AKCİĞER KANSERİ VE KONTROL ÖRNEKLERİNİN GEN İFADESİ ANALİZİ

Özet:

Bu çalışma, akciğer kanserinin erken teşhisinde kullanılabilecek potansiyel genleri belirlemeyi amaçlamaktadır. Çalışmada GSE18842 numaralı veri setini analiz ederek makine öğrenme modellerinden SVM, Random Forest, Karar Ağacı ile en önemli 10 gen belirlenmiştir. Bu genlerin, akciğer kanseri tedavisinde özel bir rol oynayabileceği düşünülmektedir. Ortak genler 213453_x_at, 225602_at, 201291_s_at tespit edilmiştir. Bu genlerin kanser biyolojisi üzerindeki etkilerinin anlaşılmasına yönelik önemli bir adımdır. Random Forest yüksek doğrululuk oranı vermiştir, bu genlerin daha detaylı bir şekilde incelenmesi gerektiğini vurgulayarak akciğer kanseri tedavisinde kişiselleştirilmiş ve etkili stratejilere olanak sağlayabilir.

Anahtar Kelimeler :Akciğer Kanseri, Makine Öğrenmesi, Veri Bilimi

1.Giriş

Akciğer kanseri, akciğer hücrelerinde kontrolsüz büyüme ve bölünme sonucu oluşan bir tür kanserdir. Türkiye Kanser İstatistikleri raporuna göre akciğer kanseri erkeklerde en sık, kadınlarda ise 4. en sık görülen kanser türüdür. Akciğer kanserinin teşhisi genellikle geç evrelerde olmaktadır. En gelişmiş tedavi yöntemlerine rağmen, hastaların %86'sı beş yıl içinde kaybedilmektedir. Akciğer kanserinin toplumda yarattığı hastalık yükü ve önemli düzeyde ekonomik yükü de bulunmaktadır. Bu nedenle hastalıkla mücadelede erken tanı ve tedavi büyük önem taşımaktadır. [1]

Microdizi teknolojisi, gen ifadesi analizi için kullanılan yüksek veri kapasiteli bir biyoteknoloji yöntemidir. Bu teknolojide, mikroskopik boyutta bir çip üzerine yerleştirilmiş binlerce küçük prob, belirli genlerin veya gen bölgelerinin ifade düzeyini ölçmek üzere tasarlanmıştır.

Öztekin ve diğerlerinin (2019) yaptığı bu çalışma, akciğer kanserini non-invazif yöntemlerle teşhis edebilme amacını taşımaktadır. Sekiz hasta ve on altı kontrol olmak üzere toplam 24 kişiden alınan nefes örnekleri, gaz sensör dizileri tarafından ölçülmüş ve sensör verileri analiz edilmiştir. Doğrusal olmayan sınıflandırma yöntemleri ile yüksek duyarlılık elde edilebileceği ve yapay sinir ağı gibi yapılarla bütün özelliklerin birlikte kullanılmasının performansı artırabileceği belirtilmiştir.[1]

Sebik ve Bülbül'ün (2018) yaptığı çalışmada, Dünya Sağlık Örgütü verilerine göre yaygın olan akciğer kanserinin daha hassas bir şekilde teşhis edilmesi amaçlanmıştır. Çalışmada, akciğer kanserinin erken teşhisine katkı sağlamak amacıyla veri madenciliği yöntemleri kullanılmıştır. Sağlık veritabanındaki anonim akciğer kanseri vakalarının verileri WEKA veri madenciliği yazılımında çeşitli algoritmalarla analiz edilmiştir. Elde edilen veri seti üzerinde Naive Bayes algoritması en yüksek başarı oranını elde etmiştir (%91.1). [2]

Bu çalışma, insan akciğer kanseri ve kontrol örneklerinde gen ekspresyon analizi yapmayı amaçlamaktadır. GSE18842 veri seti, insan akciğer kanseri ve kontrol örneklerinde gen ekspresyon analizi yapmak amacıyla kullanılmıştır. Random Forest, SVM ve karar ağacı makine öğrenme algoritmaları kullanılarak yapılan analizler ile en önemli genlerin belirlenerek yüksek doğruluk oranları elde edilmiştir. Random forest algoritmasından elde edilen sonuçlar oldukça yüksek çıkmıştır.

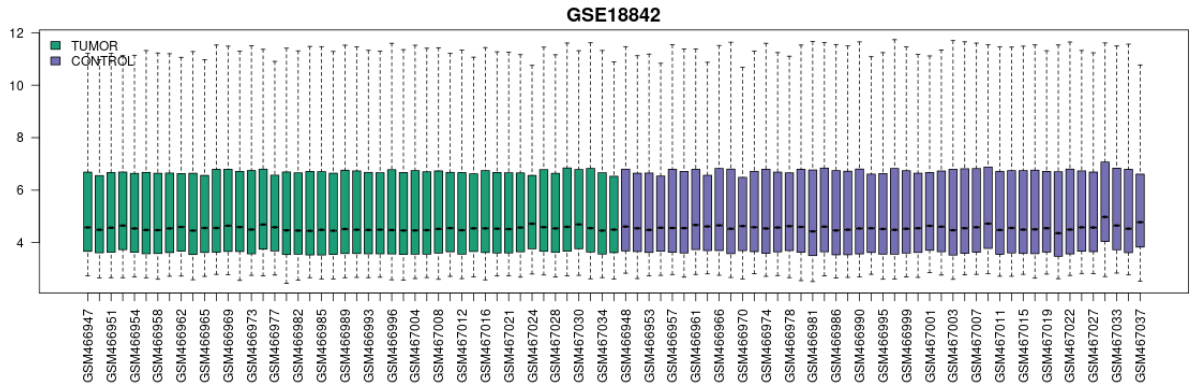
2. Materyaller ve Metodlar

2.1 Veri Seti

Bu veri seti, insan akciğer kanseri ve kontrol örneklerinde gen ekspresyon analizi yapmak üzere oluşturulmuştur. Bu veri seti GEO deposundan alınan GSE18842 numaralı GEOquery paketindeki getGEO fonksiyonu kullanılarak kullanıma hazırlanmıştır. Veri setinde 91 örnek incelenmiştir. Veri seti control ve tumor adında 2 sınıfa ayrılmıştır. 46 adet örnek akciğer kanseri olan bireylerin tümör dokularından elde edilen örneklerdir. 45 adet örnek ise akciğer kanseri bulunmayan bireylerin normal dokularından elde edilen örneklerdir. 54675 farklı gene filtreleme işlemi uygulanarak sayısı 5468'e düşürülmüştür.

2.2 Veri Setinin Kurulumu ve Görselleştirilmesi

Çalışmanın genel tasarımı, toplamda 91 örnek içerirken, bunların 46'sının tümör ve 45'inin kontrol örneği olduğu belirtilmiştir. Bu veri setinin dağılımı kapsamlı bir görselleştirme yöntemi olan kutu grafikleri ile etkili bir şekilde sunulmuştur.



Şekil1: Kutu grafiği veri seti dağılımı

2.3 Gen İfadesi ve Aday Genlerin Tanımlanması

Gen seçimi için, düşük varyans gösteren genleri elemek amacıyla GeneFilter paketi kullanılarak varFilter fonksiyonu kullanılmalıdır. Bu işlem, gen ifadelerindeki varyasyonu dikkate alarak, özellikle örnekler arasında çok az değişiklik gösteren genleri filtreleyerek gerçekleştirilir. Düşük varyans gösteren genler, genellikle sınıflar arasında anlamlı bir fark ortaya koymayan genlerdir. Bu filtreleme işlemi, çalışmanın daha anlamlı ve bilgilendirici genlere odaklanmasına olanak tanır.

Veri setinde 54675 adet özellik bulunmaktadır. Filtreleme işleminde varFilterdaki c değeri 0.9 seçilerek özellik sayısı 5468'e İndirgenmiştir. Daha sonra Random Forest, Karar Ağacı ve SVM kullanılarak özellik seçimi işlemi yapılmıştır.

3. Deneysel Değerlendirmeler

Özellik seçimi için SVM, Random Forest ve Karar Ağacı algoritmaları kullanıldı. Tabloda 10 önemli gen gösterilmektedir.

SVM bir makine öğrenimi algoritmasıdır ve genellikle sınıflandırma ve regresyon problemleri için kullanılır.

Tablo 1: SVM modeli için en önemli 10 aday gen tablosu

	PROB_ID	SYMBOL	GENE
1	206030_at	ASPA	aspartoacylase
2	240189_at	ACOXL	acyl-CoA oxidase-like
3	203362_s_at	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)
4	209408_at	KIF2C	kinesin family member 2C
5	224061_at	INMT	ndolethylamine N-methyltransferase
6	209642_at	BUB1	BUB1 mitotic checkpoint serine/threonine kinase
7	223307_at	CDCA3	cell division cycle associated 3
8	204931_at	TCF21	transcription factor 21
9	213453_x_at	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
10	225602_at	GLIPR2	GLI pathogenesis related 2

Tablo 2. SVM modeli değerlendirme tablosu

Accuracy (Doğruluk)	**Kappa**
0.96	**0.93**

Random Forest, birden çok karar ağacını bir araya getirerek güçlü bir model oluşturan bir makine öğrenimi algoritmasıdır.

Tablo 1: Random Forest modeli için en önemli 10 aday gen tablosu

	PROB_ID	SYMBOL	GENE
1	204962_s_at	SLC35F6///CENPA	solute carrier family 35 member F6///centromere protein A
2	219787_s_at	ECT2	epithelial cell transforming 2
3	209612_s_at	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide
4	212021_s_at	MKI67	marker of proliferation Ki-67
5	201291_s_at	TOP2A	topoisomerase (DNA) II alpha
6	205200_at	EXOSC7///CLEC3B	exosome component 7///C-type lectin domain family 3 member B
7	212581_x_at	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
8	204641_at	NEK2	NIMA related kinase 2
9	204768_s_at	FEN1	flap structure-specific endonuclease 1
10	204825_at	MELK	maternal embryonic leucine zipper kinase

Tablo 2. Random Forest modeli değerlendirme tablosu

Accuracy (Doğruluk)	Kappa
0.98	0.96

Karar ağacı, veri setindeki özelliklerin belirli koşullara göre sınıflandırılmasını sağlayan bir algoritmadır.

Tablo 1: Karar Ağacı modeli için en önemli 10 aday gen tablosu

	PROB_ID	SYMBOL	GENE
1	1552619_a_at	ANLN	anillin actin binding protein
2	1554408_a_at	TK1	thymidine kinase 1
3	1554768_a_at	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)
4	200822_x_at	TPI1	triosephosphate isomerase 1
5	201250_s_at	SLC2A1	solute carrier family 2 member 1
6	201291_s_at	TOP2A	topoisomerase (DNA) II alpha
7	226992_at	NOSTRIN	nitric oxide synthase trafficking
8	232578_at	CLDN18	claudin 18
9	213453_x_at	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
10	225602_at	GLIPR2	GLI pathogenesis related 2

Tablo 2. Karar Ağacı modeli değerlendirme tablosu

Accuracy (Doğruluk)	Kappa
0.94	0.91

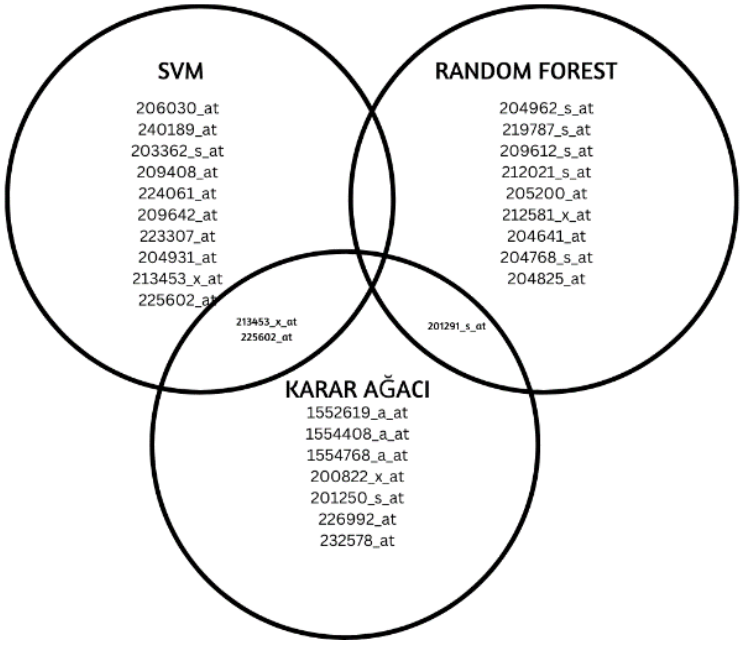
4. Sonuç

Çalışmada GSE18842 numaralı veri setinde yer alan toplam 54675 farklı gene filtreleme işlemi uygulanarak sayısı 5468'e düşürülmüştür. Filtreleme işleminden sonra makine öğrenme modellerinden random forest, svm ve karar ağaçların kullanılarak önemli 10 gen seçimi yapılmıştır. Modellerin her biri için accuracy ve kappa değerleri belirlenerek sonuçlar karşılaştırmalı olarak bir tabloda gösterilmiştir.

Tablo1:Model Karşılaştırma Tablosu

Model Adı	Accuracy (Doğruluk)	Kappa
SVM	0.96	0.93
Random Forest	0.98	0.96
Karar Ağacı	0.94	0.91

Her model için belirlenen en önemli 10 gen, gen ekspresyon profillerindeki potansiyel önemli değişikliklere odaklanarak kanser araştırmalarına yeni bir perspektif sunabilir. Modellerin doğruluk (accuracy) ve kappa değerleri, modelin tahminlerinin doğru ve güvenilir olduğu konusunda fikir edinilmesini sağlar. Bu sonuçlar, akciğer kanserinin teşhis edilmesine katkı sağlar. Bu çalışmada veri analizi, filtreleme ve özellik seçimi yaparak veri setinden anlamlı bilgi çıkarmayı hedeflemektedir.



Şekil2:Venn Şeması ortak genler

5. Tartışma

Veri setindeki 54675 gen ifadesi filtreleme işlemi yapıldı. Ardından özellik seçimi ile sayısı azaltıldı. Makine öğrenme modelleri uygulanarak en önemli 10 gen belirlendi. Bu genlerin bazılarının üç algoritmada da ortak olduğu görüldü. Bu genlerin, akciğer kanseri tedavisinde özel bir öneme sahip olabileceği düşünülmektedir. Bu genlerin kanser biyolojisi üzerindeki etkilerinin anlaşılması, daha etkili tedavi stratejilerinin geliştirilmesine katkıda sağlayabilir. Sonuçlara bakıldığında, bu genlerin daha ayrıntılı bir şekilde incelenmesi gerektiğini göstermektedir. Bu tür genlerin anlaşılması, akciğer kanseri tedavisi alanında bireysel ihtiyaçlara odaklanan yöntemlerin geliştirilmesine olanak sağlayarak hastalara daha etkili ve kişiselleştirilmiş tedavi seçenekleri sunabilir.

REFERANSLAR

[1] TEMURTAŞ, F., ÖZTEKİN, M., YAZDANİ, M., YÖRÜK, Y. E., vd. (2019). AKCİĞER KANSERİ TANISI İÇİN YENİ BİR YÖNTEM. Mühendislik Bilimleri Ve Araştırmaları Dergisi, 1(1), 35-48. https://doi.org/10.46387/bjesr.629166

[2] VERİ MADENCİLİĞİ MODELLERİNİN AKCİĞER KANSERİ VERİ SETİ ÜZERİNDE BAŞARILARININ İNCELENMESİ Nihat Barış SEBİK *, Halil İbrahim BÜLBÜL*