Tübitak 2242
Üniversite Öğrencileri Araştırma Projesi
2022

Projenin Tam Başlığı

TCGA Klinik ve Moleküler Verilerine Dayanarak Akciğer Adenokarsinom ve Akciğer Skuamöz Hücre Karsinomu Hastalarının Risk Tahmini İçin Mobil Uygulama Geliştirilmesi

Proje Kategorisi- Tematik Alanı

Sağlık – Yapay Zeka

Anahtar Kelimeler

Akciğer Kanseri, LUAD, LUSC, Mobil Uygulama, Risk Tahmini, Yapay Zekâ, Makine Öğrenmesi

Proje Yürtücüsü

Mehmet Cihan Sakman

Proje Danışmanı

Tuğba Önal Süzek

İçindekiler

Resimler, Figürler ve Tablolar Listesi

Özet

Akciğer kanseri dünya çapında en çok rastlanan ve ölüme en çok sebebiyet veren kanser türüdür. Dünya Sağlık Örgütü (DSÖ) akciğer kanserinin 2012 yılında toplam 1.8 milyon yeni olgu ile tüm kanserlerin %12.9'unu meydana getirdiğini bildirmiştir [1]. Ülkemizde Sağlık Bakanlığının 2012 yılı verilerine göre erkeklerde en sık; kadınlarda ise beşinci sıklıkta görülen kanser türüdür ve her yıl yaklaşık 30,000 yeni vakanın teşhis edildiği tahmin edilmektedir [2].

Ülkemizde akciğer kanseri tedavi seyri halen yaygınlıkla patoloji sonrası evreleme ile karar verilse de tümörün hastaya özgü genetik profilini belirleyen genetik testleri yapabilen kliniklerin sayısı hızla artmaktadır. Yurtdışında her hastanın kendi tükürüğünden çok düşük maliyetlerle ($99-$200) bilinen kanser mutasyonları için tarayabilen (23andme gibi) testler FDA tarafından klinik karar mekanizmasında kullanılmasına izin verilmiştir [18] . Kanser evrelemesinde ve tedavi sürecinin planlanmasında yurtdışında artık sadece evreleme değil, mutasyon verileri de göz önüne alınmaktadır. Biz de ülkemizde akciğer kanseri hastaları ve klinisyenler için klinik verilere ek olarak somatik mutasyon verilerini de göz önüne alan makine öğrenmesi modeli ve mobil uygulama geliştirmeyi amaçladık.

Bu proje sonunda klinisyenlerin akciğer kanseri hastanın risk grubuna göre hareket etmeleri doğrultusunda onlara karar-destek mekanizması (ikinci bir görüş) sağlamak ve kullanımı kolay bir mobil uygulama geliştirmek amaçlanmıştır. Bu mobil uygulama sayesinde klinisyenlerin hastalarının, hastalıklarının ne kadar ciddi olduğu ve hastalığın nasıl ilerleyebileceği hakkında bilgi sahibi olmaları amaçlanmıştır. Aynı zamanda kanser tedavisi gören hastaların da aynı uygulama üzerinden hastalıklarının nasıl seyredeceğine dair bilgi edinmeleri mümkün hale gelecektir.

Araştırmanın mobil uygulamaya dönüştürülmesine kadarki evrede izlenen başlıca adımlar: Önceki yayınımızda belirlediğimiz Akciğer adenokarsinoma ve Akciğer skuamoz karsinoma sağkalım riski ve somatik sürücü mutasyonları verilerinin makina algoritmasına hazırlanacak şekilde ön işlemeden geçirilmesi, Python "scikit-learn" kütüphanesi kullanılarak 5 farklı makine öğrenmesi algoritması ile hastanın, hastalığı ile ilgili en doğru tahminlerin yapılması, kullanıcılar tarafından kullanımı kolay olabilecek bir mobil uygulama arayüz tasarımının yapılması ve yazılımın mobil uygulamaya uyarlanması, Mobil uygulamanın Muğla Sıtkı Koçman Tıp fakültesindeki Göğüs Hastalıkları uzmanı, Tıbbi Onkoloji uzmanı ve Balıkesir Sevgi hastanesinden Genel Cerrahi Uzmanı Op. Dr. Erol Önal tarafından test edilmesi ve geri dönüş alınarak bu yönde iyileştirmeler yapılması, mobil uygulamanın Google Play Store ve App Store'da basılacak bir uygulama haline getirilip yaygınlaştırılmasıdır.

Hastaların klinik ve genetik bilgilerinin kullanıldığı projemizde her iki hastalık için de en iyi performans **Logistic Regression** algoritması ile elde edilmiştir. *Akciğer adenokarsinoma* hastaları için **0.90**, *Akciğer skuamoz karsinoma* hastaları için **0.85** F1 ve doğrulukla risk tahmin modelleri oluşturulmuştur.

Anahtar Kelimeler: Akciğer Kanseri, LUAD, LUSC, Mobil Uygulama, Risk Teşhisi, Medikal Enformatik, Yapay Zekâ, Makine Öğrenmesi

1. Giriş

Akciğer kanseri en sık görülen kanser türüdür ve dünya çapında önde gelen ölüm nedenidir. Dünya Sağlık Örgütü (WHO), akciğer kanserinin tüm kanserlerin %11.4'ünü oluşturan ikinci en sık teşhis edilen kanser türü olduğunu ve 2020 yılında kansere bağlı ölümlerin (%18) önde gelen nedeni olduğunu bildirmiştir **[3]**. Amerika Birleşik Devletleri'nde akciğer kanseri gelişen hastaların sadece %14'ü beş yıl hayatta kalmaktadır. Bu ölüm oranları (>150.000/yıl), edinilmiş immün yetmezlik sendromu salgınınınkinden çok daha fazladır. Bununla birlikte, bu hayatta kalma oranı son yirmi yılda sadece biraz artmıştır ve yakın gelecekte belirgin iyileşmelerin gerçekleşmesi olası görünmemektedir **[4]**. Ülkemizde Sağlık Bakanlığının 2012 yılı verilerine göre erkeklerde en sık; kadınlarda ise beşinci sıklıkta görülen kanser türüdür ve her yıl yaklaşık 30,000 yeni vakanın teşhis edildiği tahmin edilmektedir **[2]**.

Akciğer kanseri iki sınıfa ayrılabilir: Küçük Hücreli Akciğer Kanseri ve Küçük Hücreli Olmayan Akciğer Kanseri. Akciğer Adenokarsinom (LUAD) ve Akciğer Skuamöz Hücreli Karsinom (LUSC), akciğer kanseri hastalarının %85'inde görülen Küçük Hücreli Olmayan Akciğer Kanserinin üç alt tipinden ikisidir. Akciğer kanseri vakalarının büyük çoğunluğunu oluşturan LUAD ve LUSC kanseri hastalarının sağlık durumları erken evrede tespit edilir ve kanserin gidişatına göre bir tedavi uygulanırsa bu hastaların hayatta kalma olasılıkları artırılabilir. Mobil cihazların hayatımıza yerleşmesinden sonra sağlık alanında birçok mobil uygulama üretilmiş ve hastanın sağlığı güvence altına alınmaya çalışılmıştır. Kanser vakalarının tespiti hakkında birçok Makine Öğrenmesi projesi yapılmış ancak ülkemizde veya dünyada bu yönde hayata geçirilmiş herhangi bir mobil uygulama üretilememiştir. Bu alanda üretilen ve piyasaya sürülen ONCOassist uygulaması klinisyenler için verileri bir araya getirip kanser vakaları hakkında bir özet bilgi sunmaktadır. Ancak bu uygulamada herhangi bir Makine Öğrenmesi söz konusu değildir.

Şu anda, Makine Öğrenimi yöntemleri, tıp pratisyenlerine yardımcı olan sınıflandırma ve tahmindeki başarıları nedeniyle klinikteki karar verme süreçlerine entegre edilmektedir **[5]**. Makine öğrenimi, sağlık hizmetlerinin artan fiyatını düşürmek ve gelişmiş hasta-klinisyen iletişimi oluşturmak için alternatif bir çözüm sağlar.

Bu projede, Python sci-kit-learn kütüphanesini kullanarak Küçük Hücreli Olmayan Akciğer Kanserinin risk gruplarını belirlemek için beş farklı makine öğrenme algoritmasının tahmin gücünü karşılaştırmalı olarak değerlendirildi ve içlerinden en iyi performans gösteren öğrenme algoritmasına dayalı bir mobil uygulama geliştirildi. Sağkalım risk hesaplaması ve genetik ara yüzün tasarımı MSKU Moleküler Biyoloji ve Genetik Anabilim Dalı Araştırma Görevlisi ve Biyoinformatik Anabilim Dalı doktora öğrencisi Talip Zengin tarafından yürütülmüştür. Kullanılan beş farklı makine algoritması: Support Vector Machine, Logistic Regresyon, Naive Bayes, Random Forest ve K Neighbours Classifiers olarak belirlenmiştir.

1.1. Projenin Amacı ve Önemi

Akciğer kanseri dünya çapında en çok rastlanan ve ölüme en çok sebebiyet veren kanser türüdür. Dünya Sağlık Örgütü (DSÖ) akciğer kanserinin 2012 yılında toplam 1.8 milyon yeni olgu ile tüm kanserlerin %12.9'unu meydana getirdiğini bildirmiştir **[1]**. Ülkemizde Sağlık Bakanlığının 2012 yılı verilerine göre erkeklerde en sık; kadınlarda ise beşinci sıklıkta görülen kanser türüdür ve her yıl yaklaşık 30,000 yeni vakanın teşhis edildiği tahmin edilmektedir **[2]**. Klinisyenlerin hastalarının risk durumlarına göre hastaya uygulayacağı tedavi yöntemlerine karar verirken ikinci bir görüşe ihtiyaç duyduğu durumlar oluşabilmektedir. Bu ihtiyaca yönelik olarak klinisyenlerin hastanın risk grubuna göre hareket etmeleri doğrultusunda onlara karar-destek mekanizması (ikinci bir görüş) sağlamak ve kullanımı kolay bir mobil uygulama geliştirmek amaçlanmaktadır. LUAD ve LUSC için genetik veriler ve ilgili klinik bilgiler, halka açık The Cancer Genome Atlas (TCGA) veri tabanından indirildi. TCGA, 33 farklı kanser türünden 11.000 hastadan alınan veriler de dahil olmak üzere, halka açık bilinen en büyük kanser hastası verilerini içerir. (https://www.cancer.gov/tcga). Projemizde 522 LUAD kanser hastasını ve 504 LUSC kanser hastasını ilgili klinik ve genetik bilgilerle 5 farklı makine öğrenmesi algoritması eğitilmiş ve aralarından en iyi performansı gösteren algoritma gerçek hasta verileriyle test edilecektir ve kullanıcılardan alınan dönütlere göre kullanıma sunulacaktır.

1.2. Projenin çerdiği Yenilik (Özgünlük) Unsuru

LUAD (adenokarsinoma) ya da LUSC (skuamoz hücreli) akciğer kanseri hastası klinikte tedavisi birden çok klinisyen tarafından yönetilen bir süreçtir. Genelde hastanın şikayetiyle ulaştığı Göğüs Hastalıkları uzmanı hastayı Radyoloji Uzmanına sevki ve sonrasında cerrahi müdahale için Genel Cerrahi uzmanına sevki ile teşhis süreci başlar. Genel Cerrahi uzmanının Patoloji uzmanı tarafından konulan teşhis ile beraber hastayı Tıbbi Onkoloji uzmanına yönlendirmesiyle de tedavi süreci başlar. Bu birden çok klinisyenden çoğu zaman farklı kliniklerden randevu alınması gereken süreçte hasta kendi risk düzeyi hakkında bir belirsizlik içinde kalmaktadır. Birden çok klinisyenin ve yardımcı personelin içinde olduğu süreçte insan hatalarının ve gecikmelerinin en aza indirilmesi için hem klinisyenlerin hem de hastanın

hastaya özgü risk durumunu öğrenebileceği klinik destek yazılımları hem tedavi kalitesinin artırılması hem de randevu süreçlerinin hızlandırılmasında yardımcı olacaktır.

Yazılımımız bir klinisyenin LUAD ve LUSC akciğer kanseri hastaları için cinsiyet/stage/yaş/sigara kullanımı/ilaç gibi klinik verileri basit bir arayüzle girdiğinde risk sınıflandırması yapan bir klinik destek mobil uygulaması olarak piyasadaki ilk örneği olacaktır. Arka planda eğiteceğimiz model için sağkalım ve ilaç gibi klinik verilere ek olarak mutasyon ve gen ifadesi verileri kullanılmıştır. Geliştireceğimiz bu yazılım sayesinde klinisyen (Göğüs Hastalıkları uzmanı ya da Tıbbi Onkoloji uzmanı) teşhis alan hastanın düşük risk ya da yüksek risk grubu hakkında bir tahmin sonucunu göz önüne alarak farklı tedavi yaklaşımı planlayabilecektir. Aynı zamanda kanserin herhangi bir seviyesinde bulunan hastaların doktorlar tarafından hangi ilaçla tedavi edildiğine ve hastaların bu sürece nasıl tepki verdiklerine dair hastayı bilgilendirici bir arayüz tasarımı yapılacaktır ve mevcut klinisyenler için tasarlanmış risk-seviyesi tahmini modüllerine ek olarak hastaların kullanımına sunulacaktır.

Biyoinformatik Anabilim Dalından akademik danışmanlarımız Tuğba Süzek, PhD (Bilgisayar Mühendisi ve Biyoinformatik uzmanı) ve Talip Zengin, MSc (Moleküler Biyolog ve Biyoinformatik uzmanı) tarafından yayınlanmış ve basım aşamasında olan iki makale yayınında akciğer kanserinin iki çeşidi olan LUAD ve LUSC verilerinden risk gruplarını hesaplayan bir öğrenim modeli geliştirilmişti **[6,7]**. Çalışmamızın iki ana amacı 1) bu risk modellerine dayanarak hasta kohortlarının klinik parametrelerini derin öğrenme temelli tahmin modelleri geliştirmek 2) klinisyen ve hastalara TCGA verilerini ve risk analizlerimizi kullanıma sunmak için kullanıcı dostu bir mobil uygulama geliştirmektir.

Bu alanda yapılan benzer çalışmaları sıralayacak olursak:

- Mikroskobik patoloji görüntü özellikleriyle LUAD ve LUSC -adenokarsinoma ya da skuamoz hücreli- akciğer kanseri prognozunu tahmin eden yazılım **[8]**
- Küçük Hücreli Olmayan Akciğer Kanserinin (NSCLC) patolojik evresinin teşhisi ve tahmini için kullanılabilen görüntüleme biyobelirteçlerini, CT görüntü özellik analizine dayalı çoklu makine öğrenme algoritmaları kullanılarak keşfeden yazılım **[9]**
- TCGA'dan alınan RNA-sıralama verilerini kullanarak Ensemble Learning yardımı ile kanser hastalarının hayatta kalma sürelerini tahmin etmeye çalışan yazılım **[10]**
- Projemizle en çok benzerlik gösteren bir uygulama olan ONCOassist, tüm kanser çeşitleri için Mobil ve Web arayüzlerinde klinisyenler tarafından hastanın istenilen bilgilerinin girilmesi durumunda, klinisyene hastanın gelecek 5 yıl içerisindeki yaşama şansına dair istatiksel bilgi sunan bir yazılımdır.**[11]**

Sıraladığımız benzer yazılımlara rağmen yaptığımız araştırmalar sonucunda LUAD ve LUSC kanseri hastalarının klinik verilerinin girildiğinde, bu hastaların yaşam süreleriyle ilgili bir makine öğrenmesi algoritması kullanan ve klinisyenleri bilgilendiren herhangi bir mobil uygulama başvuru tarihi itibariyle bilgimiz dahilinde mevcut değildir. Akciğer kanserinin ülkemizde ve dünya çapında sebep olduğu ölümler göz önüne alındığında, temel klinik veriler yardımıyla hastalığın seyir riskini tahmin ederek insan hatalarını azaltma amaçlı bir uygulamanın önemi, bu yönde klinisyenler ve hastalar tarafından kullanıma hazır herhangi bir mobil uygulamanın olmaması bu projeyi özgün kılmaktadır. Böyle bir mobil uygulamanın hayata geçirilmesi kanser hastaları ve klinisyenlerin hasta açısından belirsizlikler içeren zor bir sürecin yönetiminde insan hatalarını ve gecikmeleri azaltarak tedavi sürecini kolaylaştıracak bir çözüm haline gelecektir.

Akciğer kanserinin ülkemizde ve dünya çapında sebep olduğu ölümler göz önüne alındığında, temel klinik veriler yardımıyla hastalığın seyir riskini tahmin ederek insan hatalarını azaltmaya yardımcı bir uygulamanın önemi, bu yönde klinisyenler ve hastalar tarafından kullanıma hazır herhangi bir mobil uygulamanın olmaması bu projeyi özgün kılmaktadır. Böyle bir mobil uygulamanın hayata geçirilmesi kanser hastaları ve klinisyenlerin hasta açısından belirsizlikler içeren zor bir sürecin yönetiminde insan hatalarını en aza indirmeye yardımcı olacaktır.

1.3. Projenin İlgili Olduğu Teknoloji Alanları

Proje kapsamında TCGA'den alınmış LUAD ve LUSC Akciğer Kanseri hastalarına ait klinik ve genetik bilgileri kullanılarak bir model eğitebilmek için Python Makine Öğrenmesi ve Veri Bilimi teknolojilerinden *Scikit-learn, Pandas, Numpy* kütüphanelerinden yararlanılmıştır. Bunun yanı sıra

eğitilen makine öğrenmesi modelini servera entegre etmek üzere *IBM Watson Machine Learning Cloud Server*'i kullanılmıştır. Eğitilen modeli kullanıcıların kullanabileceği bir platforma aktarmak üzere ise back-end tarafında Node.js, Front-end tarafında ise Reac Native.js kullanılmıştır.

2. Yöntem ve Teknikler

2.1. Verilerin Toplaması

LUAD ve LUSC için genetik veriler ve ilgili klinik bilgiler, halka açık The Cancer Genome Atlas (TCGA) veri tabanından indirildi. TCGA, 33 farklı kanser türünden 11.000 hastadan alınan veriler de dahil olmak üzere, halka açık bilinen en büyük kanser hastası verilerini içerir(https://www.cancer.gov/tcga). Projemizde 522 LUAD kanser hastasını ve 504 LUSC kanser hastasını ilgili klinik ve genetik bilgilerle indirdik. Eksik değerlerin filtrelenmesinden sonra, 51 hastanın kalan ömrü beş yıldan fazla olmak üzere toplam 504 LUAD kanser hastası ve 83 hastanın kalan ömrü beş yıldan fazla olmak üzere 494 LUSC kanser hastası ile çalışılmıştır.

2.2. Verilerin Ön İşlemesi

Verilerin eksik, yanlış veya hatalı bir şekilde girilmesi, biyolojik verilerle makine öğrenmesi eğitimi sırasında karşılaşılan yaygın bir sorundur. Biyomedikal alandaki verilerin çoğu düzgün değildir. Bu nedenle, değerli biyomedikal verileri kurtarmak ve korumak için birkaç ön işleme stratejisi uygulandı. Kayıp değerlerin yüksek oranı ile başa çıkmak için, %80'den az veri içeriğine sahip özellikler (%20'den fazla kayıp değer içeren sütunlar) eğitim ve test veri setlerinden kaldırıldı. submitter_id, diagnosis_id, exposure_id, demographic_id, treatment_id, ve bcr_patient_barcode sütunları rastgele sayıları temsil eden sütunlardır, bu sütunlar değerli bilgiler içermedikleri için veri setinden kaldırıldılar. year_of_birth, state, up_dated_datetime, tissue_or_organ_of_origin gibi yinelenen(duplicated) sütunlar da verisetinden kaldırıldı Hastanın yaşını days_to_birth(gün olarak) sütunundan bulabileceğimiz için year_of_birth sütunu da çıkartıldı. 'state' sütununda, tüm değerler 'released' olarak girilmiştir; bu nedenle veri setinden kaldırılmıştır. Veriler, sci-kit-learn'in model_selection paketi kullanılarak eğitim (%80) ve test (%20) veri kümelerine bölündü ve sonraki tüm keşifsel veri analizi ve model eğitimi, yalnızca eğitim veri kümesinde gerçekleştirildi.

2.3. Makine Öğrenme Algoritmalarının Uygulanması

Verilerin hazırlanmasından sonra LUAD ve LUSC veri setlerine beş farklı sınıflandırma algoritması (Logistic Regresyon, Random Forest Classifier, Naïve Bayes, SVC ve K-Neighbors Classifier) uygulanmıştır. Daha sonra öğrenme algoritmalarının performansını değerlendirmek için alıcı işletim karakteristikleri (ROC) eğrilerinin (AUC) altındaki alan çizilmiş ve hesaplanmıştır (Figür 1,2).

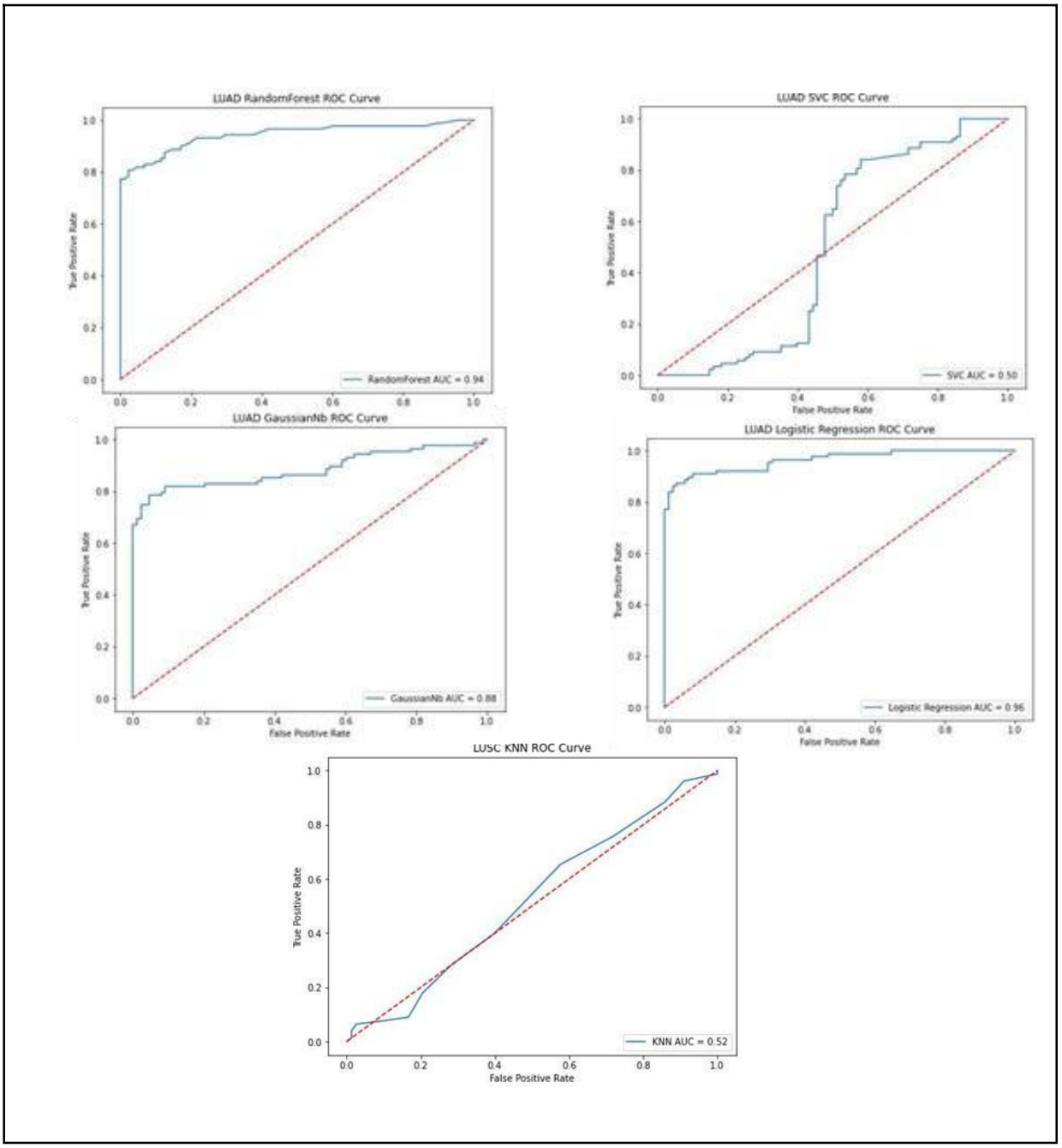


Figure 1: 5 Algoritma için AUC grafikleri LUAD

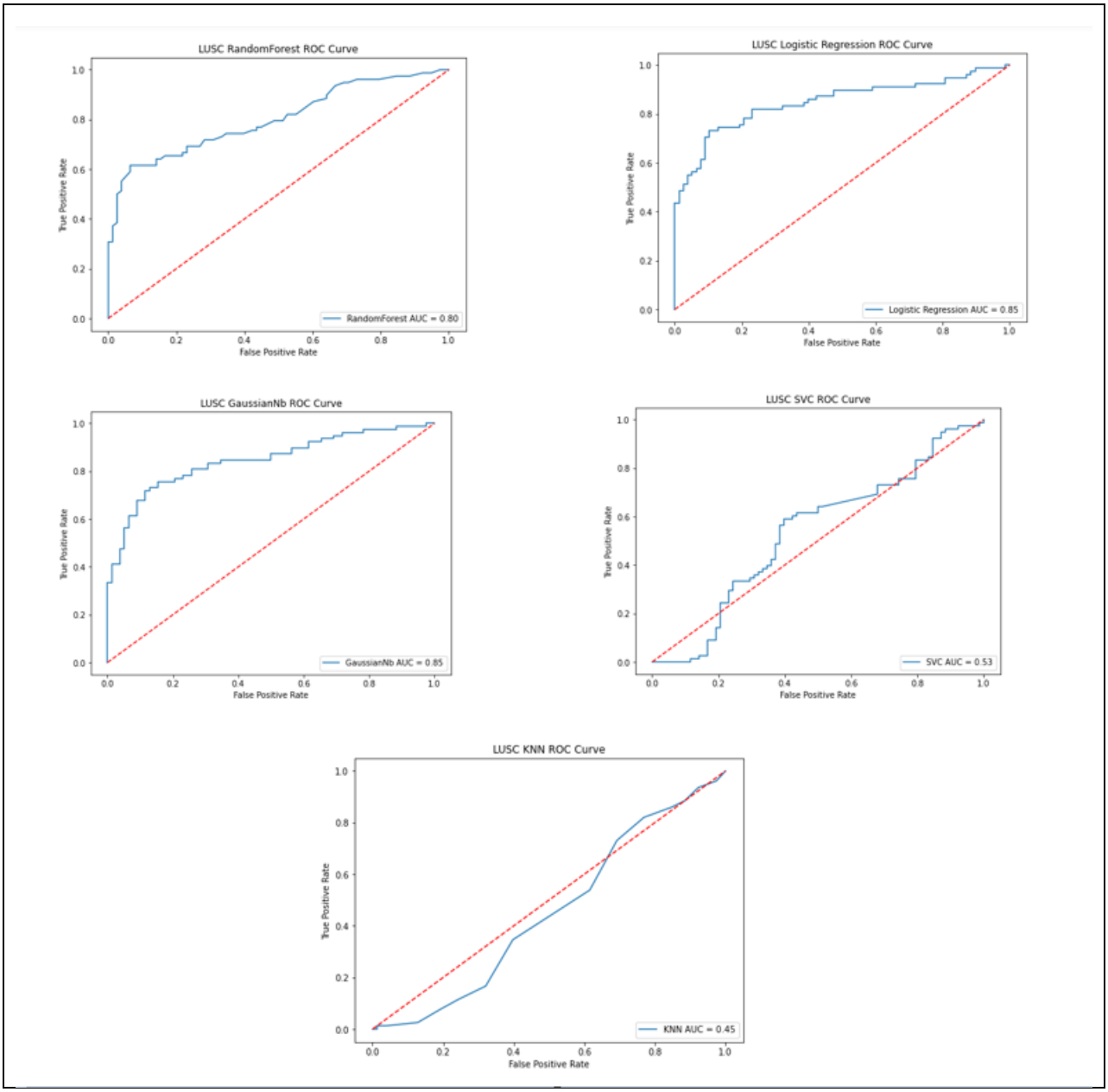


Figure 2: 5 Algoritma için AUC grafikleri LUSC

AUC, tıp alanında onlarca yıldır model seçimi için kullanılmaktadır. Ayrıca, algoritmaların performansınıdeğerlendirmek için makine öğrenme algoritmalarında da yaygın olarak kullanılmaktadır **[12,13]**. Tüm olası puan eşikleri altında, bir modelin y koordinatı olarak gerçek pozitif (TP) oranı ile x koordinatı olarak yanlış pozitif (FP) oranının grafiği olarak tanımlanır. AUC metriklerini desteklemek için F1 puanı, Precision (kesinlik) ve Recall (geriçağırma) hesaplanarak ve Tablo 1 ve Tablo 2'de gösterilmiştir.

Tablo 1: LUAD için 5 algoritmanın Precision, Recall, Accuracy ve F1-skorunun Karşılaştırılması

LUAD Modelleri Değerlendirmeleri				
Models	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression	0.90	0.89	0.89	**0.89**
Random Forest	0.82	0.81	0.81	**0.81**
Naive Bayes	0.72	0.65	0.62	**0.71**
SVC	0.53	0.52	0.45	**0.53**
KNN	0.62	0.62	0.62	**0.62**

Tablo 2: LUSC için 5 algoritmanın Precision, Recall, Accuracy ve F1-skorunun Karşılaştırılması

LUSC Modelleri Değerlendirmeleri				
Models	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression	0.75	0.74	0.73	**0.74**
Random Forest	0.71	0.69	0.68	**0.68**
Naive Bayes	0.79	0.62	0.58	**0.71**
SVC	0.56	0.53	0.48	**0.56**
KNN	0.53	0.53	0.52	**0.53**

2.4. Hyperparameter Tuning ile Öğrenme Algoritmalarının Geliştirilmesi

Beş farklı sınıflandırma algoritmasından en iyi iki puanlama algoritması **Random Forest** ve **Logistic Regression** olarak bulunmuş, böylece Random Forest ve Logistic Regression tabanlı modellere *Hyperparameter Tuning* uygulanmıştır. Bu iki algoritmanın hiperparametre ayarı için, önce eğitim setini 5 kata (5-fold cross validation) bölünür ve ardından sınıflandırma modelini eğitmek için kullanılan 4 kata rastgele aşırı örnekleme (SMOTE) uygulanır. Ardından model performans değerlendirmesi yapılan kalan 1 kat için **scikit-learn** *model_selection* kütüphanesinden **GridSearchCV** en iyi sonuçları veren hyperparametre ayarlarını kaydeder ve daha sonrasında eniyi model parametrelerini dönüdürür. En iyi iki algoritmanın Hyperparameter Tuning yapıldıktan sonra ve önceki kıyasalamalarını Tablo 3'te açıkça görebiliriz.

Scikit-learn'ün GridSearchCV (scoring='f1', cv=5) algoritmasını Random Forest'a uyguladıktan sonra LUAD için en iyi parametreler max_depth=8, max_features=log2, min_samples_leaf=3, n_estimators=50 olarak 0.93 f1- score ile bulundu. LUSC

için ise max_depth=8, max_features='auto', min_samples_leaf=3, n_estimators=200 olarak 0.87 f1-score ile bulundu. Diğer yandan, Logistic Regression'a uygulanan GridSearchCV, LUAD için en iyi parametreleri C=16.77, penalty=l2, solver=newton-cg olarak, 0.92 f1-score ile, LUSC için ise en iyi parametreleri C=109.85, penalty=l1, solver=liblinear olarak 0.85 f1-score ile buldu (Tablo 3).

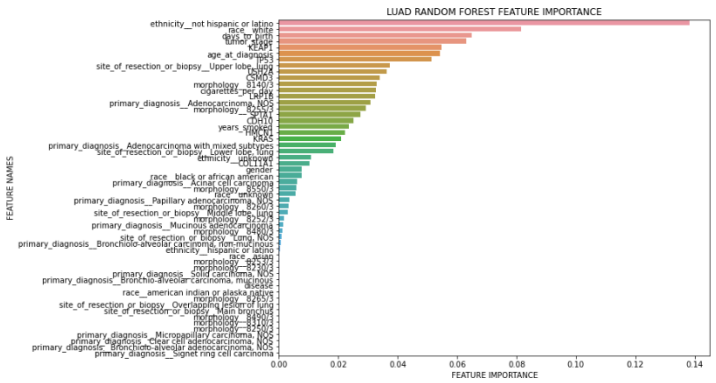
Tablo 3: Hiperparametre ayarı olan ve olmayan Random Forest ve Logistic Regression içinperformans metriklerinin karşılaştırılması

Model	Data	GridSearchCV	Precision	Recall	F1-Score	Accuracy
Logistic Regression	LUAD	No/**Yes**	0.92/**0.94**	0.92/**0.93**	0.92/**0.93**	%91.7/%**93.1**
Random Forest	LUAD	No/**Yes**	0.88/**0.90**	0.88/**0.90**	0.87/**0.90**	%87.9/%**89.7**
Logistic Regression	LUSC	No/**Yes**	0.77/**0.82**	0.74/**0.78**	0.74/**0.77**	%74.3/%**77.5**
Random Forest	LUSC	No/**Yes**	0.72/**0.77**	0.68/**0.71**	0.66/**0.71**	%66.6/%**70.5**

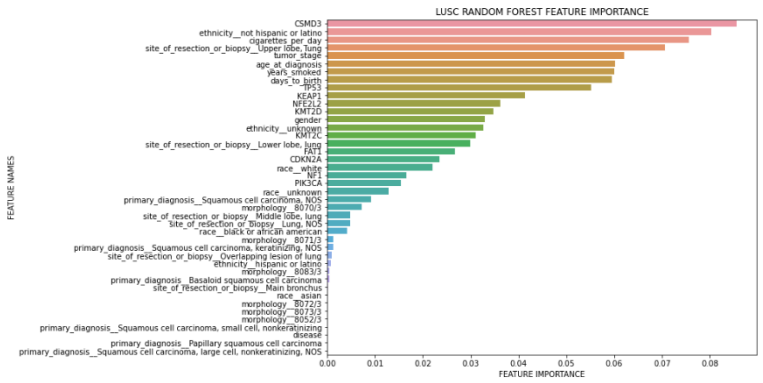
2.5. Somatik Olarak Mutasyonlu Genler

Bu bölümde, özellik mühendisliği (*feature engineering*) ile hangi somatik mutasyona uğramış genlerin risk sınıflandırmasını etkilediğini araştırıldı. Bunun için, klinikopatolojik özelliklerin yanı sıra genomik özellikler olarak en yüksek oranda mutasyona uğramış ilk 10 somatik sürücü gen birleştirildi. Bu amaçla, SominaClust**[14]** kullanılarak önceki yayınımızda**[15]** tanımlanan LUAD ve LUSC için en yüksek oranda mutasyona uğramış ilk 10somatik sürücü genini seçildi. **LUAD** hastaları için en somatik olarak mutasyona uğramış ilk 10 sürücü gen özelliği *CDH10, COL11A1, CSMD3, HMCN1, KEAP1, KRAS, LRP1B, SPTA1 TP53, USH2A* idi. Hastanın riski (yüksek-düşük) ile bu on gen arasında bir korelasyon bulmaya çalıştığımızda, ilk 10 özellik listesinde **KEAP1, TP53, USH2a** ve **CSMD3** ortaya çıktı. Ayrıca '**KEAP1**' ile hastanın riski arasında bir katsayı ilişkisi olduğu gözlemlendi. Bununla birlikte, mutasyona uğramış genlerin eklenmesinden sonra LUAD Logistic Regression ve Random Forest modelinin performansında önemli bir gelişme olmadı (Figür 3 ve Figür 5). Her beş katın performansı, ortalama ve standart sapması ile Tablo 6'da gözlemlenebilir. Ayrıca, KEAP1 mutasyonunun diğer gen mutasyonlarına kıyasla daha yüksek öneme sahip olduğu gözlemlendi ve bunu TP53, USH2A, CSMD3, LRP1B, SPTA1, CDH10, HMCN1, KRAS ve COL11A1 gen mutasyonları takip etmektedir. Mutasyona uğramış genler birçok klinik özellikten daha fazla öneme sahiptir, ancak ilginç bir şekilde etnik köken ve ırk en yüksek özellik önemine sahip olarak gözlemlenmiştir. Yaş ve tümör evresi gibi sıklıkla kullanılan klinik değişkenler, gen mutasyonlarından daha fazla öneme sahiptir. Gen mutasyonları ile birlikte site_of_resection(rezeksiyon yeri), morphology(morfolojisi), primary_diagnosis(primer tanısı), sigara içme miktarı da önemli özelliklerdir. Gen mutasyonlarının eklenmesi performansı iyileştirmese de klinik değişkenler açısından yüksek öneme sahiptir, bu nedenle birlikte kullanılabilirler. Örneğin, KEAP1 genindeki fonksiyon kaybı mutasyonları, KRAS kaynaklı akciğer tümörijenezini teşvik eder **[16]**, bu da KEAP1'in hasta riski ile korelasyonunun nedeni olabilir, bu nedenleKEAP1 ve KRAS'ın klinik değişkenlerle birlikte kullanılması düşünülebilir.

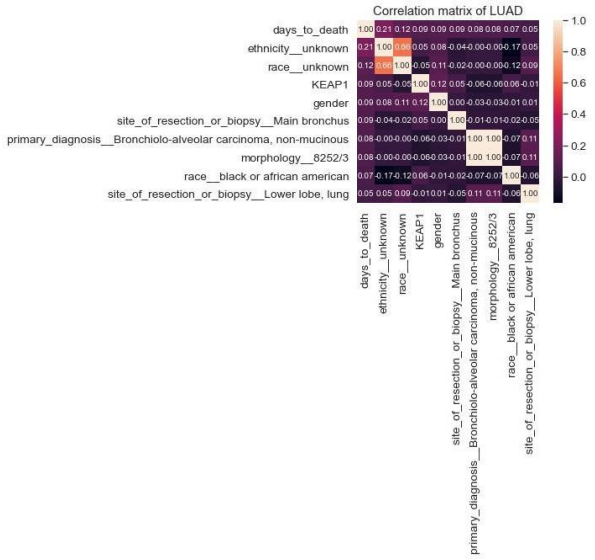
Sınıflandırma modeline en yüksek oranda mutasyona uğramış ilk 10 somatik sürücü genin eklenmesi LUAD hastalarının performansını iyileştirmese de, LUSC hastalarının sınıflandırma modelini büyük ölçüde iyileştirdiler(Tablo 5). Önceki yayınımızda rapor edilen LUSC hastalarının somatik olarak mutasyona uğramış ilk 10 geni CDKN2A, CSMD3, FAT1, KEAP1, KMT2C, KMT2D, NF1 NFE2L2, PIK3CA, TP53 idi **[15]**. Bu 10 gen ve hastanın risk (yüksek ve düşük) durumu modele birlikte verildiğinde, CSMD3 geni ile hastanın riski arasındaki korelasyon, LUSC'de hem Logistic Regression hem de Random Forest performansını iyileştirerek LUSC verisetiiçin modeli en çok etkileyen gen mutasyonu oldu (Figür 4 ve Figür 6). CSMD3, akciğer kanserinde en sık mutasyona uğrayan genlerden biridir ve potansiyel bir tümör baskılayıcıdır **[17]**. Etnik köken ikinci en önemli özelliktir ve sigara içme miktarı, yaş ve tümör evresi ise yüksek öneme sahiptir ve bunu TP53, KEAP1, NFE2L2, KMT2D, KMT2C, FAT1, CDKN2A, NF1 ve PIK3CA gen mutasyonları izlemektedir.



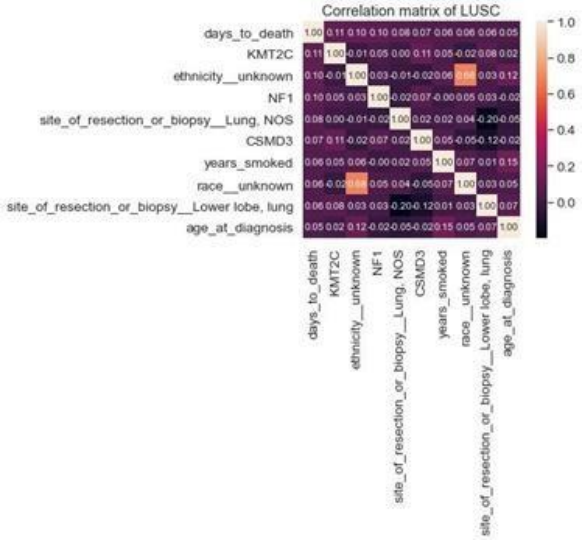
Figür3: LUAD Random Forest Özellik Önemi



Figür4: LUSC Random Forest Özellik Önemi



Figür5: LUAD days_to_death ile en çok korelasyonu olan Top 9 özellik



Figür6: LUSC days_to_death ile en çok korelasyonu olan Top 9 özellik

Tablo4: Klinik özellikler ve somatik olarak mutasyona uğramış ilk 10 gen ile LUAD modeli için her foldun doğruluğu

5-Fold Cross Validation for Clinical features and top 10 most mutated genes inLUAD									
Model	**Metrics**	Data Type	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	**mean**	**std**
Logistic Regression	F1-Score	Clinical/**Mutation**	0.91/**0.89**	0.93/**0.90**	0.89/**0.90**	0.91/**0.88**	0.88/**0.89**	0.90/**0.89**	0.017/**0.009**
	Precision	Clinical/**Mutation**	0.91/**0.90**	0.94/**0.91**	0.90/**0.90**	0.91/**0.88**	0.90/**0.89**	0.91/**0.90**	0.014/**0.011**
	Recall	Clinical/**Mutation**	0.91/**0.89**	0.94/**0.90**	0.89/**0.90**	0.91/**0.88**	0.89/**0.89**	0.91/**0.89**	0.018/**0.009**
	Accuracy	Clinical/**Mutation**	0.91/**0.89**	0.93/**0.90**	0.89/**0.90**	0.91/**0.88**	0.89/**0.89**	0.90/**0.89**	0.015/**0.009**
	AUC	Clinical/**Mutation**	0.95/**0.91**	0.95/**0.98**	0.94/**0.93**	0.95/**0.92**	0.89/**0.95**	0.93/**0.94**	0.023/**0.028**
Random Forest	F1-Score	Clinical/**Mutation**	0.86/**0.81**	0.86/**0.90**	0.87/**0.77**	0.87/**0.80**	0.73/**0.81**	0.84/**0.82**	0.054/**0.052**
	Precision	Clinical/**Mutation**	0.86/**0.85**	0.88/**0.91**	0.88/**0.85**	0.88/**0.82**	0.79/**0.87**	0.86/**0.86**	0.035/**0.031**
	Recall	Clinical/**Mutation**	0.86/**0.82**	0.86/**0.90**	0.87/**0.78**	0.87/**0.79**	0.74/**0.81**	0.84/**0.82**	0.050/**0.049**
	Accuracy	Clinical/**Mutation**	0.86/**0.81**	0.86/**0.91**	0.87/**0.78**	0.87/**0.79**	0.74/**0.81**	0.84/**0.82**	0.050/**0.049**
	AUC	Clinical/**Mutation**	0.94/**0.93**	0.91/**0.96**	0.95/**0.88**	0.91/**0.89**	0.82/**0.92**	0.90/**0.92**	0.046/**0.031**

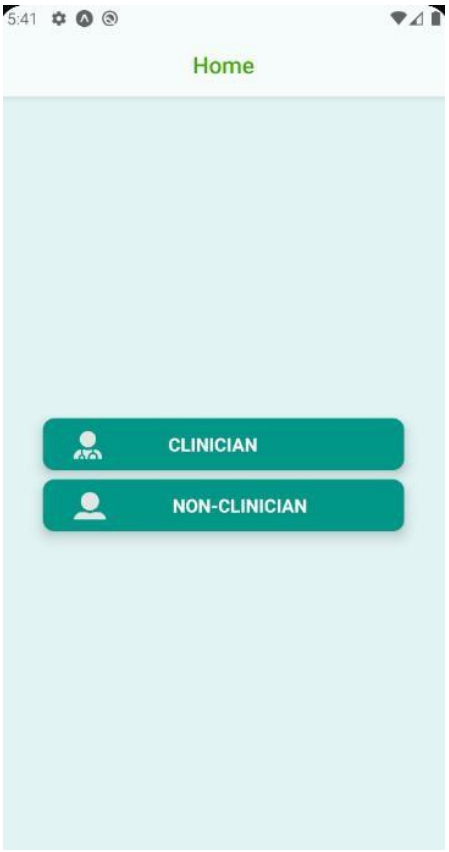
Tablo5: Klinik özellikler ve somatik olarak mutasyona uğramış ilk 10 gen ile LUSC modeli için her foldun doğruluğu

5-Fold Cross Validation for Clinical features and top 10 most mutated genes									
Model	**Metrics**	Data Type	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	**mean**	**std**
Logistic Regression	F1-Score	Clinical/**Mutation**	0.90/**0.89**	0.85/**0.80**	0.80/**0.87**	0.83/**0.82**	0.81/**0.85**	0.84/**0.85**	0.040/**0.036**
	Precision	Clinical/**Mutation**	0.91/**0.90**	0.86/**0.82**	0.84/**0.90**	0.86/**0.83**	0.81/**0.86**	0.86/**0.87**	0.038/**0.037**
	Recall	Clinical/**Mutation**	0.90/**0.89**	0.85/**0.80**	0.81/**0.87**	0.83/**0.82**	0.81/**0.85**	0.84/**0.84**	0.039/**0.036**
	Accuracy	Clinical/**Mutation**	0.90/**0.89**	0.85/**0.80**	0.81/**0.87**	0.83/**0.83**	0.81/**0.85**	0.84/**0.85**	0.039/**0.034**
	AUC	Clinical/**Mutation**	0.93/**0.93**	0.89/**0.88**	0.86/**0.93**	0.88/**0.90**	0.87/**0.91**	0.89/**0.91**	0.025/**0.021**
Random Forest	F1-Score	Clinical/**Mutation**	0.78/**0.85**	0.83/**0.75**	0.82/**0.76**	0.69/**0.81**	0.62/**0.73**	0.75/**0.78**	0.090/**0.048**
	Precision	Clinical/**Mutation**	0.79/**0.87**	0.84/**0.81**	0.83/**0.80**	0.71/**0.82**	0.67/**0.78**	0.77/**0.82**	0.074/**0.032**
	Recall	Clinical/**Mutation**	0.78/**0.85**	0.83/**0.76**	0.82/**0.77**	0.70/**0.81**	0.64/**0.74**	0.75/**0.79**	0.083/**0.044**
	Accuracy	Clinical/**Mutation**	0.78/**0.85**	0.83/**0.76**	0.82/**0.77**	0.70/**0.81**	0.64/**0.74**	0.75/**0.79**	0.083/**0.044**
	AUC	Clinical/**Mutation**	0.79/**0.91**	0.92/**0.86**	0.92/**0.82**	0.76/**0.88**	0.72/**0.84**	0.82/**0.86**	0.092/**0.037**

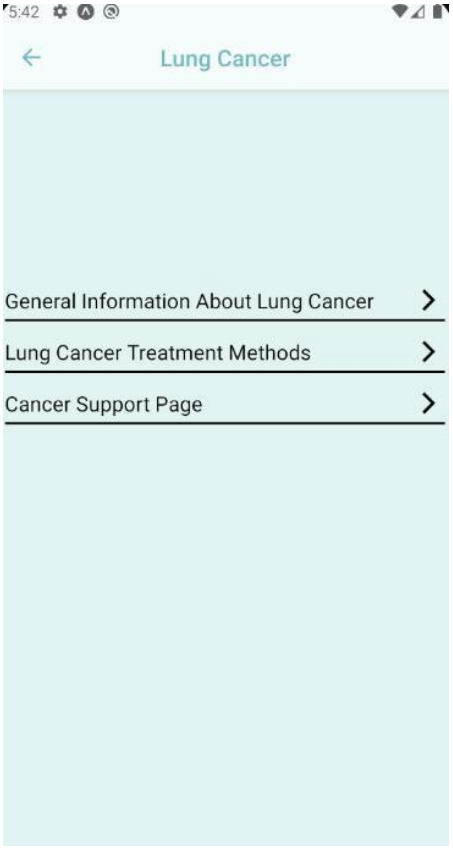
2.6. Mobil Uygulama Entegrasyonu

LUAD ve LUSC test verileri üzerinde en yüksek performansı veren Logistic Regression makine öğrenme algoritmaları, React Native kullanılarak iOS ve Android platformlarında kullanıcı ile birleştirilmek üzere bir mobiluygulama içine entegre edilmiştir. Modellerin entegrasyonu için IBM Watson Machine Learning sunucusu kullanılmıştır. IBM Watson sunucusuna entegre edilen makine öğrenmesi algoritmalarının React Native tarafında Node.js ile network bağlantıları gerçekleştirilmiştir. Daha sonra kullanıcılara sunulan ara yüz üzerinden makine öğrenme algoritmaları başarıyla çalıştırılabilmektedir. Tasarlanan mobil uygulama üzerinden gerçekleştirilebilecek işlemler aşağıdaki fotoğraflarda adım adım gösterilmiştir.

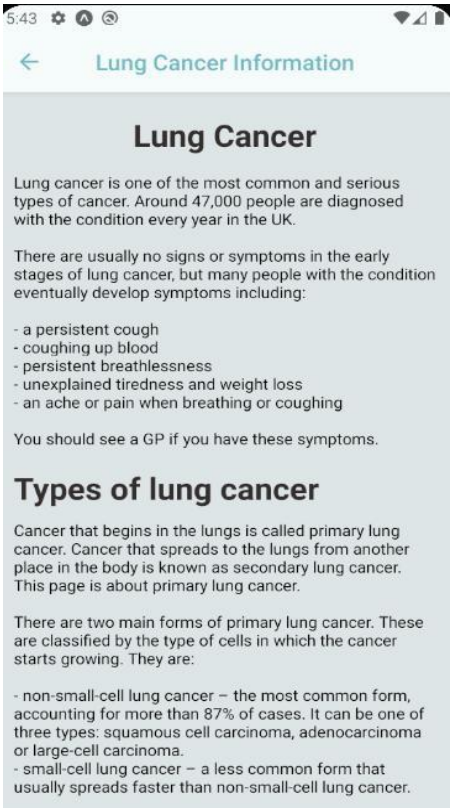
Son olarak, projemizin bitmiş hali Muğla Sıtkı Koçman Üniversitesi' ndeki sunucuya yüklenecektir ve mobil uygulama üzerinden klinisyen ve hastaların kullanımına açılacaktır. Hasta verileri hiçbir şekilde SüzekLab sunucularında ya da bulut ortamında saklanmayacak, işlenmeyecektir. Klinisyenin değerlendirme için veriler istemcinin (client) kendi cihazında KVKK'ya uygun olarak işlenecektir. Hiçbir şekilde model eğitiminde halka açık veriler dışında veri kullanılmamıştır.



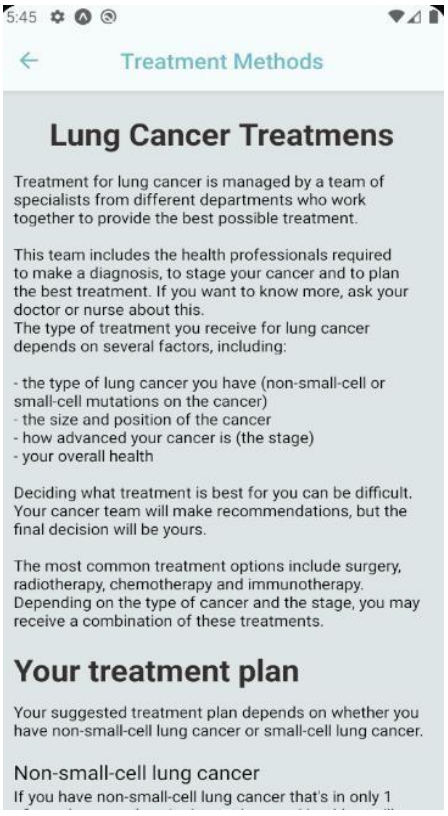
Resim1: *Mobil Uygulama Anasayfası*



Resim2: *Klinisyen Olmayan Girişi*

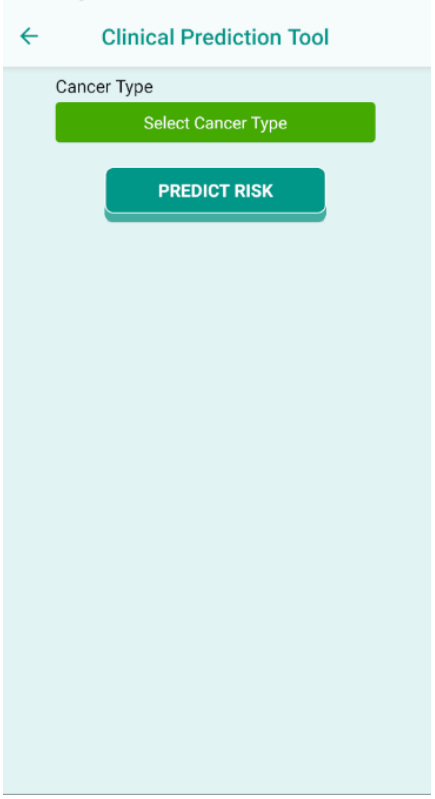


Resim3*: Akciğer Kanseri Hakkında Bilgi Sayfası*

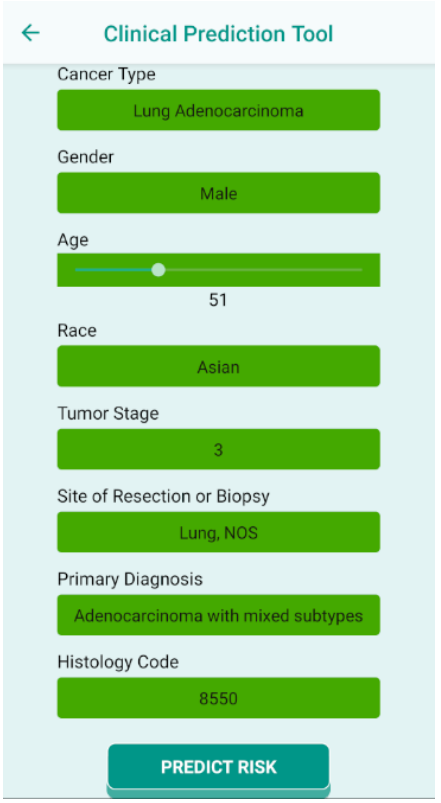


Resim4: *Akciğer Kanseri Tedavileri Sayfası*

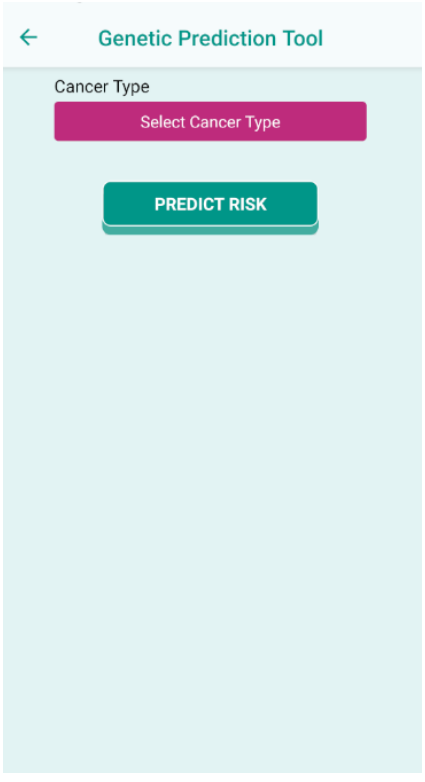
***Resim5**: Kanser Tahmin Model Seçenek Sayfası*



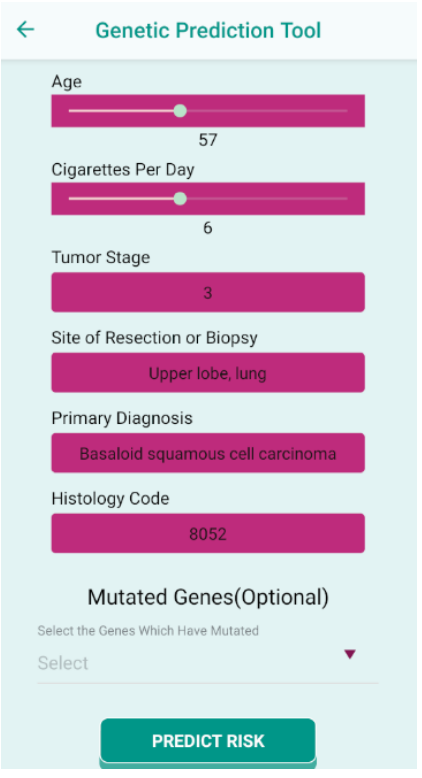
***Resim6**: Klinik Veri Tahmin Sayfası*



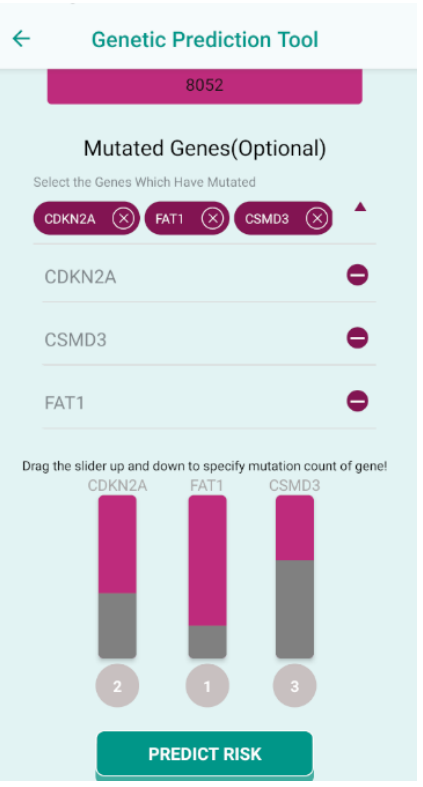
***Resim7**: Klinik Veri Tahmin Formu*



***Resim8**: Genetik Veri Tahmin Sayfası*



***Resim9**: Genetik Veri Tahmin Formu-1*



***Resim10**: Genetik Veri Tahmin Formu-2*

Resim11: Risk Sonuç Sayfası

3. Bulgular

Bu projede LUAD ve LUSC kanser hastalarının klinik ve genetik veriler ile eğitilen 5 farklı Makine Öğrenimi algoritması geliştirilmiş ve en iyi sonucu veren algoritmanın kullanıldığı bir mobil aplikasyon ortayaçıkartılmıştır.

Proje boyunca kullanılan Makine Öğrenmesi algoritmalarından Logistic Regression ve Random Forest algoritmaları hem LUAD hem de LUSC için en doğru tahminleri üreten algoritmalar olarak öne çıkmıştır. Bunun sonucunda bu algoritmaların geliştirilmesi ve doğruluklarının tutarlılıklarının test edilmesi için GrdiSearchCV ve K-Fold Cross validation yöntemleri uygulanmıştır. Tablo 3-4 ve 5'te de görüleceği üzere en başarılı algoritma LUAD için ortalama %90.6, LUSC için ortalama %84 doğruluk oranı ile Logistic Regression olmuştur. Böylelikle Logistic Regression iOS ve Android platformlarında kullanıcıya sunulmak üzere bir mobil uygulama içerisine IBM Watson Machine Learning sunucusu üzerinden entegre edilmiştir.

Bunun yanı sıra klinik verilere ek olarak en fazla mutasyona uğramış 10 geni de eklediğimizde, LUAD için hastanın riski (yüksek-düşük) ile bu on gen arasında bir korelasyon bulmaya çalıştığımızda, ilk 10 özellik listesinde **KEAP1, TP53, USH2A** ve **CSMD3** ortaya çıktı. Ayrıca '**KEAP1**' ile hastanın riski arasında bir katsayıilişkisi olduğu gözlemlendi. LUSC hastalarında bu 10 gen ve hastanın risk (yüksek ve düşük) durumu modele birlikte verildiğinde, **CSMD3** geni ile hastanın riski arasındaki korelasyon, LUSC'de hem Logistic Regression hem de Random Forest performansını iyileştirerek LUSC veriseti için modeli en çok etkileyen gen mutasyonu oldu.

4. Sonuç ve Tartışma

Çalışmamızın temel amacı, akciğer adenokarsinomu(LUAD) ve akciğer skuamöz karsinomu(LUSC) hastalarının risk sınıflandırmasının tahmininde en çok yardımcı olan klinik özellikleri veya biyobelirteç genleri araştırmak ve bunları kullanıcı dostu bir mobil uygulamaya entegre etmekti. Bu amaçla, TCGA akciğer adenokarsinomu ve akciğer skuamöz karsinomu hastalarının geniş klinik özellik setini ve somatik olarak en çok mutasyona uğramış ilk 10 gen setini kullanarak eğitilmiş birkaç makine öğrenimi çalışması kullandık ve risk sınıflandırmasına en çok katkıda bulunan özellikleri sıraladık. Genel olarak, analize gen mutasyonları eklendiğinde ve klinik verilerin gen mutasyonlarıyla birlikte kullanılması düşünüldüğünde bile klinik özellikler hala önem taşımaktadır. Bu analiz sonucunda, klinik karar süreçlerine rezeksiyon bölgesi(site_of_resection) gibi yeni klinikopatolojik özelliklere sahip LUAD için KEAP1 ve LUSC için CSMD3 gibi yeni genler eklenebilir. Modelimizin gelecekteki çalışması, bu geliştirilmiş makine öğrenimi modellerine, iki akciğer kanseri modülüne ek olarak üç yeni kanser tipi modülünü de ekleyerek uygulamanın daha büyük bir kitleye hitap etmesini sağlamaktır

5. Kaynaklar

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015 Mar 1;136(5):E359-86

[2] Akciğer Kanseri Yol Haritası – Türk Akciğer Kanseri Derneği

[3] LOBOCAN 2020: Cancer Today. Available online:https://gco.iarc.fr/today/home (accessed on 03 November 2021).

[4] De Vita VT, Lawrence TS, Rosenberg SA. De Vita, Hellman & Rosenberg's Cancer: Principles & Practice of Oncology. Philadelphia: Lippincott Williams & Wilkins; 2015.

[5] N. Bhargava, S. Sharma, R. Purohit and P. S. Rathore, "Prediction of recurrence cancer using J48 algorithm," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), 2017, pp. 386-390, doi: 10.1109/CESYS.2017.8321306.

[6] Zengin T, Önal-Süzek T. Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. BMC Bioinformatics. 2020 Sep 30;21(Suppl 14):368. doi: 10.1186/s12859-020-03691-3. PMID: 32998690; PMCID: PMC7526001.

[7] Zengin T, Önal-Süzek T. Comprehensive profiling of genomic and transcriptomic differences between risk groups of lung adenocarcinoma and lung squamous cell carcinoma, under review, Jan 2021, Journal of Personalized Medicine.

[8] Yu, KH., Zhang, C., Berry, G. et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 7, 12474 (2016).

[9] Yu, L., Tao, G., Zhu, L. et al. Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis. BMC Cancer 19, 464 (2019).

[10] Qiu, Y.L., Zheng, H., Devos, A. et al. A meta-learning approach for genomic survival analysis. Nat Commun 11, 6350 (2020).

[11] Leontina Postelnicu, Study seeks to identify key factors in adoption of mobile health solution for oncology

[12] F. Provost, T. Fawcett and R. Kohavi, "Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distribution," Proc. 3rd Int'l Conf. Knowledge Discovery and Data Mining, pp. 43-48, 1997

[13] F. Provost a n d T. Fawcett, " Robust Classification for Imprecise Environments," Machine Learning, vol. 42, pp. 203-231, 2001.

[14] Van den Eynden J, Fierro AC, Verbeke LP, Marchal K. SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinformatics. 2015 Apr 23;16:125. doi: 10.1186/s12859-015-0555-7.

[15] Zengin T, Önal-Süzek T. Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. J Pers Med. 2021 Feb 23;11(2):154. doi: 10.3390/jpm11020154. PMID: 33672117; PMCID: PMC7926392.

[16] Romero R, Sayin VI, Davidson SM, et al. Keap1 loss promotes Kras- driven lung cancer and results in dependence on glutaminolysis Nat Med. 2017;23(11):1362-1368. doi:10.1038/nm.4407

[17] Pengyuan Liu, Carl Morrison, Liang Wang, Donghai Xiong, Peter Vedell, Peng Cui, Xing Hua, Feng Ding, Yan Lu, Michael James, John D. Ebben, Haiming Xu, Alex A. Adjei, Karen Head, Jaime W. Andrae, Michael R. Tschannen, Howard Jacob, Jing Pan, Qi Zhang, Francoise Van den Bergh, Haijie Xiao, Ken C. Lo, Jigar Patel, Todd Richmond, Mary-Anne Watt, Thomas Albert, Rebecca Selzer, Marshall Anderson, Jiang Wang, Yian Wang, Sandra Starnes, Ping Yang, Ming You, Identification of somatic mutations in non-small cell lung carcinomas using whole- exome sequencing, Carcinogenesis, Volume 33, Issue 7, July 2012, Pages 1270–1276, https://doi.org/10.1093/carcin/bgs1

[18] https://investors.23andme.com/news-releases/news-release-details/23andme-receives-fda-clearance-direct-consumer-genetic-test

6. Ekler

EK-1: LUAD ve LUSC Hastalarının Klinik Özelliklerinin Dağılımını Gösteren Tablo

LUAD Category	Count	LUSC Category	Count
Age at diagnosis (median; range)	Median: 67 Range: 33-89	Age at diagnosis (median; range)	Median: 68 Range: 39-90
Gender		*Gender*	
Female	280	Female	131
Male	242	Male	373
Number of cigarattes per day (mean;range)	Mean: 2 Range: 0-9	Number of cigarattes per day(mean; range)	Mean: 3 Range: 0-13
Number of years smoked (mean; range)	Mean: 32 Rane: 2-64	Number of years smoked (mean; range)	Mean: 40 Range: 8-63
Tumor Stage		*Tumor Stage*	
I	279	I	245
II	124	II	163
III	85	III	85
IV	26	IV	7
NA	8	NA	4
Vital Status		*Vital Status*	
Alive	334	Alive	284
Dead	188	Dead	220
Ethnicity		*Ethnicity*	
Hispanic or Latino	7	Hispanic or Latino	8
Not Hispanic or Latino	389	Not Hispanic or Latino	319
NA	126	NA	177

EK-2: Klinik özellikler ve somatik olarak mutasyona uğramış ilk 10 gen ile LUAD modeli için her foldun doğruluğu

5-Fold Cross Validation for Clinical features and top 10 most mutated genes inLUAD									
Model	**Metrics**	Data Type	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	**mean**	**std**
Logistic Regression	F1-Score	Clinical/**Mutation**	0.91/**0.89**	0.93/**0.90**	0.89/**0.90**	0.91/**0.88**	0.88/**0.89**	0.90/**0.89**	0.017/**0.009**
	Precision	Clinical/**Mutation**	0.91/**0.90**	0.94/**0.91**	0.90/**0.90**	0.91/**0.88**	0.90/**0.89**	0.91/**0.90**	0.014/**0.011**
	Recall	Clinical/**Mutation**	0.91/**0.89**	0.94/**0.90**	0.89/**0.90**	0.91/**0.88**	0.89/**0.89**	0.91/**0.89**	0.018/**0.009**
	Accuracy	Clinical/**Mutation**	0.91/**0.89**	0.93/**0.90**	0.89/**0.90**	0.91/**0.88**	0.89/**0.89**	0.90/**0.89**	0.015/**0.009**
	AUC	Clinical/**Mutation**	0.95/**0.91**	0.95/**0.98**	0.94/**0.93**	0.95/**0.92**	0.89/**0.95**	0.93/**0.94**	0.023/**0.028**
Random Forest	F1-Score	Clinical/**Mutation**	0.86/**0.81**	0.86/**0.90**	0.87/**0.77**	0.87/**0.80**	0.73/**0.81**	0.84/**0.82**	0.054/**0.052**
	Precision	Clinical/**Mutation**	0.86/**0.85**	0.88/**0.91**	0.88/**0.85**	0.88/**0.82**	0.79/**0.87**	0.86/**0.86**	0.035/**0.031**
	Recall	Clinical/**Mutation**	0.86/**0.82**	0.86/**0.90**	0.87/**0.78**	0.87/**0.79**	0.74/**0.81**	0.84/**0.82**	0.050/**0.049**
	Accuracy	Clinical/**Mutation**	0.86/**0.81**	0.86/**0.91**	0.87/**0.78**	0.87/**0.79**	0.74/**0.81**	0.84/**0.82**	0.050/**0.049**
	AUC	Clinical/**Mutation**	0.94/**0.93**	0.91/**0.96**	0.95/**0.88**	0.91/**0.89**	0.82/**0.92**	0.90/**0.92**	0.046/**0.031**

EK-3: Klinik özellikler ve somatik olarak mutasyona uğramış ilk 10 gen ile LUSC modeli için her foldun doğruluğu

5-Fold Cross Validation for Clinical features and top 10 most mutated genes									
Model	**Metrics**	Data Type	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	**mean**	**std**
Logistic Regression	F1-Score	Clinical/**Mutation**	0.90/**0.89**	0.85/**0.80**	0.80/**0.87**	0.83/**0.82**	0.81/**0.85**	0.84/**0.85**	0.040/**0.036**
	Precision	Clinical/**Mutation**	0.91/**0.90**	0.86/**0.82**	0.84/**0.90**	0.86/**0.83**	0.81/**0.86**	0.86/**0.87**	0.038/**0.037**
	Recall	Clinical/**Mutation**	0.90/**0.89**	0.85/**0.80**	0.81/**0.87**	0.83/**0.82**	0.81/**0.85**	0.84/**0.84**	0.039/**0.036**
	Accuracy	Clinical/**Mutation**	0.90/**0.89**	0.85/**0.80**	0.81/**0.87**	0.83/**0.83**	0.81/**0.85**	0.84/**0.85**	0.039/**0.034**
	AUC	Clinical/**Mutation**	0.93/**0.93**	0.89/**0.88**	0.86/**0.93**	0.88/**0.90**	0.87/**0.91**	0.89/**0.91**	0.025/**0.021**
Random Forest	F1-Score	Clinical/**Mutation**	0.78/**0.85**	0.83/**0.75**	0.82/**0.76**	0.69/**0.81**	0.62/**0.73**	0.75/**0.78**	0.090/**0.048**
	Precision	Clinical/**Mutation**	0.79/**0.87**	0.84/**0.81**	0.83/**0.80**	0.71/**0.82**	0.67**/0.7 8**	0.77/**0.82**	0.074/**0.032**
	Recall	Clinical/**Mutation**	0.78/**0.85**	0.83/**0.76**	0.82/**0.77**	0.70/**0.81**	0.64/**0.74**	0.75/**0.79**	0.083/**0.044**
	Accuracy	Clinical/**Mutation**	0.78/**0.85**	0.83/**0.76**	0.82/**0.77**	0.70/**0.81**	0.64/**0.74**	0.75/**0.79**	0.083/**0.044**
	AUC	Clinical/**Mutation**	0.79/**0.91**	0.92/**0.86**	0.92/**0.82**	0.76/**0.88**	0.72/**0.84**	0.82/**0.86**	0.092/**0.037**