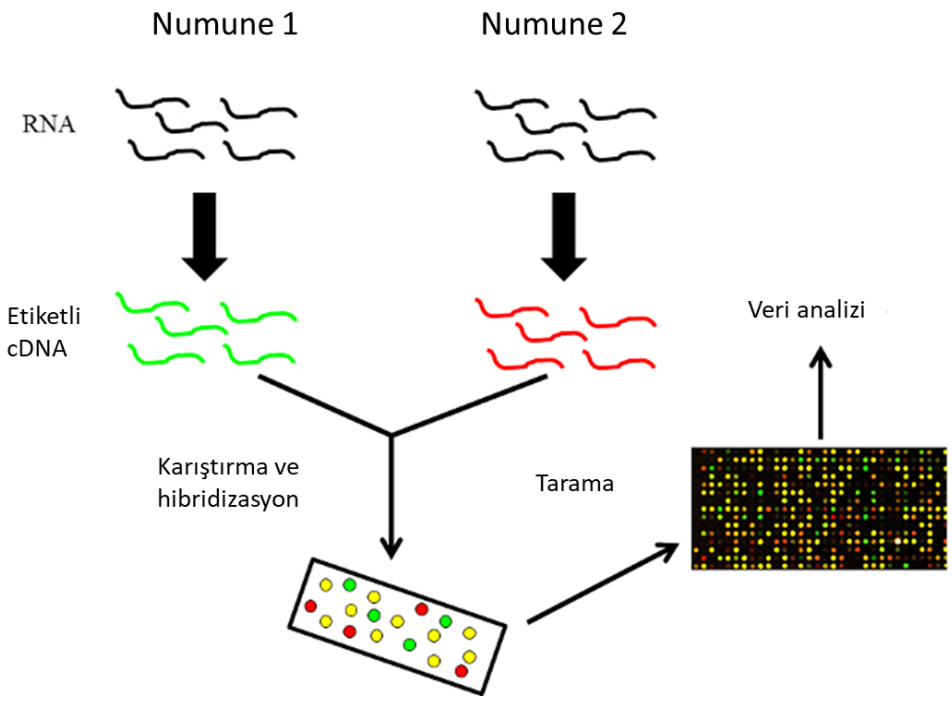
İlk gen ifade miktarı ölçüm çalışmalarında kullanılan Northern blotlar ve kantitatif polimeraz zincir reaksiyonu gibi yöntemler tek transkript ölçümü ile sınırlıydı. Son 20 yıldır gen ifadesinin daha kantitatif ve daha detaylı transkriptomik ölçüme yönelik araştırmalar devam etmektedir. İlk transkriptomik çalışmalar Schena ve ekibi tarafından 1995 yılında ortaya konmuş ve daha sonra yerini mikrodizi teknolojisine bırakmıştır. Mikrodizi teknolojisinin yaygınlaşması, hem moleküler biyoloji alanında hem de biyoinformatik alanında önemli çalışmaların literatüre kazandırılmasında bir dönüm noktası olmuştur. Transkriptomiks (transcriptomics) olarak tanımlanan çok sayıda gen ifade miktarı ölçümünü sağlayan mikrodizi teknolojisinin de bazı kısıtları bulunmaktadır. Tarama yapılacak dizilerin önceden bilinmesi gerektiği, çok benzer dizilerin analizinde çapraz hibridizasyon gürültülerinin olması, çok az veya çok fazla miktarda ifade veren genlerin kantitatif tayinindeki zorluklar bu kısıtlardan bazılarıdır. Hibridizasyona dayalı metotların aksine transkriptomu açıklayabilmek için transkript sekansını doğrudan belirleyen dizileme dayalı yaklaşımlar geliştirilmiştir. İlk olarak, tamamlayıcı DNA'nın (cDNA) Sanger dizilimi ile ifade veren sekans etiketi kütüphanelerinin üretilmesi, gen ifade çalışmalarında kullanılmıştır, ancak bu yaklaşım, nispeten düşük verimlidir ve transkriptleri ölçmek için ideal değildir. Bu yöntemde kantitatif ölçümü yapılan etiketli dizi miktarının mRNA transkript miktarına karşılık gelmesi önemli bir avantaj sağlarken gen keşfinde kullanışlı değildir. Ayrıca, dizi etiketlerinin zahmetli şekilde klonlanması, yüksek otomatik Sanger dizilimi maliyeti ve büyük miktarlarda başlangıç RNA gerekliliği bu

yöntemin kullanımını büyük ölçüde sınırlamaktadır. Bu kısıtlar nedeniyle yüksek verimli yeni nesil dizileme olarak bilinen RNA-seq teknolojisi ortaya çıkmıştır [30].

2.2.2.1. Mikrodizi teknolojisi

Moleküler biyoloji ve genetik alanındaki araştırmalar ilerledikçe ve moleküller arası etkileşimin önemi fark edildikçe aynı anda birden fazla genin aktivitesine bakabilmenin daha faydalı olabileceği düşüncesi yaygınlaşmaya başlamıştır. SAGE (Serial Analysis of Gene Expression) yöntemi bu arayışlar doğrultusunda 1995 yılında ortaya çıkmıştır [31]. Burada, bir hastalık durumunda gen ifade düzeylerinin (expression level) sağlıklı bireylerin gen ifade düzeylerine göre nasıl değiştiği ve böylelikle hastalığın genomik nedeni veya hastalığın neleri etkilediği/değiştirdiğinin anlaşılması hedeflenmiştir. SAGE yönteminin bir diğer avantajı ise hücredeki transkriptlerin ne olduğunu önceden bilmenizi gerektirmeyen ve yeni genlerin keşfine olanak sağlayan bir yaklaşıma sahip olmasıdır. SAGE metodu DNA dizilim işlemine dayanır ve bu metodun keşfedildiği dönemde en iyi DNA dizilim yaklaşımı Sanger yöntemidir. Bu yöntemde, eğer dizilim işlemi yapılmak istenen DNA bölgesi fazla uzunsa dizilim işlemi hem uzun süreler hem de yüksek maliyetler gerektirmektedir. Bu nedenle, yine aynı dönemde geliştirilen mikrodizi teknolojisi daha düşük maliyetler vadettiği için bir anda popüler hale gelmiş ve SAGE teknolojisinin yerini almıştır. Her iki teknoloji karşılaştırıldığında, SAGE yönteminin mikrodizi teknolojisine göre çok daha kesin ve nicel sonuçlar verebildiği görülmektedir.

Mikrodizi teknolojisi; binlerce farklı genin ifade miktarlarının aynı anda ölçülebilmesi, hızlı bir yöntem olması, hasta ve sağlıklı hücrelerdeki genlerin ifade miktarlarının karşılaştırılmasına olanak vermesi ve hastalıkların alt-gruplar halinde kategorize edilebilmesi gibi avantajlara sahiptir. Bunun yanında tek bir seferde çok fazla veri analizi yapıldığından, tüm sonuçların analizinin zaman alması, gen ifade profillerinin yorumlamak için oldukça kompleks olabilmesi, sonuçların yeterince kantitatif olmaması ve hala oldukça pahalı bir teknoloji olması gibi bazı dezavantajlara da sahiptir [32]. Bu teknolojide numunenin konulduğu çipler cam, silikon veya plastik malzemeden yapılmaktadır. Şekil 2.4'te mikro dizi teknolojisinin basamaklarının şematik bir gösterimi yer almaktadır [33].



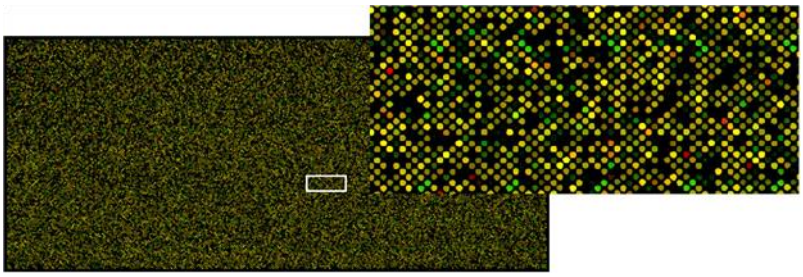
Şekil 2.4 Mikrodizi teknolojisi

Mikrodizi teknolojisinde, DNA probları ile immobilize edilmiş diziler, tamamlayıcısı olan hedef dizilere yönlendirilmekte ve hibridizasyon derecesi ölçülmektedir. Bu teknoloji ile enzim-substrat, DNA-protein, protein-protein etkileşimleri araştırılmaktadır [34].

DNA çipleri bakteriler, mayalar, bitkiler ve insanlar dâhil olmak üzere pek çok organizmadaki farklı genlerin ifade seviyelerinin izlenmesi için kullanılmaktadır. Nöropsikiyatri alanında hastalıklı ve sağlıklı bireylerin karşılaştırılmasına yönelik çalışmalar da vardır [35]. Mikrodizi teknolojisinin geliştirilmesinden [32] sonra ilaç keşfi ve tanıları, mutasyon analizleri, farmogenomik uygulamalar, moleküler etkileşimler ve kanser gibi hastalıklar ile ilgili literatürde çok sayıda derleme çalışmaları mevcuttur [25].

Bir hücre içindeki hangi genin aktif veya pasif olduğunu belirlemek için öncelikle hücre içindeki mRNA'lar toplanır. Toplanan bu mRNA'lardan ters transkriptaz (reverse transcriptase) enzimleri ile tamamlayıcı DNA (complementary DNA, cDNA) elde edilir. Bu süreçte floresan ile işaretlenmiş nükleotidler cDNA'ya bağlanır. Her

farklı örnek farklı renkteki floresan boya ile etiketlenir. Sonra etiketlenmiş olan cDNA'lar DNA mikrodizi üzerine yerleştirilir. mRNA'ları gösteren her bir etiketlenmiş cDNA; mikrodizi üzerinde bulunan suni olarak hazırlanmış tamamlayıcı DNA'lara bağlanırlar. Böylece floresan etiketlerini bırakırlar. Bir gen çok aktif ise fazla mRNA üretir, fazla sayıda etiketlenmiş cDNA elde edilir ve çok parlak floresan noktaları tespit edilir. Eğer hiç floresan nokta yoksa genin hiç mRNA zinciri üretmediği yani pasif olduğu anlamına gelmektedir [36].



Şekil 2.5 Mikrodizi floresan görüntüsü

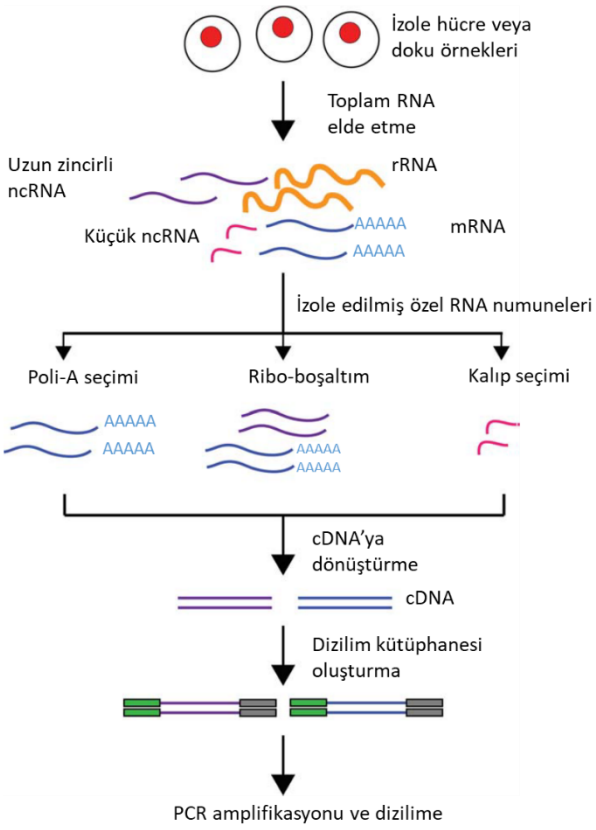
2.2.2.2. Yeni Nesil Dizileme Teknolojisi

Yeni Nesil Dizileme (New Generation Sequencing, NGS) Teknolojisi, gen aktivitesinin ölçümünde önceki yöntemlere kıyasla çok daha fazla kantitatif bilgiler sunmaktadır. Ayrıca alternatif zincirleşme ve gen çiftine (allel) özel ifade verme gibi detaylı bilgiler sunar. RNA dizileme (RNA-seq) olarak adlandırılan bu yöntem, önceki yaklaşımlara göre belirgin avantajlara sahiptir ve transkriptomun karmaşık ve dinamik doğasını anlamada devrim yaratmıştır. RNA-seq, gen ifadesi, alternatif ekleme ve alel özel ifadenin daha ayrıntılı ve niceliksel bir görünümünü sağlar. Bu yöntem daha önce anlatılan mikrodizi teknolojisi ve Sanger dizileme yaklaşımındaki birçok zorluğu ortadan kaldırmaktadır.

Tipik bir RNA-seq işleminde; öncelikle RNA örnekten izole edilir. Bu yöntemin başarılı bir şekilde tamamlanması, dizilim kütüphanesinin oluşturulacağı RNA'nın yeterli kalitede olması ile mümkündür. RNA kalitesi biyoanalizör cihazı ile ölçülmektedir. Bu cihaz 1 ila 10 arasında RNA Bütünlük Numarası (RNA Integrity Number, RIN) üretmektedir. Bu cihaz en az bozulmuş RNA için 10 sayısını

vermektedir. RIN, jel elektroforezi ve 28S ila 18S ribozomal bantlarının oranlarının analizini kullanarak numune bütünlüğünü tahmin eder. RIN ölçümleri memeli organizmalar için geçerli olup anormal ribozomal oranlarına sahip canlılar için hatalı değerler içermektedir. Düşük kaliteli RNA (RIN <6), dizileme sonuçlarını (örneğin düzensiz gen kapsamı, 3′ – 5 ′ transkript önyargısı, vb.) büyük ölçüde etkileyebilir ve hatalı biyolojik sonuçlara yol açabilir.

RNA-seq yönteminde, öncelikle doku ve hücre gibi biyolojik materyalden RNA elde edilir. Sonra RNA alt molekülleri, poli-A ve ribo-depletion gibi protokollere göre izole edilir. Daha sonra RNA, ters transkripsiyon ve dizilim adaptörlerinin cDNA fragmanların sonuna bağlanması suretiyle tamamlayıcı DNA (cDNA) 'ya dönüştürülür. PCR ile amplifiye edilmesinin ardından bir RNA-seq kütüphanesi dizilim işlemi için elde edilmiş olur [30] (Şekil 2.6).



Şekil 2.6 Yeni nesil dizileme adımları

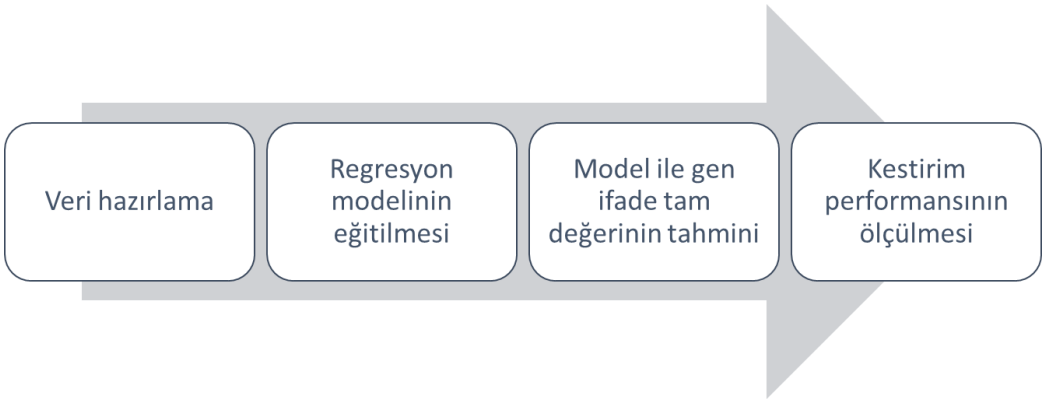
2.2.3. Gene Expression Omnibus Veritabanı

Ulusal Biyoteknoloji Bilgi Merkezi (NCBI - The National Center for Biotechnology Information), Sağlık Ulusal Enstitülerinin (National Institutes of Health) bir kolu olan Birleşmiş Devletler Ulusal Tıp Kütüphanesi'nin (NLM – United States National Library of Medicine) birimlerinden biridir. Gene Expression Omnibus (GEO) veritabanı, NCBI bünyesinde yer almaktadır [37]. Bu çalışmada kullanılan tüm gen ifade matrisleri GEO veritabanından elde edilmiştir. Bu veritabanında toplam 4.348 veri seti, gen ölçümünde kullanılan 18.317 platform ve 2.443.376 ölçüm yapılan örnek yer almakta olup her geçen gün bu sayılar artmaktadır. Biyoinformatik alanında yaygın kullanılan ve halka açık olan bu genomik veritabanı mikrodizi ile yeni nesil dizileme yöntemleri kullanılarak elde edilen verileri içermektedir. Bu veritabanında, farklı hastalık türlerine ait dokulardan elde edilen ve diğer kriterlere (ilaç etkileşimi, cerrahi işlem vb.) ilişkin elde edilen gen ifade matrisleri ve diğer genomik veriler yer almaktadır [38]. Bu tez çalışmasında hazırlanan veri kümeleri, sadece insan dokularından elde edilen veri setlerinden alınmıştır.

2.3. Veri Hazırlama

Çalışmada kullanılan tüm veriler bir önceki bölümde anlatılan GEO veritabanından elde edilmiştir. Bu veritabanında veri setleri GSE ile başlayıp numerik olarak devam eden bir tanımlayıcı kod ile saklanmaktadır. Her veri seti içinde verinin özelliklerini özetleyen metadata, verinin nasıl ve hangi koşullarda elde edildiğini anlatan tanımlayıcı metin ve text dosya biçiminde veri matrisi yer alır. Bu text dosyasında, açıklama bölümünden sonra gen ifade matrisleri yer almaktadır. Bu matriste satır başlarında ölçüm probu tanımlayıcı numaraları ve sütun başlıklarında ise deney tanımlayıcı numaraları yer almaktadır. Bu text dosyasının Notepad+ ile açılması verilerin düzgün bir şekilde kopyalanmasını kolaylaştırmaktadır. Ayrı bir Excel dosyasına kopyalanan veri matrisinde düşey eksendeki probların hangi genlere karşılık geldiğini bulmak gerekir. Bunun için bazı ara yüzlerin veya araçların kullanılması gerekmektedir. Birden fazla probun aynı gene karşılık geldiği durumlarda prob ölçümlerinin ortalaması alınarak karşılık gelen genin ifade vektörü olarak kaydedilmiştir. Bir gen ifade vektöründe kayıp verilerin yerine aynı genin ifade vektöründeki diğer deneylere ait ifade değerlerinin ortalaması atanmıştır. Böylece bu kayıp verilerin aykırılık yaratarak veri bütünlüğünün bozmaması ve ortalama gen

ifade değerini etkilememesi sağlanmıştır. Literatürde uygulanan diğer bir yaklaşım ise bu kayıp verilere sıfır değerinin atanmasıdır. Böylece o örnek için ilgili genin ifade vermediği kabul edilmiş olur. Mikrodizi teknolojisi optik ölçüme dayalı olduğundan farklı aralıklarda ölçümler içerebilmektedir. Bu nedenle literatürde çoğunlukla analize başlanmadan önce bir normalizasyon işleminin yapıldığı görülmektedir. Şekil 2.7'de gen ifade tahmini uygulama adımları gösterilmektedir.



Şekil 2.7 Gen ifade tahmini için temel uygulama adımları

Yukarıda anlatıldığı gibi öncelikle veritabanından elde edilen veri kümeleri, uygun dönüştürücü ve normalizasyon işlemleri ile işlenmeye hazır hale getirilmektedir. Daha sonra eğitim ve test verileri oluşturulur. Eğitim ve test verilerinin seçimi literatürde yer aldığı gibi farklı koşullara ve miktarlara bağlı olarak değişebilmektedir. Örneğin bazı çalışmalarda verinin %80'i eğitim ve %20'si test amaçlı kullanılırken bu çalışmada olduğu gibi bazı uygulamalarda leave-one-out (jackknife) prosedürü uygulanır. Bu prosedürde her defasından bir gen ifade vektörü test amacıyla dışarıda bırakılır ve geri kalan gen ifade vektörleri ile model eğitilir. Bu işlem toplam veri sayısınca tekrarlanır. Bu çalışmada her bir genin ifade vektörü test ve diğer gen ifade vektörleri eğitim için kullanılmış olup bu işlem toplam gen sayısınca tekrarlanmıştır. Kestirim performansının ölçümü ve gösterimi için farklı yöntemler kullanılabilir. Bu çalışmada; kestirim performansının değerlendirilmesi için Spearman benzerlik katsayısı, Pearson benzerlik katsayısı ve ortalama hata kareleri toplamının karekökü (Root Mean Squared Error-RMSE) kullanılmıştır. Kestirim performansının gösterimi için ise literatürde yeni sayılabilen bir gösterim biçimi olan

Spearman, Pearson ve RMSE değişim eğrileri ve saçılım grafiği (scatter plot) kullanılmıştır.

Bilgisayar ortamında yapılan tüm kestirim çalışmalarında regresyon modellerinin uygulanması için MATLAB programının Olasılıksal Modelleme Aracı (Probabilistic Modeling Toolkit) [39] ve RVM regresyon modeli için Sparse Bayes paketi [40] kullanılmıştır.

3. MİKRODİZİ KAYIP VERİ KESTİRİMİ

3.1. Giriş

Protein kodlayan genlerin ve bunları düzenleyen diğer nükleotid dizilerinin ifade miktarlarının ölçülebilmesi için farklı teknolojileri kullanan ölçüm sistemleri geliştirilmiştir. Bu sistemler, hücrede gen ve genleri düzenleyen diğer moleküllerin aktivitelerini kantitatif ölçümlerle analiz edilmesinde ve fenotip ile ilişkilendirilmesinde önemli yere sahiptir. Mikrodizi teknolojisi, Schena ve diğerleri [32, 41] tarafından icat edilen, binlerce gene ait ifade düzeylerinin izlenmesi, klinik sonuçların ve kansere bağlı hücresel süreçlerin anlaşılması için gen ifadesi ölçümünde kullanılan yaygın bir yöntemdir. Bu teknoloji ile elde edilen gen ifade ölçümlerinde sinyal gürültüsü ve deneysel hatalardan kaynaklanan veri kaybı söz konusu olabilmektedir. Bu yöntem, tek bir numunede binlerce genin aynı anda analiz edilmesini sağlar, ancak deneysel hataların meydana gelmesi durumunda maliyet etkinlik açısından tekrar edilmesi mümkün değildir [41]. Bu kayıp verilerin tahmin edilmesine yönelik literatürde farklı yaklaşımlar söz konusu olup bu problem kayıp veri atama (missing value imputation) olarak adlandırılmaktadır. Son yıllarda, gen ifade tahmininde yeni hesaplama yaklaşımları ile kestirim modelleri oluşturmak biyoinformatik alanında başlı başına bir konu haline gelmiştir. Literatürde farklı kanser tiplerinin tespiti ve prognozu için sınıflandırma ve kümeleme gibi çeşitli yaklaşımlar bulunmaktadır. Bu yaklaşımlar arasında gen ifade profillerinden kanser türlerini tahmin etmek için gözetimsiz (unspervised) [42,43] veya gözetimli (supervised) [44,45] yöntemlerin önerildiği çalışmalar mevcuttur.

1970'li yıllardan itibaren kayıp gen ifade değerlerini tahmin etmek için birçok istatistiksel yöntem geliştirilmiştir. Liew ve diğerleri tarafından kestirim performansı, veri yapısı, uygulanan metodoloji ve öğrenme yöntemi gibi çeşitli koşullar değerlendirilerek yaygın kullanılan bazı algoritmalar kapsamlı olarak karşılaştırılmıştır. Gen ifade matrislerinde satırlar, ifade miktarı ölçülen genleri ve sütunlar ise farklı durum veya hastalıklara ait deneyleri temsil etmektedir. Satırların temsil ettiği M farklı genin N farklı durum veya hastalık için ölçülen N adet ifade miktarından oluşan gen ifade vektörleri arasındaki korelasyon hücredeki moleküler süreçler arasındaki benzerliği gösterirken, sütunlarda yer alan farklı durum veya hastalıklara ait M adet gen ifade miktarından oluşan düşey vektörler arasındaki

korelasyon ise aynı çevresel koşullara veya hastalığa farklı genlerin verdiği yanıtların benzerliğini göstermektedir [46].

Beer ve Tavazoie tarafından; mRNA'ların ifade verip vermediğini belirlemek için Bayes tabanlı bir sınıflandırma modeli geliştirilmiştir [27]. Başka bir çalışmada ise Yuan ve diğerleri tarafından Naïve Bayes tabanlı daha az karmaşık bir model kullanılarak aynı sonuçlara ulaşılmıştır [28]. Bir genin tam değerinin tespit edilmesi daha ileri düzey meta analiz çalışmaları için önemlidir. Öte yandan, gen ifadesi tam değerini tahmin etmek için literatürde daha önce yapılmış bir çalışmaya rastlanılmamıştır. Ogul ve Tuncer tarafından farklı kanser tiplerine ait dokulardan elde edilen miRNA ifade tam değerlerini tahmin etmek için farklı regresyon modellerinin karşılaştırıldığı bir çalışma yapılmıştır [29]. Bunun dışında, gen ifade matrisinde bir genin ifade verip vermediğini tespit eden birçok çalışma vardır [47, 48]. Literatürde, model öğrenmede küresel (global), yerel (local), karma (hybrid) ve bilgi tabanlı (knowledge-based) olmak üzere dört tip yaklaşım vardır. Model öğrenmede tüm gen veri kümesinin kullanıldığı küresel yaklaşımda hesaplama süreleri çok daha uzundur. Bu nedenle, gen ifade değerinin tahmin edilmesinde düşük zaman maliyetli olan yerel yaklaşımların kullanılması daha uygun görülmektedir. Yerel yaklaşımda; kayıp gen ifade miktarlarını tahmin etmek için tüm gen ifade matrisindeki verilerin kullanılması yerine bir alt gen kümesine ait ifade verileri model öğrenmede kullanılır [46]. Burada alt gen kümesinin tespiti için kullanılacak öznitelik azaltma (reduction) veya seçme (selection) yöntemlerinin iyi belirlenmesi gerekmektedir. Bir diğer çalışmada, Destek Vektör Makinesi (Support Vector Machine - SVM) regresyonu ve genetik algoritmanın birleştirildiği bir karma model sunulmaktadır. Bu çalışmada kestirim performans değerlendirme ölçütü olarak Hata Karelerinin Ortalamasının Karekökü (Root Mean Squared Error-RMSE) ve kayıp oranı (missing rate) parametreleri kullanılmıştır [49].

Tez çalışmasının bu bölümünde, farklı regresyon modelleri kullanılarak 55 meme, 53 kolon ve 11 prostat kanseri olmak üzere toplam 119 hastaya ait kanser dokusundan elde edilen mRNA ifade tam değerleri tahmin edilmiştir. Daha önce de belirtildiği gibi bir genin ifade miktarının tam değerinin tahmin edilmesi regresyon problemi olarak ele alınmıştır. Bu kapsamda doğrusal, k en yakın komşu (k-NN) ve

RVM regresyon modelleri kullanılmıştır. Burada tüm genlerin her bir örneğe ait ifade miktarları sırasıyla kayıp veri olarak varsayılarak kestirim çalışmaları yapılmıştır. Bu sayede her bir gen ifade değeri kayıp veri gibi düşünülerek tahmin edilmiş olup deneylerde ölçülen gerçek değerlerle karşılaştırılmıştır.

3.2. Materyal ve Yöntem

3.2.1. Veri

Çalışmada 55 meme kanseri, 53 kolon kanseri ve 11 prostat kanseri hastalarına ait veri kümeleri kullanılmıştır. Bu veri kümeleri GSE75285, GSE18088 ve GSE45016 erişim numaraları ile GEO veritabanından ulaşılan veri setlerinden elde edilmiştir [50-52]. Bu deneylerdeki gen ifade miktarı ölçümleri Affymetrix Human Genome U133 Plus 2.0 Array'inin biyoçipi kullanılarak GPL570 platformu ile yapılmıştır [53]. Bu mikrodizi veri setlerinde 54675 gen tanımlayıcı probu ile yapılan ölçümler yer almakta ve birden fazla gen probu mikrodizi teknolojisinde sadece bir mRNA'yı ifade edebilmektedir. Bu nedenle, dönüştürücü araçları (DAVID Bioinformatics Resources 6.0037, www.ensemble.org) kullanılarak problara karşılık gelen mRNA'lar tespit edilmiş ve rastgele seçilen 1000 gene ait ifade değerlerinden oluşan bir veri kümesi hazırlanmıştır [54, 55].

Çok sayıda prob ölçümü olması ve veritabanı sorgu kısıtlamalarından dolayı veri tabanlarından daha hızlı veri çekebilmek için R yazılımı (versiyon 3.2.5) kullanılmıştır. Gen profili çıkarımında kullanılan biyoçipin referans noktasına göre farklı deneylerin farklı ölçüm aralıkları olabilir. Bunun için ifade değerleri, min-max normalizasyonu kullanılarak 0 ve 1 aralığında normalize edilmiştir. Çizelge 3.1'de veri kümelerinin elde edildiği veri setlerine ilişkin GEO veritabanı erişim numaraları, örnek sayıları ve deney yapılan ölçüm platformları yer almaktadır.

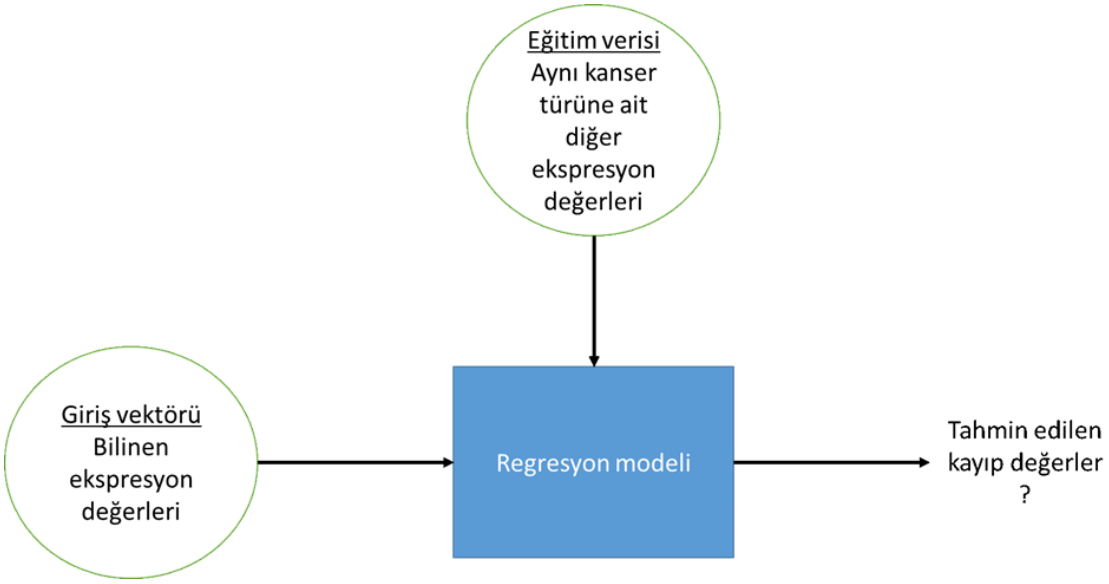
Çizelge 3.1 Kullanılan veri setleri

Veriseti Erişim Numarası	Hasta Sayısı	Platform
GSE18088 (Kolon Kanseri)	53	Affymetrix Human Genome U133 Plus 2.0 Array
GSE75285 (Meme Kanseri)	55	Affymetrix Human Genome U133 Plus 2.0 Array
GSE45016 (Prostat Kanseri)	11	Affymetrix Human Genome U133 Plus 2.0 Array

3.2.2. Yöntem

3.2.2.1. Genel Çerçeve

Gen ifadesi tam değerinin tahmin edilmesi için doğrusal, k-NN ve İlişkisel Vektör Makinesi regresyon modelleri kullanılmıştır. Şekil 3.1'deki şemada bu bölümde izlenen yöntem özetlenmektedir. Buradaki amaç bir kanser tipinin gen ifadesinin bilinen diğer ifade değerleri kullanılarak tahmin edilip edilmeyeceğini göstermektir. Giriş vektörü, aynı kanser tipine ait bilinen mRNA ifadelerini içerir.



Şekil 3.1 Gen ifade tahmini için genel çerçeve

$G\ (g_i, s_j)$ i. gen ve j. örneğe ait ifade miktarı olsun. Örneğin; gen ifade matrisinde 1. genin 1. örneğe ait ifade miktarı $G\ (g_1, s_1)$ olur. Bu durumda; 1.iterasyonda $G\ (g_1, s_1)$ ifade miktarının tahmin edilmesi için $G\ (g_1, s_2) \ldots G\ (g_1, s_N)$ ifade vektörü test verisi olarak kullanılmaktadır. Bu ilk vektör dışında kalan $(M-1)\ x\ (N-1)$ boyutlu $G\ (g_2, s_N) \ldots G\ (g_M, s_N)$ veri matrisi ise model öğrenmede kullanılmaktadır. Bu uygulama biçimi leave-one-out (Jackknife) prosedürü için geçerlidir. k-fold veya 10-fold geçerli kılma yöntemleri için test ve eğitim veri setlerinin büyüklüğü değişmektedir.

3.2.2.2. Doğrusal Regresyon

Regresyon yönteminde, açıklanmış bağımlı değişken (ifade miktarı bilinen mRNA) ve bağımsız değişken (ifade miktarı tahmin edilecek mRNA) arasındaki ilişki bir model ile ortaya konulmaktadır. Doğrusal regresyon, bilinen ifadeler arasında doğrusal bir denklem oluşturur ve modelin parametrelerini eğitim verilerini kullanarak elde eder. Eğitim verisinde i. örneğin j. mRNA ifade değeri ile diğer örneklerin j. mRNA ifade değerleri arasında bir doğrusal denklem oluşturulur. N adet farklı koşul veya hasta sayısı için regresyon modelinin giriş vektörü $x = [x_1, x_2, ..., x_N]$ ile ifade edilebilir. Burada x vektörü aynı kanseri taşıyan farklı hastalara ait dokuların diğer mRNA ifadelerinden oluşur. Bilinmeyen bir x girdisi için $y = W^T x + c$ değeri tahmin edilir. Burada W bir parametre vektörüdür ve c, W'ye eklenen bir ofset sabitidir. Görüldüğü gibi doğrusal regresyon yönteminin ilk amacı, eğitim verisini kullanarak bir regresyon doğrusu uydurmaktır. Bu doğru, yukarıda belirtilen parametrelerden oluşmakta ve diğer örneklere ait mRNA ifade değerleri oluşturulan bu modele girdi olarak verilerek hedef mRNA değeri tahmin edilmektedir. Doğrusal bir regresyon modeli oluşturabilmek için genellikle en küçük kareler yaklaşımı kullanılır. Bu yaklaşım, her bir veri noktasının doğruya olan dikey sapmalarının karelerinin toplamını en aza indirerek giriş verileri için en uygun doğruyu hesaplar [29].

3.2.2.3. k-NN Regresyonu

Bu regresyon modelinde bir genin ifade değeri; eğitim setindeki k en yakın gen ifadesinin ortalaması alınarak uzaklığın hesaplanması ile tahmin edilmektedir. Gen ifade dizileri arasındaki uzaklık her bir p_i ve p_j dizi elemanlarının farkının alındığı Öklid denklemi (Eşitlik 1) ile hesaplanır. Bu yöntem k-NN regresyonu olarak adlandırılır.

$$d = \sqrt{\sum (p_i - p_j)^2} \qquad (1)$$

3.2.2.4. İlişkisel Vektör Makinesi Regresyonu

İlişkisel Vektör Makinesi (RVM) bir öğrenme metodu olarak ilk defa Tipping [56] tarafından örüntü tanımada regresyon ve sınıflama problemlerini çözmek için Bayesian çıkarımı kullanılarak ortaya konulmuştur. Doğrusal olmayan bir ilişkiyi göstermek için $y = W^T \emptyset(x)$ şeklinde bir fonksiyon tanımlanabilir. Burada $\emptyset(x)$, girdi değişkenler arasında doğrusal olmayan bir haritalama yapar ve doğrusal, radyal temelli veya polinom gibi bir çekirdek fonksiyonlar ile birlikte uygulanır. Temel nokta W parametresini tahmin etmektir. Her bir hedef değer $t_i = w^T \phi(x_i) + \varepsilon_i$ ile ifade edilir. Burada ε_i gürültü veya hata, x_i ise giriş değerleridir. ε_i gürültüsü, Gauss gürültüsünden bağımsız olarak kabul edilir. RVM, w parametresi üzerinden sıfır ortalamalı Gauss dağılımını tercih eder;

$$P(w, \alpha_i) \sim N(0, \alpha_i^{-1}) \quad (2)$$

Burada α_i; her bir w_i'nin ters varyansını ifade eder. Olasılık aşağıdaki denklemle ifade edilebilir;

$$P(w, \alpha, \sigma^2 | t) = P(w | t, \alpha, \sigma^2) P(\alpha, \sigma^2 | t) \quad (3)$$

Burada eşitliğin solundaki kısım $P(w|t, \alpha, \sigma^2) = N(m, \Sigma)$ şeklinde tekrar yazılabilir. Burada m ortalama ve kovaryans Σ ; $\mathrm{m} = \sum \Phi^{\mathrm{T}} \mathrm{t} / \sigma^2$ ve $\Sigma = (\mathrm{A} + \Phi^{\mathrm{T}} \Phi / \sigma^2)^{-1}$ denklemleri ile hesaplanabilir. Buradaki A değeri $\mathrm{A} = \mathrm{diag}(\alpha)$ ile bulunur.

RVM regresyon ile durma kriterine ulaşılıncaya kadar σ^2 değeri sürekli tahmin edilir. Yeni bir giriş vektörüne göre $t = m^T \phi(x')$ ile kestirim yapılır. Bu çalışma kapsamında doğrusal çekirdek fonksiyonu $K = X_1 \cdot X_2$ ile $K = \left(1/\sqrt{2pi\sigma^2}\right) e^{-\frac{\|x_1, x_2\|}{2\sigma^2}}$ (RBF-1) ve $K = e^{-\gamma \|x_1, x_2\|}$ (RBF-2) radyal tabanlı fonksiyonlar kullanılmıştır. Burada σ Gauss dağılımın genişliğini ve $\boldsymbol{\gamma}$ ise standart sapmanın tersini ifade etmektedir. En iyi kestirime erişebilmek için bu değerler ile optimizasyon sağlanır.

3.2.2.5. Performans değerlendirme

Geçerli kılma yöntemi olarak Jackknife (leave-one-out, LOO) prosedürü kullanılmıştır. Kestirim yapan bir modelin rastgele etki performansının tahmin edilmesi; çok sayıda değişken içeren veri tabanları için önemli bir sorun haline gelmektedir. Bu problemden dolayı, istatistik tabanlı modeller aşırı öğrenmeye (over-fitting) karşı eğilimlidir. LOO prosedürünün amacı, bir modelin yeni bir öznitelik vektörüne uygulandığında nasıl performans göstereceğini tahmin etmektir. Bu yöntemde, sırasıyla her mRNA ifade değeri dışarıda bırakılarak model öğrenmede geriye kalan diğer mRNA ifade değerleri kullanılır. Model daha sonra dışarıda bırakılan mRNA ifade değerini tahmin etmek için kullanılır. Bu yöntemle, her mRNA ifade değeri tahmin edilir. Bu nedenle, tüm verilerin tamamı alınarak uygulanan en iyi geçerli kılma yöntemi olarak kabul edilir.

Kullanılan regresyon modellerinin kestirim performanslarını ölçmek için Spearman ve Pearson korelasyon katsayıları kullanılmıştır. Bu katsayılar istatistik tabanlı olup gerçek ve tahmin değerlerinden oluşan A ve B vektörleri arasındaki korelasyonu göstermektedir. Spearman korelasyon katsayısı; iki vektör sıralaması arasındaki korelasyonu gösterirken (Eşitlik 4); Pearson korelasyon katsayısı ise iki vektörün korelasyonunun doğrudan ölçümünü ifade eder ve Eşitlik 5'teki formül ile hesaplanır.

$$d_{Spearman}\left(A,B\right)=d_{Pearson}\left(Rank\left(A\right),Rank\left(B\right)\right) \quad (4)$$

$$d_{Pearson}\left(A,B\right)=1-\frac{1}{n-1}\sqrt{\frac{\left(A_i-\mu_A\right)\left(B_i-\mu_B\right)}{\sigma_A\sigma_B}} \quad (5)$$

Burada n gen sayısını, μ gen ifadesi değerlerinin ortalamasını ve σ standart sapmayı göstermektedir. Spearman korelasyon katsayısının hesaplanabilmesi için öncelikle Pearson korelasyon katsayısının hesaplanması gerekmektedir.

Her iki korelasyon katsayısı -1 ile 1 arasında değişmektedir. Korelasyon katsayısı 0 değeri olursa vektörler arasında hiçbir korelasyon olmadığı anlamına gelmektedir. Bu korelasyon katsayılarının değerlendirme kriteri Çizelge 3.2'de gösterilmektedir [57].

Çizelge 3.2 Korelasyon katsayılarının değerlendirme kriteri

Aralık	Değerlendirme
0.00-0.19	Çok zayıf
0.20-0.39	Zayıf
0.40-0.59	Orta
0.60-0.79	Güçlü
0.80-1.0	Çok güçlü

Regresyon modellerinin uygulanmasında MATLAB ortamında Probabilistic Modeling Toolkit kütüphanesinden yararlanılmıştır [39].

3.3. Sonuçlar

RVM kullanılırken en iyi kestirim performansına erişmek amacıyla σ ve γ parametreleri 0.1-7.0 arasında değerler alacak şekilde kestirim tekrarlanmıştır. Deney sayıları meme kanseri için 55, kolon kanseri için 53 ve prostat kanseri için 11'dir. Her kanser türünde kayıp veri kestirimi yapılmış olup ortalama Pearson ve Sperman KK değerleri Çizelge 3.3 ve Çizelge 3.4'te sunulmaktadır.

Çizelge 3.3 Kayıp veri atama Pearson KK değerleri

Regresyon modeli	Ortalama Pearson KK		
	Prostat kanseri	Meme kanseri	Kolon kanseri
Doğrusal Regresyon	0,933	**0,740**	**0,980**
RVM (doğrusal kernel)	**0,934**	0,643	0,971
RVM (RBF 1)	0,874	0,662	0,979
RVM (RBF 2)	0,895	0,680	0,965
k-NN	0,898	0,523	0,975

Pearson KK değerlerine bakıldığında prostat kanseri için en iyi yöntemin doğrusal çekirdek fonksiyonlu RVM (0,934), meme kanseri için doğrusal regresyon (0,740) ve kolon kanseri için yine doğrusal regresyon (0,980) olduğu görülmektedir.

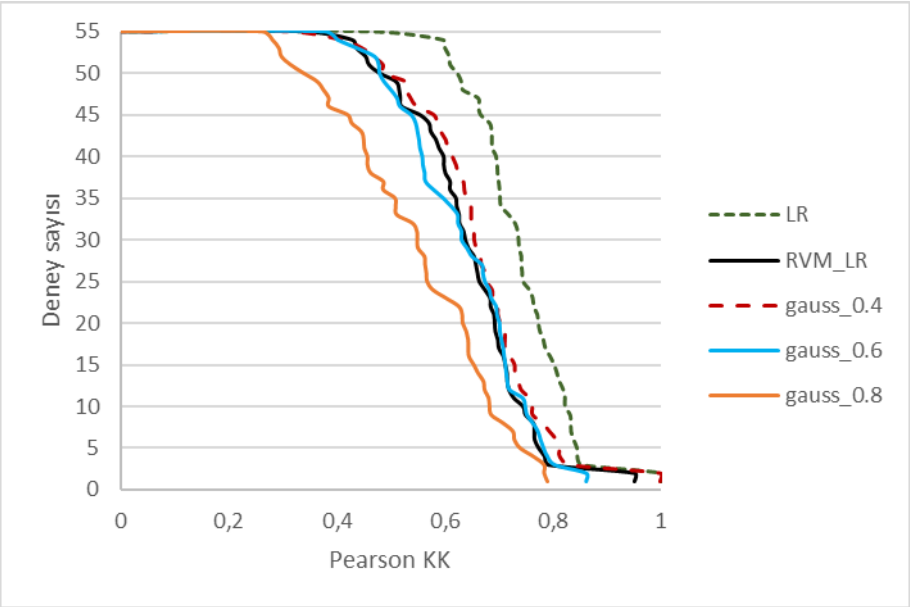
Çizelge 3.4 Kayıp veri atama Spearman KK değerleri

Regresyon modeli	**Ortalama Spearman KK**		
	Prostat kanseri	**Meme kanseri**	**Kolon kanseri**
Doğrusal Regresyon	0,892	**0,739**	0,979
RVM (doğrusal kernel)	0,895	0,648	0,971
RVM (RBF 1)	0,904	0,665	**0,980**
RVM (RBF 2)	0,897	0,640	0,979
k-NN	**0,906**	0,539	0,976

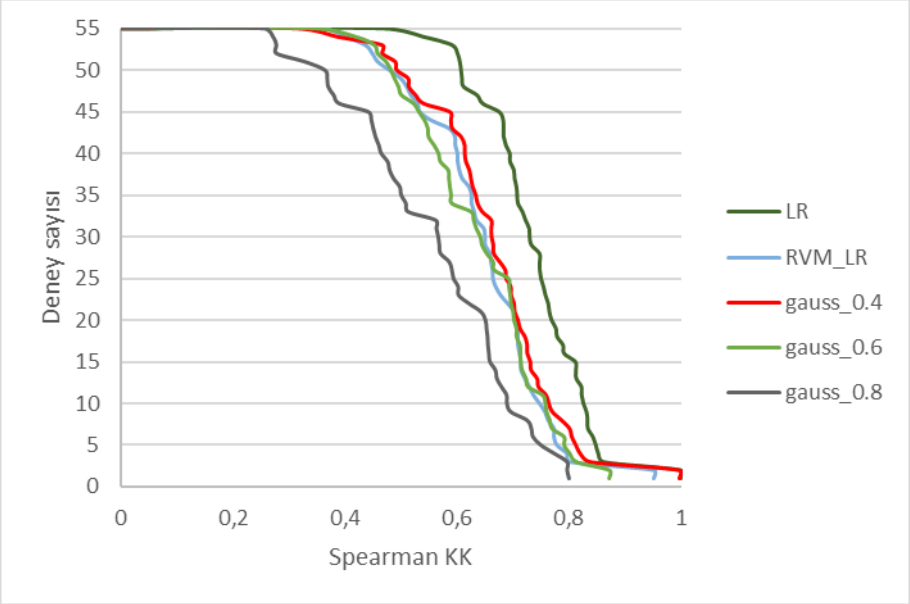
Spearman KK değerlerine bakıldığında prostat kanseri için en iyi yöntemin k-NN (0,906), meme kanseri için doğrusal regresyon (0,739) ve kolon kanseri için yine RBF-1 çekirdek fonksiyonlu RVM (0,980) olduğu görülmektedir.

Pearson KK ve Sperman KK ölçütlerine göre daha iyi bir karşılaştırma yapabilmek için hesaplanan bu katsayılar büyükten küçüğe doğru sıralanmıştır. Örneğin; 55 meme kanseri deneyi arasında en yüksek benzerlik katsayısına sahip olanın gerisinde kalan deney sayısı 54 olur. Böylece her yöntem için en yüksek deney sayısından sıfıra doğru uzanan eğriler oluşur. Bu eğri altında kalan alan ne kadar büyükse yöntemin performansının o kadar iyi olduğu sonucu ortaya çıkar. Grafiklere bakıldığında en üstte kalan eğri en iyi performansa sahip yöntemi göstermektedir.

Şekil 3.2 ve Şekil 3.3.'te sırasıyla meme kanseri verisinde en iyi kestirim performansına sahip beş modele ait Pearson KK ve Spearman KK eğrileri gösterilmektedir. Her iki korelasyon katsayısı eğrilerine göre en iyi model doğrusal regresyon olmakla birlikte ikinci sırada RVM'in RBF-1 çekirdek fonksiyonu ($\sigma = 0.4$) yer almaktadır.

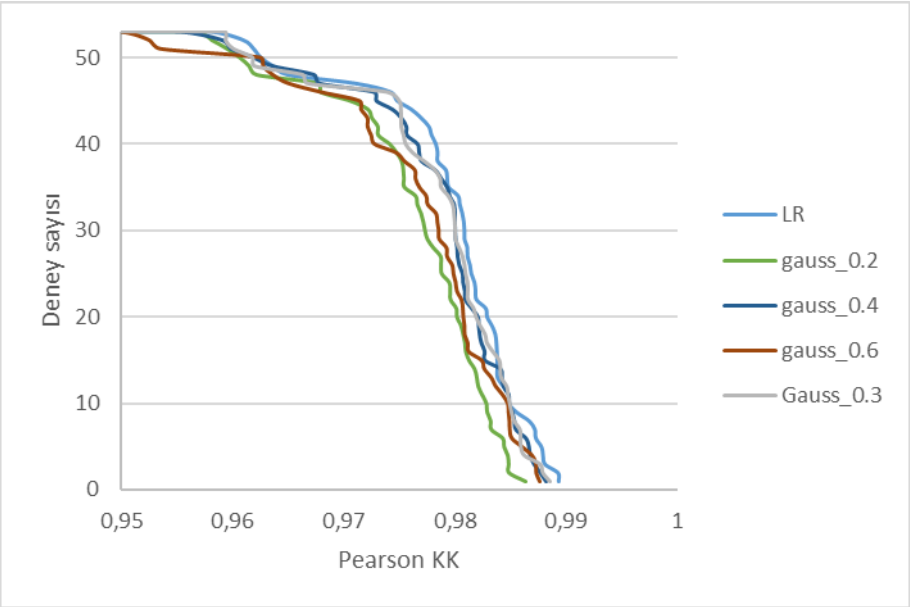


Şekil 3.2 Meme kanseri verisi için kestirim performansı (Pearson KK eğrileri)

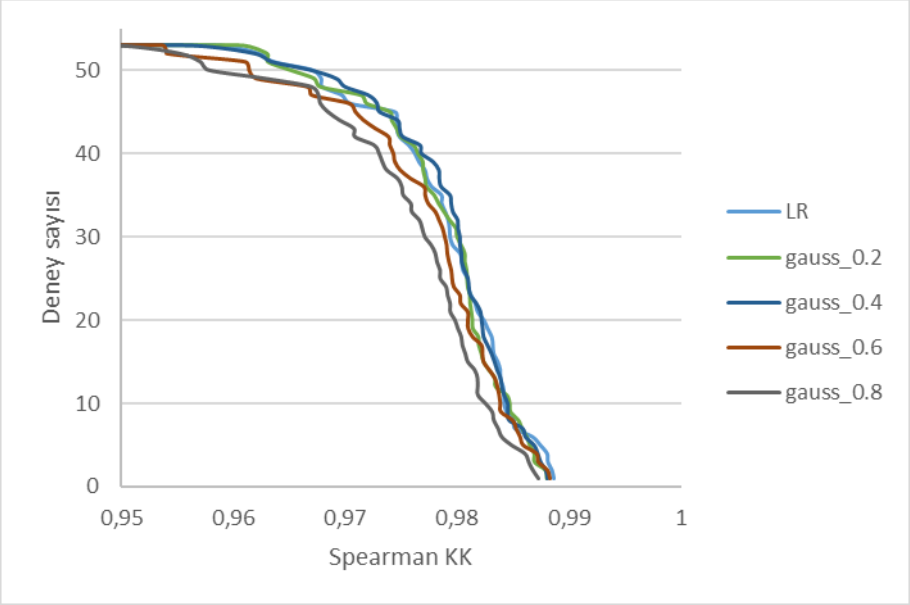


Şekil 3.3 Meme kanseri verisi için kestirim performansı (Spearman KK eğrileri)

Şekil 3.4 ve Şekil 3.5'te sırasıyla kolon kanseri verisinde en iyi kestirim performansına sahip beş modele ait Pearson KK ve Spearman KK eğrileri gösterilmektedir. Pearson KK eğrilerine göre en iyi model doğrusal regresyon olmakla birlikte ikinci sırada RVM'in RBF-1 çekirdek fonksiyonu ($\sigma = 0.3$) yer almaktadır. Spearman KK eğrilerine göre ise en iyi model RBF-1 çekirdek fonksiyonu olup sırasıyla $\sigma = 0.3$ ve $\sigma = 0.4$ parametreleri ile elde edilmektedir.



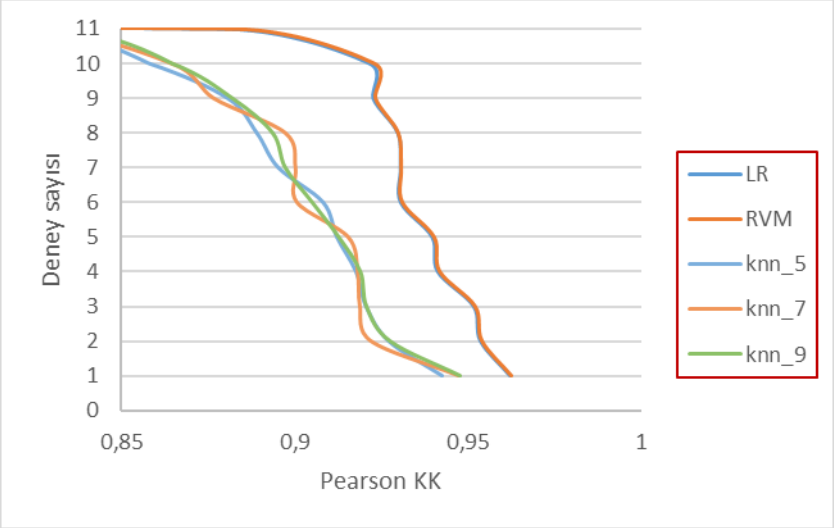
Şekil 3.4 Kolon kanseri verisi için kestirim performansı (Pearson KK eğrileri)



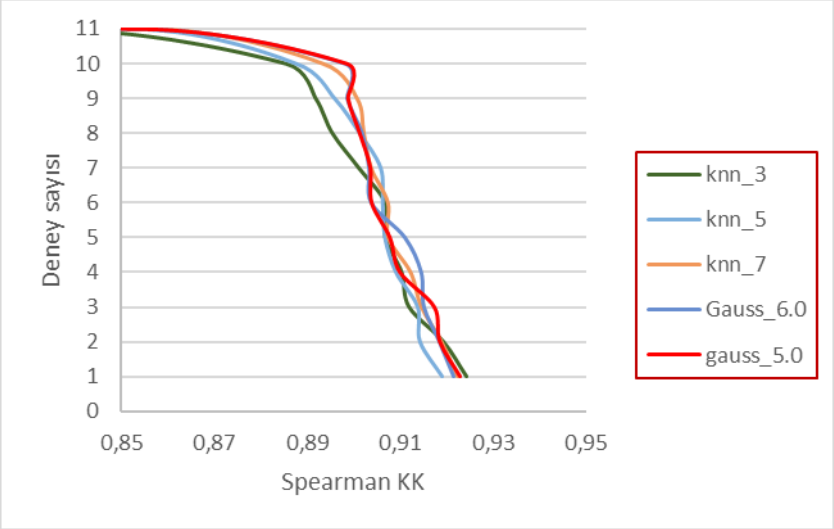
Şekil 3.5 Kolon kanseri verisi için kestirim performansı (Spearman KK eğrileri)

Şekil 3.6 ve Şekil 3.7'de sırasıyla prostat kanseri verisinde en iyi kestirim performansına sahip beş modele ait Pearson KK ve Spearman KK eğrileri gösterilmektedir. Pearson KK eğrilerine göre en iyi model RVM'in doğrusal çekirdek fonksiyonu olmakla birlikte ikinci sırada doğrusal regresyon yer almaktadır.

Spearman KK eğrilerine göre ise en iyi model RBF-1 çekirdek fonksiyonu ($\boldsymbol{\sigma = 5.0}$) ve ikinci sırada k-NN regresyonu (k=7) yer almaktadır.



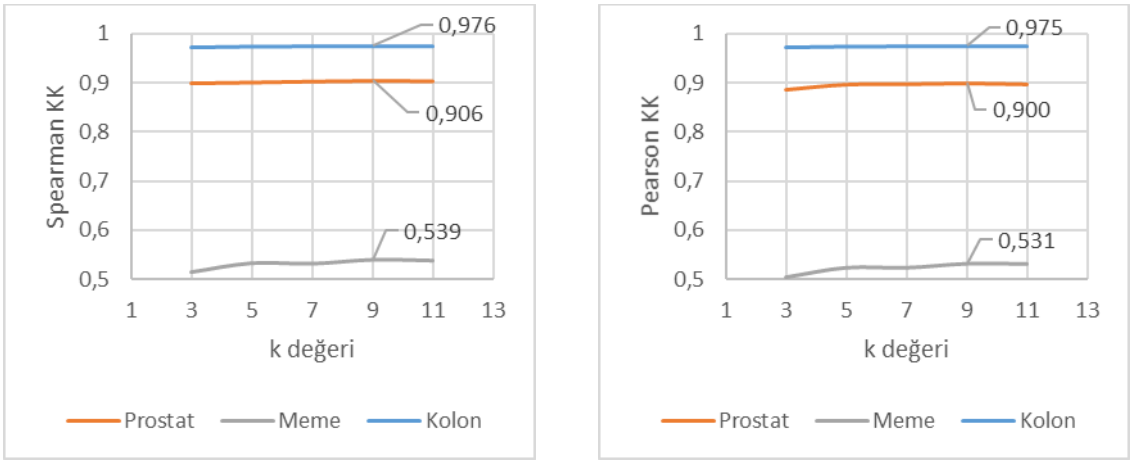
Şekil 3.6 Prostat kanseri verisi için kestirim performansı (Pearson KK eğrileri)



Şekil 3.7 Prostat kanseri verisi için kestirim performansı (Spearman KK eğrileri)

Bu bölümde uygulanan regresyon modellerinde parametre değişikliklerine göre performans değişimleri de incelenmiştir. Her yöntemin ilgili optimizasyon parametresindeki değişimlerin kestirim performansına olan etkisi grafikler ile gösterilmiştir. Burada yöntemlerin ilgili parametrelerinin değişiminden etkilenmeden

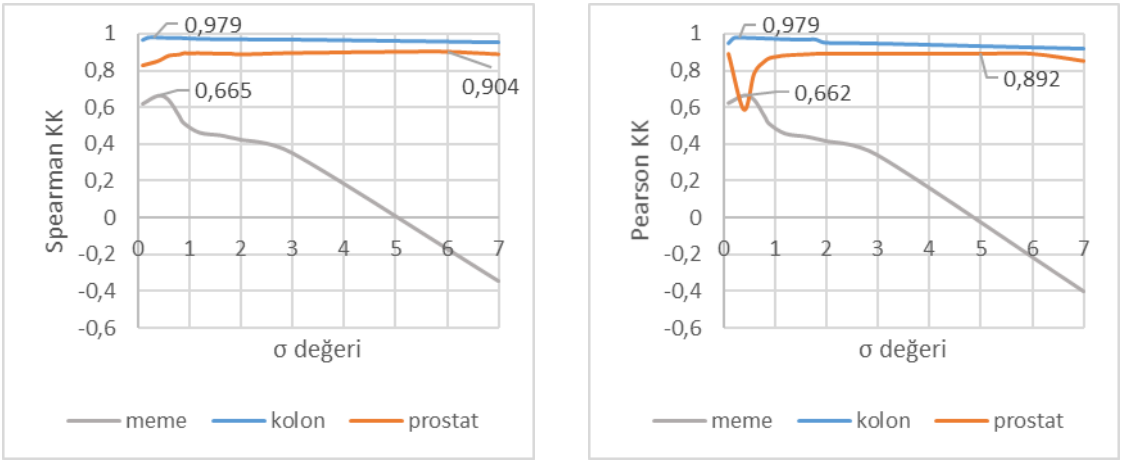
yüksek ve kararlı performans sergilemesi oldukça önemlidir. Bu nedenle doğrusal regresyon dışında k-NN ve RVM regresyon yöntemlerinin ilgili parametreleri değiştirilerek uygulamalar tekrarlanmış ve sonuçlar gösterilmiştir. Bu karşılaştırma eğrilerinde en az bir adet maksimum noktası yer almaktadır. Bu maksimum nokta, modelin en iyi kestirim performansına ulaştığı noktayı temsil etmektedir. Bunun için yöntemler ilgili parametrelerinin belirli bir değer aralığında (k için 3-11, $\boldsymbol{\sigma}$ ve γ için 0.1-7) test edilerek en iyi performans gösterdiği parametre değeri tespit edilmiştir. Bu bağlamda; k-NN regresyonunun $k = 3,5,7,9 \text{ ve } 11$ durumları için Spearman KK ve Pearson KK performans değişimleri Şekil 3.8'de görülmektedir. Şekilde meme, kolon ve prostat kanser türleri için k-NN regresyon modelinin farklı k değerleri için kararlı bir performans gösterdiği söylenebilmektedir. En iyi kestirim performansına kolon kanseri verisi kullanılarak ulaşıldığı görülmektedir. Tüm kanser türlerinde k-NN regresyon modelinin $k = 9$ olduğunda en iyi kestirim sağladığı görülmektedir.



a. b.

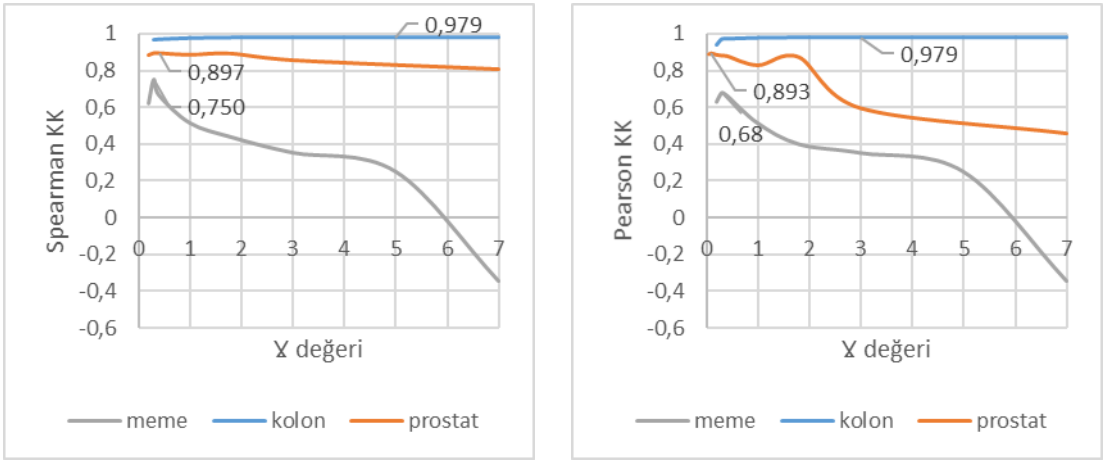
Şekil 3.8 k-NN regresyonu performans değişimi a. Spearman KK b. Pearson KK

Daha önce de ifade edildiği gibi RVM regresyonunda RBF-1 ve RBF-2 olmak üzere iki farklı çekirdek fonksiyonu kullanılmıştır. Bu çekirdek fonksiyonlarının ilgili parametrelerindeki değişimlere göre performans değişimleri Şekil 3.9 ve Şekil 3.10'da gösterilmektedir. Kolon ve prostat kanseri verileri için her iki çekirdek fonksiyonunun daha kararlı performans gösterdiği görülmektedir.



a. b.

Şekil 3.9 RVM RBF-1 kernel fonksiyonunun performans değişimi a. Spearman KK b. Pearson KK



a. b.

Şekil 3.10 RVM RBF-2 kernel fonksiyonunun performans değişimi a. Spearman KK b. Pearson KK

Yukarıdaki şekiller incelendiğinde RVM regresyon yönteminin farklı kanser türlerinde en iyi performansı sağlarken σ ve ɣ parametrelerinin birbirinden farklılaştığı görülebilmektedir. Yalnız en iyi kestirim performansına her üç kanser verisi içinde RBF-1 çekirdek fonksiyonunun $\sigma = 0.2 - 0.4$ aralığında olduğu durumda elde edildiği gözlemlenmektedir. RVM regresyon modeli σ ve ɣ parametreleri ile optimize edilerek kestirim performansı artırılabilir.

Bu bölümde ayrıca bir kanser türüne ait gen ifade değerlerinin tahmin edilmesinde farklı bir kanser türüne ait gen ifade değerlerinin kullanılmasının kestirim performansına etkisi araştırılmıştır. Çizelge 3.5'te yer alan korelasyon katsayıları

incelendiğinde kolon ve prostat kanseri verilerinin model öğrenmede kullanımının bu kanser türlerine ait gen ifade değerlerinin tahmin edilmesinde kullanılabileceği görülmektedir. Buna karşılık meme kanserine ait gen ifade değerlerinin kestiriminde kolon ve prostat kanser verisinin kullanılmasının kestirim performansının artırmadığı görülmektedir.

Çizelge 3.5 Farklı kanser türüne ait verilerin bütünleştirilmesi

			Kestirim yapılan kanser türü		
			Meme	**Kolon**	**Prostat**
Eğitim verisi	*Meme ve Kolon*	Spearman	-	-	0,65
		Pearson			0,525
	Meme ve Prostat	Spearman	-	0,543	-
		Pearson		0,527	
	Kolon ve prostat	Spearman	0,004	-	-
		Pearson	0,001		

3.4. Tartışma

Bu bölümde literatürde yaygın bir problemin regresyon tabanlı modeller ile çözümüne yönelik çalışmalar sunulmuştur. Diğer yandan veri hazırlama sürecinde kullanılabilecek yazılım araçları anlatılmıştır. Normalizasyon haricinde herhangi bir ön işleme (outlier yok etme, veri dönüştürme vb.) çalışması yapılmayarak orijinal veriden uzaklaşılmamıştır. Literatürde yer alan diğer çalışmaların birçoğunda veri ön işleme sürecinde, verinin dağılımında sapmalar meydana getirebilecek değerlerin çıkarıldığı görülmektedir. Mikrodizi verilerinin işlem yapılmadan veya sadece bir normalizasyon işleminden geçirilerek ele alınması hem kullanılan yöntemlerin gerçek performansının ölçülmesi hem de gen araştırmalarında kullanılacak yazılım araçlarının geliştirilmesi açısından önemlidir.

Kayıp veri atama probleminin çözümüne yönelik doğrusal, k-NN ve RVM regresyon modelleri kullanılmıştır. Farklı regresyon modellerinin aynı veri üzerinde kestirim performansları karşılaştırılmıştır. Elde edilen sonuçlara göre, RVM regresyon modeli ile uygun çekirdek fonksiyonları ve parametreleri kullanılarak yüksek kestirim performansına ulaşılabileceği görülmüştür. Doğrusal veya k-NN regresyon

modellerine kıyasla RVM regresyon modelinin işlem zamanı açısından maliyetli olduğu görülmektedir. Özellikle doğrusal olmayan çekirdek fonksiyonlarının kullanılması kestirim süresini uzatmaktadır. Ancak kaynak kodunun optimizasyonu ve daha düşük seviye (low-level) bir programlama diline dönüştürülmesi ile zaman maliyeti düşürülebilir.

Şekil 3.8, Şekil 3.9 ve Şekil 3.10'da k-NN ile RVM regresyon modellerinin farklı k, σ ve γ parametreleri için elde edilen Spearman KK ve Pearson KK değerleri görülmektedir. k-NN regresyonunda kullanılan üç kanser türü (meme, kolon ve prostat) için doğrusal performans değişimi elde edilmektedir. Yalnız meme kanseri verisinde bu doğrusal değişimde düzensizlikler olduğu görülmektedir. RVM regresyonunda kolon ve prostat kanseri verisi için daha doğrusal bir performans değişimi elde edilmektedir. Meme kanseri verisi için ise k-NN regresyonuna göre düzensiz bir performans değişimine sahiptir. Her iki regresyon modelinde Pearson KK değişiminin Spearman KK değişimine göre daha düzensiz olduğu söylenebilir.

Bu bölümde ayrıca farklı kanser türlerini içeren deneylerin kullanılmasının gen ifade miktarı kestirim performansına etkisi analiz edilmiştir. Elde edilen sonuçlara göre prostat ve kolon kanseri verisinin aynı modelde kullanılabileceği görülmüştür. Her iki kanser türüne ait gen ifade miktarlarının aynı modelde kullanılabilmesi bu kanserlerin oluşumunda hücrede meydana gelen süreçler benzerliklerin olduğu anlamına gelebilir. Bazı kaynaklarda prostat kanseri teşhisi konmuş hastaların büyük çoğunluğunda anormal kolon poliplerinin olduğu ve bu nedenle prostat kanseri teşhisi konmuş hastaların mutlaka kolonoskopi tetkiki yaptırmaları gerektiği bilgisi yer almaktadır [58].

4. İKİ YÖNLÜ İŞBİRLİKÇİ FİLTRELEME İLE GEN İFADE TAHMİNİ

4.1. Giriş

Tavsiye sistemlerinin (recommender system) amacı izleyiciye veya kullanıcıya kişisel tercihlerinden yola çıkarak içerik önerilerinde bulunmaktır. Günümüzde izleyiciye içerik önerisinde bulunan Netflix ve ürün önerisinde bulunan Amazon gibi platformlarda tavsiye sistemlerinin kullanıldığı bilinmektedir. Bu tavsiye sistemlerinde kullanılan yöntemler temel olarak üç kategori altında ele alınmaktadır. Bunlar içerik tabanlı (content based), işbirlikçi (collaborative) ve karma filtreleme yöntemleridir. İçerik tabanlı filtrelemenin temelinde kullanıcı ve ürün profilleri analiz edilerek model oluşturulurken işbirlikçi filtrelemede ise kullanıcının diğer ürünlere ilişkin değerlendirmeleri ile aynı ürüne farklı kullanıcıların yapmış olduğu değerlendirmelerin model öğrenmede birlikte kullanıldığı görülmektedir. İşbirlikçi filtreleme yönteminin tavsiye sistemlerinde yaygın bir kullanımı bulunmaktadır [59-62].

İşbirlikçi filtreleme yönteminde, komşuluk (neighborhood) ve model tabanlı yaklaşımlar yer almaktadır. Komşuluk tabanlı işbirlikçi filtreleme yönteminde, bir ürünün bir kullanıcıya önerilmesinde ürünün diğer benzer kullanıcılar tarafından tercih edilip edilmediği ve kullanıcının diğer tercih ettiği ürünler ile benzerliği analiz edilmektedir. Model tabanlı işbirlikçi filtreleme yönteminde, eğitim verisinden makine öğrenme algoritmaları ile model oluşturularak kullanıcıya ürün önerilmektedir. Model tabanlı yaklaşımda tüm kullanıcı değerlendirme verilerine ihtiyaç duyulmadığı için daha az bir veri ile işlemler tamamlanabilmektedir [62]. Komşuluk tabanlı işbirlikçi filtreleme yönteminin, model öğrenme işlemi bulunmadığı ve model tabanlı yaklaşıma kıyasla daha az parametre içerdiği için Amazon ve Netflix gibi platformlar tarafından daha çok tercih edildiği bilinmektedir[63].

Karma filtreleme yönteminde, işbirlikçi filtreleme algoritması ile kullanıcı demografik özellikleri bir küme ağırlıklı (cluster-weighted) mekanizmada birleştirilerek izleyiciye öneriler sunar. Bu yöntem; zaman ve işlem kolaylığı açısından önemli avantajlara sahiptir. Ancak kullanıcının beğendiği içeriklere benzer içerikleri önermede iyi olması ile birlikte kullanıcılara ilişkin diğer bilgileri ve diğer kullanıcı

değerlendirmelerini kullanamaz. Bu nedenle, birbirinden çok farklı kullanıcı profillerinde ve geniş dağılıma sahip içerik değerlendirmelerinde iyi sonuçlar verdiği söylenemez [62].

Ogul ve Ekmekciler tarafından geliştirilen film önerme sisteminde, yukarıda anlatılan komşuluk ve model tabanlı işbirlikçi filtreleme yöntemleri birleştirilerek geliştirilen iki yönlü işbirlikçi filtreleme yaklaşımı kullanılmıştır. Bu yeni yaklaşımda, bir izleyicinin farklı filmlere verdiği puanların yanında farklı izleyicilerin aynı filme verdiği puanlar da model eğitiminde kullanılmaktadır. Bu yaklaşım ile oluşturulan yeni öznitelik uzayının sistem performansını artırdığı görülmüştür [59].

Bilinen model öğrenme yaklaşımında, ifade tam değeri tahmin edilecek bir genin diğer deneylere ait ifade değerleri model eğitiminde kullanılmakta ve oluşturulan model test verisinde kullanılmaktadır. Bu klasik yaklaşımda, model parametrelerinin belirlenmesinde sadece hedef genin farklı deneylerdeki ifade değerleri hesaba katılmaktadır. Protein sentezi sürecinde genlerin birbirleriyle ve diğer moleküller ile karmaşık bir etkileşim içinde olduğu ve bu süreçte çok fazla değişkenin rol oynadığı bilinmektedir. Hücrede meydana gelen bu karmaşık etkileşimler; genlerin ifade değerlerine ve dolayısıyla nihai ürün olan proteinin fonksiyonuna bağlı olarak fenotipe kadar yansımaktadır. Yukarıda anlatılan tek yönlü öznitelik sunum biçiminde aynı deneye ait farklı genlerin ifade değerleri model eğitiminde kullanılamamaktadır. Bir genin hücre içindeki etkinliği ifade değerleri ile anlaşılmakta ve bir genin ifade miktarının nispeten büyük olması protein sürecinde daha aktif görev aldığı anlamına gelmektedir. Bu nedenle, bir genin ifade değeri ile bu genle etkileşim halinde olan veya çalışmasında rol alan başka bir genin ifade değerinin aynı modelde kullanılmasının kestirim performansını artıracağı hipotezi oluşturulmuştur. Bu hipotezin kestirim çalışmaları ile desteklenmesi için regresyon modelinin eğitiminde kullanılan tek yönlü gen ifade matrisinin amaçlanan çalışmaya uygun olarak iki yönlü gen ifade matrisine dönüştürülmesi gerekmektedir. İki Yönlü İşbirlikçi Filtreleme işlemi bu tez çalışmasında gen ifade tam değerinin tahmini için uyarlanmış olup bu kapsamda ilk defa kullanılmıştır.

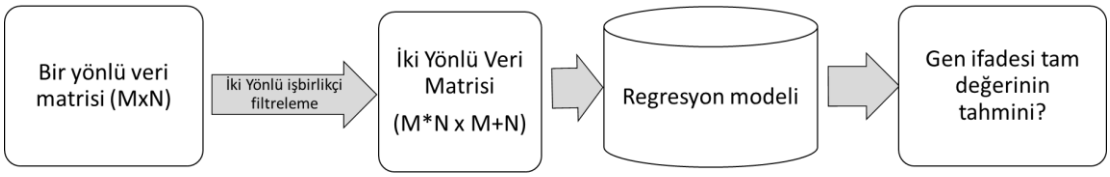
4.2. Materyal ve Yöntem

4.2.1. Veri

Bu bölümde kullanılan veri kümeleri, GEO veritabanında GSE18088 ve GSE45016 erişim numaraları ile erişime açık olan [51,52] prostat ve kolon kanseri mikrodizi veri setlerinden elde edilmiştir. Bu veri setlerindeki gen ifade miktarı ölçümleri GPL570 platformunda Affymetrix Human Genome U133 Plus 2.0 Array biyoçipi kullanılarak gerçekleştirilmiştir [53]. Genel olarak, bu biyoçip binlerce gen tanımlayıcı probu içerir ve birden fazla prob, mikrodizi teknolojisinde sadece bir mRNA'yı ifade edebilir. Her veri setinde uygulama yapabilmek için birden fazla probun karşılık geldiği ortak mRNA'ların belirlenmesi gerekir. 54675 proba ait ölçüm verilerinden ortak mRNA'lar tespit edildikten sonra çalışma kapsamında rastgele 1600 mRNA seçilmiştir. Bir mRNA için eğer birden fazla probdan ölçüm alınmışsa bunların ortalamaları alınarak tekilleştirilmiştir. Veri kümesindeki prob tanımlayıcı numaralarının mRNA sembollerine dönüşümü için ücretsiz erişim sağlanan bazı araçlar kullanılmıştır [54,55]. Çoğu durumda, mikrodizi ölçüm platformları, kullanılan biyoçipin referans noktası nedeniyle farklı ölçüm aralıkları içerebilir. Verinin değişim aralığını aynı yapmak amacıyla ifade değerleri min-max normalizasyonu kullanılarak 0 ve 1 aralığında normalize edilmiştir.

4.2.2. Yöntem

İki Yönlü İşbirlikçi Filtrenin gen ifadesi tam değerinin tahmin edilmesi işleminde kullanıldığı işlem basamakları Şekil 4.1'de gösterilmektedir. Tek yönlü veri matrisinde, satırlar genlerin farklı deneylere ait ifade değerlerini ve sütunlar ise deneyleri temsil etmektedir. İki yönlü işbirlikçi filtreleme yaklaşımında; M adet gen ve N adet örnekten oluşan tek yönlü veri matrisi, satır sayısı M * N ve sütun sayısı M + N olan iki yönlü veri matrisine dönüştürülür.



Şekil 4.1 İki Yönlü İşbirlikçi Filtrenin kestirim işlemindeki yeri

Çalışmanın bu bölümünde uygulamalar ve sonuçlar sistematik bir şekilde sunulmuştur. Uygulama adımları ve her bir adımdaki en iyi performans çıktısı Şekil 4.2'de gösterilmiştir. Bu bölümde ortaya konulan sistematik yaklaşım aşağıdaki sorular ile açıklanmaktadır:

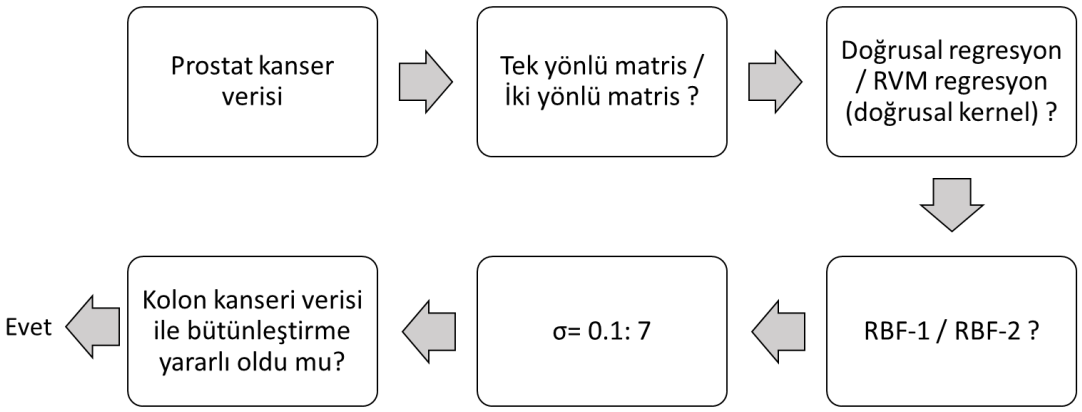
I. Bir yönlü veri matrisine kıyasla İki Yönlü İşbirlikçi Filtreleme ile elde edilen iki yönlü veri matrisinin kullanılması kestirim performansını artırmakta mıdır?
(Öznitelik sunum yönteminin kestirim performansına etkisi)

II. Doğrusal regresyon ile doğrusal çekirdek fonksiyonlu RVM regresyon modellerinden hangisi kestirim performansını daha fazla artırmaktadır?
(Regresyon modelinin kestirim performansına etkisi)

III. RVM regresyon modeli için hangi doğrusal olmayan çekirdek fonksiyonu ile daha yüksek kestirim performansı elde edilmektedir?
(Çekirdek fonksiyonunun kestirim performansına etkisi)

IV. En iyi kestirim performansı elde edilen çekirdek fonksiyonuna ait parametre hangi değer seçildiğinde maksimum kestirim performansı elde edilmektedir?
(Çekirdek fonksiyonu parametresinin kestirim performansına etkisi)

V. Prostat kanseri ile kolon kanseri gen ifade verilerinin bütünleştirilmesi kestirim performansını artırmakta mıdır?
(Birden fazla farklı kanser verisinin bütünleştirilmesinin kestirim performansına etkisi)



Şekil 4.2 İki Yönlü İşbirlikçi Filtreleme ve uygulama adımları

Spearman KK ve Pearson KK ölçütlerine ek olarak Hata Kareleri Ortalamasının Karekökü (Root Mean Squared Error) parametresi ayrı bir performans değerlendirme ölçütü olarak kullanılmıştır. Bu değer sıfıra yaklaştıkça gerçek ve tahmin edilen iki verinin birbirine yaklaştığı anlamı ortaya çıkmaktadır. RMSE ölçütü Eşitlik 6'daki denklem ile hesaplanır [64]. Burada n gen sayısını, M_i gerçek gen ifade tam değerini ve P_i ise tahmin edilen değeri göstermektedir.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \|M_i - P_i\|^2}{n}} \quad (6)$$

Geçerli kılma yöntemi olarak LOO prosedürü uygulanmıştır. Bu prosedürde M adet gene ait ifade vektörlerinden M-1 tanesi eğitim ve geriye kalan biri test amaçlı kullanılır. Bu kestirim işlemi M defa tekrarlanır. Sonuç olarak N adet deney için toplam işlem sayısı M x N olur. Her bir deney için hesaplanan kestirim performans ölçütleri hesaplanır.

4.2.2.1. İki Yönlü İşbirlikçi Filtreleme

İşbirlikçi filtreleme yönteminde m adet kullanıcı $\{k_1, k_2, \dots, k_m\}$ ile gösterilsin ve n adet içerik $\{i_1, i_2, \dots, i_n\}$ ile ifade edilsin. i. kullanıcı k_i'nin ise L_{k_i}şeklinde bir içerik izleme listesi olsun. Kullanıcının bu listedeki içerikleri beğenme ölçütü; 1-5 arasında

puan veya tıklanma sayısı gibi farklı şekillerde olabilir. m adet kullanıcının n adet içeriğe verdiği puanlama "iyi-kötü" gibi kategorik veya sınırları belirlenmiş numerik bir puanlama şeklinde de olabilir. Her bir kullanıcının bu listede puanlama yapmadığı içerikler olabilir. Bu şekilde hazırlanan bir matris örneği Çizelge 4.1'de görülmektedir. Bu çizelgede k_4'ün i_4'e verdiği puanlamayı tahmin etmek ve böylece bu içeriği önerip önermemek konusunda değerlendirme yapabilmek için diğer içeriklere verilen puanlamalar hesaba katılır. Bazı kullanıcıların bazı içerikleri değerlendirmediği durumlar da söz konusu olabilir. Bu veriler kayıp veri olarak ifade edilir. Bu tez çalışmasının birinci bölümünde tek yönlü veri matrisi kullanılarak kayıp veri kestirimine ilişkin yapılan çalışmalar sunulmuştur. Çizelge 4.1'de görüldüğü gibi iyi-kötü değerlendirmelerinin tahmin edilmesi sınıflandırma problemi olarak ele alınmaktadır. Diğer yandan bu değerlendirmelerdeki puanlama (örneğin iyi->5 kötü->1) tahmin edilecekse, bu tezde olduğu gibi regresyon problemi olarak ele alınmaktadır.

Çizelge 4.1 Örnek kullanıcı-içerik değerlendirme matrisi

	i_1	i_2	i_3	i_4
k_1	İyi	İyi		kötü
k_2		İyi	kötü	İyi
k_3		kötü	İyi	
k_4	İyi		kötü	?

Tek yönlü veri matrisinin kullanıldığı durumda, Çizelge 4.1'de k_4'ün i_4 için yapabileceği değerlendirmeyi tahmin etmek için her içerik için yapılan tüm değerlendirmeler (çizelgedeki sütun değerleri) model eğitiminde kullanılmaktadır.

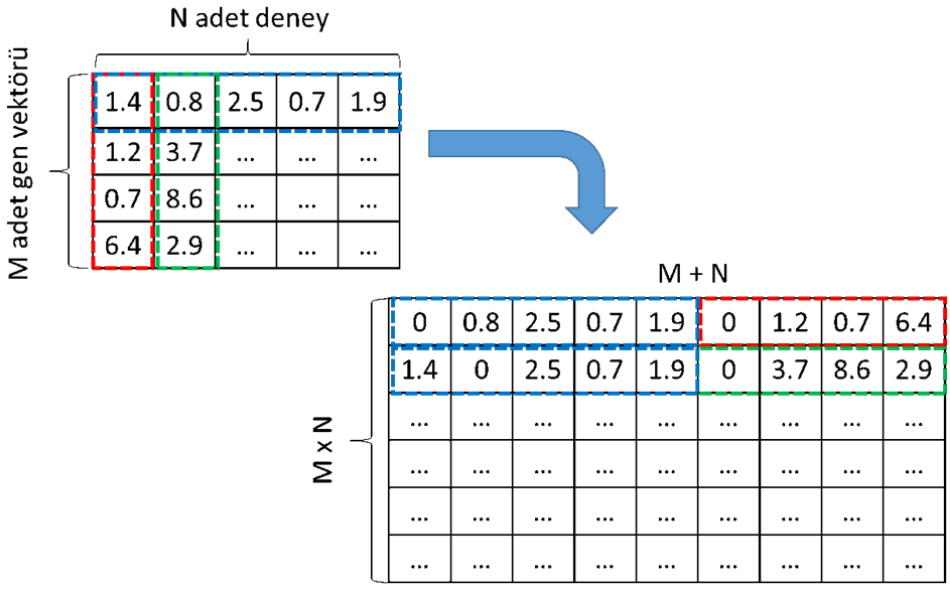
İşbirlikçi filtreleme yöntemleri temelde komşuluk tabanlı ve model tabanlı modeller olmak üzere ikiye ayrılmaktadır. **Komşuluk tabanlı yaklaşım diğer ismi ile bellek tabanlı yaklaşımda;** model eğitim aşaması bulunmaz ve benzerlik ölçütü kullanılır. Bu yaklaşımda, kolay ve hızlı uygulamalar yapılması mümkündür. Hızlı ve efektif sonuçlar alınması nedeniyle ticari amaçlı kullanımı yaygındır. Ancak bu avantajlarının yanında bazı kısıtlamaları da mevcuttur. Özellikle eksik değerlerin çok olduğu "sparse" olarak tanımlanan veri kümelerinde güvenilir sonuçlar

vermemektedir. Bu yaklaşımda tüm kullanıcı ve içerik bilgileri kullanılarak kullanıcılar ve içerikler arasındaki benzerlik ölçütlerinin yer aldığı bir matris elde edilir. Amaç bu benzerlik ölçütlerine göre daha önce içerik izleyerek puanlama yapmış aktif bir kullanıcıya N adet içerik önermektir. Burada kullanılan benzerlik ölçütleri Pearson KK veya kosinüs dönüşüm gibi yöntemler olabilir. **Model tabanlı yaklaşımda;** daha karmaşık öznitelikler ve daha az veri ile eğitilmiş modeller kullanılarak daha doğru kestirimler yapılabilmektedir. Bu yaklaşımda da veri tipi kategorik ise sınıflandırma veya numerik ise regresyon tabanlı modeller kullanılmaktadır [61]. Bu çalışma; sadece genlerin ifade değerleri kullanıldığı için model tabanlı yaklaşıma daha uygundur. Çizelge 4.1'de gösterilen veri matrisi tek yönlü öznitelik sunumu olarak tanımlanabilir. Bu çalışmada kullanılan iki yönlü işbirlikçi filtreleme yöntemi ise bu tek yönlü matrisini iki yönlü veri matrisine dönüştürme işlemidir.

İki Yönlü İşbirlikçi Filtreleme yöntemi ilk olarak film önerme sistemlerini geliştirmek için önerilen yeni bir öznitelik gösterim biçimi olarak ortaya çıkmıştır. Klasik yaklaşımının kullanıldığı film önerme sisteminde bir izleyicinin farklı filmlere verdiği puanlar model öğrenme sürecinde kullanılırken İki Yönlü İşbirlikçi Filtreleme yönteminde ise aynı film için farklı izleyiciler tarafından verilen puanlar da model öğrenme sürecine dâhil edilmektedir [59]. Bu tez çalışmasında ise İki Yönlü İşbirlikçi Filtreleme yöntemi gen ifade tam değeri tahmininde kullanılmak üzere adapte edilmiştir. Bir genin ifade tam değerinin tespitinde genel yaklaşım, farklı deneylere ait ifade değerlerinin bir vektör olarak model öğrenmede kullanılmasıdır. Aynı genin farklı deneylere ait ifade değerlerinden oluşan bu vektör tek yönlü vektör olarak adlandırılabilir. Bu durumda Şekil 4.3'te $M\ (gen)\ x\ N\ (deney)$ boyutundaki matris tek yönlü veri matrisi olarak tanımlanabilir. İki Yönlü İşbirlikçi Filtreleme ile her bir deneye ait diğer gen ifade değerlerinden oluşan düşey vektörlerin transpozu alınarak tek yönlü vektörlerin devamına eklenir. Böylece $(M * N)\ x\ (M + N)$ boyutunda yeni bir matris elde edilir. Bu matris iki yönlü veri matrisi olarak adlandırılabilir. Burada dikkat edilmesi gereken önemli nokta, yeni oluşan $(M + N)$ uzunluğundaki vektörde tahmin edilecek hedef genin ifade değerine sıfır atanmasıdır. Böylece tahmin edilecek hedef genin gerçek gen ifade değeri model öğrenmede kullanılmayacaktır. Aksi durumda modelin doğru tahmin etmesi gereken

gen ifade değeri eğitim setinde kullanılmış olacaktır ve bu durum kestirim performansının manipüle edilmesi anlamına gelmektedir. Yerine sıfır atanan bu gen ifade değerleri ayrı bir değişkende saklanarak regresyon modelinin oluşturulmasında hedef çıktılar olarak kullanılmaktadır.

Şekil 4.3Şekil 4.3'te tek yönlü veri matrisi ve İki Yönlü İşbirlikçi Filtreleme yöntemi ile elde edilen iki yönlü veri matrisi gösterilmektedir. Soldaki küçük matris $M\ (gen)\ x\ N\ (deney)$ boyutlu tek yönlü veri matrisidir. Sağdaki matris ise önerilen İki Yönlü İşbirlikçi Filtre işlemi sonrası elde edilen $(M * N)\ x\ (M + N)$ boyutlu iki yönlü veri matrisidir.



Şekil 4.3 İki yönlü işbirlikçi filtreleme yöntemin ile matris dönüşümü

G genleri, N deneyleri (numuneleri) ve $G\ (g_i, s_j)$ ise j. örneğe ait i. genin ifade değerini göstersin. Soldaki model öğrenmede kullanılan MxN boyutlu matriste "$g(i,1), g(i,2), \ldots, g(i,N)$" i. genin ifade değerlerinden oluşan yatay vektörü (soldaki matriste yer alan mavi çerçeveli vektör) ve "$s(1,j), s(2,j), \ldots, s(M,j)$" j. deneyin diğer gen ifade değerlerinden oluşan düşey vektörü (soldaki matriste yer alan kırmızı ve yeşil çerçeveli matrisler) göstermektedir. İki Yönlü İşbirlikçi Filtreleme yönteminin kullanıldığı yeni yaklaşımda; aynı deneye ait farklı genlerin ifade değerlerinden oluşan düşey vektörün transpozu alınarak yatay vektörün devamına

eklenmektedir (sağdaki matris) ve ifade tam değeri tahmin edilecek gen değerleri sıfır yapılarak ayrı bir vektörde saklanmaktadır. Böylece tahmin edilecek gerçek veri model öğrenmesinde kullanılmamış olacaktır. Bu durumda elde edilen yeni vektör "$g\ (i,1), g(i,2), \ldots, g(i,N), s(1,j), s(2,j), \ldots, s(M,j)$" şeklinde tanımlanabilir.

4.3. Sonuçlar

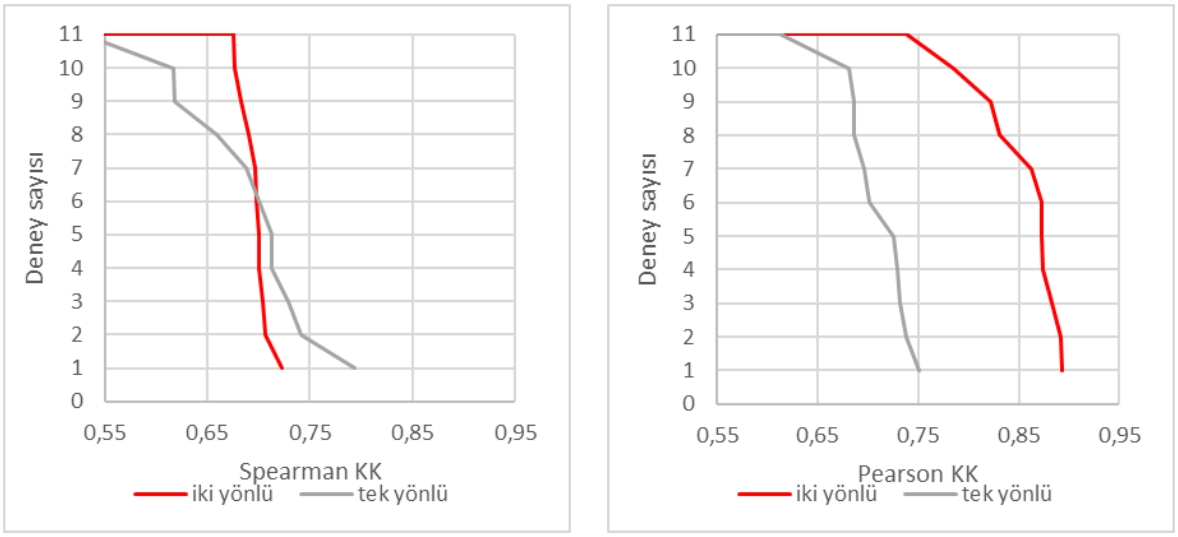
Şekil 4.2'deki şemada gösterildiği gibi sistematik bir yaklaşım ortaya konularak en iyi kestirim performansının elde edildiği model konfigürasyonu tespit edilmiştir. Her bir aşamanın sonunda en iyi kestirim performansına ulaşılan model parametreleri yer almaktadır. Birinci adımda, tek yönlü veri matrisi ile karşılaştırıldığında iki yönlü veri matrisinin kestirim performansını artırdığı görülmüştür. İkinci adımda, iki yönlü veri matrisi doğrusal regresyon ve doğrusal çekirdek fonksiyonlu RVM regresyon modelleri ile ayrı ayrı test edilmiş ve RVM regresyon modelinin daha iyi kestirim performansına sahip olduğu görülmüştür. Üçüncü adımda, iki yönlü veri matrisi üzerinde RVM regresyon modelinin RBF-1 ve RBF-2 doğrusal olmayan çekirdek fonksiyonları test edilmiştir. RBF-1 çekirdek fonksiyonu ile en iyi kestirim performansına ulaşıldığı görülmüştür. Son adımda prostat ve kolon kanseri verilerinin bütünleştirilmesi işleminin kestirim performansını artırdığı görülmüştür. Özetle belirtilen prostat kanseri veri seti için sırasıyla İki Yönlü İşbirlikçi Filtreleme yöntemi, RVM regresyon modeli, RBF-1 çekirdek fonksiyonu ($\sigma = 2$) ve kolon kanseri veri seti ile birleştirme işlemi sonrası en iyi kestirim performansının elde edildiği tespit edilmiştir.

11 deneyden oluşan prostat kanseri verisinden rastgele seçilen 1600 gen için tek yönlü veri matrisi $1600x11$ boyutundadır ve İki Yönlü İşbirlikçi Filtrenin uygulanması sonrası elde edilen yeni veri matrisi $17600\ (11 * 1600)\ x\ 1611\ (11 + 1600)$ boyutunda olmaktadır. İki yönlü veri matrisi için RVM (RBF-1, $\sigma = 2$) ile en iyi kestirim performansı elde edilmiş olup tek yönlü veri matrisi için RVM (RBF-2, $\gamma = 0.2$) ile en iyi kestirim performansına erişilmiştir. Şekil 4.4'te performans eğrileri gösterilmektedir. Her iki korelasyon katsayısına göre iki yönlü veri matrisi için elde edilen performans eğrisi altında kalan alan daha büyüktür. Yani iki yönlü veri matrisinin daha iyi sonuç verdiği görülmektedir. 11 adet prostat kanser deneyinden

elde edilen tek yönlü ve iki yönlü veri matrisleri için ortalama Pearson KK ve Spearman KK değerleri Çizelge 4.2'de yer almaktadır.

Çizelge 4.2 Tek yönlü ve iki yönlü veri matrisleri için kestirim performans değerleri

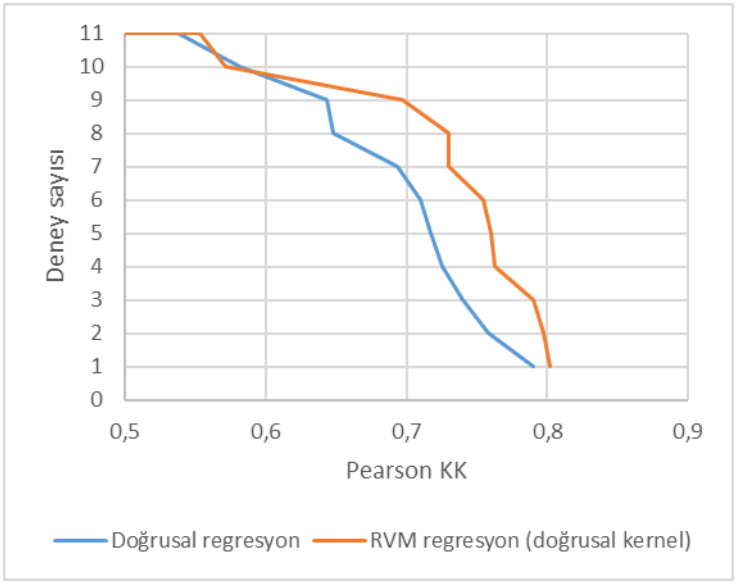
	Tek yönlü veri matrisi	**İki yönlü veri matrisi**
Ortalama Spearman KK	0,682	0,696
Ortalama Pearson KK	0,704	0,847
Maksimum Spearman KK	0,793	0,733
Maksimum Pearson KK	0,751	0,893



a. b.

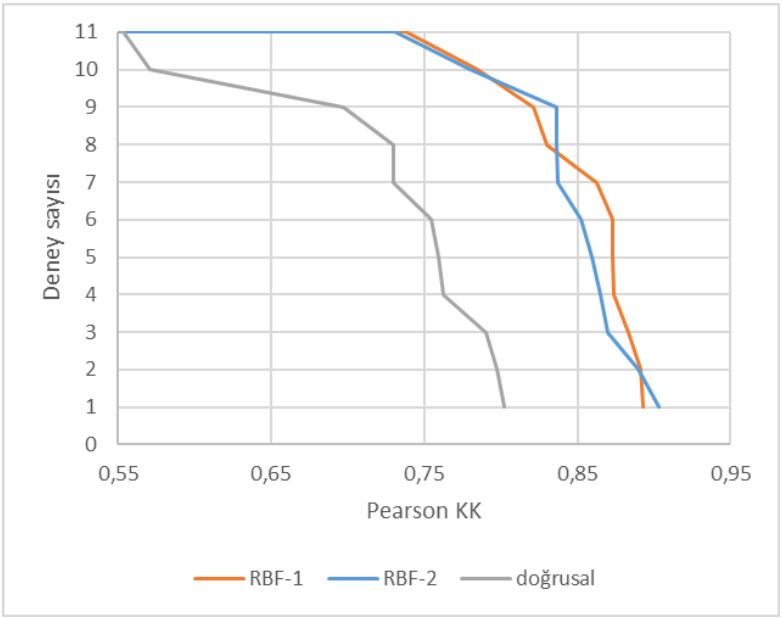
Şekil 4.4 İki Yönlü İşbirlikçi Filtrenin kestirim performansına etkisi a. Spearman KK b. Pearson KK

Doğrusal ve RVM regresyon modellerinin sonuçları karşılaştırıldığında; ortalama Pearson KK değerinin doğrusal regresyon modeli için 0.686 ± 0.076 ve RVM regresyon modeli için ise 0.723 ± 0.081 olduğu gözlemlenmiştir. Şekil 4.5'te performans eğrileri gösterilmektedir.



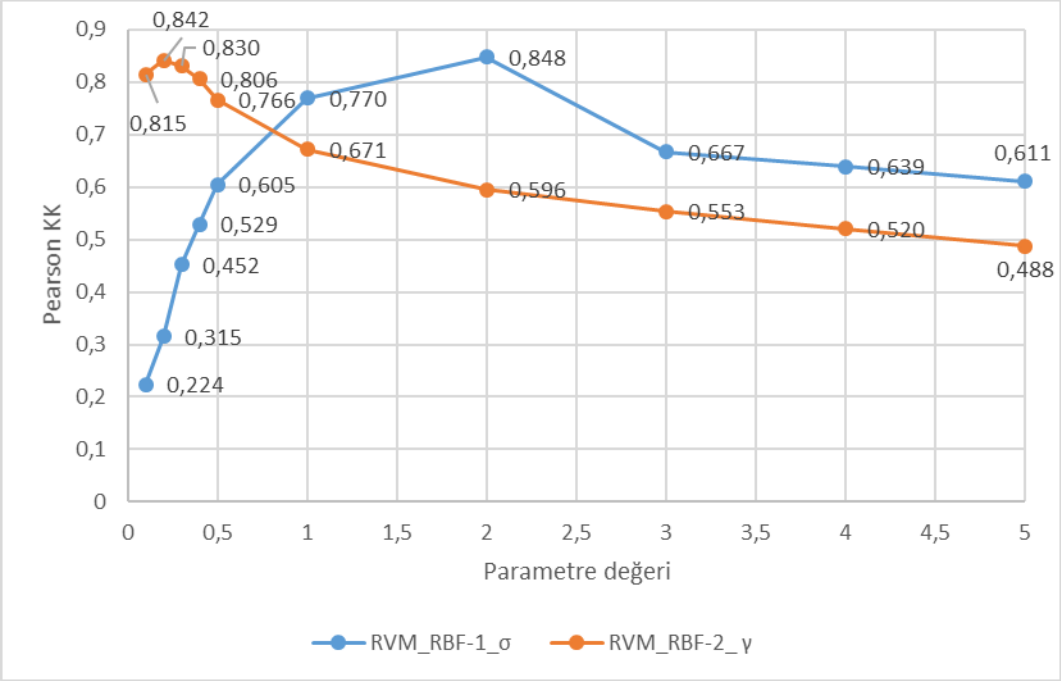
Şekil 4.5 Regresyon modelinin kestirim performansına etkisi

RVM regresyon modelinin daha iyi kestirim sağladığı görüldükten sonra en iyi kestirim performansını elde etmek için farklı çekirdek fonksiyonları farklı σ ve γ parametreleri için test edilmiştir. Kullanılan doğrusal, RBF-1 ve RBF-2 çekirdek fonksiyonları için ortalama Pearson KK değerleri sırasıyla 0.723 ±0.09, 0.848 ± 0.05 ve 0.842 ± 0.05'tir. Şekil 4.6'da farklı çekirdek fonksiyonları için Pearson KK performans eğrileri gösterilmektedir. Şekle göre en iyi kestirim performansına sahip çekirdek fonksiyonu RBF-1'dir.

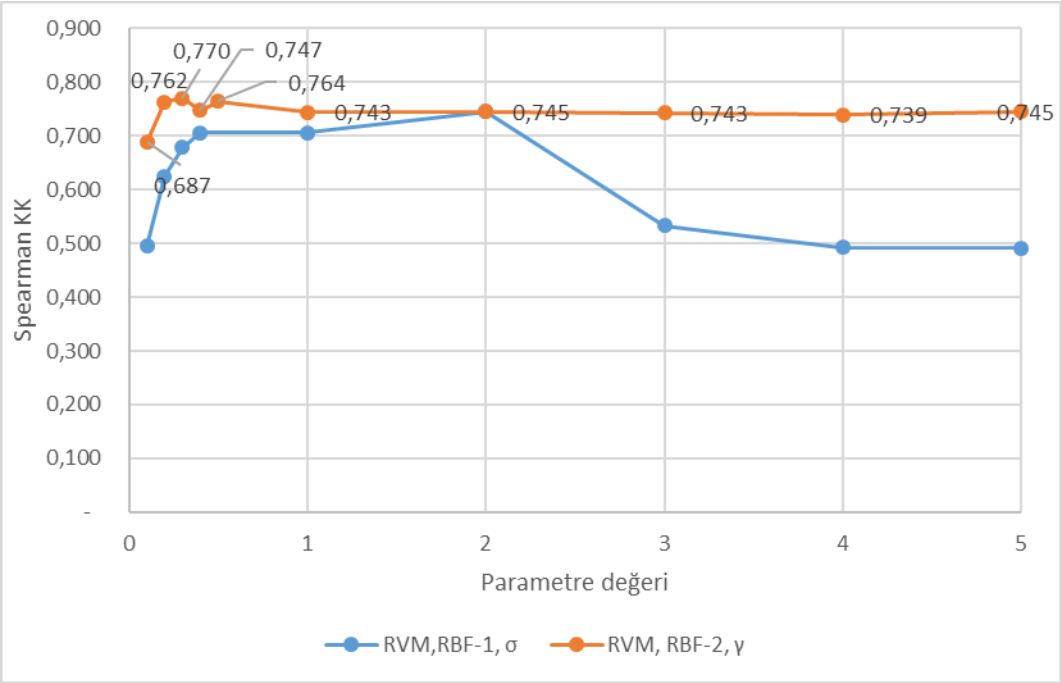


Şekil 4.6 RVM çekirdek fonksiyonlarının kestirime etkisi

Doğrusal çekirdek fonksiyonun çok daha kötü performans gösterdiği görüldükten sonra diğer iki doğrusal olmayan çekirdek fonksiyonları için 0.1-5.0 arasında değişen σ ve γ parametreleri kullanılarak uygulamalar yapılmış ve bu parametrelerin değişiminin kestirim performansına etkisi incelenmiştir. Şekil 4.7 ve Şekil 4.8'de RVM regresyon modelinin RBF-1 ve RBF-2 çekirdek fonksiyonlarına ait parametre değişimlerinin kestirim performansına etkisi görülmektedir. Her iki şekilde gösterilen Pearson KK ve Speaman KK eğrilerine göre en iyi iki kestirim performansına sırasıyla RBF-1 $(\sigma = 2.0)$ ve RBF-2 $(\gamma = 0.2)$ ile ulaşılmaktadır. Bunun yanında Spearman KK ve Pearson KK eğrilerine göre en kötü kestirim performansı RBF-1 için $\sigma = 0.1$ ve RBF-2 için ise $\gamma = 5.0$ ile elde edilmiştir.



Şekil 4.7 Çekirdek fonksiyon parametrelerinin kestirim performansına etkisi (Pearson KK)



Şekil 4.8 Çekirdek fonksiyon parametrelerinin kestirim performansına etkisi (Spearman KK)

Bu bölümde ayrıca birden fazla kanser türüne ait gen ifade verilerinin bütünleştirilmesinin regresyon tabanlı bir kestirim işleminin performansına etkisi araştırılmıştır. Bu amaçla çalışma kapsamında, prostat kanseri gen ifade verileri ile kolon kanseri gen ifade verileri aynı eğitim setinde bütünleştirilmiştir.

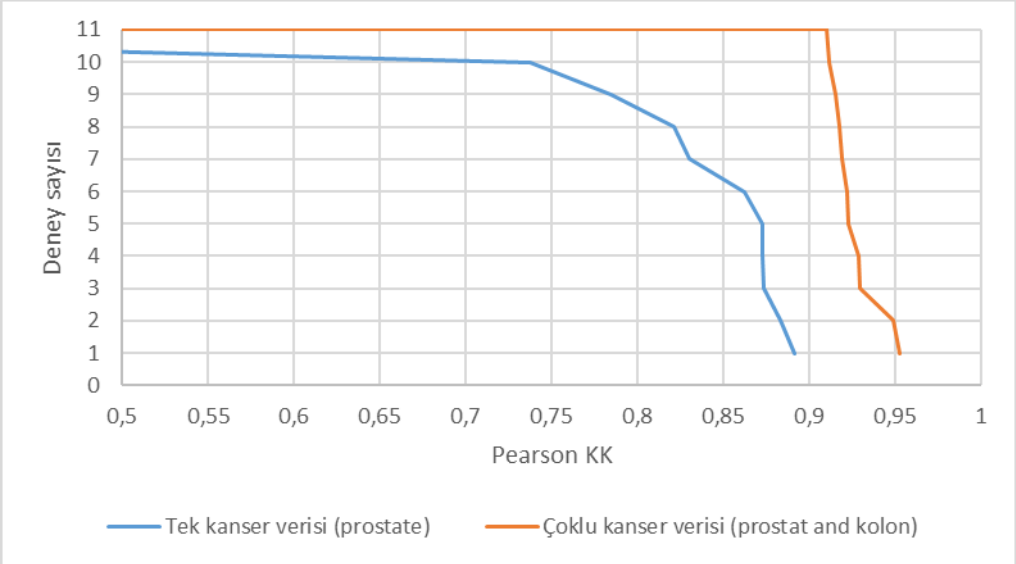
Mikrodizi ölçüm teknolojisinde ölçüm yönteminden kaynaklı olarak farklı platformlardan elde edilen gen ifade matrisleri; veri aralıkları ve referans noktaları bakımından farklı olabilir. Bu durumun performans sonuçlarına olası etkisini önlemek için;

1. İki veri kümesi de GPL570 platformu kullanılarak elde edilen veri setleri arasından seçilmiştir.
2. Veri kümeleri 0-1 arasında normalize edilmiştir.

Prostat kanseri veri kümesinde 11 deney ve kolon kanseri veri kümesinde 6 deney vardır. Toplam deney sayısı buradan $N = 17$ ve gen sayısı $M = 1600$'dür. İki yönlü işbirlikçi filtreleme ile elde edilen 27200 $(M * N)$ satır ve 1617 $(M + N)$ sütundan

oluşan iki yönlü veri matrisi, daha önce en iyi kestirim sonuçlarını elde ettiğimiz kestirim konfigürasyonu (RBF-1, $\sigma = 2$) çerçevesinde test edilmiştir.

Şekil 4.9'daki Pearson KK eğrilerinin altında kalan alanlar karşılaştırıldığında prostat kanseri gen ifade değerlerinin tahmin edilmesinde prostat kanseri gen ifade değerleri ile kolon kanseri ifade değerlerinin bütünleştirilmesinin kestirim performansını yaklaşık %10 oranında artırdığı görülmüştür. Tek bir kanser verisi ve birden fazla kanserin bütünleştirilmiş verisinden eğitilen regresyon modeli ile elde edilen kestirim performansı ortalama Pearson KK değerleri sırasıyla 0,848 ± 0,05 ve 0,925 ± 0,01'dir.



Şekil 4.9 Birden fazla farklı kanser verisi kullanımının kestirim performansına etkisi

Çizelge 4.3'te farklı kanser türlerine ait gen ifade değerlerinin bütünleştirilmesinin kestirim performansına etkisi Spearman KK ve Pearson KK ölçütleri ile gösterilmektedir. Çizelge incelendiğinde meme kanseri gen ifade değerlerinin diğer kanser verileri ile bütünleştirilmesi sonrası eğitilen regresyon modelinin prostat ve kolon kanserine ait gen ifade değerlerinin tahmin edilmesinde daha düşük performansa sahip olduğu görülmektedir. Bunun yanında sadece kolon ve prostat kanseri verilerinin regresyon modelinin eğitiminde kullanılması durumunda meme kanserine ait gen ifade değerlerinin kestirim performansının çok daha düşük olduğu görülmektedir.

Çizelge 4.3 Farklı kanser verilerinin bütünleştirilmesinin kestirim performansına etkisi

Eğitim verisi	Korelasyon katsayısı	Kestirim yapılan kanser türü		
		Meme	Kolon	Prostat
Meme ve Kolon	Spearman	-	-	0,650
	Pearson			0,525
Meme ve Prostat	Spearman	-	0,543	-
	Pearson		0,527	
Kolon ve prostat	Spearman	0,004	-	-
	Pearson	0,001		

Gen ifade miktarı ölçümünde en yaygın ve uzun zamandır kullanılan yöntemin mikrodizi teknolojisi olduğu daha önce ifade edilmiştir. Literatürde mikrodizi verileri kullanılarak yapılan çalışma sayısının fazla olmasına rağmen son yıllarda geliştirilen yeni nesil dizileme yöntemi ile elde edilen gen ifade verilerinin kullanıldığı çalışmalar da mevcuttur. Daha maliyetli bu yöntem ile genlerin çalışma ve etkileşimleri hakkında daha fazla bilgiye erişilmekte olup özellikle araştırma amaçlı kullanımları artmaktadır.

Ayrıca mikrodizi verileri kullanılarak en iyi kestirim performansının elde edildiği model konfigürasyonunun RNAseq verisi için de etkin olup olmadığı araştırılmıştır. Bu kapsamda Illumina HiSeq 2500 biyoçipi kullanılarak GPL16791 platformundan elde edilen GEO veritabanında GSE89134 erişim numaralı 14 deneyden oluşan RNAseq verisi kullanılmıştır. Mikrodizi verilerinde en iyi kestirim performansına ulaşılan model konfigürasyonu RVM regresyon (RBF-1, $\sigma = 2.0$) rastgele seçilen 2800 adet gen için uygulanmıştır. Çizelge 4.4'te 2800 gene ait mikrodizi ve RNAseq verilerinde uygulanan kestirim performans sonuçları karşılaştırılmaktadır. RVM regresyon modelinin RNAseq verisi için de yüksek kestirim performansı sağladığı gözlemlenmiştir.

Çizelge 4.4 Mikrodizi ve RNAseq verileri için elde edilen ortalama performans ölçütleri

Veri	Spearman KK	Pearson KK	RMSE
Mikrodizi	0,733	0,896	0,082
RNAseq	0,790	0,909	0,011

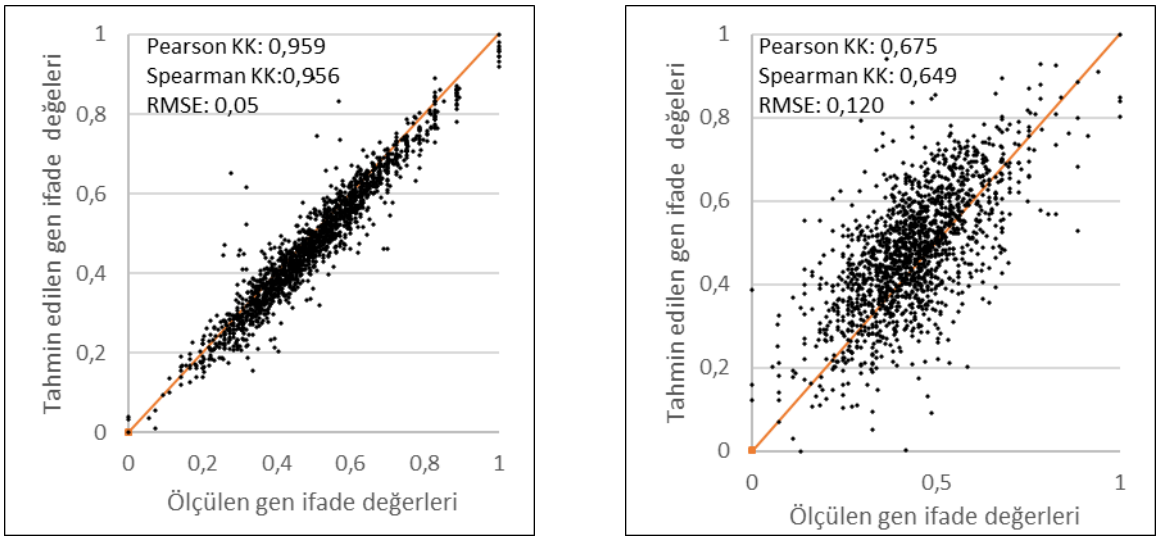
Bu bölümde elde edilen sonuçlar; RVM regresyon modeli, İki Yönlü İşbirlikçi Filtreleme ve çoklu kanser verisinin bütünleştirilmesi işlemlerinin kestirim performansını artırdığını göstermektedir. Kestirim performansındaki bu artışın istatistiksel olarak anlamlı olduğunu göstermek için tüm karşılaştırma durumları için eşleştirilmiş t-testi (Paired t-Test) ve Wilcoxon signed rank testi analizleri gerçekleştirilmiştir. Çizelge 4.5'te yer alan sonuçlara göre kestirim performanslarının istatiksel olarak birbirinden farklı olduğu görülmektedir.

Çizelge 4.5 Her bir durum için karşılaştırmalı istatistiksel analizler

Karşılaştırma	Eşleştirilmiş t-testi	Wilcoxon signed rank testi
Bir yönlü vs. İki yönlü gösterim	$1{,}11\text{x}10^{-10}$	$9{,}77\text{x}10^{-4}$
RBF1 vs. Doğrusal çek. fonk.	$9{,}95\text{x}10^{-7}$	$9{,}77\text{x}10^{-4}$
RBF2 vs. Doğrusal çek. fonk.	$1{,}66\text{x}10^{-6}$	$9{,}77\text{x}10^{-4}$
RVM Reg. vs. Doğrusal Reg.	$5{,}07\text{x}10^{-4}$	$1{,}95\text{x}10^{-3}$
Tekli kanser verisi vs. Çoklu kanser verisi	$3{,}28\text{x}10^{-4}$	$9{,}77\text{x}10^{-4}$

Kestirim performansındaki değişimi göstermek için kullanılan bir diğer gösterim biçimi ise saçılım grafiğidir. Bu grafik biçiminde eksenlerden biri deneyde ölçülen gerçek gen ifade miktarlarını gösterirken diğer eksen ise tahmin edilen değerleri göstermektedir. Saçılım grafiğinde ölçülen ve tahmin edilen veriler diyagonal eksene olan uzaklığı kestirim performansını göstermektedir. Şekil 4.10'da mikrodizi verisi için en iyi ve en kötü kestirim performanslarına ait saçılım grafikleri yer almaktadır. Kestirimin iyi olduğu grafikte, ölçülen ve tahmin edilen gen ifade değerlerinin diyagonal eksene daha yakın olduğu görülmektedir. Buna karşılık kötü kestirim performansına ait saçılım grafiğinde ölçülen ve tahmin edilen gen ifade

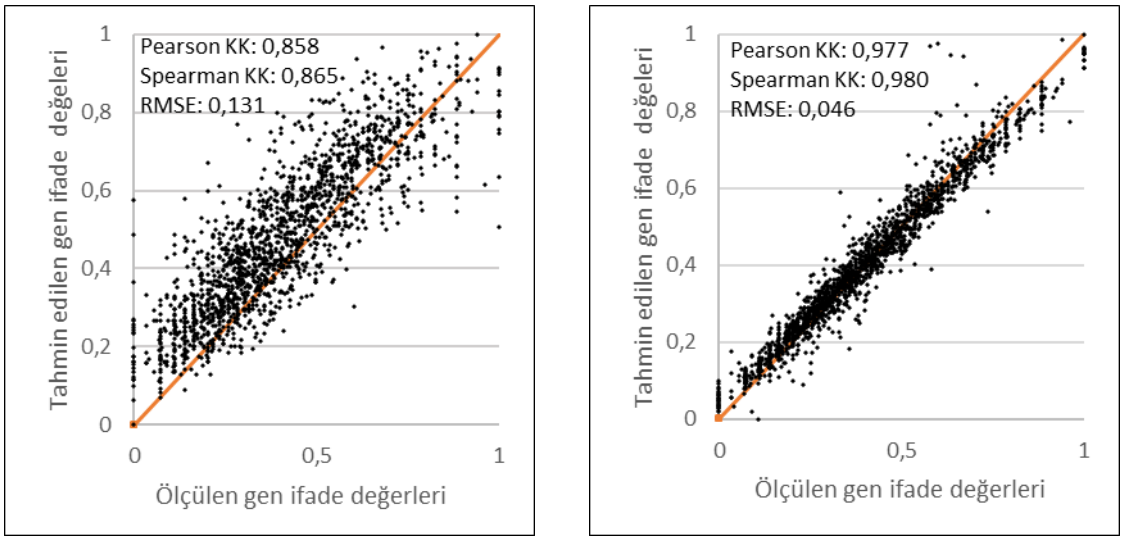
değerlerinin daha dağınık ve diyagonal eksenden daha uzakta olduğu görülmektedir. En iyi kestirim için Pearson KK, Spearman KK ve RMSE değerleri sırasıyla 0.959, 0.956 ve 0.050'dir. En kötü kestirim için Pearson KK, Spearman KK ve RMSE değerleri sırasıyla 0.675, 0.649 ve 0.120'dir.



a. b.

Şekil 4.10 Mikrodizi verisi için saçılım grafiği a) En iyi kestirim b) En kötü kestirim

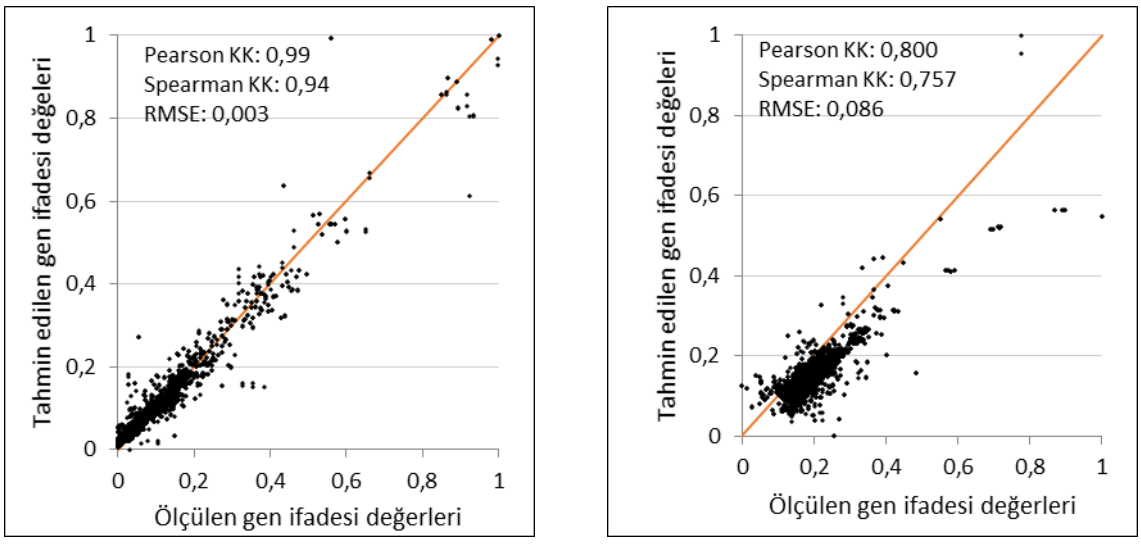
Şekil 4.11(a)'daki grafikte model eğitiminde sadece prostat kanser verisinin kullanılması durumunda elde edilen kestirim değerleri ve gerçek gen ifade değerleri yer almaktadır. Pearson KK, Spearman KK ve RMSE değerleri sırasıyla 0.858, 0.865 ve 0.131'dir. Şekil 4.11.b'de ise prostat ile kolon kanseri verilerinin bütünleştirilmesi sonrası elde edilen kestirim değerleri ve gerçek gen ifade değerleri yer almaktadır. Pearson KK, Spearman KK ve RMSE değerleri sırasıyla 0.977, 0.980 ve 0.046'dır. Her iki grafikteki gerçek gen ifade değerleri aynıdır. Saçılım grafikleri incelendiğinde prostat ve kolon kanseri gen ifade değerlerinin bütünleştirilmesi sonrası elde edilen yeni gen ifade matrisinin modelde kullanılmasının kestirim performansını artırdığı açık bir şekilde anlaşılmaktadır. Sonuç olarak birden fazla kanser türüne ait gen ifade verilerinin kullanılmasının kestirim performansını artırdığı gözlemlenmiştir.



a. b.

Şekil 4.11 Farklı kanser verilerinin bütünleştirilmesi a. Tek kanser çeşidi b. Birden fazla kanser çeşidi

Şekil 4.12'de ise RNAseq verisi için en iyi ve en kötü kestirimlere ait saçılım grafikleri yer almaktadır. En iyi kestirim için Pearson KK, Spearman KK ve RMSE ortalama değerleri sırasıyla 0.990, 0.940 ve 0.003'tür. En kötü kestirim için Pearson KK, Spearman KK ve RMSE ortalama değerleri sırasıyla 0.800, 0.757 ve 0.086'dır. Şekil incelendiğinde aykırı (outlier) olarak değerlendirilebilecek verilerin var olduğu görülmektedir. Çalışmanın hiçbir kısmında aykırı veri silme işlemi uygulanmamıştır.



a. b.

Şekil 4.12 RNAseq verisi için saçılım grafiği a. En iyi kestirim b. En kötü kestirim

4.4. Tartışma

Bu bölümde, gen ifadesi değerinin tahmin edilmesinde başarımı artıracak birçok yeni yöntem ortaya konulmuştur. Bu yeni yöntemler; RVM regresyon modeli, İki Yönlü İşbirlikçi Filtreleme ile yeni öznitelik sunum yaklaşımı ve birden fazla kanser türüne ait gen ifade verilerinin bütünleştirilmesidir. Ölçüm yönteminden bağımsız olarak gen ifade miktarlarının hücredeki moleküler düzeyde meydana gelen olayların en önemli yansıması olduğu bilinmektedir. Gen ifade matrisi üzerinden yapılan çalışmaların çoğu bu nedenle geriye dönük olarak hücrede genler ile diğer moleküller arasındaki ilişkilerin keşfedilmesini amaçlamaktadır. Hücrede, genler arası veya genler ile diğer moleküller arası etkileşimlerin oldukça karmaşık bir düzende gerçekleştiği bilinmekte olup bu etkileşimlerin keşfine yönelik araştırmalar devam etmektedir. Bu karmaşık düzenin en iyi temsil edildiği modelin oluşturulması özellikle gen ifade tahmini gibi çalışmalarda oldukça önemlidir. RVM regresyonunda doğrusal olmayan çekirdek fonksiyonların kullanılması ile farklı bir düzlemde haritalama yapıldığı ile hücredeki karmaşık düzenin daha iyi temsil edildiği düşünülmektedir. Bunu desteklemek amacıyla bu çalışmada RVM ile doğrusal regresyon modelleri karşılaştırılmıştır. Öznitelik sunum biçimi her iki yöntem için de aynı olmak koşuluyla RVM regresyon modelinin daha iyi kestirim başarımına sahip olduğu gösterilmiştir.

Sonuçlar incelendiğinde; İki Yönlü İşbirlikçi Filtreme ile ortaya konulan yeni öznitelik sunum yönteminin kestirim performansını artırdığı görülmüştür. Daha önce ifade edildiği gibi hücrede meydana gelen olaylar ve genler arası etkileşimin kantitatif sonucu ölçülen gen ifade miktarlarıdır. İki yönlü veri matrisinin kullanılması ile farklı genlerin aynı deneye ait ifade miktarlarının model öğrenmede kullanılması ise genler arası etkileşimlerin gen ifade tahmini hesabına katılması anlamına gelmektedir. Bu yaklaşım ile daha iyi kestirim performansına ulaşılmasının biyolojik süreçlerin hesaba katıldığı modellerin kestirim performansını artırmak amacıyla kullanılabileceği sonucunu ortaya koymaktadır. Bu yaklaşım ile elde edilen kestirim performansları, genlerin birbiriyle olan ilişkilerini ortaya koymaktadır. Buradaki biyolojik bilgiye benzer olarak miRNA ve mRNA arasındaki ilişki de kestirim performansının artırılması amacıyla da kullanılabilir.

Tek yönlü ve iki yönlü veri matrislerini RVM regresyon modeli kullanarak test ettiğimizde; iki yönlü öznitelik sunum biçiminin kestirim performansını %5 - %7 oranında artırdığı görülmüştür. Diğer birçok parametreye ek olarak, gen ifade değerlerinin dağılım aralığının korelasyon katsayıları üzerinde bir etkisi olduğu açıkça görülmektedir. Bu yüzden genel olarak önerilen yöntemin başarımını ayırt etmek için birbirinden bağımsız birden fazla hastanın verisi kullanılmıştır. RNAseq verilerinin, mikrodizi verilerinin aksine daha heterojen dağılım gösterdiği görülmektedir (Şekil 4.12). Tüm verilerde kullanılabilecek bir yöntemin geliştirilmesi amacıyla gen ifade verilerinin normalizasyonu dışında herhangi bir ön işleme prosedürü (aykırı verinin silinmesi vb.) uygulanmadan kullanılmıştır.

Elde edilen sonuçlar, İki Yönlü İşbirlikçi Filtreleme yöntemi ile RVM regresyon modelinin, mikrodizi ve RNAseq verileri üzerinde makine öğrenme problemlerinde kullanılabileceğini göstermektedir. RVM regresyon modelinin diğer modellere kıyasla daha kararsız performans göstermesine rağmen, ilgili çekirdek fonksiyonu parametreleri ile optimize edilebilir bir avantaja sahiptir.

Birden fazla kanser türüne ait gen ifade değerlerinin aynı modelde kullanılması ile kestirim performansının artırılabileceği görülmüştür. Ancak bu kanser türlerinin patofizyolojik açıdan benzer olması önemlidir. Örneğin bu çalışmada; prostat veya kolon kanseri verisinin meme kanserine ait gen ifadesi değerlerinin kestirim performansını artırmadığı görülürken kolon kanseri verisinin prostat kanserine ait gen ifadesi değerlerinin kestirim performansını artırdığı görülmektedir. Yurt dışındaki bazı tıp merkezleri tarafından prostat kanseri olan bir hastanın mutlaka belirli periyotlarda kolonoskopi tetkiki yaptırması gerektiği ifade edilmektedir. Bunun nedeni anatomik, fizyolojik ve patofizyolojik olarak prostat ve kolon kanser türlerinin ilişkili olması gösterilebilir. Bölüm 4.3'te yer alan sonuçlar ile prostat ve kolon kanseri arasındaki ilişkinin bu kanser türlerinin oluşmasında benzer gen aktivitelerinin bulunduğunun bir göstergesi olduğu düşünülmektedir.

5. VERİ BÜTÜNLEŞTİRME

5.1. Giriş

Bu bölümde; miRNA ve TF regülasyon bilgisi kullanılarak gen ifade tam değerinin kestirim başarımını artırmak için yapılan çalışmalar sunulmaktadır. Daha önce gen translasyonu sürecinde miRNA, TF ve mRNA arasındaki ilişkiler Bölüm 2'de anlatılmıştı.

Teknolojinin hızla gelişmesi ile birlikte yeni ölçüm platformları ortaya çıkmaktadır. Çok fazla tipte verinin bütünsel bir çerçevede birleştirilerek işlenmesi gittikçe önem kazanmaktadır. Bu nedenle "veri bütünleştirme (data integration)" olarak bilinen bir araştırma alanı ortaya çıkmıştır. Literatürde "veri bütünleştirme" terimi genellikle daha iyi eğitilmiş modellerle daha anlamlı sonuçlar elde etmek amacıyla farklı kaynaklardan elde edilen aynı amaçla toplanmış verilerin aynı formata dönüştürülerek kullanılması anlamını taşımaktadır. Genel olarak, genlerin, proteinlerin ve moleküler düzeyde meydana gelen olayların keşfedilmesi ve daha iyi anlaşılması için farklı platformlardan veya çalışmalardan elde edilen çeşitli veri türleri entegre edilebilir. Bunun için benzer veya heterojen veri tiplerini bütünleştirmek için birçok yöntem literatürde mevcuttur [65].

Bir çalışmada, 7 kolon kanseri dokusundan ve 4 sağlıklı dokudan alınan miRNA ve mRNA ifade profillerinden elde edilen birleştirilmiş yeni ifade profillerine Kısmi En Küçük Kareler (partial least squares) Regresyonu ve bootstrap tabanlı testler uygulanmıştır. Burada yapılan analiz ile miRNA hedefleri ve ilgili miRNA-mRNA ağları tespit edilmiştir. Pathway analizi yapılarak miRNA'lar ve ilişkili mRNA hedeflerinin tespiti ile biyolojik süreçler tanımlanmaya çalışılmıştır. Bilinen miRNA ve mRNA ağı ile ifade matrisleri birleştirilerek miRNA'ların hedef mRNA'larının tespit edilebileceği gösterilmiştir [65].

Bir miRNA çok sayıda mRNA'yı düzenleyebilir. Tek bir mRNA üzerinde her bir miRNA'nın etkisi sınırlı olduğundan hedef mRNA'nın düzenlenmesinde birden fazla miRNA birlikte çalışabilmektedir. Bu durum gen düzenleme sürecini daha da karmaşık hale getirmektedir. miRNA'ların ifade değerleri deneysel olarak ölçülmekle

birlikte hedef mRNA'ların tespiti daha zor bir konudur. Bu kapsamda miRNA ve mRNA ifade profillerini bütünleştiren korelasyon ve regresyon tabanlı çalışmalar mevcuttur. Bu çalışmaların genelinde ifade vektörleri üzerinden doğrudan yapılan olasılıksal yaklaşımlarda miRNA'ların bütünleşik bir şekilde gen düzenlemeye etkisi göz ardı edilmektedir. Yumurtalık kanseri verisi üzerinde test edilen PIMiM yazılımı; mikrodizi ifade değerleri ile miRNA-mRNA regülasyon bilgisi kullanılarak oluşturulmuş bir sistem olup miRNA tarafından düzenlenen mRNA gruplarının tespitinde kullanılabilmektedir [66]. Ayrıca miRNA'ların protein sentezindeki işlevini analiz etmek için farklı biyokimyasal ve deneysel kaynaklardan elde edilen verilerin bütünleştirildiği çalışmalar da mevcuttur [77].

Gelişen teknoloji ile birlikte farklı yöntemlerin ortaya çıkması hücre içindeki biyolojik süreçlerin daha iyi anlaşılması için farklı türde ve sayıda kullanışlı bilgiler sunabilmektedir. Bu nedenle gen düzenleyici ağları yerine fonksiyonel bağlantı ağlarının daha popüler hale geldiği görülmektedir. Örneğin mikrodizi teknolojisi yerine son birkaç yılda kullanılmaya başlanılan yeni nesil RNA dizilime yönteminde protein sentezi sürecine ilişkin daha fazla bilgi elde edilmektedir. Bu bilgilerin geliştirilen bir modelde birleştirilerek kullanılması ise protein sentezi sürecinin moleküler seviyede daha iyi anlaşılmasını sağlamaktadır. Bu kapsamda veri bütünleştirme işlemlerinde karşılaşılan hesaplama zorluklarının başında; farklı veri boyutları, formatları, büyüklükleri ve farklı veri tabanlarına özgü gürültüler gelmektedir. Ayrıca farklı veri tabanlarına ait veri kümesi seçimi de bütünleştirmenin bir parçası olarak değerlendirilebilir. Farklı veri bütünleştirme tekniklerini karşılaştıran bir çalışmaya göre bazı yöntemler küçük boyuttaki veri kümelerinde başarılı iken bazıları daha büyük boyutlardaki verilerde belirgin performanslar göstermektedir. Seçilen veri kümesinin homojen veya heterojen dağılım göstermesi de kullanılan yönteme göre farklı performans düzeylerinde sonuçların elde edilmesinde önemli yer tutmaktadır [67].

Literatürde yaygın olarak kullanılan veri bütünleştirme çalışmaları çok fazla verinin var olduğu durumlarda önemli görülmektedir. Ancak hâlihazırda veri tabanlarında yer alan verilerin büyük çoğunluğu anlatılan veri bütünleştirme işlemleri için yeterli boyutta değildir. Örneğin yeni nesil dizilime yöntemi oldukça zengin veriler sunmakla

birlikte pratikte her klinikte veya araştırma merkezinde yer alan bir teknoloji değildir. Özellikle sağlık hizmeti veren tesislerde çoğunluklar mikrodizi teknolojileri yeterli gelmektedir. Araştırma amaçlı kullanılan yeni nesil dizilime cihazları yerine yaygın olarak kullanılan mikrodizi teknolojisinden elde edilen verilerden maksimum faydanın elde edilmesinin de ayrı bir önemi vardır. Bu bağlamda en yaygın elde edilen verilerden klinik bulguları veya diğer tanı-tedavi süreçlerine katkı sağlayan çıktılar üretecek modellerin oluşturulması ve yazılımların geliştirilmesi oldukça önemlidir.

Bu tez çalışmasının amaçlarından biri de miRNA ve TF moleküllerinin düzenleyici işlevlerini kullanarak mRNA ifade tam değerinin kestirim performansını artıracak bir yaklaşım ortaya koymaktır. Bu bölümde gen ifade tahmininde kullanılan regresyon tabanlı modellerin; farklı veri türlerinin bütünleştirilmesi ile oluşturulması amaçlanmaktadır. Şekil 5.1'de bütünleştirilecek olan veri yapıları görülmektedir. Gen ifadesi tam değeri numerik veri tipine sahip iken miRNA regülasyon bilgisi ikili kodlama (binary) biçimindedir. Numerik değerin tahmin edildiği bir regresyon modeline ikili kodlama yapısındaki verilerin doğrudan girdi olarak verilmesi matematiksel olarak doğru bir yaklaşım değildir. Şekil 5.1'de A çerçevesi ile gösterilen verilerin kullanımı numerik değerler üzerinden bağıntı kurmak anlamına gelirken; B çerçevesi içindeki verilerin aynı modele entegre edilmesi ise translasyon ve transkripsiyon süreçlerindeki biyolojik bilginin bilgisayar hesaplamalarında birlikte kullanılması anlamına gelmektedir.

	Ölçülen gen ifadesi değeri	**A Diğer deneylerdeki gen ifadesi değerleri**				**B miRNA regülasyon bilgisi**				**Tahmin edilen gen ifadesi değeri**
Gen		Ör1	Ör2	…	Ör**N**	mi1	mi2	…	mi**M**	
1	0,90	0,36	0,58	1,88	2,45	1	0	1	1	?
2	0,78	0,55	0,24	1,22	1,11	0	1	1	0	
3	0,86	1,62	1,25	0,87	0,45	1	0	1	0	
4	1,43	0.95	2,48	0,41	0,69	1	1	1	0	
5	0,56	0,75	0,69	0,19	0,78	0	1	0	1	
…	…	…	…	…	…	…	…	…	…	

Şekil 5.1 Veri bütünleştirmedeki veri yapıları

Literatüre bakıldığında, veri bütünleştirme olarak yapılan çalışmaların büyük çoğunluğunun; farklı veri tabanlarından veya dokulardan elde edilen gen ifadesi değerlerinin birleştirilmesi üzerine yoğunlaştığı görülmektedir. Bunun nedeni farklı tipte ve miktarda genomik ve klinik verinin sürekli artmasıdır. Biyolojik süreçler oldukça karmaşıktır ve uygulanan tekniklerle elde edilen ölçümler ışığında bu karmaşık süreçler anlaşılmaya çalışılmaktadır. Aynı bireylerden elde edilen heterojen verilerin entegre edilmesi ihtiyacı, geniş çapta klinik uygulamalarda da ortaya çıkmaktadır. Buna en iyi örnek belki de kanser araştırmacılarının ve onkologların kanser gibi karmaşık bir hastalığın teşhisi, tedavisi ve prognozunda karşılaştıkları zorluklardır. Kanserin klinik takibi şu anda büyük ölçüde klinik çalışmalardan elde edilen bilgiler üzerine dayanmaktadır. Bununla birlikte, kanser oluşumunda genetik mutasyonların önemli etken olduğu düşünülmekte olup gen ifadesi, regülasyon ve protein sentezi gibi moleküler düzeyde elde edilen veriler, tümör sınıflandırılmasında ve kanser prognozunda kullanılabilir. Dolayısıyla bu alandaki çalışmalar kanserin klinik açıdan izlenmesine de katkı verebilir. Farklı veri türlerinin birleştirilmesine yönelik veri bütünleştirici yaklaşımlar karmaşık hastalık süreçlerinin daha iyi anlaşılmasında yararlı olabileceği düşünülmektedir.

5.2. Materyal ve Yöntem

5.2.1. Veri

Veri bütünleştirme çalışmalarında gen ve miRNA ifade profillerini içeren GEO veritabanında yer alan GSE75285 erişim kodlu 54 hastaya ait meme kanseri verisi kullanılmıştır [37]. Bu veri kümesinde daha önceki bölümlerde anlatılan gen tanımlayıcı numaraların gen sembollerine dönüşümleri sonrası 5082 gen ifade vektörü elde edilmiştir. Bu bölümde 705 miRNA'ya ait ifade miktarları ve regülasyon bilgisi kullanılmıştır. miRNA-mRNA regülasyon bilgisi için üç farklı veritabanından elde edilen veriler kullanılmıştır. Bunlar mirDB, miRTarBase ve mirConnX veri tabanlarıdır. Her bir veritabanındaki miRNA sayısı ve seçilen veri kümesindeki düzenledikleri genlerin sayısı farklıdır. mirDB ve miRTarBase veri tabanlarındaki bilgiler birleştirilerek tek bir miRNA-mRNA regülasyon matrisi oluşturulmuştur. İfade değeri bulunan 705 miRNA içinden 368 tanesinin 4707 geni regüle ettiği bilgisi bu veritabanından elde edilmiştir. mirConnX veritabanında ise 705 miRNA içinden 104

tanesinin 3335 geni regüle ettiği bilgisi yer almaktadır. Bu nedenle çalışma bir veya birden fazla miRNA tarafından regüle edilen genlerin ifade değerlerinin tahmin edilmesi yönünde ilerlemiştir. Tüm bu veri tabanlarında yer alan veriler 0-1 (binary) veri formatındadır. mRNA-miRNA regülasyon matrisinde herhangi bir hücredeki değer 0 ise bu değere karşılık gelen satırdaki miRNA'nın sütundaki mRNA'yı regüle etmediği, 1 ise regüle ettiği anlamına gelmektedir.

5.2.1.1. mirTarBase veritabanı

miRTarBase veritabanı; ilgili deney, western blot, mikrodizi ve yeni nesil dizilime yöntemleri ile doğrulanan üç yüz altmış bin miRNA-hedef etkileşim verisinden oluşan literatürdeki en güncel miRNA regülasyon bilgisini sunan bir veritabanıdır. Tayvan'da National Chia Tung Üniversitesi Biyoinformatik ve Sistem Biyolojisi Enstitüsü Biyolojik Bilimler ve Teknoloji Bölümü bünyesindeki ISBLab tarafından kullanıma açılmıştır. En güncel versiyonu 15.09.2017 tarihinde miRTarBase 7.0 olarak yayınlanmıştır. Bu versiyonunda 23 farklı türden elde edilen 4.076 miRNA etkileşimi 8.510 bilimsel yayın ile doğrulanmış olarak sunulmaktadır. Bu veritabanında toplam miRNA-hedef etkileşim sayısı 422.517'dir [68]. Bu çalışmada da son versiyondaki veriler kullanılmıştır [38].

5.2.1.2. mirDB veritabanı

mirDB veritabanında insan, fare, rat, köpek ve tavuk türlerine ait genomik veriler yer almaktadır. Bu veritabanı Washington Üniversitesi Tıp Fakültesi Radyasyon Onkolojisi Bölümü'ndeki Xiaowei Wang'ın laboratuvarı tarafından kullanıma açılmıştır. İnsan dokusuna ait deneylerden elde edilen miRNA sayısı 2.588'dir ve bu miRNA'ların 17.925 geni regüle ettiği bilgisi yer almaktadır. Tüm türlere ait toplam 6.709 miRNA 78.114 geni regüle ettiği bilgisi yer almaktadır. mirDB veritabanının 5.0 güncel versiyonu 2014 yılında yayınlanmıştır [69]. Bu çalışmada güncel versiyonu kullanılmıştır [69].

5.2.1.3. mirConnX veritabanı

mirConnX veritabanı ise diğerlerine göre daha eski ve dar kapsamlı veriler sunmaktadır. 2011 yılında Pittsburgh Üniversitesi Tıp Fakültesi Hesaplamalı ve Sistem Biyolojisi Bölümü tarafından kullanıma açılmış olan bu veritabanında sınırlı sayıda türe ait miRNA-gen regülasyon ağı bilgisi yer almaktadır [70]. Daha geniş kapsamlı regülasyon bilgisinin yanında daha dar çerçevede hazırlanan veritabanının önerilen sistemde olumlu sonuçlar vermesi yaklaşımların başarımı açısından önemli görülmektedir.

5.2.1.4. TRANSFAC veritabanı

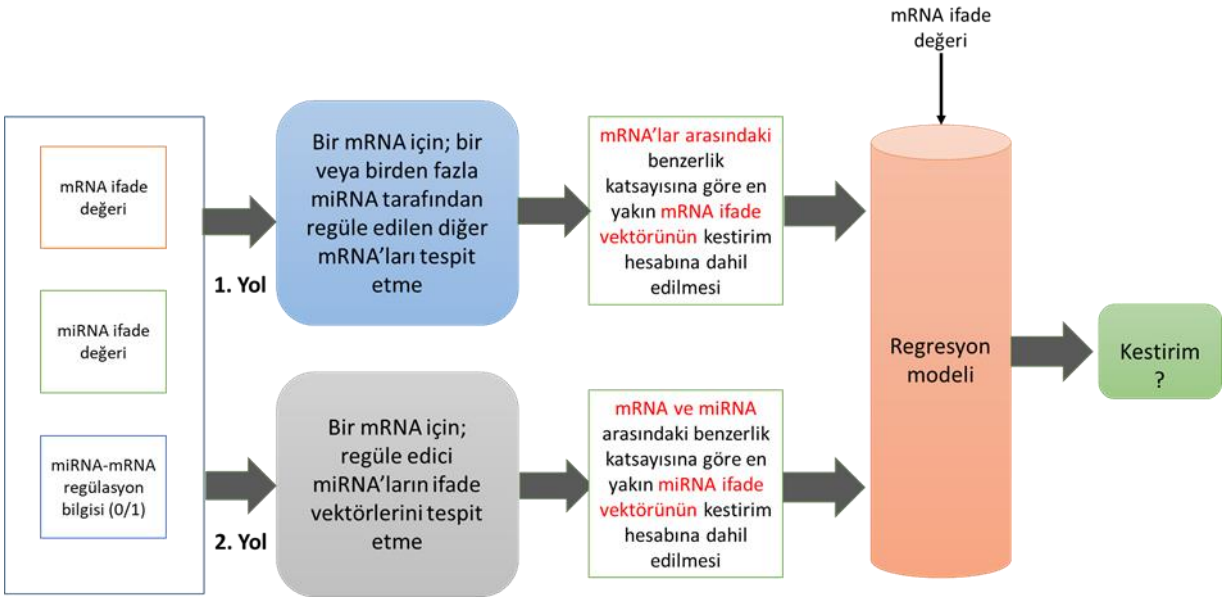
TRANSFAC (TRANScription FACtor database) ökaryotik transkripsiyon faktörleri, deneysel olarak kanıtlanmış bağlanma bölgeleri, pozisyonel ağırlık matrislerinden sağlanan konsensüs bağlanma dizileri ve düzenlenmiş genler hakkında veriler sağlar. Bu veritabanında 1988 yılından beri toplanan veriler yer almaktadır. TRANSFAC veritabanının ilk versiyonu Almanya Ulusal Biyoteknoloji Araştırma Enstitüsü (şimdiki adı Helmholtz Enfeksiyon Araştırma Merkezi) tarafından yayınlanmıştır. Daha sonra 1993 yılında ilk kamu teşvikli biyoinformatik projesi ile TRANSFAC veritabanı internet üzerinden erişilebilir bir yapıya kavuşmuştur. Son olarak Almanya'da geneXplain firması bünyesinde yer almaktadır [71].

5.2.2. Yöntem

Veri bütünleştirme yapılmayan yaklaşımda; bir örneğe ait mRNA ifade tam değerinin kestiriminde diğer örneklere ait ifade tam değerleri regresyon modeline girdi olarak verilmekte ve oluşan model eğrisi ile kestirim yapılmaktadır. miRNA-mRNA regülasyon bilgisi kullanılarak kestirim performansının iyileştirilebildiğini göstermek için veri bütünleştirme işleminin kestirim performansına etkisi analiz edilmiştir.

Şekil 5.2'de miRNA regülasyon bilgisi kullanılan veri bütünleştirme yaklaşımının genel çerçevesi gösterilmektedir. Veri bütünleştirme işleminde iki farklı yol izlenmiştir. Birinci yolda; kestirimi yapılacak gen ifade vektörleri ile düzenleyici

miRNA'ların ifade vektörleri birleştirilmektedir; ikinci yolda ise aynı miRNA tarafından düzenlenen genlerin ifade vektörleri birleştirilmektedir.



Şekil 5.2 miRNA regülasyon bilgisi kullanılan veri bütünleştirme genel çerçevesi

Şekil 5.2'de üçüncü kısımda bütünleştirilecek ifade vektörleri arasındaki uzaklık tespitinde *Affine dönüşüm, Öklid* ve *Bhattacharyya* ölçütleri kullanılmıştır. Regresyon modeli olarak önceki bölümlerde anlatılmış olan doğrusal ve RVM regresyon modelleri kullanılmıştır. Kestirim performansının değerlendirilmesi için yine önceki bölümlerde anlatılan Pearson ve Spearman benzerlik katsayıları ile Hata Kareleri Ortalamasının Karekökü (RMSE) ölçütü kullanılmıştır.

5.2.2.1. Bhattacharyya uzaklık ölçütü

Bhattacharyya ölçütü; genellikle sinyal ve görüntü işleme alanında kullanım alanına sahip olup [72] sinyal işlemede iki farklı dizideki hata olasılıklarını karşılaştırmaya dayalı hesaplanan bir benzerlik katsayısıdır [73]. Bu ölçüt değeri arttıkça iki dizi arasındaki benzerliğin arttığı anlamına gelmektedir. Bu uzaklık ölçüsü sinyal işleme, görüntü işleme ve örüntü tanıma alanlarında sıklıkla kullanılmaktadır. $p_1\ (x)$ ve $p_2(x)$ iki olasılık dağılımı olsun; bunlar arasındaki Bhattacharyya benzerlik katsayısı (BC) sürekli zaman uzayında aşağıdaki denklem ile hesaplanır:

$$BC(p_1, p_2) = \int \sqrt{p_1(x)\, p_2(x)}\, dx \qquad (12)$$

Kesikli zaman uzayında ise aşağıdaki denklem ile hesaplanır:

$$BC(p_1, p_2) = \sum_{x \in X} \sqrt{p_1(x)\, p_2(x)} \qquad (13)$$

Denklemlerde yer alan $p_1(x)$ ve $p_2(x)$ olasılık dağılımları, bu çalışmada aralarında uzaklık ölçümü hesaplanacak olan iki vektördeki ifade değerlerine ait frekans vektörlerinden elde edilmektedir. Frekans vektörü; bir gen ifade vektöründeki her bir gen ifadesinin tekrarlanma sayılarından oluşmaktadır. $\widehat{p_\iota}$ ve $\widehat{p_J}$ iki farklı frekans vektörünün olasılık dağılımı olsun. BC uzaklık ölçütü aşağıdaki denklem ile hesaplanır:

$$BC(i, j) = BC\left(\widehat{p_\iota}, \widehat{p_J}\right) = \sum_{h=1}^{m} \sqrt{\widehat{p_{\iota h}}, \widehat{p_{Jh}}} \qquad (14)$$

Burada m aynı ifade (gen veya miRNA) değerlerini içeren grup sayısını göstermektedir. $\widehat{p_{\iota h}} = \frac{\#h}{\#i}$ olup $\#i$, sıfırdan farklı ifade değerleri için toplam tekrarlanma sayısını gösterirken; $\#h$ ise h ifade değerinin tekrarlanma sayısını göstermektedir. $\sum_{h=1}^{m} \widehat{p_{\iota h}} = \sum_{h=1}^{m} \widehat{p_{Jh}} = 1$ olmaktadır [74].

I ve J, $I = (1.8, 0, 3.4, 4.8, 3.4, 2.5, 4.8, 3.3)^{\mathrm{T}}$ ve $J = (0, 4.8, 1.8, 3.4, 3.4, 2.5, 3.3, 4.8)^{\mathrm{T}}$ olmak üzere iki ifade vektörü olsun. Tekil olarak ifade değerleri (1.8, 2.5, 3.3, 3.4, 4.8) şeklindedir. I ve J arasındaki BC değeri ise;

$$BC(I, J) = \sum_{h=1}^{5} \sqrt{\widehat{I_h}\, \widehat{J_h}}$$

$$= \sqrt{\left(\frac{1}{7}\right)\left(\frac{1}{7}\right)} + \sqrt{\left(\frac{1}{7}\right)\left(\frac{1}{7}\right)} + \sqrt{\left(\frac{1}{7}\right)\left(\frac{1}{7}\right)} + \sqrt{\left(\frac{2}{7}\right)\left(\frac{2}{7}\right)} + \sqrt{\left(\frac{2}{7}\right)\left(\frac{2}{7}\right)} = 1$$

şeklinde hesaplanır. Bu örnekte dikkat edilirse her iki vektördeki dağılımın aynı olduğu kolaylıkla görülebilir. Birbirinden farklı dağılımlara sahip vektörler için bu katsayı değişim gösterecektir. Yukarıdaki I ve J ifade vektörleri iki farklı gen ifade vektörü ya da gen ve miRNA ifade vektörleri olabilir. Bu örnekte hasta sayısı sekizdir.

Bhattacharyya ölçütünün literatürde öznitelik ve veri boyutunu azaltma (feature and data dimension reduction) amacıyla kullanıldığı çalışmalar da mevcuttur. Örneğin kanserli ve kanserli olmayan örneklerle ilişkili gen ifade vektörlerinin Bhattacharyya ölçütü aracılığıyla belirlenmesine yönelik uygulamalar bulunmaktadır. Bir i geni için kolon kanseri ve normal dokulara ait ifade değerleri kullanılarak hesaplanan Bhattacharyya katsayısı aşağıdaki denklem ile bulunur:

$$B(i) = \frac{1}{4}\frac{(\mu_+(i)-\mu_-(i))^2}{(\sigma^2{}_+(i)+\sigma^2{}_-(i))^2} + \frac{1}{2} ln\left(\frac{\sigma^2{}_+(i)+\sigma^2{}_-(i)}{2\sigma_+(i)\sigma_-(i)}\right) \quad (15)$$

Burada $\mu_+(i)$ ve $\sigma^2{}_+(i)$ kolon kanseri dokularına ait gen ifade değerlerinin sırasıyla ortalamasını ve varyansını göstermektedir. $\mu_-(i)$ ve $\sigma^2{}_-(i)$ ise normal dokulara ait gen ifade değerlerinin sırasıyla ortalamasını ve varyansını göstermektedir. $B(i)$ değeri ne kadar büyük olursa gen ile kolon kanseri arasında yakın ilişki olduğu anlamına gelmektedir. Bu sayede binlerce gene ait Bhattacharyya değerleri için bir eşik değer belirlenerek analiz çalışmalarında kullanılacak gen sayısı azaltılabilmektedir [63].

Öklid ve Affine dönüşümü gibi diğer uzaklık ölçütlerinden farklı olarak Bhattacharyya ölçütü iki vektörünün dağılımlarının benzerliğini ölçmektedir. Bu nedenle bu değerin büyük olduğu vektörlerin aslında birbirine yakın olduğu bilinmelidir. Sadece benzer dağılımlara ve karakteristiklere sahip verilerle eğitilen modellerin farklı karakteristik özelliklere sahip test verilerinde başarımı düşmektedir. Bu nedenle birbirinden farklı verilerin eğitimde kullanılması modelin daha doğruluğunu ve hassasiyetinin artırdığı unutulmamalıdır.

5.2.2.2. Affine dönüşümü uzaklık ölçütü

Affine dönüşümünde; gen ifade vektörleri arasındaki uzaklık normalizasyon ve korelasyon işlemleri kullanılarak hesaplanmaktadır. Örneğin; gen ve miRNA ifade vektörleri sırasıyla m ve mi olsun. Affine dönüşümü ile bu vektörler arasındaki benzerliğin hesaplanması için öncelikle bu iki vektör arasındaki korelasyon katsayısı hesaplanmakta ve aşağıdaki denklem kullanılmaktadır [75].

$$D(m, mi) = \sqrt{2 - 2corr(m, mi)} \quad (16)$$

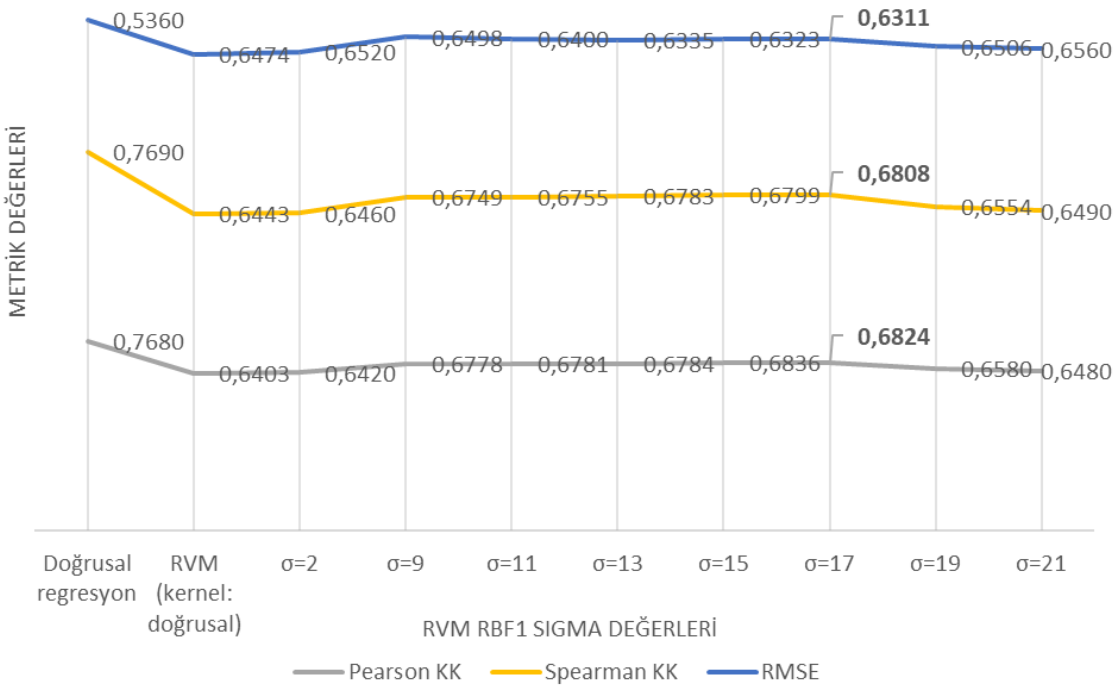
5.3. Sonuçlar

Bu bölümde, sonuçlar karşılaştırmalı olarak sunulmuştur. Bütünleştirme işlemi yapılarak veya yapılmayarak oluşturulan verilerin aynı model ile elde edilen kestirim sonuçları aynı ölçütlerle değerlendirilmiştir. Böylece bütünleştirme yaklaşımlarının ve regresyon modellerinin kestirim performansına etkisi görülmüştür. Farklı yöntemlerin kestirim performansını değerlendirmek için herhangi bir bütünleştirme işlemi yapılmadan bir gen ifadesi değerinin kestiriminde aynı gene ait diğer örneklerdeki ifade miktarlarının kullanıldığı klasik yaklaşım referans olarak alınmıştır. Çalışmada doğrusal ve RVM regresyon modelleri kullanılmıştır.

En basit bütünleştirme işlemi olarak gen ifade vektörleri ile miRNA-mRNA regülasyon bilgileri aynı matriste hiçbir işlem yapılmadan doğrudan birleştirilip regresyon modelinde kullanılarak kestirim çalışmaları da yapılmıştır. Şekil 5.1'de yer alan gen ifade miktarları (float) ve regülasyon bilgisi (ikili değer) aynı matriste birleştirilerek doğrusal regresyon modeline verilmiştir.

GSE75285 erişim numaralı veri setinden elde edilen 54 meme kanseri verisine ait 5082 gene ait ifade verileri doğrusal ve RVM regresyon modelleri ile tahmin edildiğinde ölçülen ve kestirim yapılan gen ifade değerlerinden hesaplanan ortalama Pearson KK, Spearman KK ve RMSE değerleri Şekil 5.3'te yer almaktadır. Burada doğrusal regresyon, RVM regresyon modelinin doğrusal çekirdek fonksiyonu ve RBF-1 fonksiyonundaki farklı sigma (σ) değerleri için hiçbir veri bütünleştirme işlemi

yapılmadan gen ifade tahmini yapılmıştır. Eğriler incelendiğinde doğrusal regresyon modelinin en iyi kestirim performansına sahip olduğu görülmektedir (Spearman KK: 0.769, Pearson KK: 0.768, RMSE: 0.536). RVM regresyon modeli ile yapılan çalışmalarda genel itibariyle kestirim performansının azaldığı görülmektedir. Daha önceki bölümlerde RVM regresyon modeli kullanımında RBF-1 çekirdek fonksiyonunun daha iyi kestirim sonucu verdiği görülmüştü. Bu nedenle bu bölümde RBF-1 çekirdek fonksiyonunun $\sigma = 2, \sigma = 9, \sigma = 11, \sigma = 13, \sigma = 15, \sigma = 17, \sigma = 19$ ve $\sigma = 21$ değerleri için kestirim işlemleri yapılmıştır. Burada en iyi kestirim performansına $\sigma = 17$ ile ulaşıldığı görülmektedir. Veri bütünleştirme işlemi olmaksızın gerçekleştirilen kestirim çalışmasında doğrusal regresyon ile en iyi performansa ulaşılmaktadır. Doğrusal regresyonun yapısı gereği model girdilerini RVM gibi başka bir düzleme düşürerek işlem yapmamaktadır. Bu nedenden dolayı doğrusal regresyon modelinin kullanılmasının, veri bütünleştirme yöntemlerinin kestirim performansına olan etkisinin incelenmesi açısından daha uygun olduğu düşünülmektedir.



Şekil 5.3 Veri bütünleştirme işlemi yapılmadan elde edilen ortalama kestirim performans değerleri

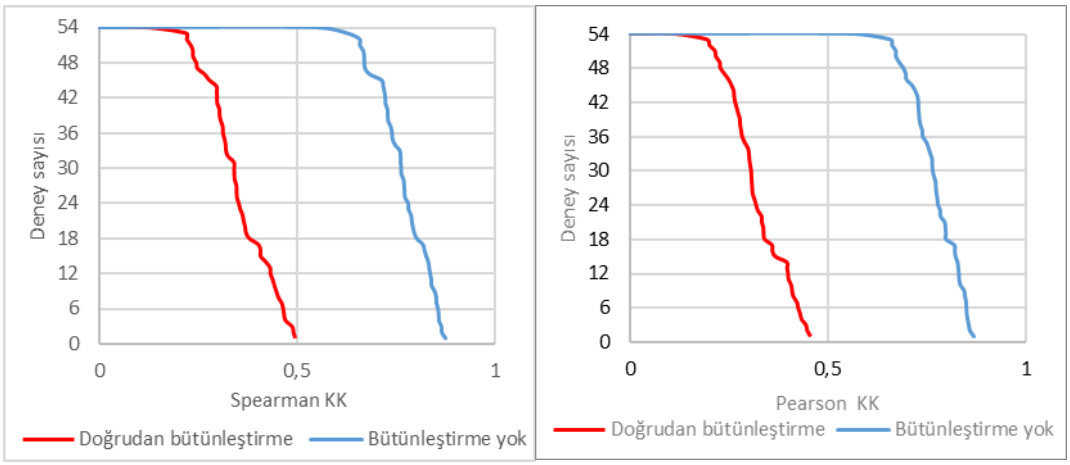
Yukarıdaki şekilde yer alan kestirim sonuçlarını iyileştirmek için öncelikle hiçbir bütünleştirme işlemi yapılmadan miRNA-mRNA regülasyon bilgisi doğrudan regresyon modeline dahil edilmiştir. Bu işlem gen ifade değeri kestiriminde veri bütünleştirme çalışmasının gerekli olup olmadığını ortaya koymak için gerçekleştirilmiştir. Burada regresyon modelinde kullanılan matris; mRNA ifade matrisinin (float) devamına regülasyon bilgisi (ikili değer) matrisinin eklenmesi ile elde edilir. $5082\ gen\ x\ 705\ miRNA$ regülasyon matrisi $5082\ gen\ x\ 54\ deney$ ifade matrisinin devamına eklendiğinde regresyon modeline verilen her bir vektör $(54\ float\ + 705\ binary)$ değerden oluşan bir vektör olmaktadır. Buna istinaden doğrusal regresyon modeli kullanılarak elde edilen ortalama kestirim performans değerleri Çizelge 5.1'de gösterilmiştir.

Çizelge 5.1 Regülasyon bilgisinin kestirim işlemine doğrudan dâhil edilmesi ile elde edilen performans sonuçları

Regresyon modeli	**Bütünleştirme işlemi**	**Pearson KK**	**Spearman KK**	**RMSE**
Doğrusal	Doğrudan	0.319	0.351	1.60
Doğrusal	Yok	0.769	0.768	0.536

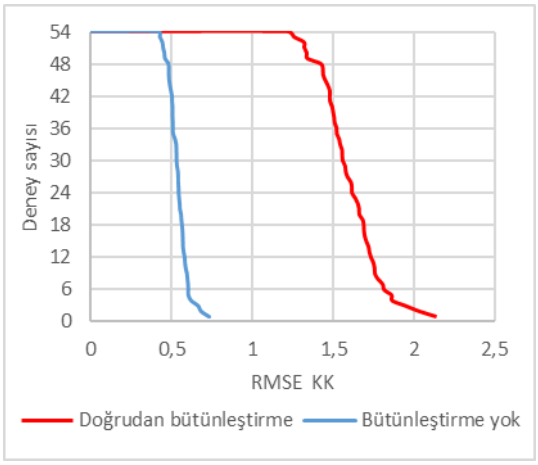
Şekil 5.4'te bütünleştirme işlemi olmadan ve doğrudan bütünleştirme işlemi sonrası yapılan kestirim işlemlerinin performansları Pearson KK, Spearman KK ve RMSE ölçütleri ile karşılaştırılmaktadır. Bu ölçütlerden elde edilen eğrilere göre; miRNA regülasyon bilgisinin kestirim modeline doğrudan entegre edilmesi neticesinde kestirim performansı düşmektedir.

miRNA-mRNA regülasyon bilgisinin regresyon modeline hiçbir işlem uygulamadan doğrudan entegre edilmesi Pearson KK ve Spearman KK ölçütlerinde azalma meydana getirmektedir. Ayrıca Şekil 5.4'teki eğriler altında kalan alanlar incelendiğinde aynı sonuca ulaşılmaktadır. Bunun yanında doğrudan bütünleştirme işlemi sonrası yapılan kestirim işleminde ölçülen gen ifade değerleri ile kestirim değerleri arasındaki hatanın arttığı RMSE eğrisinde belirgin şekilde görülmektedir. Doğrudan bütünleştirme işlemi sonrası yapılan kestirim değerleri ile gerçek değerler arasındaki hata farklarındaki artışın şekilde yer alan RMSE eğrisi altındaki alanın da artmasına neden olduğu görülmektedir.



a.

b.



c.

Şekil 5.4 Doğrudan bütünleştirme işleminin kestirime etkisi a. Spearman KK b. Pearson KK c. RMSE

miRNA'ların regülasyon bilgisi kullanılarak çalışma kapsamında iki farklı bütünleştirme yaklaşımı sunulmuştur. *Birinci yaklaşım*; mRNA'lara ait ifade vektörlerinin devamına bu mRNAları düzenleyen en yakın miRNA ifade vektörünün eklenmesidir. *İkinci yaklaşım* ise mRNA'lara ait ifade vektörlerinin devamına en fazla sayıda ortak miRNA tarafından düzenlenen en yakın mRNA ifade vektörünün eklenmesidir.

Çizelge 5.2'de doğrusal regresyon modeli ile farklı uzaklık ölçütleri kullanılarak elde edilen kestirim sonuçları yer almaktadır. Bhattacharyya uzaklık ölçütü ile doğrusal regresyon kullanılarak mRNA ifadesi bütünleştirme işlemi kullanılarak (1. Yol) diğer ölçütlere (Öklid ve Affine) kıyasla en iyi kestirim sonuçlarına ulaşıldığı görülmektedir. miRNA ifade vektörünün bütünleştirilmesi (2. Yol) işlemi ile bütünleştirme işlemi olmadan yapılan kestirimlere kıyasla daha iyi performans gösterdiği ve kullanılan

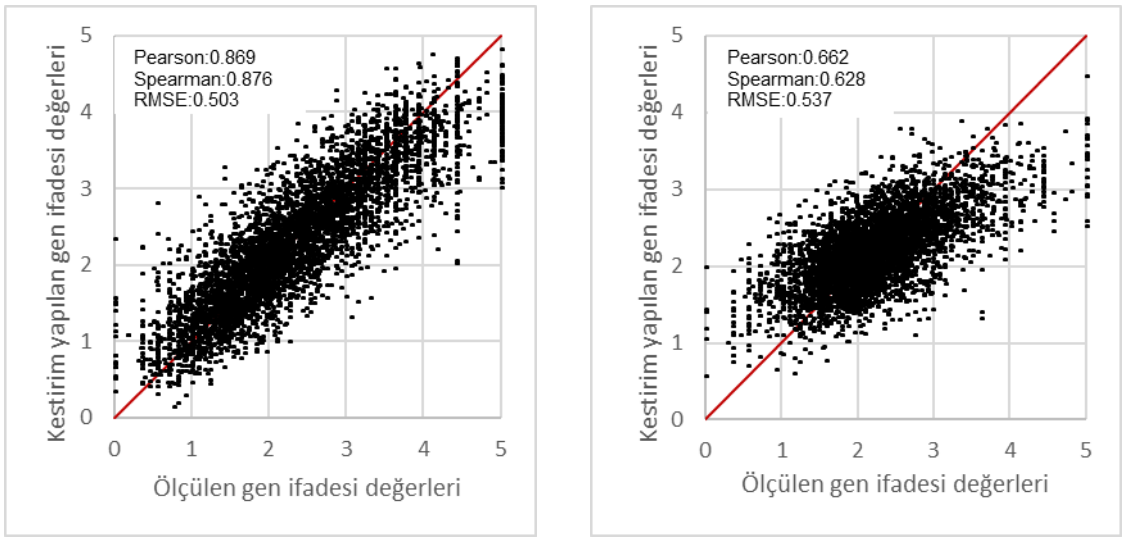
uzaklık ölçütlerinin birbirine çok yakında performans sergilediği görülmektedir. Sonuç olarak izlenen 2. Yol kapsamında her üç uzaklık ölçütünün veri bütünleştirme çerçevesinde kestirim performansını artırdığı söylenebilir.

Çizelge 5.2 Doğrusal regresyon kullanılarak veri bütünleştirme ile elde edilen ortalama kestirim performansları

miRNA–mRNA regülasyon bilgisi kullanılıyor mu?	**İzlenen yol**	**Uzaklık ölçütü**	**Pearson KK**	**Spearman KK**	**RMSE**
Hayır	-	-	0.769	0.768	0.536
Evet	1. Yol (mRNA ifade vektörünün bütünleştirilmesi)	Öklid	0.785	0.781	0.518
		Affine Dönüşüm	0.774	0.774	0.528
		Bhattacharyya	0.959	0.955	0.232
	2. Yol (miRNA ifade vektörünün bütünleştirilmesi)	Öklid	0.828	0.816	0.326
		Affine Dönüşüm	0.829	0.816	0.325
		Bhattacharyya	0.828	0.817	0.326

Şekil 5.5'te veri bütünleştirme işlemi olmadan doğrusal regresyon modeli kullanılarak elde edilen en iyi ve en kötü kestirimlere ait saçılım grafikleri gösterilmektedir. Yatay eksende ölçülen gerçek gen ifade miktarları ve düşey eksende kestirim değerleri yer almaktadır. En iyi kestirimin elde edildiği örneğe ait grafikte (Şekil 5.5.a) ölçülen miktarlar ile kestirim değerlerinin diyagonal eksene yaklaştığı görülmektedir. Buna karşın en kötü kestirim grafiğinde (Şekil 5.5.b) ise diyagonal eksenden uzaklaşıldığı görülmektedir.

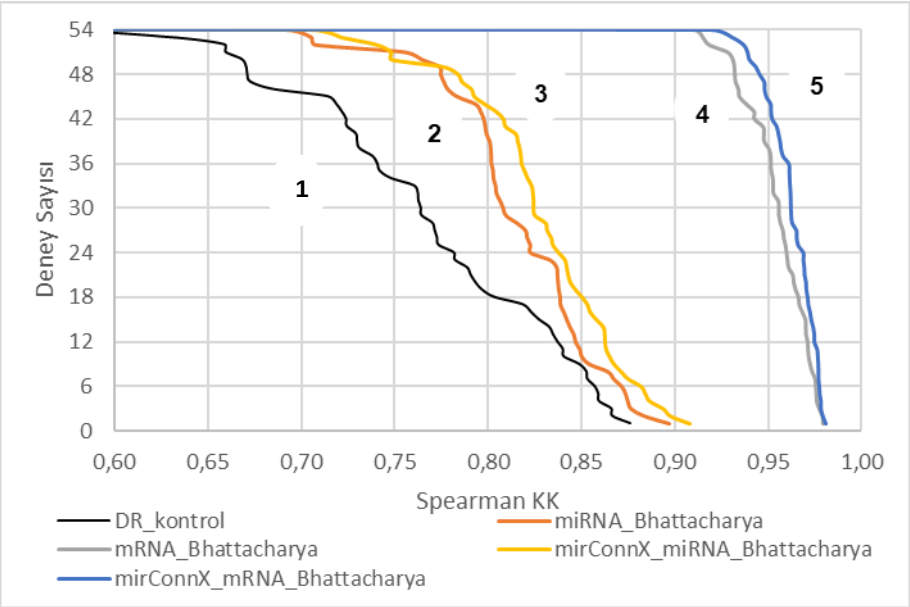
Şekil 5.6'da veri bütünleştirme işleminin kestirim performansına etkisi Spearman KK eğrileri ile gösterilmiştir. Bu gösterim biçiminde, altında en fazla alan olan eğriye ait yöntem en iyi kestirim performansına sahiptir. Buna göre veri bütünleştirme işlemi yapılmadan uygulanan doğrusal regresyon yöntemi (DR_kontrol eğrisi) en kötü performansa sahipken veri bütünleştirme ile daha iyi kestirim performansları elde edilmektedir.



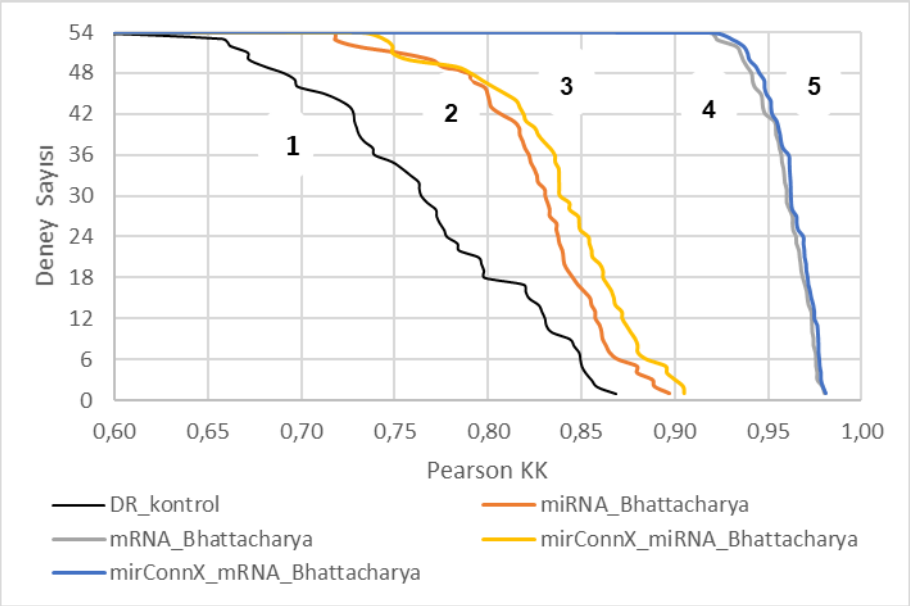
a. b.

Şekil 5.5 Veri bütünleştirme olmadan doğrusal regresyon saçılım grafiği a.En iyi kestirim b. En kötü kestirim

Şekil 5.6 ve Şekil 5.7'deki eğriler altında kalan alanlar hesaplandığında; miRNA-mRNA regülasyon bilgisine göre birbiri ile ilişkili mRNA ifade vektörlerinin bütünleştirilmesi işleminin (4. ve 5. eğriler) miRNA ifade vektörü ile yapılan bütünleştirme işlemine göre daha iyi kestirim performansı sağladığı görülmektedir (2. ve 3. eğriler). Burada 2. ve 4. eğriler miRNA-mRNA regülasyon bilgisinin mirTarBase veritabanından elde edilmesi ile ulaşılan sonuçlardır. Şekildeki 3. ve 5. eğriler ise miRNA-mRNA regülasyon bilgisinin mirConnX veritabanından elde edilmesi ile ulaşılan sonuçlardır. Şekil 5.7'de Pearson KK ölçütüne ait eğriler yer almaktadır. Spearman KK ölçütünde olduğu gibi burada da altındaki alan en büyük olan eğri en iyi kestirim performansını göstermektedir. Yine aynı şekilde veri bütünleştirme işlemi yapılmadan uygulanan doğrusal regresyon modeli en kötü kestirim (DR_kontrol eğrisi) performansına sahiptir. Buna karşın veri bütünleştirme ile daha iyi kestirim performanslarının elde edildiği görülmektedir.

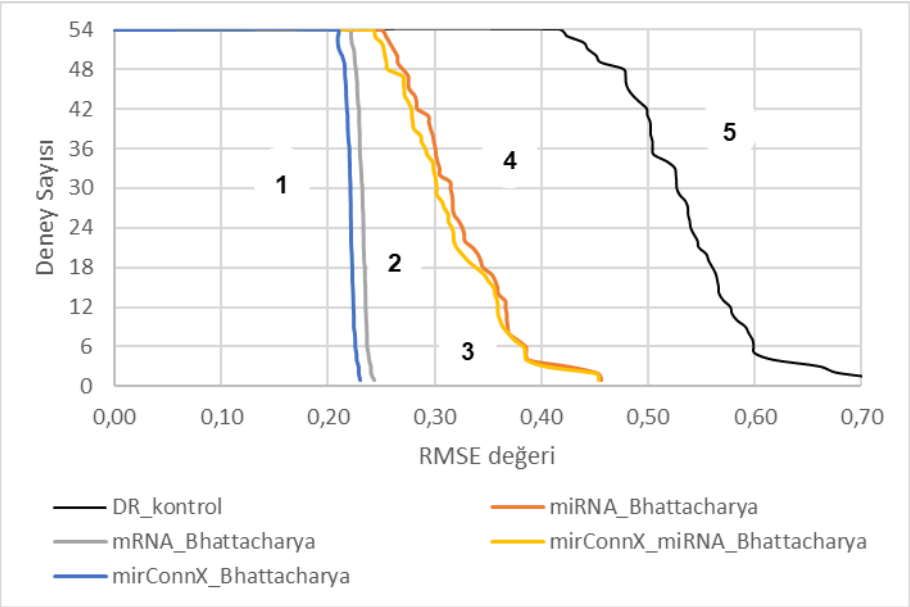


Şekil 5.6 miRNA temelli bütünleştirme işlemi ile doğrusal regresyon için Spearman KK eğrileri



Şekil 5.7 miRNA temelli bütünleştirme işlemi ile doğrusal regresyon için Pearson KK eğrileri

RMSE eğrisi altındaki alan büyük ise kestirim performansı daha kötüdür. Şekil 5.8'deki eğriler incelendiğinde hiçbir bütünleştirme işlemi yapılmadan gerçekleştirilen kestirime (5. eğri) ait RMSE değerlerinin diğerlerine kıyasla daha fazla olduğu görülmektedir.



Şekil 5.8 miRNA temelli bütünleştirme işlemi ile doğrusal regresyon için RMSE eğrileri

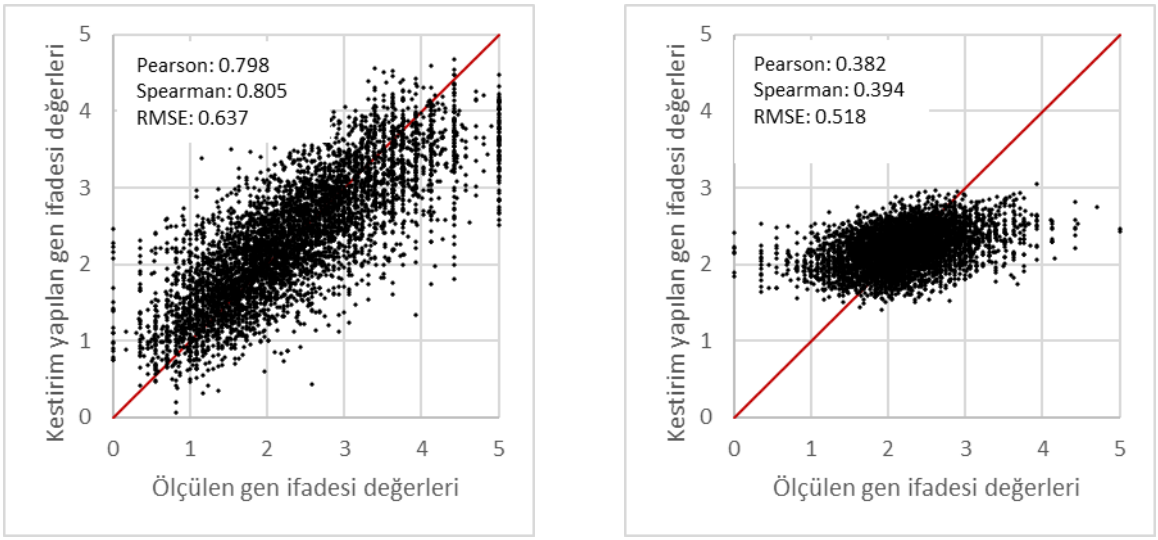
Bhattacharyya uzaklık ölçütü diğerlerine kıyasla daha iyi sonuçlar vermektedir. Bu nedenle RVM ile yapılan uygulamalarda Bhattacharyya ölçütü kullanılmış olup performans grafiklerinde sadece Bhattacharyya ile elde edilen sonuçlar gösterilmiştir. RVM regresyon modelinde RBF-1 çekirdek fonksiyonu ve $\sigma = 17$ parametreleri kullanılarak kestirim yapılmıştır.

*Çizelge 5.3*Çizelge 5.3'te RVM regresyon modeline ait kestirim performanslarının ortalama değerleri yer almaktadır. Veri bütünleştirme işleminin ortalama kestirim performansını artırdığı görülmektedir.

Çizelge 5.3 miRNA temelli bütünleştirme ve RVM regresyon ile elde edilen ortalama kestirim performansları

Veri Kümesi	Bütünleştirilen ifade değerleri	Pearson KK	Spearman KK	RMSE
-	-	0.682	0.681	0.631
mirTarBase	miRNA	**0.803**	**0.786**	**0.345**
mirTarBase	mRNA	0.653	0.655	0.635
mirConnX	miRNA	**0.816**	**0.799**	**0.338**
mirConnX	mRNA	0.669	0.672	0.612

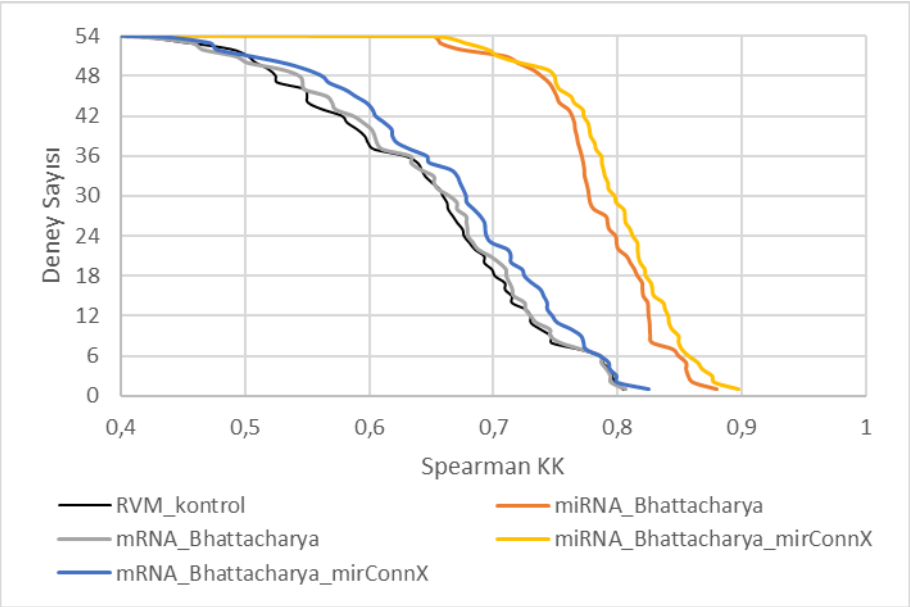
Şekil 5.9'da veri bütünleştirme işlemi olmadan uygulanan RVM regresyon modeli ile elde edilen en iyi ve en kötü kestirimlere ait saçılım grafikleri gösterilmektedir. En iyi kestirim performansı için Pearson KK: 0.798, Spearman KK: 0.805 ve RMSE: 0.637'dir. En kötü kestirim performansı için Pearson KK 0.382, Spearman KK 0.394 ve RMSE 0.518'dir.



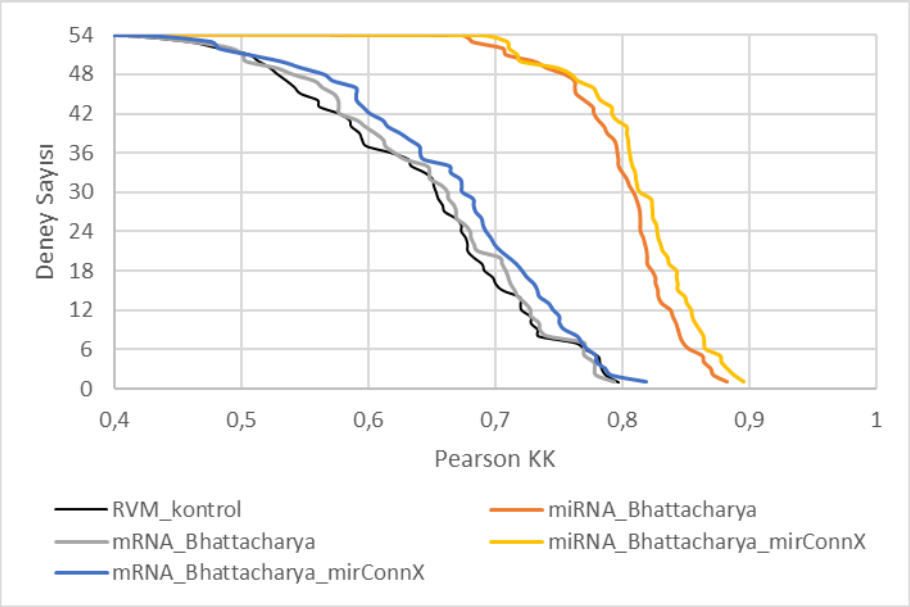
a. b.

Şekil 5.9 Veri bütünleştirme işlemi olman RVM regresyon kestirim sonuçları saçılım grafiği a. En iyi kestirim b. En kötü kestirim

Şekil 5.10 ve Şekil 5.11'de RVM regresyon modeline ait Spearman KK ve Pearson KK grafikleri yer almaktadır. Bu grafiklerde düşey eksende deney sayısı ve yatay eksende benzerlik katsayısı değerleri yer almaktadır. RVM'in bütünleştirme işlemi olmaksızın kullanıldığı duruma ait Spearman KK ve Pearson KK eğrilerinin altındaki alanların diğerlerinden daha küçük olduğu görülmektedir.

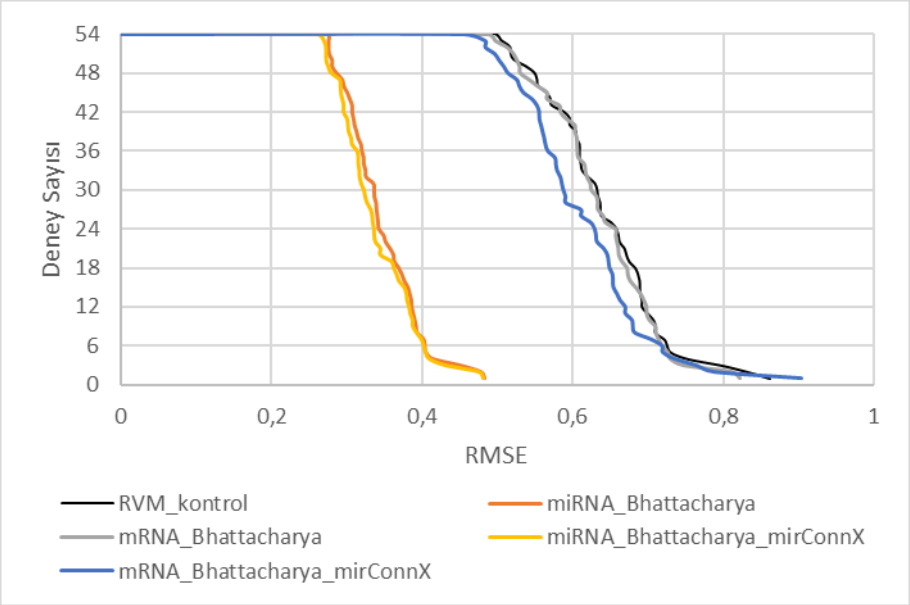


Şekil 5.10 RVM regresyon için Spearman KK eğrileri



Şekil 5.11 RVM regresyon için Pearson KK eğrileri

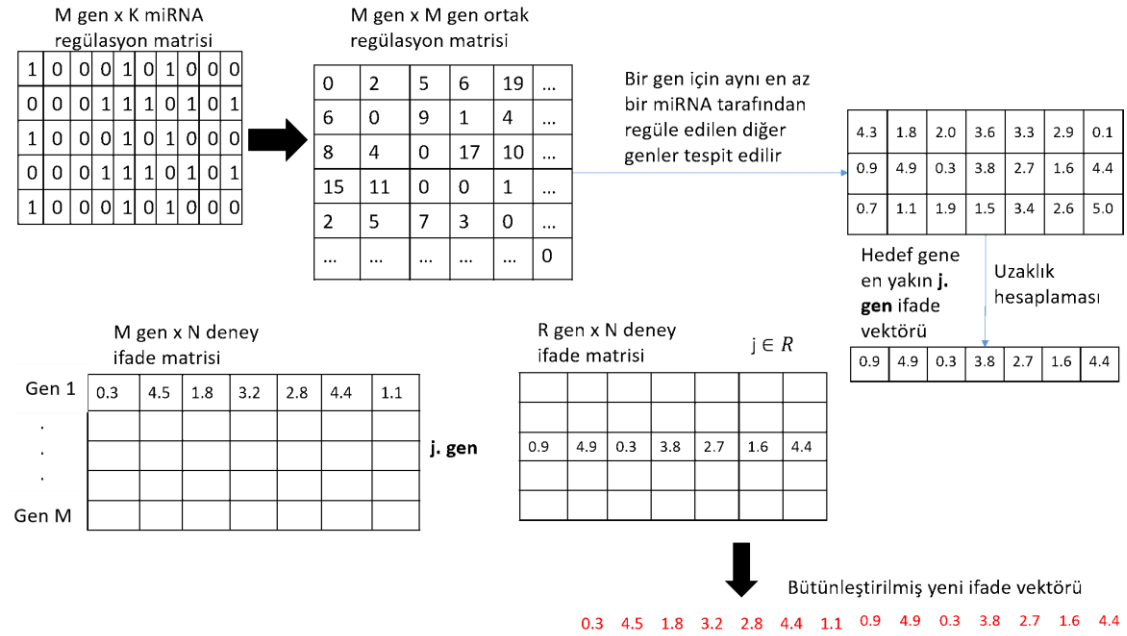
Şekil 5.12'de ise RVM regresyon modeline ait RMSE grafiği yer almaktadır. RVM'in veri bütünleştirme olmadan kullanıldığı durumda elde edilen kestirimlerin daha büyük RMSE değerlerine sahip olduğu görülmektedir.



Şekil 5.12 RVM regresyon için RMSE eğrileri

5.3.1. miRNA regülasyon bilgisi kullanılarak mRNA ifade vektörlerinin bütünleştirilmesi

Boyutları $M\ adet\ gen\ x\ K\ adet\ miRNA$'dan oluşan $M\ x\ K$ regülasyon matrisi düşünelim. Bu matris kullanılarak öncelikle her bir mRNA için aynı miRNA'lar tarafından düzenlenen diğer mRNA'lar tespit edilir. Böylece $M\ x\ M$ boyutunda matris elde edilir. Bu matristeki her bir değer düşey ve yatay eksendeki genleri düzenleyen miRNA sayısını vermektedir. Bir gen kendi kendini düzenleyemeyeceği için matristeki diyagonal değerler sıfırdır. Daha sonra her bir gen için aynı miRNA'lar tarafından (en az 1 adet) düzenlenen R tane mRNA tespit edilir. Bu tespit edilen R adet mRNA içinden ifade tam değeri tahmin edilecek gene en yakın olanı daha önceki bölümlerde ifade edilen uzaklık ölçütleri ile hesaplanır ve en yakın j. gen belirlenir. Bu j. gen'e ait N adet örneğe ait ifade değerlerinden oluşan bir vektör bulunmaktadır. Bu vektör ile kestirim yapılacak genin ifade vektörü birleştirilir. Bu işlem ilk matriste yer alan M adet gen için tekrarlanır. Bu işlemler Şekil 5.13'te gösterilmektedir.



Şekil 5.13 miRNA regülasyon bilgisi kullanılarak mRNA ifade vektörlerinin bütünleştirilmesi

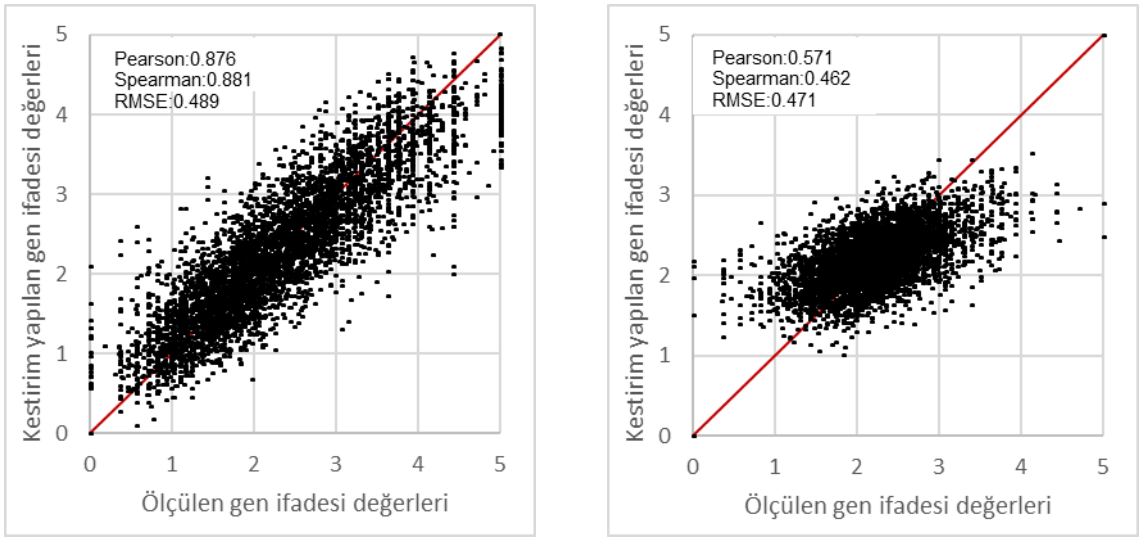
Bu çalışmada kullanılan veri setindeki deney sayısı $54\ (N)$ ve bütünleştirilmiş yeni ifade vektörünün eleman sayısı 108'dir $(N + N)$. Bütünleştirilmiş ifade vektörlerinden oluşan yeni matris $M\ x\ (2 * N)$ boyutundadır.

5.3.1.1. mirTarBase veritabanının kullanılması

Bu bölümde mirTarBase ve mirDB veritabanlarından alınan regülasyon bilgisi birleştirilerek oluşturulan yeni regülasyon matrisi kullanılarak elde edilen kestirim performans sonuçları sunulmuştur. Şekil 5.14, Şekil 5.15 ve Şekil 5.16'da Öklid, Affine dönüşüm ve Bhattacharyya uzaklık ölçütleri kullanılarak yapılan bütünleştirme işlemleri sonrası doğrusal regresyon modeli kullanılarak gerçekleştirilen kestirimlere ait saçılım grafikleri ve performans ölçüm değerleri gösterilmektedir.

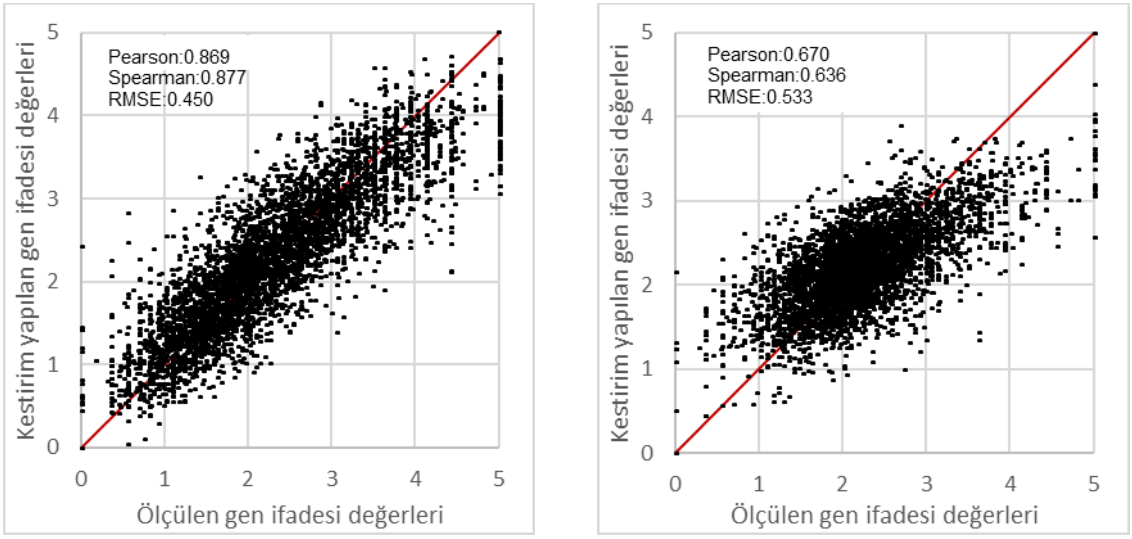
Öklid uzaklık ölçütü kullanılarak elde edilen en iyi kestirim performansı için Pearson KK: 0.876, Spearman KK: 0.881 ve RMSE: 0.489 iken en kötü kestirim performansı için Pearson KK: 0.571, Spearman KK: 0.462 ve RMSE: 0.471'dir. Affine dönüşüm uzaklık ölçütü kullanılarak elde edilen en iyi kestirim performansı için Pearson KK:

0.869, Spearman KK: 0.877 ve RMSE: 0.450 iken en kötü kestirim performansı için Pearson KK: 0.670, Spearman KK: 0.636 ve RMSE: 0.533'tür. Bhattacharyya uzaklık ölçütü kullanılarak elde edilen en iyi kestirim performansı için Pearson KK: 0.980, Spearman KK: 0.979 ve RMSE: 0.241 iken en kötü kestirim performansı için Pearson KK: 0.918, Spearman KK: 0.910 ve RMSE: 0.227'dir.



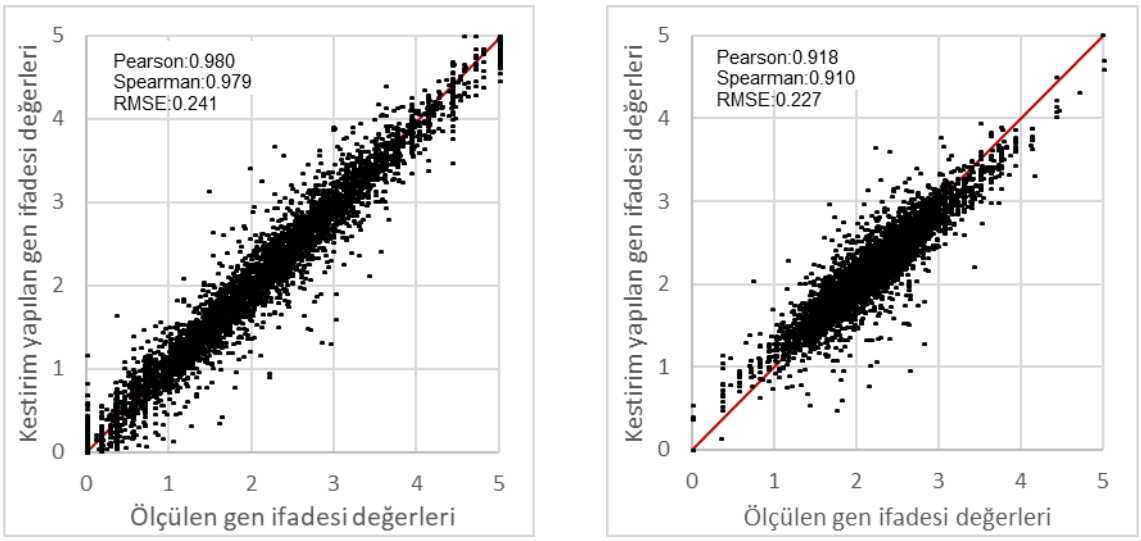
a. b.

Şekil 5.14 Doğrusal regresyon ve Öklid ile bütünleştirme a. En iyi kestirim b. En kötü kestirim



a. b.

Şekil 5.15 Doğrusal regresyon ve Affine dönüşüm ile bütünleştirme a. En iyi kestirim b. En kötü kestirim

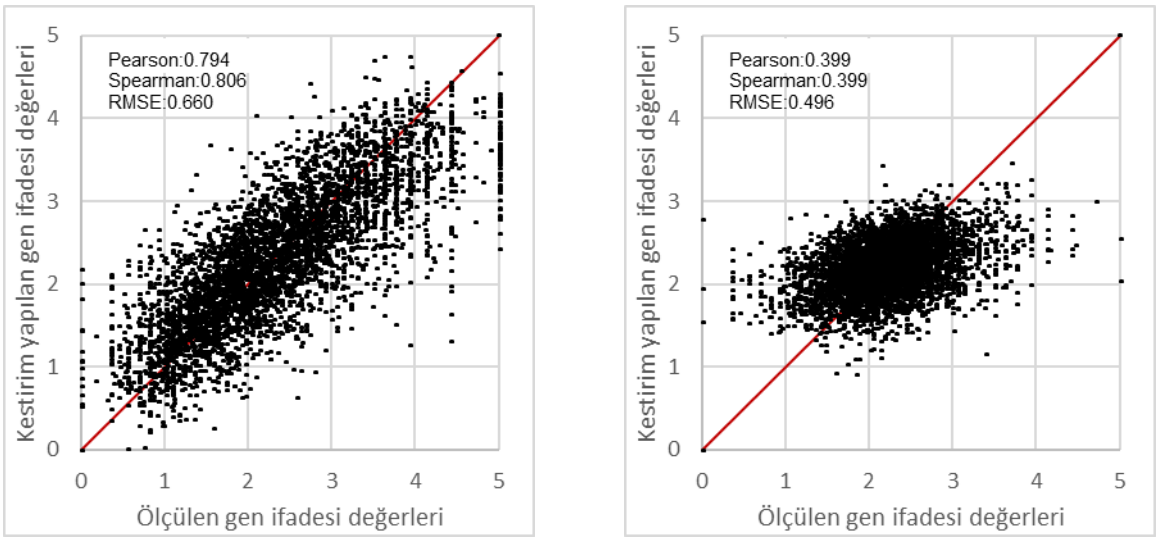


a. b.

Şekil 5.16 Doğrusal regresyon ve Bhattacharyya ile bütünleştirme a. En iyi kestirim b. En kötü kestirim

Şekil 5.17'de RVM regresyon modeli kullanılarak elde edilen en iyi ve en kötü kestirim performanslarına ait saçılım grafikleri gösterilmektedir. Kestirim performansı en iyi olan örneğe ait ölçülen gen ifade miktarları ile kestirim değerlerinin diyagonal eksene daha yakın dağılım gösterdiği görülmektedir (Şekil 5.17.a). Bu örnek için elde edilen Pearson KK 0.794, Spearman KK 0.806 ve RMSE 0.660'tır. Buna karşın en kötü kestirim performansı elde edilen örneğe ait saçılım grafiğinde ise orta kısımda düzensiz bir dağılım oluştuğu görülmektedir (Şekil 5.17.b). Bu örnek için Pearson KK 0.399, Spearman KK 0.399 ve RMSE 0.496'dır.

Daha önceki bölümlerde RMSE ölçütünün iki değerin farkından elde edildiği anlatılmıştı. Ölçülen gen ifade miktarı ile kestirim değeri arasındaki hata farkından elde edilen bu ölçütün küçük olması daha iyi bir kestirim yapıldığı anlamına gelmektedir. Ölçülen gen ifade miktarları ve kestirim değerlerinden elde edilen korelasyon katsayıları ile hesaplanan RMSE değerleri arasındaki ters orantı olduğu daha önce verilen korelasyon katsayısı eğrileri ile gösterilmişti. Buna ek olarak en iyi ve en kötü kestirim yapılan örneklere ait bazı saçılım grafiklerinde (örneğin Şekil 5.16) korelasyon katsayıları ile RMSE arasında doğru orantı olduğu görülmektedir.



a. b.

Şekil 5.17 RVM ve Bhattacharya ile bütünleştirme ve RVM a. En iyi kestirim b. En kötü kestirim

5.3.1.2. mirConnX veritabanının kullanılması

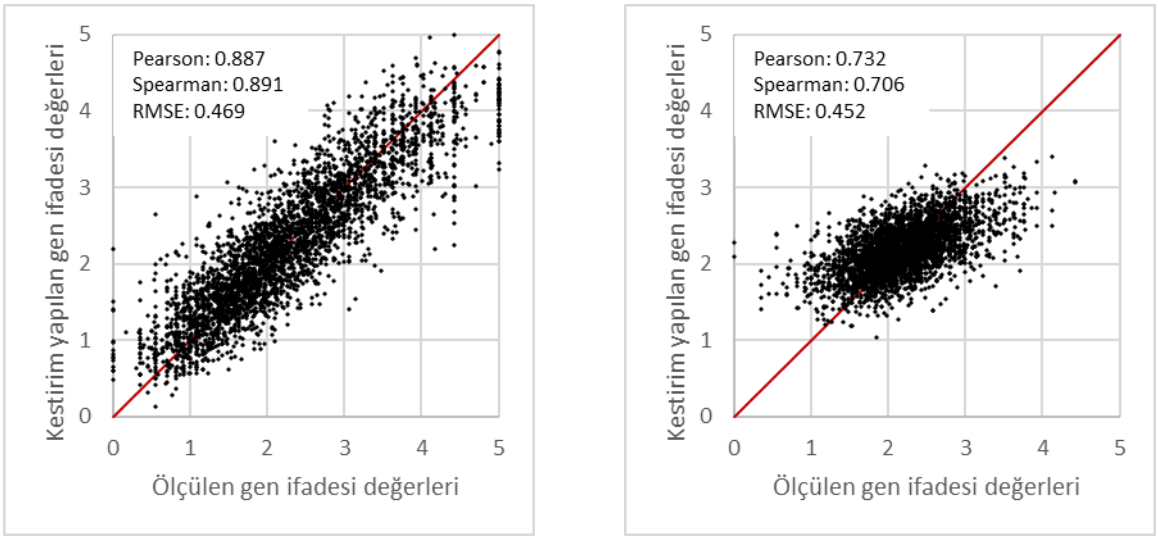
Bu bölümde mirConnX veritabanından alınan mirNA-mRNA regülasyon bilgisi kullanılarak farklı uzaklık ölçütleri ile elde edilen bütünleştirilmiş matrisler için doğrusal ve RVM regresyon modelleri kullanılarak elde edilen kestirim sonuçları sunulmuştur.

Şekil 5.18, Şekil 5.19 ve Şekil 5.20'de Öklid, Affine dönüşüm ve Bhattacharyya uzaklık ölçütleri kullanılarak yapılan bütünleştirme işlemleri sonrası doğrusal regresyon modeli kullanılarak gerçekleştirilen kestirimlere ait saçılım grafikleri ve performans ölçüm değerleri gösterilmektedir.

Öklid uzaklık ölçütü kullanılarak elde edilen en iyi kestirim performansı için Pearson KK: 0.887, Spearman KK: 0.891 ve RMSE: 0.469 iken en kötü kestirim performansı için Pearson KK: 0.732, Spearman KK: 0.706 ve RMSE: 0.452'dir.

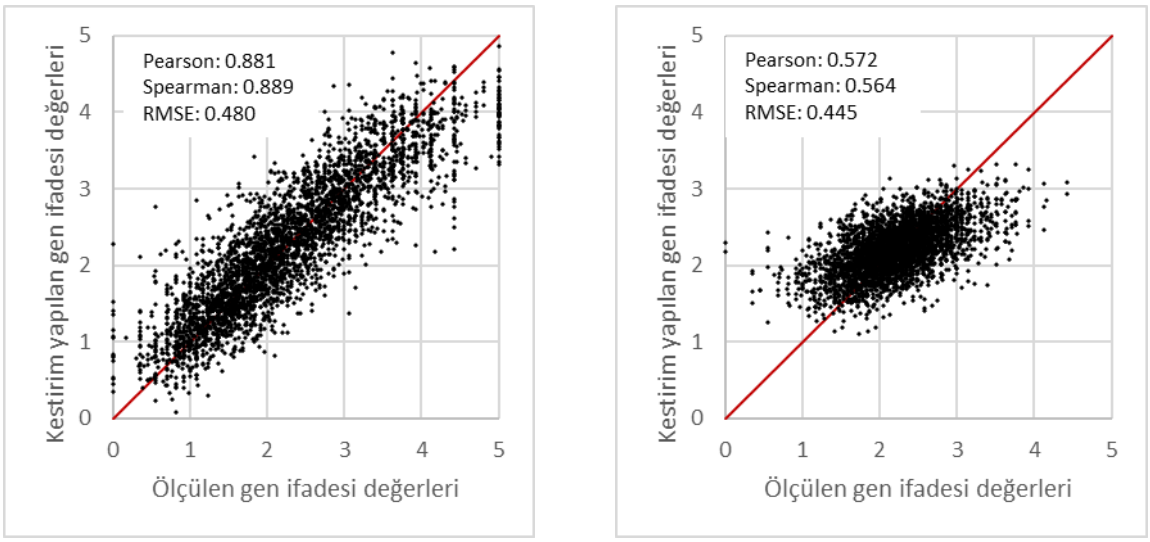
Affine dönüşüm uzaklık ölçütü kullanılarak elde edilen en iyi kestirim performansı için Pearson KK: 0.881, Spearman KK: 0.889 ve RMSE: 0.480 iken en kötü kestirim performansı için Pearson KK: 0.562, Spearman KK: 0.564 ve RMSE: 0.445'tir.

Bhattacharyya uzaklık ölçütü kullanılarak elde edilen en iyi kestirim performansı için Pearson KK: 0.981, Spearman KK: 0.981 ve RMSE: 0.280 iken en kötü kestirim performansı için Pearson KK: 0.920, Spearman KK: 0.920 ve RMSE: 0.217'dir.



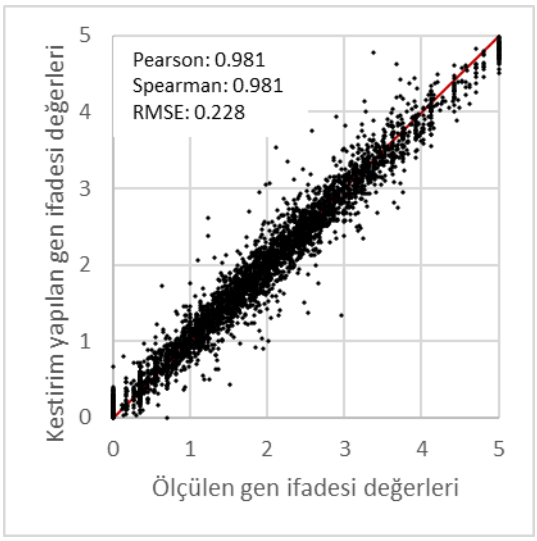
a. b.

Şekil 5.18 Öklid ile bütünleştirme a. En iyi kestirim b. En kötü kestirim

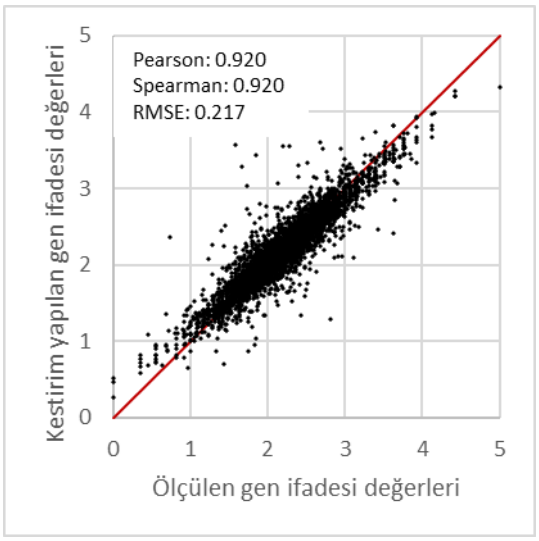


a. b.

Şekil 5.19 Affine dönüşüm ile bütünleştirme a. En iyi kestirim b. En kötü kestirim



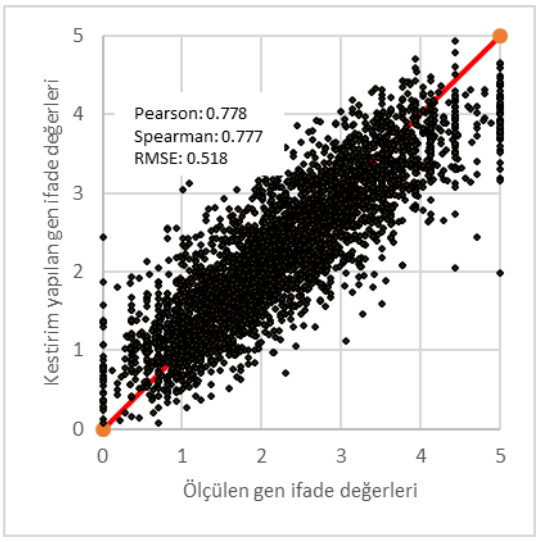
a.



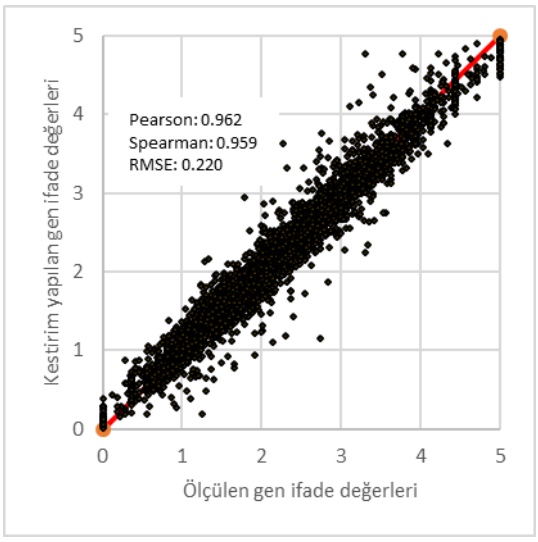
b.

Şekil 5.20 Bhattacharyya ile bütünleştirme a. En iyi kestirim b. En kötü kestirim

Şekil 5.21'de herhangi bir hastaya ait genler için bütünleştirme işlemi yapılmadan ve yapılarak doğrusal regresyon modeli kullanılarak gerçekleştirilen kestirim değerleri ile ölçülen gerçek gen ifade miktarlarının aynı düzlemde gösterilen saçılım grafikleri yer almaktadır. Veri bütünleştirme işlemi yapıldıktan sonra elde edilen kestirim değerlerinin ölçülen gerçek gen ifade değerleri ile aynı diyagonal eksende yakınsadıkları görülmektedir.



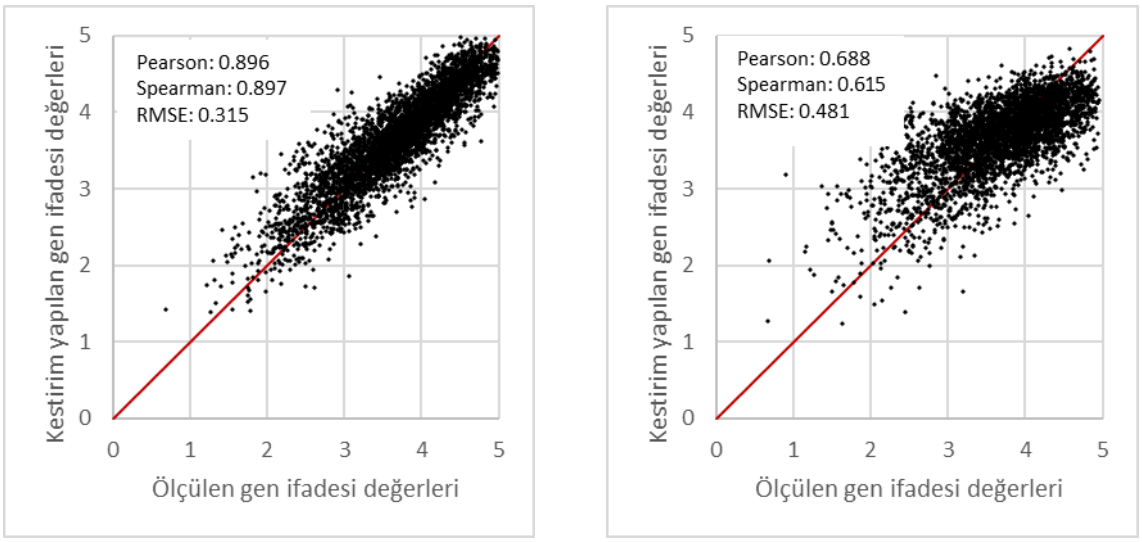
a.



b.

Şekil 5.21 Herhangi bir hasta için bütünleştirme işleminin etkisi a. bütünleştirme öncesi saçılım grafiği b. bütünleştirme sonrası saçılım grafiği

Şekil 5.22'de RVM regresyon modeli kullanılarak elde edilen en iyi ve en kötü kestirim performanslarına ilişkin saçılım grafikleri gösterilmektedir. Kestirim performansı en iyi olan örneğe ait ölçülen gen ifade miktarları ile kestirim değerlerinin diyagonal eksene yakın dağılım gösterdikleri görülmektedir (Şekil 5.22.a). Bu örnek için elde edilen Pearson KK: 0.896, Spearman KK: 0.897 ve RMSE: 0.315'tir. Buna karşın en kötü kestirim performansı elde edilen örneğe ait saçılım grafiğinde ise orta kısımda düzensiz bir dağılım oluştuğu görülmektedir (Şekil 5.22.b). Bu örnek için Pearson KK: 0.688, Spearman KK: 0.615 ve RMSE: 0.481'dir.



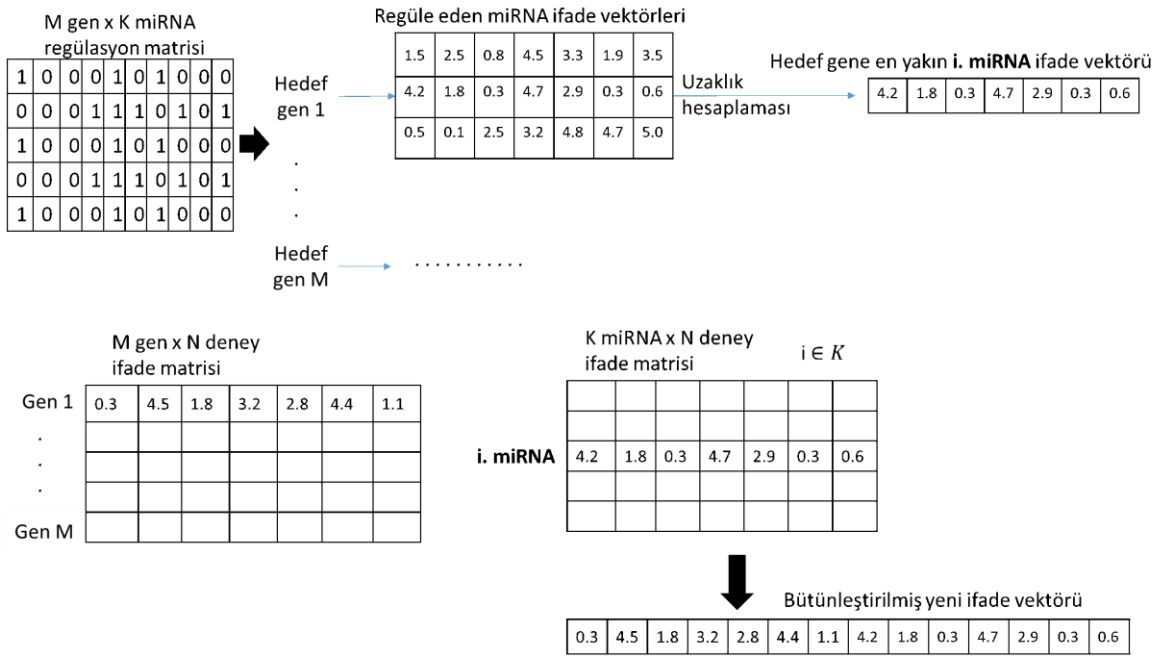
a. b.

Şekil 5.22 Bhattacharyya ile bütünleştirme ve RVM a. En iyi kestirim b. En kötü kestirim

5.3.2. miRNA ve mRNA ifade vektörülerinin bütünleştirilmesi

İfade tam değeri tahmin edilecek bir genin düzenlendiği miRNA'lar mirDB, mirTarBase ve mirConnX veri tabanlarından elde edilmiş olup farklı uzaklık ölçütleri kullanılarak en yakın miRNA'nın ifade vektörü hedef genin ifade vektörü ile birleştirilmiştir. Şekil 5.23'te mRNA ve miRNA bütünleştirme işleminin şematik gösterimi yer almaktadır. İlk olarak, bir geni düzenleyen miRNA'lar sıralanır ve uzaklık hesabına göre en yakın olanı tespit edilir. Tespit edilen en yakın miRNA'ya ait ifade vektörü hedef gen ifadesi vektörünün devamına eklenir. Bu çalışmada kullanılan deney sayısı $54\,(N)$ ve bütünleştirilmiş yeni ifade vektörünün eleman

sayısı $108'dir$ $(N + N)$. Elde edilen bu bütünleştirilmiş ifade vektörleri regresyon modeline uygulanmıştır.

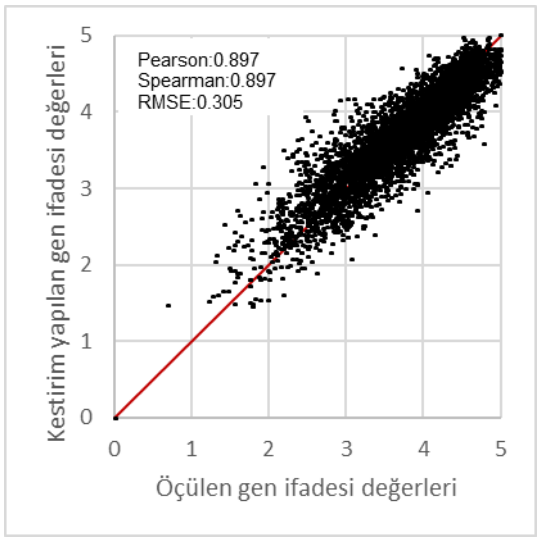


Şekil 5.23 mRNA ifade vektörü ile miRNA ifade vektörünün bütünleştirilmesi

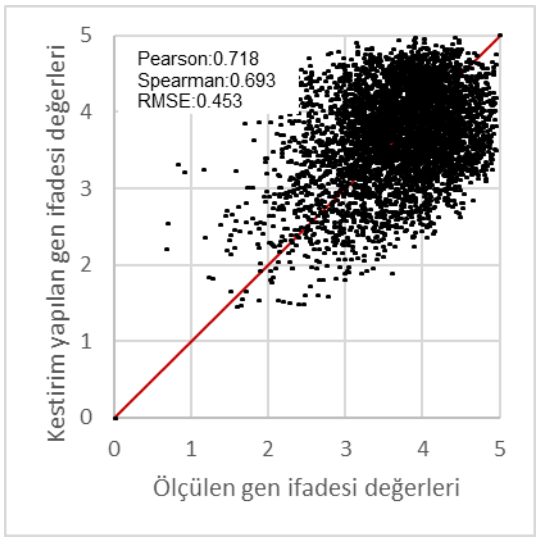
5.3.2.1. mirTarBase veritabanının kullanılması

mirTarBase veritabanından elde edilen regülasyon bilgisine göre Şekil 5.13'teki adımlar uygulandığında Öklid, Affine ve Bhattacharyya uzaklık ölçütleri ile bütünleştirme işlemi yapılmayan kestirim performansına kıyasla daha iyi sonuçlar elde edilmiştir. Şekil 5.24, Şekil 5.25 ve Şekil 5.26'da doğrusal regresyon modeli kullanılarak elde edilen en iyi ve en kötü kestirim sonuçlarına ait saçılım grafikleri yer almaktadır.

Grafiklerde yatay eksende ölçülen (gerçek) gen ifade miktarları ve düşey eksende ise kestirim değerleri yer almaktadır. Kestirim performansları iyi olan grafiklerde örneklerin diyagonal eksene daha da yakın olduğu görülmektedir. Her üç uzaklık ölçütü için en iyi kestirim işlemine ilişkin Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.897, 0.897 ve 0.305'tir. Bunlar 54 deneye ait ortalama değerlerdir.

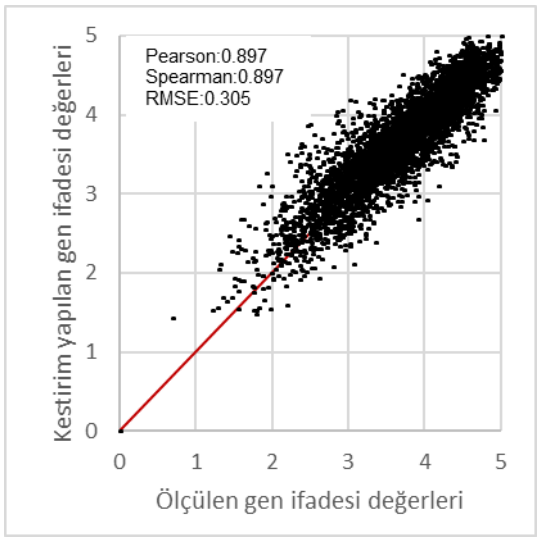


a.

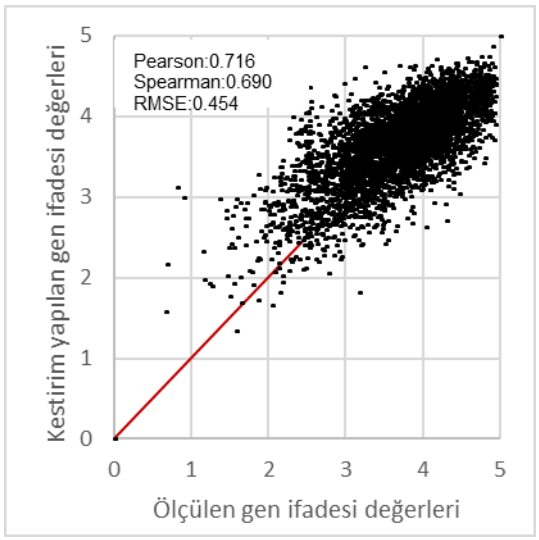


b.

Şekil 5.24 mirTarBase veritabanı, Öklid ile veri bütünleştirme ve doğrusal regresyon a. En iyi kestirim b. En kötü kestirim

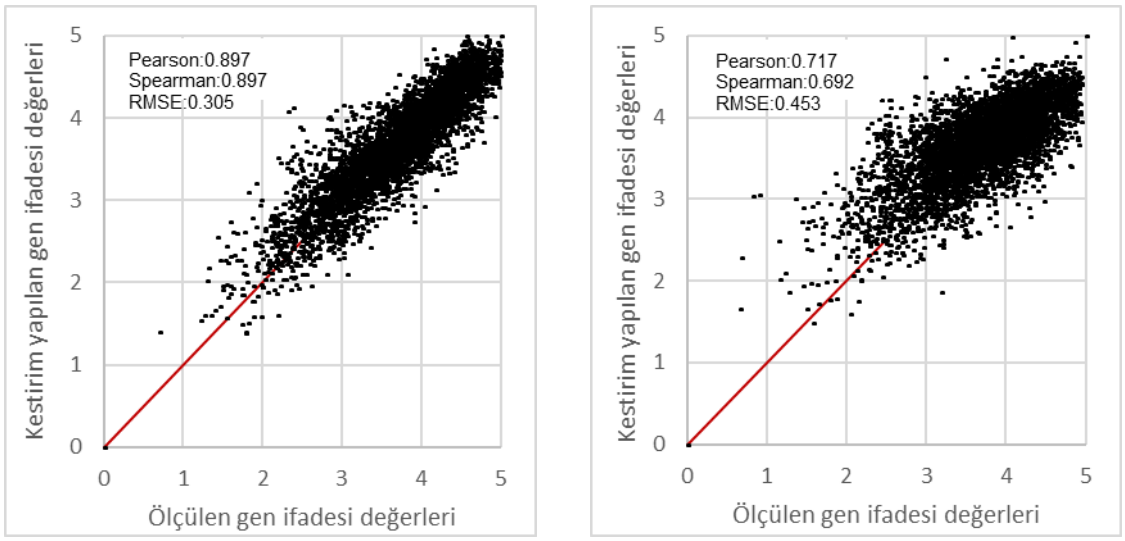


a.



b.

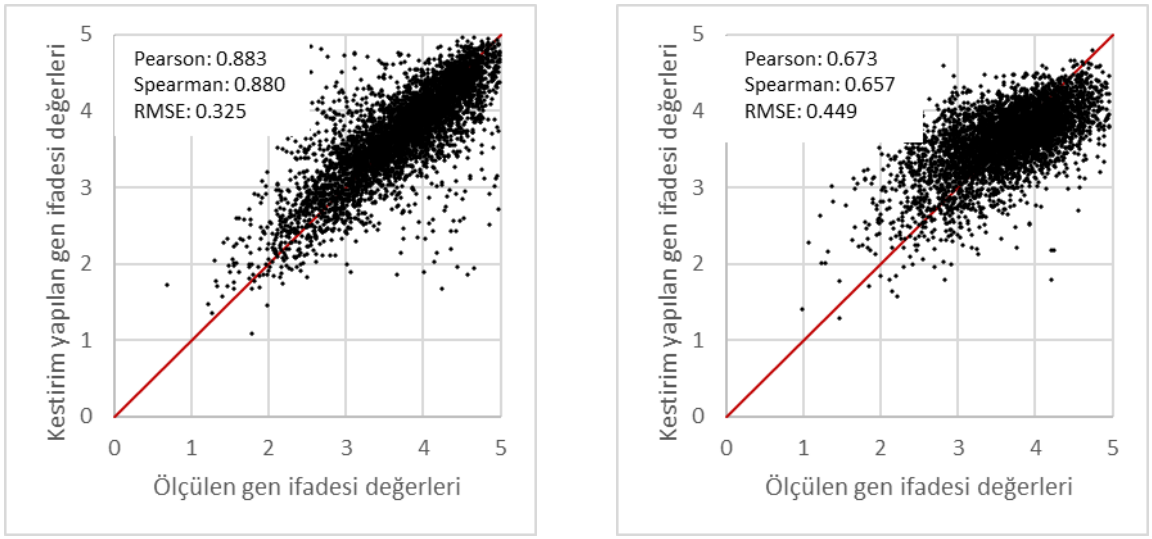
Şekil 5.25 mirTarBase veritabanı, Affine dönüşüm ile veri bütünleştirme ve doğrusal regresyon a. En iyi kestirim b. En kötü kestirim



a. b.

Şekil 5.26 mirTarBase veritabanı, Bhattacharya ile veri bütünleştirme ve doğrusal regresyon a. En iyi kestirim b. En kötü kestirim

Şekil 5.27'de RVM regresyon modeli kullanılarak gerçekleştirilen kestirimlere ait en iyi ve en kötü saçılım grafikleri yer almaktadır. En iyi saçılım grafiğinde ölçülen gen ifadesi değerleri ile kestirim değerlerinin diyagonal eksene daha yakın konumlandıkları görülmektedir (Şekil 5.27.a). Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.883, 0.880 ve 0.325'tir. Diğer saçılım grafiğinde ise değerlerin diyagonal eksenden uzak bir şekilde daha dağınık bir düzende konumlandığı görülmektedir (Şekil 5.27.b).

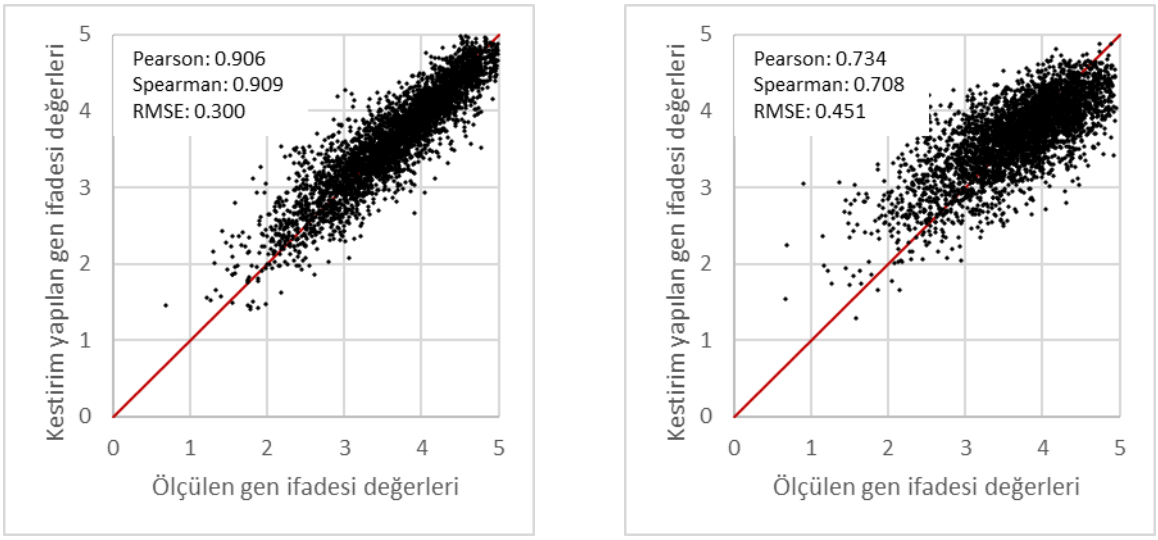


a. b.

Şekil 5.27 mirTarBase veritabanı, Bhattacharya ile veri bütünleştirme ve RVM regresyon a. En iyi kestirim b. En kötü kestirim

5.3.2.2. mirConnX veritabanının kullanılması

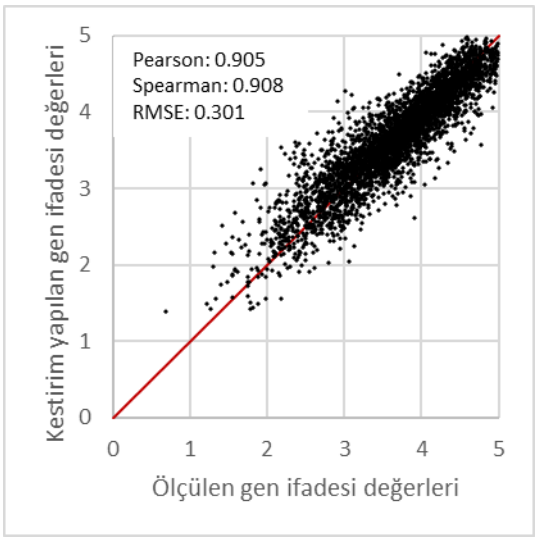
mirConnX veritabanından elde edilen gen-miRNA regülasyon ilişkisi kullanılarak gerçekleştirilen veri bütünleştirme işlemi sonrası elde edilen bütünleştirilmiş matrisler doğrusal ve RVM regresyon modellerinde kullanılmıştır. Şekil 5.28'de Öklid uzaklık ölçütü ile bütünleştirilmiş matris ve doğrusal regresyon kullanılarak elde edilen kestirim performansına ait en iyi ve en kötü saçılım grafikleri yer almaktadır. En iyi kestirim performansı için Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.906, 0.909 ve 0.300'tür. En kötü kestirim performansı için ise Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.734, 0.708 ve 0.451'dir.



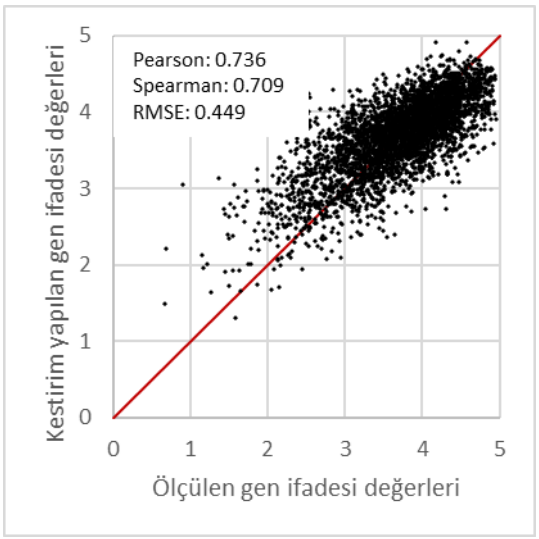
a. b.

Şekil 5.28 mirConnX veritabanı, Öklid ile veri bütünleştirme ve doğrusal regresyon a. En iyi kestirim b. En kötü kestirim

Şekil 5.29'da Affine dönüşümü ile bütünleştirilmiş matris ve doğrusal regresyon kullanılarak elde edilen kestirim performansına ait en iyi ve en kötü saçılım grafikleri yer almaktadır. En iyi kestirim performansı için Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.905, 0.908 ve 0.301'dir. En kötü kestirim performansı için ise Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.736, 0.709 ve 0.449'dur.



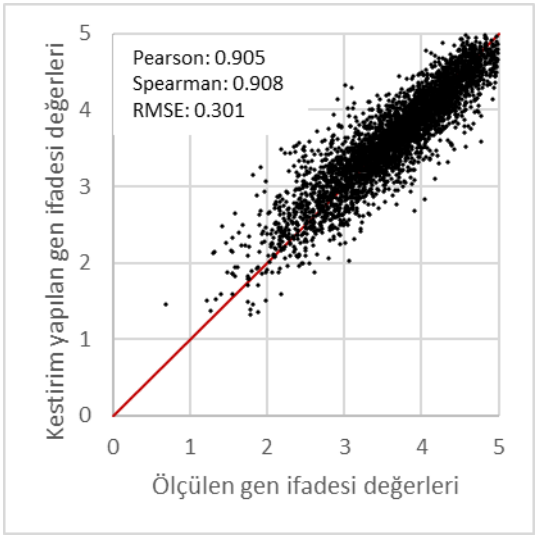
a.



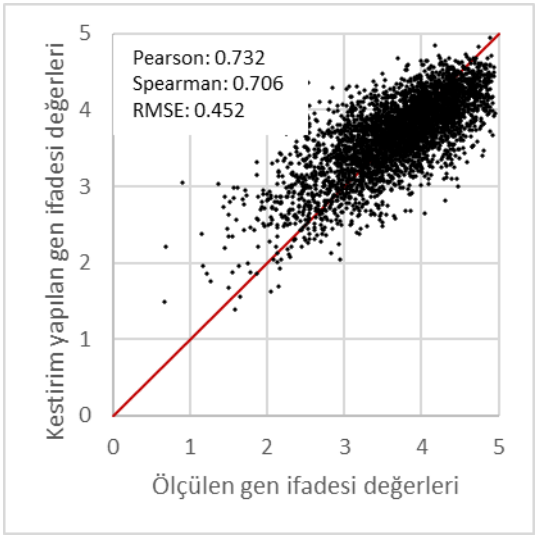
b.

Şekil 5.29 mirConnX veritabanı, Affine dönüşüm ile veri bütünleştirme ve doğrusal regresyon a. En iyi kestirim b. En kötü kestirim

Şekil 5.30'da Bhattacharyya uzaklık ölçütü ile bütünleştirilmiş matris ve doğrusal regresyon kullanılarak elde edilen kestirim performansına ait en iyi ve en kötü saçılım grafikleri yer almaktadır. En iyi kestirim performansı için Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.905, 0.908 ve 0.301'dir. En kötü kestirim performansı için ise Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.732, 0.706 ve 0.452'dir.



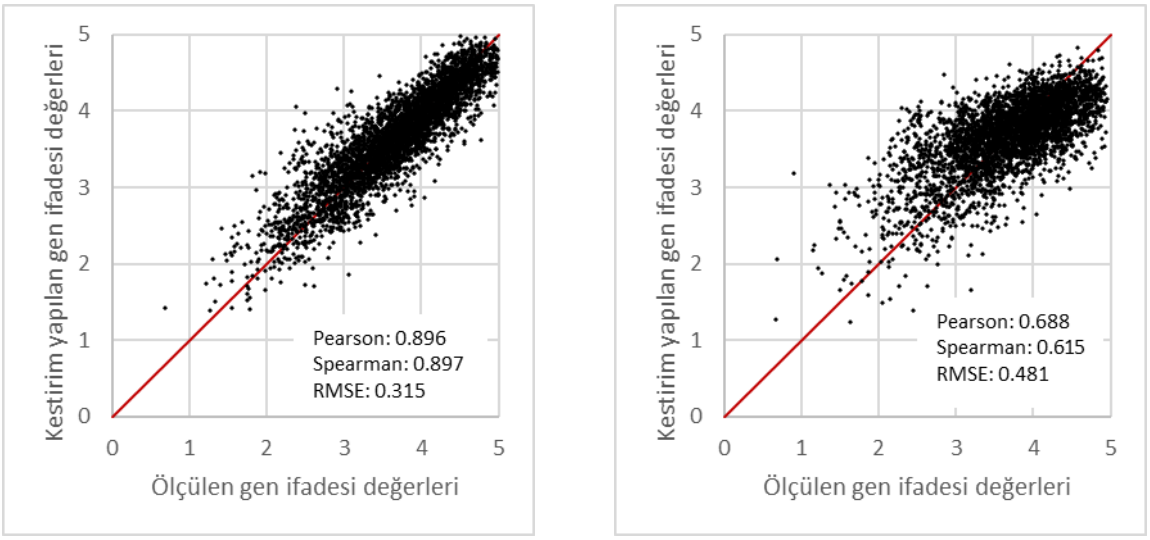
a.



b.

Şekil 5.30 mirConnX veritabanı, Bhattacharyya ile veri bütünleştirme ve doğrusal regresyon a. En iyi kestirim b. En kötü kestirim

Şekil 5.31'de Bhattacharyya uzaklık ölçütü ile bütünleştirilmiş matris ve RVM regresyon kullanılarak elde edilen kestirim performansına ait en iyi ve en kötü saçılım grafikleri yer almaktadır. En iyi kestirim performansı için Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.896, 0.897 ve 0.315'dir. En kötü kestirim performansı için ise Pearson KK, Spearman KK ve RMSE ölçütleri sırasıyla 0.688, 0.615 ve 0.481'dir.

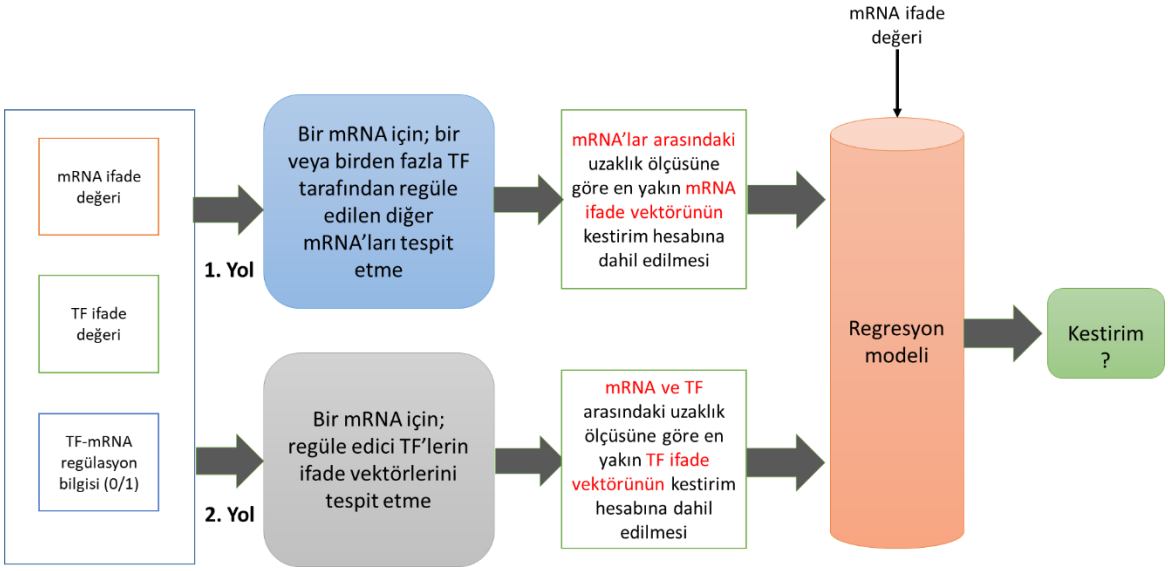


a. b.

Şekil 5.31 mirConnX veritabanı, Bhattacharyya ile veri bütünleştirme ve RVM regresyon a. En iyi kestirim b. En kötü kestirim

5.3.3. Transkripsiyon faktör regülasyon bilgisi kullanılarak veri bütünleştirme

Bu bölüme kadar miRNA-mRNA regülasyon bilgisi kullanılarak gen ifade tam değerinin kestirim performansı artırılmaya çalışılmıştır. Protein sentezi sürecinde miRNA dışında genleri düzenleyen transkripsiyon faktör (TF) adı verilen ve yönetici olarak değerlendirilen genlerin var olduğu daha önceki bölümlerde anlatılmıştır. Bu bölümde TF-mRNA regülasyon bilgisi kullanılarak gen ifade tam değerinin tespitinde kestirim performansı iyileştirilmeye çalışılmıştır. Bölüm 5.3.1 altında anlatılan miRNA regülasyon bilgisinin kullanımında ortaya konulan yaklaşımlar (Şekil 5.32) bu bölümde de uygulanmıştır. TF-mRNA regülasyon bilgisi kullanılarak gerçekleştirilen kestirimlerde performansın artmadığı görülmüştür. 54 farklı göğüs kanseri hastasına ait aynı TF'ler tarafından regüle edilen 1040 mRNA, bunları regüle eden ve TRANSFAC veritabanında kayıtlı olup veri setinde yer alan 896 TF için uygulamalar yapılmıştır.



Şekil 5.32 TF-mRNA regülasyon bilgisi kullanılan veri bütünleştirme genel çerçevesi

5.3.3.1. Doğrusal regresyon modeli kullanılarak elde edilen sonuçlar

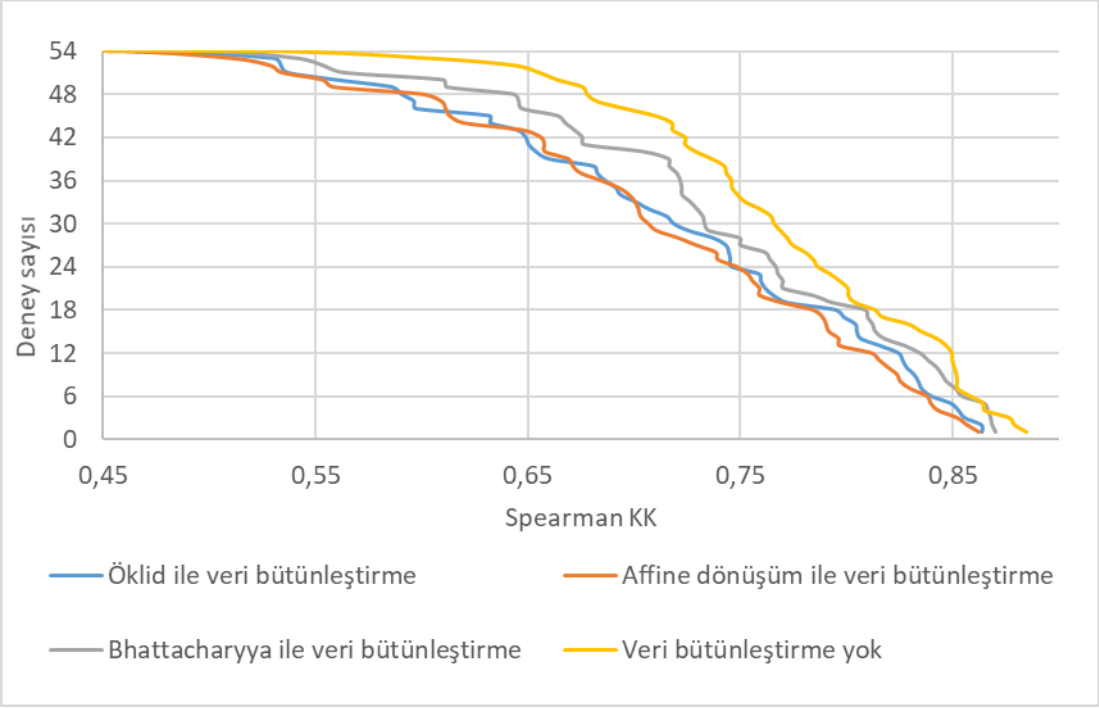
Kestirim performansının ölçümünde Spearman KK, Pearson KK ve RMSE ölçütleri kullanılmıştır. Çizelge 5.4'te hiçbir veri bütünleştirme işlemi olmadan doğrusal regresyon modeli ile elde edilen kestirim performans değerleri ve TF-mRNA regülasyon bilgisi kullanılarak Şekil 5.32'de gösterilen her iki veri bütünleştirme yaklaşımı sonrası elde edilen ortalama kestirim performans değerleri görülmektedir. Çizelge 5.4 incelendiğinde TF regülasyon bilgisi kullanılarak gerçekleştirilen veri bütünleştirme işleminin gen ifade kestirim performansını artırmadığı görülmektedir.

TF-mRNA regülasyon bilgisinin Şekil 5.32'deki birinci yaklaşımda gen ifade matrisi ile bütünleştirilmesi sonrası doğrusal regresyon modeli uygulanarak elde edilen Spearman KK, Pearson KK ve RMSE kestirim performans eğrileri sırasıyla Şekil 5.33, Şekil 5.34 ve Şekil 5.35'te gösterilmektedir. Bu eğrilerin altında kalan alanlar incelendiğinde veri bütünleştirme işlemi olmadan yapılan gen ifade kestirim performansının veri bütünleştirme sonrası elde edilen kestirim performansından daha kötü olduğu görülmektedir. Veri bütünleştirmede kullanılan uzaklık ölçütleri arasında Bhattacharyya ölçütünün ise Öklid ve Affine dönüşüme kıyasla daha iyi sonuç verdiği görülmektedir. Yalnız bu fark miRNA ile veri bütünleştirme

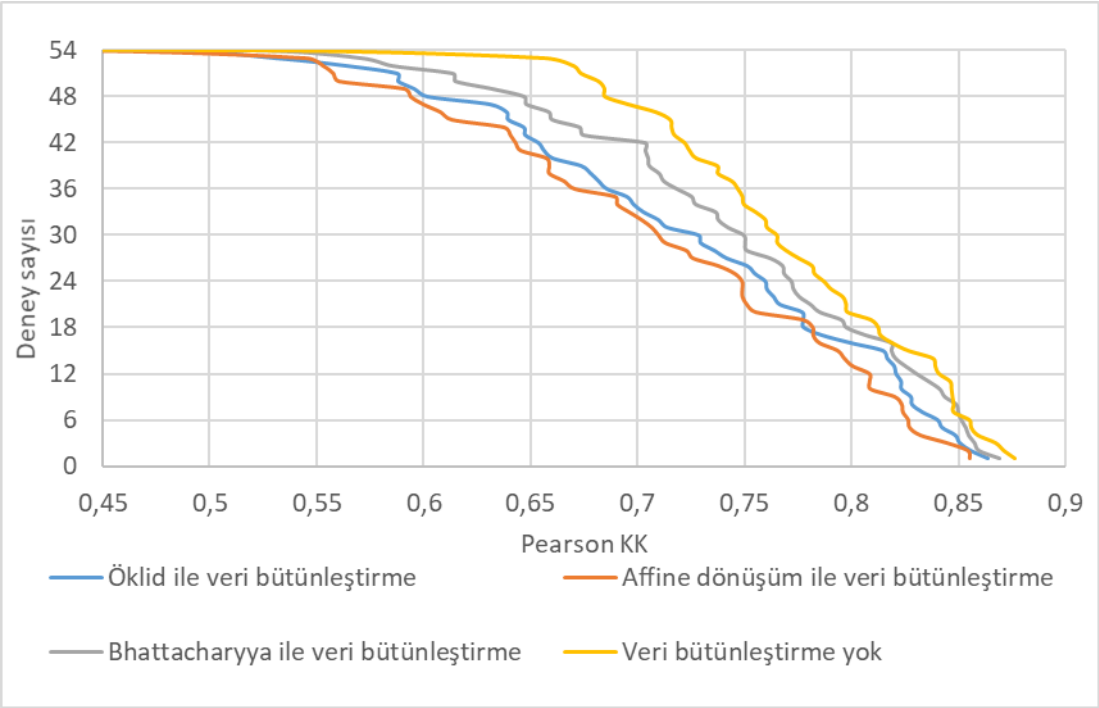
yaklaşımlarındaki kadar belirgin değil ve veri bütünleştirme olmadan yapılan gen ifade tahmini performans sonuçlarına kıyasla daha iyi olduğu söylenemez.

Çizelge 5.4 TF-mRNA regülasyon bilgisi temelli veri bütünleştirme işlemi ile elde edilen kestirim sonuçları

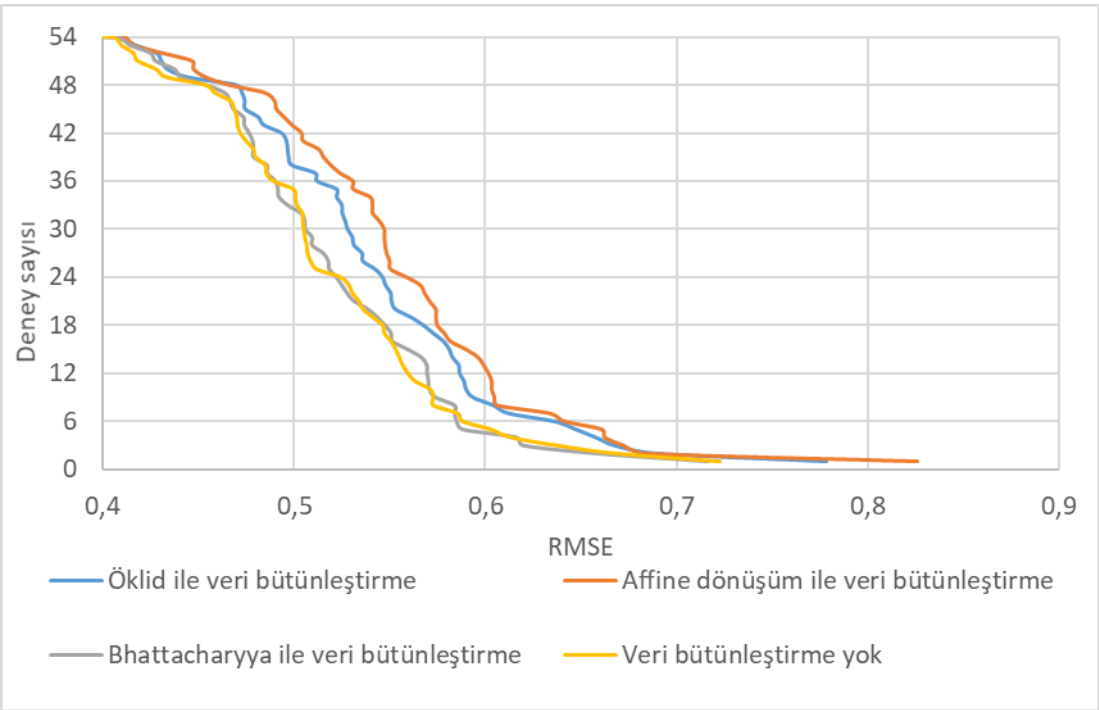
TF–mRNA regülasyon bilgisi kullanılıyor mu?	İzlenen yol	Uzaklık ölçütü	Pearson KK	Spearman KK	RMSE
Hayır	-	-	0,772	0,771	0,518
Evet	1. Yol (mRNA ifade vektörünün bütünleştirilmesi)	Öklid	0,727	0,722	0,540
		Affine Dönüşüm	0,714	0,716	0,553
		Bhattacharyya	0,762	0,760	0,522
	2. Yol (TF ifade vektörünün bütünleştirilmesi)	Öklid	0,752	0,753	0,533
		Affine Dönüşüm	0,754	0,753	0,531
		Bhattacharyya	0,753	0,752	0,532



Şekil 5.33 TF-mRNA veri bütünleştirme 1. yaklaşım kestirim sonuçları (Spearman KK)

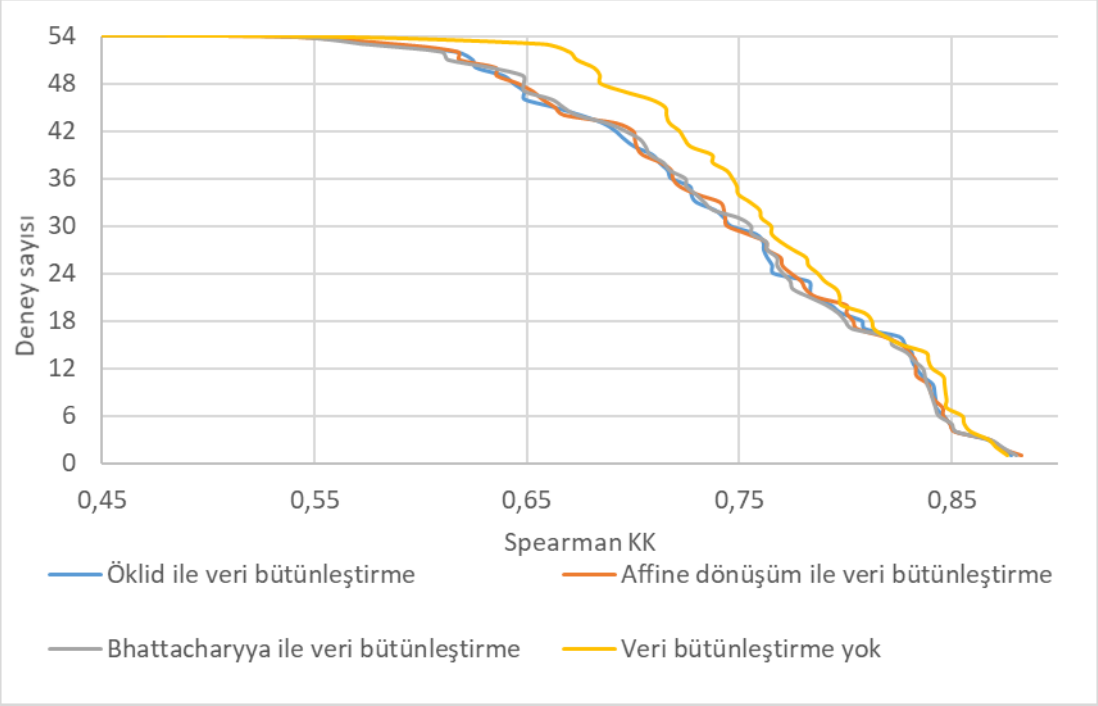


Şekil 5.34 TF-mRNA veri bütünleştirme 1. yaklaşım kestirim sonuçları (Pearson KK)

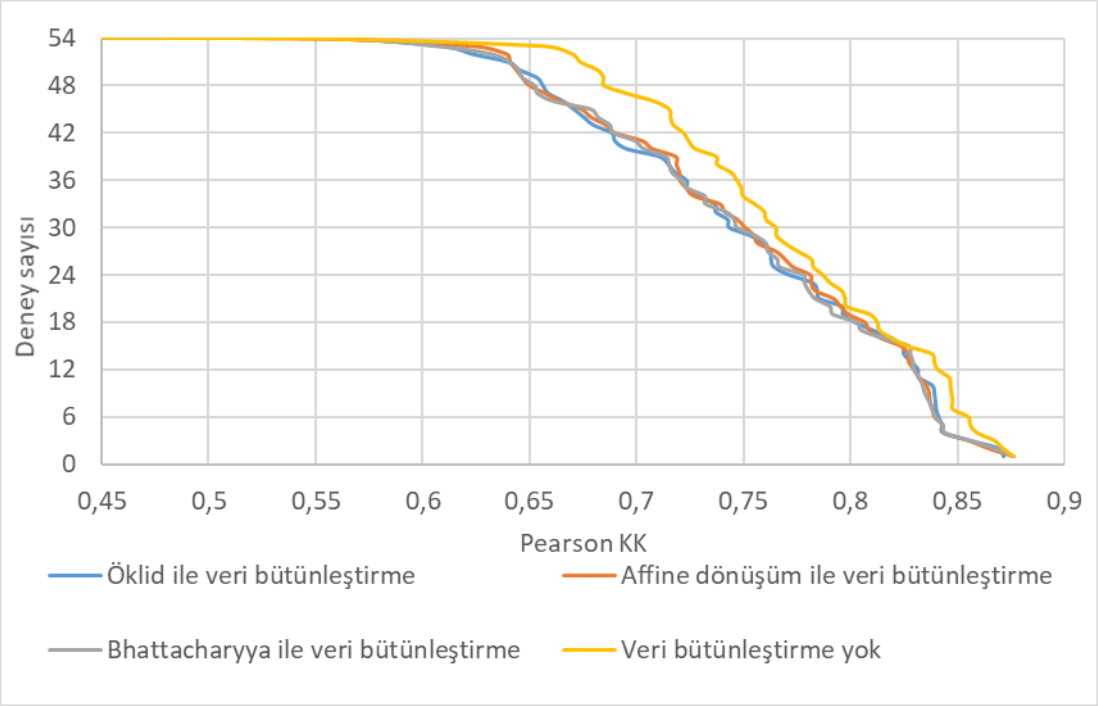


Şekil 5.35 TF-mRNA veri bütünleştirme 1. yaklaşım kestirim sonuçları (RMSE)

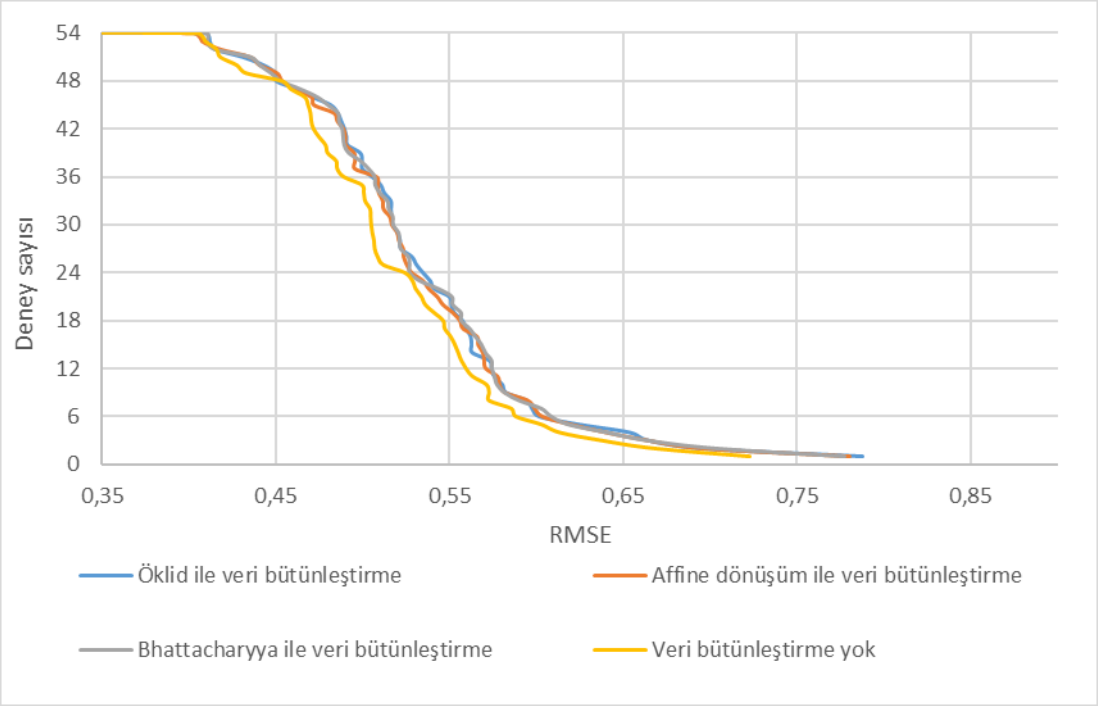
TF-mRNA regülasyon bilgisinin Şekil 5.32'deki ikinci yaklaşım ile elde edilen bütünleştirilmiş gen ifade matrisi kullanılarak yapılan kestirim işlemlerine ilişkin Spearman KK, Pearson KK ve RMSE performans eğrileri sırasıyla Şekil 5.36, Şekil 5.37 ve Şekil 5.38'de gösterilmektedir. Bu eğrilerin altında kalan alanlar incelendiğinde veri bütünleştirme işlemi olmadan yapılan gen ifade kestirim performansının veri bütünleştirme sonrası yapılan kestirim performansından daha kötü olduğu görülmektedir. İkinci yaklaşımda uzaklık ölçütlerinin kestirim performansını artırma açısından birbirinde farkı bulunmamaktadır. Bazı grafiklerde daha iyi bir görünüm açısından yatay eksen değer aralıkları daraltılmıştır.



Şekil 5.36 TF-mRNA veri bütünleştirme 2. yaklaşım kestirim sonuçları (Spearman KK)



Şekil 5.37 TF-mRNA veri bütünleştirme 2. yaklaşım kestirim sonuçları (Pearson KK)



Şekil 5.38 TF-mRNA veri bütünleştirme 2. yaklaşım kestirim sonuçları (RMSE)

5.3.3.2. RVM regresyon modeli kullanılarak elde edilen sonuçlar

Veri bütünleştirme yapılan ve yapılmayan durumlar için RVM regresyon modeli kullanılarak birden fazla TF tarafından düzenlenen 1040 mRNA'ya ait ifade miktarları tahmin edilmiştir. RVM regresyon modelinin uygulanan çekirdek fonksiyonu olarak doğrusal ve önceki bölümlerde en iyi kestirim performansının elde edildiği RBF-1 fonksiyonu ($\sigma = 17$) kullanılmıştır.

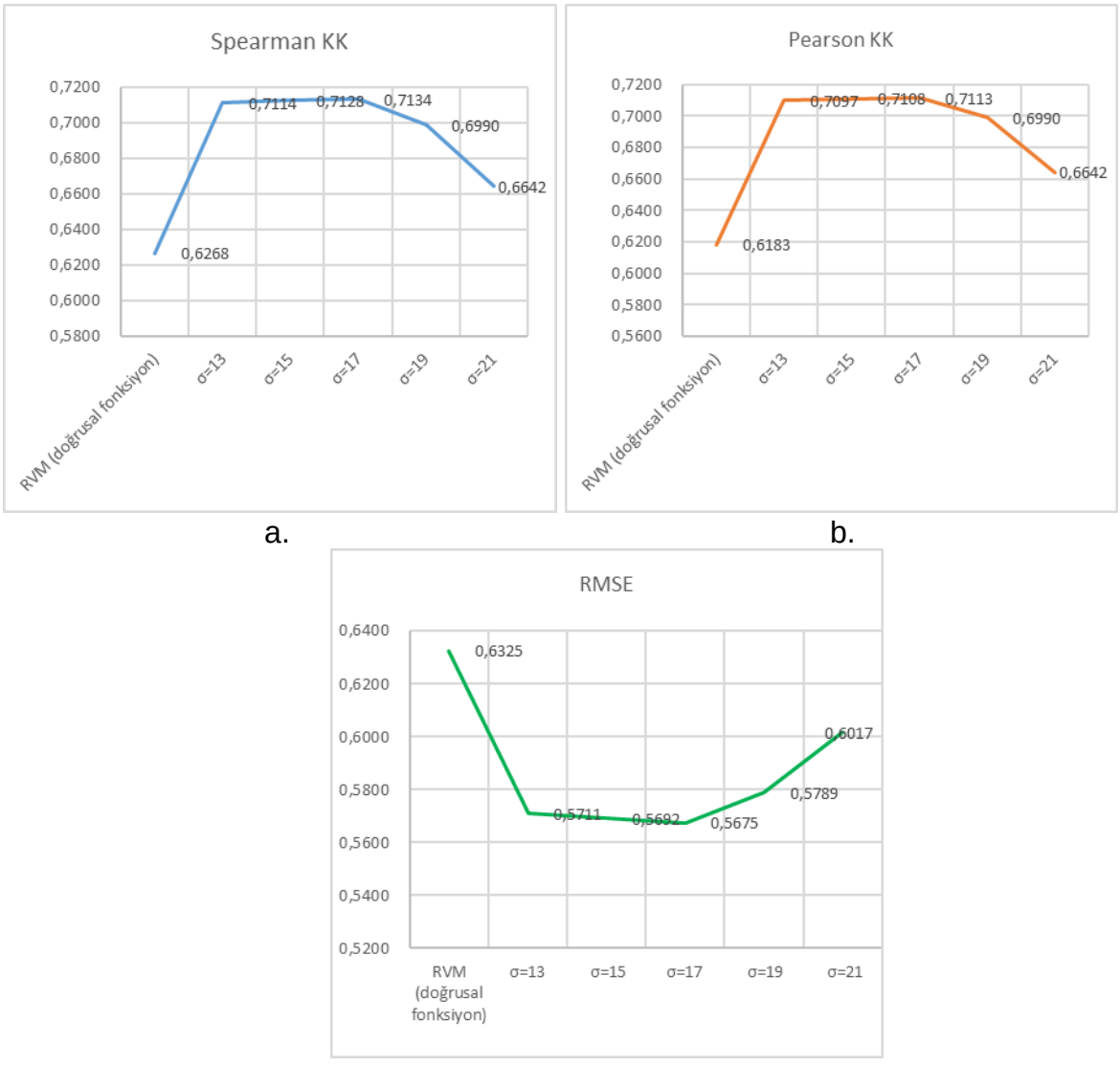
Şekil 5.32'de yer alan TF-gen regülasyon verisi bütünleştirme yaklaşımları ile elde edilen kestirim performans sonuçları Çizelge 5.5'te gösterilmektedir. Önceki bölümlerde en iyi performansa erişildiği durumda kullanılan Bhattacharyya uzaklık ölçütü bu bölümde kullanılmıştır. Buna göre hiçbir bütünleştirme işlemi olmadan doğrusal çekirdek fonksiyonlu RVM regresyon ile yapılan kestirimlere ilişkin 54 hasta için ortalama Pearson KK, Spearman KK ve RMSE sırasıyla 0.688, 0.684 ve 0.598'dir. RVM regresyonun RBF-1 çekirdek fonksiyonu ($\sigma = 17$) için Pearson KK, Spearman KK ve RMSE sırasıyla 0.691, 0.695 ve 0.590'dır.

Çizelge 5.5 RVM regresyon kullanılarak veri bütünleştirme ile elde edilen ortalama kestirim performansları

TF–mRNA regülasyon bilgisi kullanılıyor mu?	**İzlenen yol**	**Regresyon modeli**	**Pearson KK**	**Spearman KK**	**RMSE**
Hayır	-	RVM (doğrusal çekirdek fonksiyonu)	0.688	0.684	0.598
		RVM (σ=17)	0,691	0,695	0,590
Evet	1. Yol (mRNA ifade vektörünün bütünleştirilmesi)	RVM (doğrusal çekirdek fonksiyonu)	0,618	0,625	0,634
		RVM (σ=17)	**0,705**	**0,708**	**0,573**
	2. Yol (TF ifade vektörünün bütünleştirilmesi)	RVM (doğrusal çekirdek fonksiyonu)	0.618	0.627	0.632
		RVM (σ=17)	**0,711**	**0,713**	**0,567**

Şekil 5.32'deki 2. Yaklaşım uygulanarak RVM regresyon modelinin doğrusal ve RBF-1 çekirdek fonksiyonlarının kestirim performansına etkisi incelenmiştir. RBF-1 çekirdek fonksiyonunun optimizasyonu sağlayan σ parametresi için

13,15,17,19 ve 21 değerleri ayrı ayrı test edilmiştir. Şekil 5.39'da doğrusal fonksiyon ve farklı σ değerleri için Spearman KK, Pearson KK ve RMSE değişimleri görülmektedir. Spearman KK ile Pearson KK değerlerinin maksimum ve RMSE değerinin minimum olduğu en iyi kestirim performansına $\sigma = 17$ seçildiğinde erişilmektedir. Bölüm 5.3'te bütünleştirme işlemi yapılmadan en iyi kestirim performansına RVM regresyon modelinin RBF-1 çekirdek fonksiyonu $\sigma = 17$ ile ulaşıldığı bilgisi de yer almaktadır.



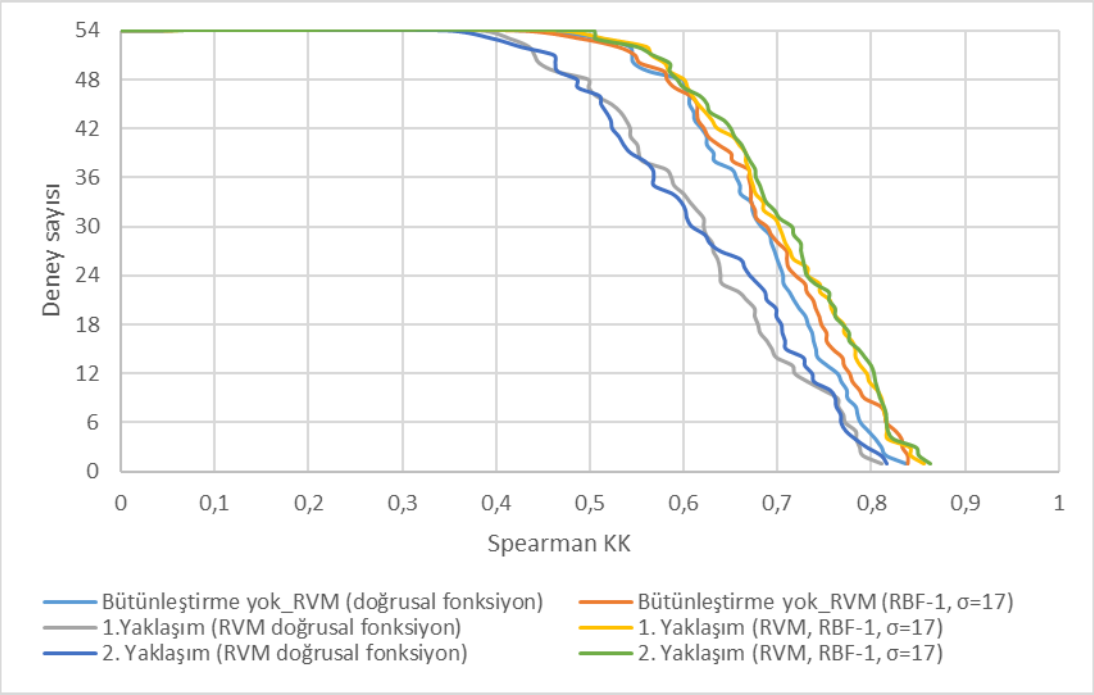
a. b.

c.

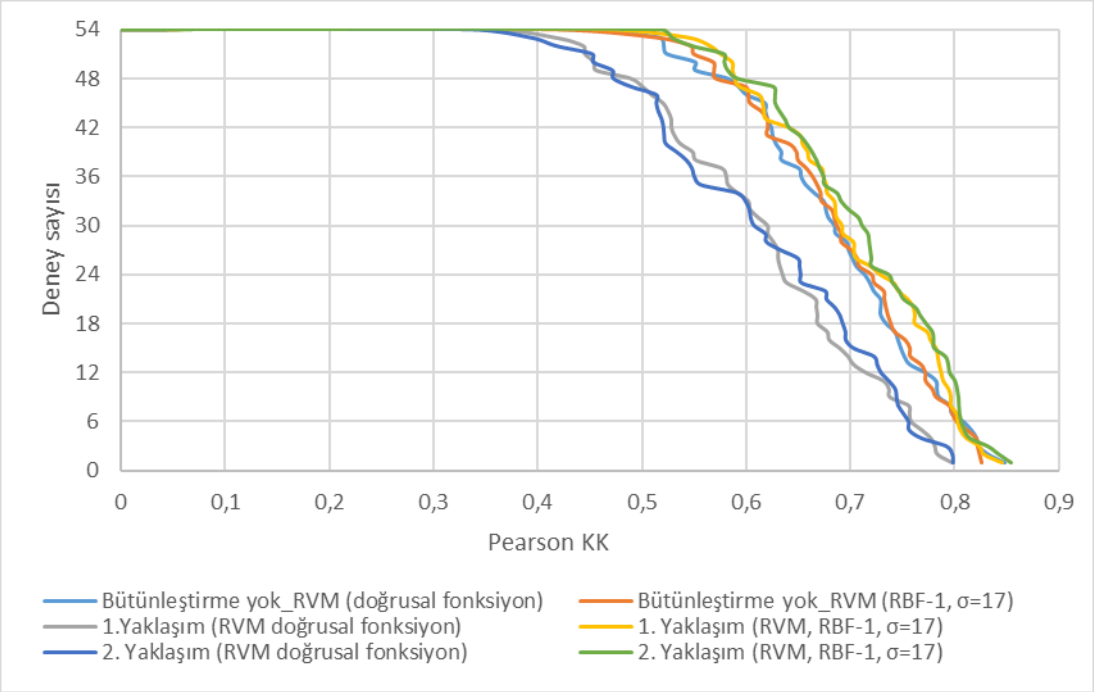
Şekil 5.39 Veri bütünleştirme 2. yaklaşımına σ parametresinin etkisi a. Spearman KK b. Pearson KK c. RMSE

Şekil 5.32'de belirtilen TF regülasyon bilgisi temelli veri bütünleştirme yaklaşımları ile RVM regresyon modeli kullanılarak elde edilen kestirim sonuçlarına ilişkin

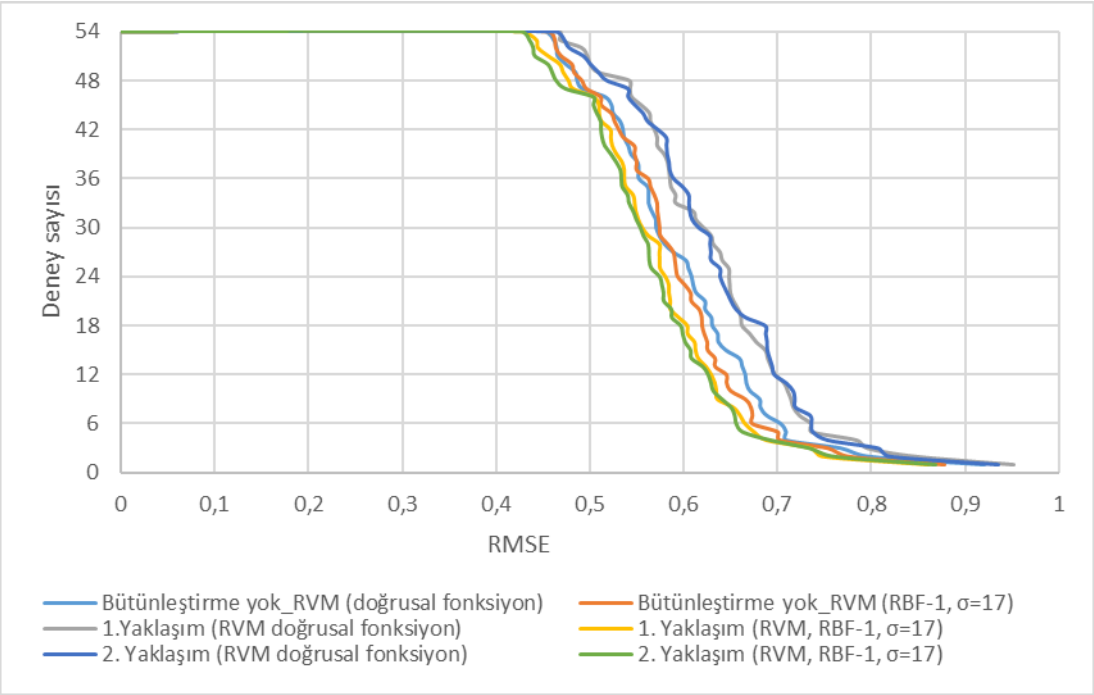
sırasıyla Spearman KK, Pearson KK ve RMSE eğrileri Şekil 5.40, Şekil 5.41 ve Şekil 5.42'de gösterilmektedir.



Şekil 5.40 Veri bütünleştirme RVM regresyon sonuçları(Spearman KK)



Şekil 5.41 Veri bütünleştirme RVM regresyon sonuçları(Pearson KK)



Şekil 5.42 Veri bütünleştirme RVM regresyon sonuçları(RMSE)

RVM regresyonunun RBF-1 çekirdek fonksiyonu ($\sigma = 17$) için elde edilen Pearson KK, Spearman KK ve RMSE değerleri incelendiğinde daha iyi kestirim performansına ulaşıldığı görülmektedir. Şekil 5.42'deki RMSE eğrileri incelendiğinde altında kalan alanın diğerlerine göre daha küçük olmasından dolayı RBF-1 çekirdek fonksiyonunun daha iyi kestirim performansı gösterdiği anlaşılmaktadır.

5.4. Tartışma

Literatürde, gen profili analizleri kullanılarak gen ve miRNA'lar arasındaki ilişkinin tahmin edilmesine yönelik çok sayıda çalışma mevcuttur. Ancak gen ve miRNA arasındaki regülasyon ilişkisi kullanılarak gen ifadesi tahmini ile ilgili çalışmaya rastlanılmamıştır. Burada gen ve miRNA arasındaki ilişkiyi ifade eden ikili yapıdaki değerler ile float yapıda olan ifade değerlerinin aynı modelde kullanılması amaçlanmıştır. Bu nedenle bu veri yapısı farklılığından kurtularak bir bütünleştirme işlemine ihtiyaç duyulmaktadır. Önceki bölümlerde float ve binary yapılarının aynı modelde doğrudan kullanılmasının kestirim performansının düşmesine neden olduğu gösterilmiştir. Buna karşın miRNA regülasyon bilgisi kullanılarak

gerçekleştirilen bütünleştirme işlemlerinin kestirim performansını artırdığı gözlemlenmiştir.

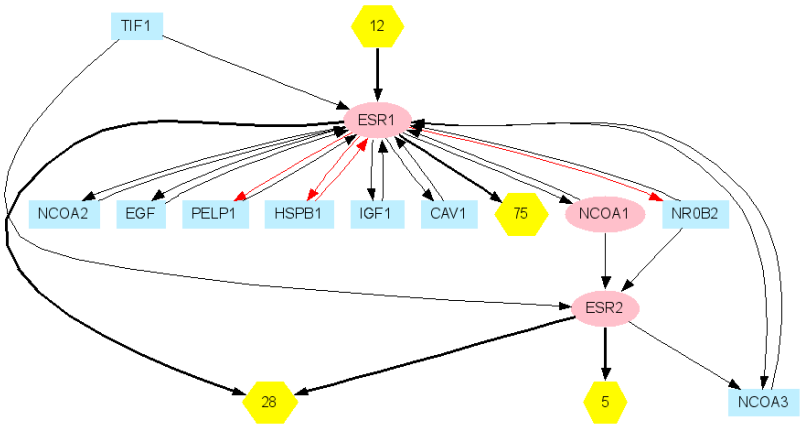
Özellikle miRNA ve mRNA ifade değerlerinden oluşan vektörlerin tek bir vektörde birleştirildiği regresyon modeli ile daha iyi kestirim performansının elde edilmesi; hücrede meydana gelen gen ve miRNA etkileşimlerinin gen ifade profilinde önemli yansımalarının bulunduğunun bir göstergesi olarak değerlendirilebilir. Diğer yandan bütünleştirilecek gen veya miRNA ifade vektörlerinin seçiminde tüm uzaklık ölçütlerinin aynı performansı sağlamadığı gözlemlenmiştir. Öklid gibi farka dayalı ve Affine dönüşümü gibi korelasyon temelli uzaklık ölçütlerine kıyasla Bhattacharyya uzaklık ölçütünün daha iyi sonuçlar vermesi; bu ölçütün sinyal işleme alanında farklı dağılımlara sahip vektörler arasındaki uzaklığın ölçülmesindeki başarımının etkili olduğu düşünülmektedir.

mirTarBase ve mirConnX olmak üzere her iki veritabanından alınan miRNA-mRNA regülasyon bilgilerinin ayrı ayrı gen ifade tahmininde kullanılması ile kestirim performansının arttığı gözlemlenmiştir. Böylece veri bütünleştirme işleminin kestirim performansına olan olumlu etkisinin farklı veri tabanlarından etkilenmediği görülmektedir.

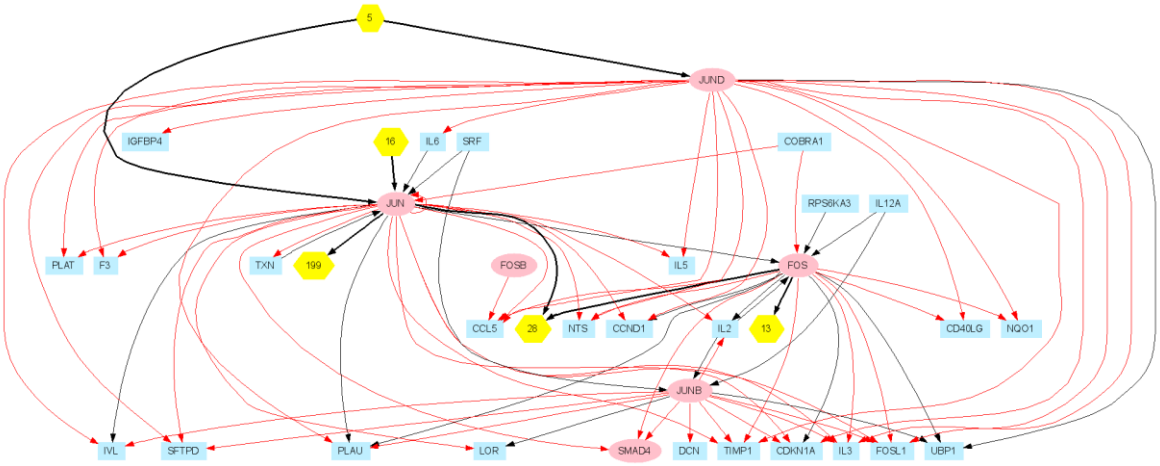
RVM regresyon modelindeki çekirdek fonksiyonlarının ilgili parametreler ile optimize edilebileceği önceki bölümlerde ifade edilmişti. Ancak en iyi kestirim performansına ulaşmak için çekirdek fonksiyonu parametrelerinin çalışılan veriye göre farklılık gösterebilmesi söz konusudur. Bu nedenle optimum parametre değerini tahmin edecek veya bulacak bir tarama modülü geliştirilebilir. Bu modülde sisteme verilen bir veri matrisi için belirli aralıkta tarama yaparak en iyi kestirim sistematik olarak elde edilebilir. RVM regresyon modelinin zaman maliyeti; doğrusal regresyona göre daha fazladır ve çalışılan verinin büyüklüğü ile bu maliyet artmaktadır. Bu nedenle MATLAB ortamında yazılan kodun farklı bir yazılım platformunda optimize edilmesi de düşünülebilir.

Bölüm 5.3'te bazı örnekler için benzerlik katsayıları ile RMSE arasında doğru orantı olabildiği ifade edilmiştir. Bu durum performans değerlendirme işleminde bir karmaşıklık meydana getirse de literatürdeki kestirim çalışmalarında en çok gösterim biçimi olarak görülen saçılım grafiklerinden kestirim yönteminin performansı açık bir şekilde görülebilmektedir. Bu bağlamda; ortaya konulan yaklaşımlar ve kullanılan yöntemler için performans değerlendirme ölçütlerinin yanında saçılım grafiği kullanılmasına dikkat edilmiştir.

TF olarak bilinen genlerin regülasyon ağı miRNA regülasyon ağına göre daha karmaşıktır. Şekil 5.43 ve Şekil 5.44'te ER1 ve AP1 TF ailesine ait regülasyon ağları gösterilmektedir [76]. Örneğin ER1 TF ailesinde, ESR1, NCOA1 ve ESR2 TF moleküleridir ve diğerleri düzenleyici genlerdir. Bir TF başka bir TF'i düzenlediği gibi düzenlediği TF'lerin düzenlediği genleri de düzenleyebilmektedir. Bu iç içe geçmiş karmaşık ağ yapısında tüm yolağı hesaba katmadan TF-gen regülasyon bilgisinin gen ifade tahmininde kullanımı kestirim performansını artırmada yetersiz kaldığı düşünülebilir.



Şekil 5.43 ER1 TF ailesinin genlerle etkileşimi



Şekil 5.44 AP1 TF ailesinin genlerle etkileşimi

TF-mRNA regülasyon bilgisinin gen ifade matrisi ile bütünleştirilmesi işleminin kestirim performansını artırmadığı görülmüştür. Bunun nedeni olarak 54 hastanın miRNA ifade değerlerinin TF ifade değerlerine kıyasla daha fazla değişkenlik göstermesi düşünülebilir. Bu değerlendirmenin TF olarak bilinen genlerin ifade değerlerinin mRNA ifade değerleri ile bütünleştirildiği Şekil 5.32'deki ikinci yaklaşım için daha anlamlı olduğu düşünülmektedir. Diğer yandan birinci yaklaşımda bütünleştirilen ifade vektörleri aynı TF'ler tarafından düzenlenen diğer mRNA ifade vektörleridir. Burada TF ifade miktarlarındaki değişkenliğin kestirim performansını etkilememesi gerektiği düşünülmektedir. Ancak miRNA-mRNA regülasyon bilgisinin TF-mRNA regülasyon bilgisine kıyasla kestirim performansını artırmasının muhtemel nedenlerinden biri olarak miRNA moleküllerinin protein sentezi sürecinde mRNA'lardan istatistiksel olarak farklı bir şekilde ifade vermesi olarak görülebilir. Diğer yandan TF olarak bilinen yönetici genler ise yine ifade tahmini yapılacak diğer genlerle istatistiksel olarak aynı şekilde ifade vermektedir. Zaten TF ile mRNA'ların moleküler yapı ve fonksiyonel açıdan benzer olduğu bilinmektedir. Bunu göstermek amacıyla, 705 miRNA ile 5082 mRNA ifade vektörleri ve 896 TF ile 1040 mRNA ifade vektörleri için Eşleştirilmiş t-testi (Paired t-test) ve Wilcoxon testi uygulanmıştır. Her bir miRNA ifade vektörünün kestirimi yapılacak mRNA ifade vektöründen istatistiksel olarak farklı olduğu ($p<0.05$) ve her bir TF ifade vektörünün kestirimi yapılacak mRNA ifade vektöründen istatistiksel olarak farklı olmadığı ($p>0.05$) görülmüştür.

Bütünleştirme işleminde, mRNA'lardan istatistiksel, hücre içinde moleküler yapısı ve fonksiyonu açısından farklı olan miRNA ifade vektörlerinin kullanılmasının oluşturulacak regresyon modelinin kestirim performansını artırdığı görülmektedir. Diğer yandan TF molekülleri farklı regresyon modelini etkileyebilecek şekilde mRNA'lardan farklı bir bilgi formuna sahip değildir.

6. SONUÇ VE TARTIŞMA

Bu çalışmada test edilen yöntemler ve elde edilen sonuçlar açısından literatüre birçok açıdan katkı sağlandığı düşünülmektedir. Çalışmanın nihai hedefi veri bütünleştirme işlemi ile kestirim performansını artırmak olsa da İki Yönlü İşbirlikçi Filtreleme ve RVM regresyonu gibi yöntemlerin de gen ifade tahmininde kullanılabileceği gösterilmiştir.

Gen ifade tam değerinin tahmin edilmesi problemi sistematik bir şekilde ele alınmış ve sonuçlar karşılaştırmalı olarak sunulmuştur. İlk olarak literatürde sıklıkla çalışılmış olan kayıp veri atama problemi için RVM regresyon modeli önerilmiş ve diğer regresyon yöntemlerine kıyasla daha iyi sonuçların elde edilebileceği görülmüştür. Daha sonra bilinen bir boyutlu veri matrisi yerine İki Yönlü İşbirlikçi Filtreleme ile elde edilen bu matristen elde edilen iki yönlü veri matrisi kullanılarak daha iyi kestirim performansları elde edilmiştir. Ayrıca farklı kanser türlerine ait gen ifade verileri aynı modelde kullanılarak farklı kanser türlerinin genomik düzeyde ilişkisi gösterilmiştir. Bu çalışmada, TF ve miRNA regülasyon bilgisi kullanılarak iki farklı bütünleştirme yaklaşımı önerilmiştir. Bütünleştirme yaklaşımında uzaklık ölçütü olarak Öklid, Affine dönüşümü ve Bhattacharyya kullanılmıştır. Sinyal işleme alanında yaygın olarak kullanılan Bhattacharyya ölçütünün daha önce gen ifadesi tahmininde kullanımına literatürde rastlanmamıştır.

Çalışmanın literatüre katkıları aşağıdaki maddelerde sıralanmaktadır:

- ✓ Kayıp veri kestirimi (missing value imputation) problemi doğrusal, k-NN ve RVM regresyon modelleri ile ele alınmıştır. RVM regresyon modelinin daha iyi kestirim performansına sahip olabileceği gösterilmiştir.
- ✓ İki yönlü işbirlikçi filtreleme yöntemi ilk defa kullanılarak gen ifadesi kestirim performansının artırılabileceği gösterilmiştir.
- ✓ Farklı kanser türlerini içeren deneylerin kullanılmasının gen ifadesi tahmininde performansı artırdığı gösterilmiştir.
- ✓ Mikrodizi verilerinin yanında yeni nesil sekanslama deneylerinden elde edilen verilerde de uygulamalar yapılarak kestirim performansı analiz edilmiştir.

- ✓ Sinyal işleme alanında sıklıkla kullanılan Bhattacharyya uzaklık ölçütü gen ifadesi tahmininde kullanılmış olup kestirim performansını artırdığı gösterilmiştir.
- ✓ miRNA regülasyon bilgisi kullanılarak farklı veri bütünleştirme yaklaşımları ile gen ifadesi tahmini performansının artırılabileceği gösterilmiştir.
- ✓ miRNA ifade vektörü ile gen ifade vektörü aynı modelde bütünleştirilerek kestirim performansının artırılabileceği gösterilmiştir.

Çalışmada; deneylerden elde edilen verilere sadece normalizasyon işlemi uygulanmış olup farklı türden veriler ile çalışılarak veriden bağımsız bir modelin oluşturulması üzerine gayret gösterilmiştir. Literatürde yer alan çok sayıda çalışmada model performansını olumsuz etkileyen "outlier" adı verilen uç verilerin veri matrisinden silinmesi işleminin yapıldığı görülmüştür. Bu çalışmada herhangi bir uç veri silme işlemi yapılmayıp deneylerden elde edilen veriler aynen kullanılmıştır. Sadece işlem zaman maliyetini düşürmek için bazı durumlarda tüm genler yerine rastgele seçilen genler ile çalışma yapılmıştır. Kayıp veri kestirimi ve İki Yönlü İşbirlikçi Filtreleme için uygulanan modelin performansını göstermek için zaman maliyetinden kaçınmak adına bu rastgele seçim yapılırken; veri bütünleştirme çalışmasında tüm gen havuzu üzerinde çalışma yapılmıştır. Çünkü bütünleştirme işleminde miRNA regülasyon bilgisi kullanılarak kestirim yapıldığı için her genin içinde bulunduğu bir yaklaşım ile bütünleştirme işleminin kestirim performansına olumlu etkisinin daha iyi görülebileceği düşünülmektedir.

Hücrede moleküler düzeyde meydana gelen karmaşık olayların fizyolojik, patofizyolojik ve fonksiyonel sonuçları canlının işlevselliğini oluşturmaktadır. Hücrede meydana gelen olayların; moleküllerin birbirine olan etkisinin yanında patofizyolojik etmenler açısından araştırılması oldukça önemlidir. Ülkemizin genom haritasının çıkarılarak; gen, miRNA ve patolojik bulgular arasındaki ilişkinin ortaya konulması ile birlikte genomik düzeyde tedavi yöntemleri geliştirilebilir ve eğilimli olunan hastalıklara yönelik önleyici çalışmalar yapılabilir. Örneğin Spinal Muscular Atrophy (SMA) hastalığı ülkemizde görülebilen ve ölümlerin meydana geldiği nadir görünen bir kas hastalığıdır. SMN1 ve SMN2 genlerindeki mutasyonlara bağlı olarak meydana gelen bu hastalığın tedavi maliyetleri oldukça yüksektir. Daha ileri

moleküler etkileşimler üzerine yapılacak araştırmalar ile bu genlerin çalışmasında etkili diğer gen veya miRNA gibi moleküllerin baskılanması gibi tedaviye yönelik çalışmalar yapılabilir.

Literatürde veri bütünleştirme çalışmaları iki farklı platformdan elde edilen aynı nitelikteki verilerin birleştirilmesi ile veri havuzunun genişletilmesi temelinde şekillenmektedir. Bu sayede daha geniş veri havuzunda çalışmalar yapılarak sonuçların güvenilirliği artırılmaya çalışılmaktadır. Ancak bu çalışmadaki veri bütünleştirme işlemi daha farklı olup iki farklı molekülün birbiri ile olan ilişkisinin gen ifadesi tahmininde kullanılabilmesi için bir dizi işlemleri içermektedir. Aslında hücrede gerçekleşen gen-miRNA regülasyon sürecine ilişkin biyolojik bilgi ile gen ifadesi matrisindeki sayısal bilgi bütünleştirilmektedir.

miRNA ifade değerleri ile aynı miRNA'lar tarafından düzenlenen genlerin ifade değerlerinin aynı modelde bütünleştirilmesi, gen ifadesi tam değerinin kestirim performansını artırdığı gözlemlenmiştir. Bu sonuç; hücrede meydana gelen tüm işlemlerin birbiri ile ilişkisinin detaylı bir şekilde araştırılmasının önemli olduğunu göstermektedir.

İki Yönlü İşbirlikçi Filtreleme yönteminde aynı örneğe ait farklı genlerin ifade değerlerinin model eğitiminde kullanılması, ilk bakışta matematiksel bir manipülasyon olarak görülse de bu genler arasındaki hücre içindeki ilişkisinin kestirim hesabına katılması olarak da yorumlanabilir. Nitekim binlerce genin bir birinden bağımsız olarak çalıştığı söylenemez.

RVM regresyonunda model oluşturma süresi uzun fakat yeni bir girdiye eğitilen modelin uygulanması ve çıktı üretme süresi kısadır. Bu durum uygulamalarda da görülmüştür. Sonuç olarak; bu tez çalışmasında ilk defa ortaya konulan yaklaşımlar ve kullanılan yöntemler hücredeki moleküler etkileşimlerin daha ileri gen analizlerinde ışık tutacak nitelikte olduğu düşünülmektedir.

miRNA regülasyon bilgisinin gen ifadesi tahmininde başarılı sonuçlar vermesi bu sürecin daha detaylı ele alınması durumunda farklı hastalıkların teşhis ve

prognozunda faydalı araçların geliştirilebileceği anlamına gelmektedir. Ayrıca bu tez çalışması ile genler arasındaki ilişkinin gen aktivitesinde önemli yeri olduğu görülmüş olup tüm gen ifade matrisi yerine sadece ilişkili genlere odaklanılabilir. Özellikle prevelansı yüksek olan ve yaşam kalitesini olumsuz etkileyen uyku apnesi gibi hastalıkların moleküler düzeyde teşhis edilmesi ve hızlı tedavi süreçlerinin ortaya konulması için çalışmalar yapılabilir. Ayrıca sinyal işleme alanında kullanlan birçok yöntem biyoinformatik alanında kullanılabilir. Bu çalışma ile edinilen bilgiler ışığında; uyku apnesi teşhisi ve tedavi yöntemlerinin etkinliğinin araştırılması ve kolon kanserinde cerrahi prosedürü değiştiren tümör proksimite tahmini gelecekte yapılması planan çalışmalardan bazılarıdır.

KAYNAKLAR LİSTESİ

[1] FINCH, Megan L., MARQUARDT, Jens U., YEOH, George C. and CALLUS, Bernard A., Regulation of microRNAs and their role in liver development, regeneration and disease, The international journal of biochemistry & cell biology, vol.54, s.288-303, 2014.

[2] LI, Li, XU, Jianzhen, YANG, Deyin, TAN, Xiaorong and WANG, Hongfei, Computational approaches for microRNA studies: a review, Mammalian Genome, vol.21, s.1-12, 2010.

[3] AMBROS, Victor, The functions of animal microRNAs, Nature, vol.431, s.350-355, 2004.

[4] BARTEL, David P., MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, vol.116, s.281-297, 2004.

[5] WIENHOLDS, Erno and PLASTERK, Ronald H.A., MicroRNA function in animal development, FEBS Lett, vol.579, s.5911-5922, 2005.

[6] PILLAI, Ramesh S., BHATTACHARYYA, Suvendra N. and FILIPOWICZ, Witold, Repression of protein synthesis by miRNAs: how many mechanisms?, Trends in cell biology, vol.17, no.3, s.118-126, 2007.

[7] HAYES, Josie, PERUZZİ, Piere P. and LAWLER, Sean, MicroRNAs in cancer: biomarkers, functions and therapy, Trends in molecular medicine, vol.20, no.8, s.460-469, 2014.

[8] CHRUŚCİK, Anna and LAM, Alfred K.Y., Clinical pathological impacts of microRNAs in papillary thyroid carcinoma: A crucial review, Experimental and molecular pathology, vol.99, no.3, s.393-398, 2015.

[9] GOPALAN, Vinod, SMITH, Robert A. and LAM, Alfred K.Y., Downregulation of microRNA-498 in colorectal cancers and its cellular effects, Exp. Cell Res, vol.330, s.423-428, 2015.

[10] AMIN, Moein and LAM, Alfred K.Y., Current perspectives of mi-RNA in oesophageal adenocarcinoma: roles in predicting carcinogenesis, progression and values in clinical management, Exp. Mol. Pathol, vol.98, s.411-418, 2015.

[11] GOPALAN, Vinod et al., Regulation of microRNA-1288 in colorectal cancer: altered expression and its clinicopathological significance, Mol. Carcinog, vol.53, s.E36-E44, 2014.

[12] EBRAHIMI, Faeza, GOPALAN, Vinod., SMITH, Robert A. and LAM, Alfred K.Y., miR-126 in human cancers: clinical roles and current perspectives, Exp. Mol. Pathol, vol.96, s.98-107, 2014.

[13] MAROOF, Hamidreza, SALAJEGHEH, Ali, SMITH, Robert A. and LAM, Alfred K.Y., Role of microRNA-34 family in cancer with particular reference to cancer angiogenesis, Exp. Mol. Pathol, vol.97, s.298-304, 2014.

[14] VOSGHA, Haleh, SALAJEGHEH, Ali, SMITH, Robert A. and LAM, Alfred K.Y., The important roles of miR-205 in normal physiology, cancers and as a potential therapeutic target, Curr. Cancer Drug Targets, vol.14, s.621-637, 2014.

[15] XUAN, Yu, YANG, Huiliang, ZHAO, Linjie, LAU, Wayne B., LAU, Bonnie, REN, Ning, HU, Yuehong, YI, Tao and WEI, Yuquan, MicroRNAs in colorectal cancer: Small molecules with big functions, Cancer letters, vol.360, no.2, s.89-105, 2015.

[16] DEBERARDINIS, Ralph J., LUM, Julian J., HATZIVASSILIOU, Georgia and THOMPSON, Craig B., The biology of cancer: metabolic reprogramming fuels cell growth and proliferation, Cell metabolism, vol.7, no.1, s.11-20, 2008.

[17] VICENTINI, Caterina et al., Clinical application of microRNA testing in neuroendocrine tumors of the gastrointestinal tract, Molecules, vol.19, no.2, s.2458-2468, 2014.

[18] LU, Jun et al., MicroRNA expression profiles classify human cancers. nature, vol.435, no.7043, s.834, 2005.

[19] SUMAZIN, Pavel et al., Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups, Hepatology, vol.65, no.1, s.104-121, 2017.

[20] CALIN, George A. and CROCE, Carlo M., MicroRNA signatures in human cancers, Nature Reviews Cancer, vol.6, no.11, s.857-866, 2006.

[21] HUANG, Grace T., ATHANASSIOU, Charalambos and BENOS, Panayiotis V., mirConnX: condition-specific mRNA-microRNA network integrator, Nucleic acids research, vol.39, no.2, s.416-423, 2011.

[22] NAEEM, Haroon, KÜFFNER, Robert and ZIMMER, Ralf, MIRTFnet: analysis of miRNA regulated transcription factors, PloS one, vol.6, no.8, e22519, 2011.

[23] LE BÉCHEC, Anthony et al., MIR@ NT@ N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model, BMC bioinformatics, vol.12, no.1, s.67, 2011.

[24] JEMAL, Ahmedin, SIEGEL, Rebecca, WARD, Elizabeth, MURRAY, Taylor, XU, Jiaquan and THUN, Michael J., Cancer statistics, 2007. CA: a cancer journal for clinicians, vol.57, no.1, s.43-66, 2007.

[25] KURASHIGE, Yoshihito et al., Profiling of differentially expressed genes in porcine epithelial cells derived from periodontal ligament and gingiva by DNA microarray, archives of oral biology, vol.53, no.5, s.437-442, 2008.

[26] BARRETT, Tanya et al., NCBI GEO: mining tens of millions of expression profilesdatabase and tools update, Nucleic Acids Research, vol. 35, s.760-765, 2007.

[27] BEER, Michael A. and TAVAZOIE, Saeed, Predicting gene expression from sequence, Cell, vol.117, no.2, s.185-198, 2004.

[28] YUAN, Yuan, GUO, Lei, SHEN, Lei and LIU, Jun S., Predicting gene expression from sequence: A reexamination, PLoS Comput Biol, vol.3, no.11, s.243, 2007.

[29] OĞUL, Hasan and TUNCER, Emre M., MicroRNA expression prediction: Regression from regulatory elements, Biocybernetics and Biomedical Engineering, vol.36, no.1, s.89-94, 2016.

[30] KUKURBA, Kimberly R. and MONTGOMERY, Stephen B., RNA Sequencing and Analysis. Cold Spring Harbor Protocols, 2015.

[31] VELCULESCU, Victor E., ZHANG, Lin, VOGELSTEIN, Bert and KINZLER, Kenneth W., Serial analysis of gene expression, Science, vol.270, no.5235, s.484-487, 1995.

[32] SCHENA, Mark, SHALON, Dari, DAVIS, Ronald W. and BROWN, Patrick O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, vol.270, no.5235, s.467-470, 1995.

[33] AFZAL, Muhammad, MANZOOR, Irfan and KUIPERS, Oscar P., A fast and reliable pipeline for bacterial transcriptome analysis case study: serine-dependent gene regulation in Streptococcus pneumoniae, Journal of visualized experiments: JoVE, vol.98, 2015.

[34] TEMPLİN, Markus F., STOLL, Dieter, SCHRENK, Monika, TRAUB, Petra C., VOHRINGER, Christian F. and JOOS, Thomas O., Protein microarray technology. Drug Discovery Today, vol.7, no.15, s.815-822, 2002.

[35] KUMAR, A., GOEL, G., FEHRENBACH, E., PUNIYA, A. K. and SINGH, K., Microarrays: the technology, analysis and application, Engineering in life sciences, vol.5, no.3, s.215-222, 2005.

[36] HAYRAN, Ahmet, Gen ifade veritabanlarında içerik tabanlı arama, Yüksek Lisans Tezi, Başkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Türkiye, 2014.

[37] Gene Expression Omnibus Database, https://www.ncbi.nlm.nih.gov/geo/.

[38] miRTarBase update 2018, A resource for experimentally validated microRNA-target interactions, Nucleic Acids Research, vol.4, no.46, s.296-D302, 2018.

[39] MURPHY, Kevin, Machine learning: a probabilistic perspective, The MIT Press, 2012.

[40] Tipping, Michael E. and Faul, Anita C., Fast marginal likelihood maximisation for sparse Bayesian models, AISTATS, September, 2003.

[41] TROYANSKAYA, Olga, CANTOR, Michael, SHERLOCK, Gavin, Brown, Pat, HASTIE, Trevor, TIBSHIRANI, Robert, BOTSTEIN, David and ALTMAN, Russ B., Missing value estimation methods for DNA microarray, Bioinformatics, vol.17, no.6, s.520-525, 2001.

[42] VAN'T VEER, Laura J. et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature, vol.415, no.6871, s.530-536, 2002

[43] LEE, Ju -S., CHU, In –S., HEO, Jeonghoon, CALVISI, Diego F., SUN, Zongtang, ROSKAMS, Tania, DURNEZ, Anne, DEMETRIS, Anthony J., THORGEIRSSON, Snorri S., Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling, Hepatology, vol.40, no.3, s.667-676, 2004.

[44] GOLUB, T., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, vol.286, no.5439, s.531-537, 1999.

[45] KHAN, Javed et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine, vol.7, no.6, s.673-679, 2001.

[46] LIEW, Alan W.-C., LAW, Ngai –F., YAN, Hong, Missing value imputation for gene expression data: computational techniques to recover missing data from available information, Briefings in Bioinformatics, vol.12, no.5, s.498-513, 2011.

[47] CHEN, Ye, WANG, Aiguo, DING, Huitong, QUE, Xia, LI, Yabo, AN, Ning and JIANG, Lili, A global learning with local preservation method for microarray data imputation, Computers in Biology and Medicine, vol.77, s.76-89, 2016.

[48] TUTZ, Gerhard and RAMZAN, Shahla, Improved methods for the imputation of missing data by nearest neighbor methods, Computational Statistics & Data Analysis, vol.90, s.84-99, 2015.

[49] AYDILEK, Ibrahim B. and ARSLAN, Ahmet, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Information Sciences, vol.233, s.25-35, 2013.

[50] Breast cancer mRNA and miRNA expression data is available from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75285

[51] GRÖNE, Jörn et al., Molecular profiles and clinical outcome of stage UICC II colon cancer patients, Int J Colorectal Dis, vol.26, no.7, s.847-58, 2011.

[52] SATAKE, Hirofumi et al., The ubiquitin-like molecule interferon-stimulated gene 15 is overexpressed in human prostate cancer, Oncol Rep, vol.23, no.1, s.11-16, 2010.

[53] BARRETT, Tanya et al., NCBI GEO: archive for functional genomics data sets update, Nucleic Acids Res, vol. 41, s.991-995, 2013.

[54] HUANG, Dai W., SHERMAN, Brad T., LEMPICKI, Richard A., Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources, Nature Protoc, vol.4, no.1, s.44-57, 2009.

[55] YATES, Andrew et al., Ensembl 2016. Nucleic Acids Res. 2016 44 Database issue: D710-6.PubMed PMID: 26687719; PubMed CentralPMCID: PMC4702834. 2016

[56] TIPPING, Michael E., Sparse Bayesian learning and the relevance vector machine, J Mach Learn Res, vol.2001, no.1, s.211-244, 2001.

[57] EVANS, Geoffrey, HEATH, Anthony and LALLJEE, Mansur, Measuring left-right and libertarian-authoritarian values in the British electorate, British Journal of Sociology, vol.47, no.1, s.93-112, 1996.

[58] Prostate cancer patients are at increased risk of precancerous colon polyps, ScienceDaily. Retrieved March 16, 2018 from www.sciencedaily.com/releases/2010/10/101019121756.htm., 2019.

[59] Ogul, H., Ekmekciler, E., Two-way collaborative filtering on semantically enhanced movie ratings, Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces, Zagreb, Hırvatistan, s.361-366, 2012.

[60] ADOMAVICIUS, Gediminas and TUZHILIN, Alexander, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Trans Knowl Data Eng, vol.17, no.6, s.734-749, 2005.

[61] SU, Xiaoyuan and KHOSHGOFTAAR, Taghi M., A Survey of Collaborative Filtering Techniques, Adv Artif Intell, vol.2009, no.3, s.1–19, 2009.

[62] BOBADILLA, Jesus, ORTEGA, F, HERNANDO, Antonio and GUTIÉRREZ, Abraham, Recommender systems survey, Knowledge-Based Syst, vol.46, s.109–132, 2013.

[63] YU, Bin and ZHANG, Yan, The analysis of colon cancer gene expression profiles and the extraction of informative genes, Journal of Computational and Theoretical Nanoscience, vol.10, no.5, s.1097-1103, 2013.

[64] DONG, Xianjun et al., Modeling gene expression using chromatin features in various cellular contexts, Genome biology, vol.13, no.9, s.53, 2012.

[65] LI, Xiaohong, GILL, Ryan, COOPER, Nigel G.F., YOO, Jae K. and DATTA, Susmita, Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. BMC medical genomics, vol.4, no.1, s.44, 2011.

[66] LE, Hai –S. and BAR-JOSEPH, Ziv, Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation, Bioinformatics, vol.29, no.13, s.89-97, 2013.

[67] GLIGORIJEVIĆ, Vladimir and PRŽULJ, Natasa, Methods for biological data integration: perspectives and challenges, Journal of the Royal Society Interface, vol.12, no.112, 20150571, 2015.

[68] CHOU C.H. et al., miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic acids research, vol.46, no.D1, s.296-302, 2017.

[69] NATHAN, Wong, and XIAOWEI, Wang, miRDB: an online resource for microRNA target prediction and functional annotations, Nucleic Acids Research, vol.43, no.1, s.146-152, 2015.

[70] HUANG, Grace T., ATHANASSIOU, Charalambos, BENOS, Panayiotis V., mirConnX: condition-specific mRNA-microRNA network integrator, Nucleic acids research, vol.39, no.2, s.416-423, 2011.

[71] LIU, Zhi-Ping, WU, Canglin, MIAO, Hongyu and WU, Hulin, RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and Mouse, vol.2015, 2015.

[72] KAILATH, Thomas, The divergence and Bhattacharyya distance measures in signal selection, IEEE transactions on communication technology, vol.15, no.1, s.52-60, 1967.

[73] BHATTACHARYYA, A, On a measure of divergence between two statistical populations defined by their probability distributions, Bull. Calcutta Math. Soc, vol.35, no.1, s.99–109, 1943.

[74] PATRA, Bidyut K., LAUNONEN, Raimo, OLLIKAINEN, Ville and NANDI, Sukumar, A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data, Knowledge-Based Systems, vol.82, s.163-177, 2015.

[75] LUO, Zijun, AZENCOTT, Robert, ZHAO, Yi, Modeling miRNA-mRNA interactions: fitting chemical kinetics equations to microarray data. BMC systems biology, vol.8, no.1, s.19, 2014.

[76] JIANG, C., XUAN, Z., ZHAO, F., ZHANG, M. Q., TRED: a transcriptional regulatory element database, new entries and other development, Nucleic acids research, vol.35, no.1, s.137-140, 2007.

[77] OGUL, H. and AKKAYA, Mahinur S., Data integration in functional analysis of microRNAs. Current Bioinformatics, vol. 6, no. 4, s.462-472, 2011.