Yayın Geliş Tarihi (Submitted): **07/10/2023**

Yayın Kabul Tarihi (Accepted): **10/05/2024**

Makele Türü (Paper Type): **Araştırma Makalesi – Research Paper**

Please Cite As/Atıf için:
Uluskan, M. ve Şenli, H. D. (2024), YSA sınıflandırma modellerinde korelasyon-hipotez testi tabanlı filtreleme yoluyla girdi seçimi, *Nicel Bilimler Dergisi*, 6(1), 68-102. *doi: 10.51541/nicel.1372774*

YSA SINIFLANDIRMA MODELLERİNDE KORELASYON-HİPOTEZ TESTİ TABANLI FİLTRELEME YOLUYLA GİRDİ SEÇİMİ

Meryem Uluskan[1] ve Halil Derya Şenli[2]

ÖZ

Bu çalışmada başlıca amaç, yüksek miktardaki olası girdi değişken sayısını, bu değişkenler arasındaki korelasyonları göz önünde bulundurarak azaltarak sınıflandırma performansı yüksek Yapay Sinir Ağı (YSA) modelleri elde etmektir. Bunu gerçekleştirmek için 30 adet olası girdi değişkeni olan bir meme kanseri teşhis problemi ele alınmış ve önerilen korelasyon-hipotez testi tabanlı bir filtreleme yöntemi ile girdi değişken sayısı azaltılarak YSA modeli oluşturulmuştur. Önerilen modelin etkinliği farklı girdi değişken setlerini içeren altı YSA modeli ile karşılaştırılmıştır. Bu altı model, tüm girdi değişkenlerini içeren modelle, model tabanlı seçim yöntemlerinden aşamalı regresyon, ileri doğru seçim ve geriye doğru eleme yöntemleri ile seçilmiş girdi değişkenleriyle elde edilmiş olan modelleri kapsamaktadır. Modeller oluşturulurken veri seti farklı eğitim-test yüzdelerine bölünmüş ve gizli katmanda farklı nöron sayıları denenmiştir. Modellerin sınıflandırma performanslarını karşılaştırmak için doğruluk, duyarlılık, kesinlik ve F1-skoru ölçütleri kullanılmıştır. Sonuç olarak, önerilen korelasyon tabanlı filtreleme yöntemi ile seçilen dokuz girdi değişkenli modeller için doğruluk değeri 0,93-0,95 arasında bulunmuş olup bu değer belirgin şekilde iyidir. Duyarlılık değeri modelimiz için 0,85-0,88 aralığında ve yeterli düzeyde elde edilmiştir. Kesinlik değerinin önerilen modelimiz için 0,98-0,988 aralığında ve çok yüksek olduğu belirlenmiştir. Bu çalışmada önerilen modelin F1-skoru 0,907-0,931 arasında olup yeterince yüksek bir

[1]Sorumlu yazar, Doç. Dr., Eskişehir Osmangazi Üniversitesi, Eskişehir, Türkiye. ORCID ID: https://orcid.org/0000-0003-1287-8286
[2] Eskişehir Osmangazi Üniversitesi, Eskişehir, Türkiye. ORCID ID: https://orcid.org/0000-0001-6966-3388

değere sahiptir. Karşılaştırılan modeller içinde önerilen dokuz girdi değişkenli modelin değişken sayısının en düşük olduğu, yani en sade model olduğu ve gizli katmanda sadece 10 nöronla bile iyi bir sınıflandırma performansına sahip olduğu göz önüne alındığında bu yöntemin özellikle model tabanlı yöntemlere kıyasla kısa sürede ve düşük maliyetlerle anlaşılır sınıflandırma modelleri oluşturmada verimli olacağı belirlenmiştir.

Anahtar Kelimeler: Girdi değişken seçimi, Filtreleme yöntemi, Yapay Sinir Ağları, Sınıflandırma, Korelasyon, Hipotez testleri.

INPUT SELECTION THROUGH CORRELATION-HYPOTHESIS TESTING BASED FILTERING IN ANN CLASSIFICATION MODELS

ABSTRACT

The main goal of this study is to obtain high performing Artificial Neural Network (ANN) models for classification by reducing the large number of potential input variables using correlations between these variables. To achieve this, a breast cancer diagnosis problem with 30 potential input variables was considered and an ANN model was created by reducing the number of input variables with a proposed correlation-hypothesis test-based filtering method. The effectiveness of the proposed model was compared with six ANN models containing different sets of input variables. These six models include the model containing all input variables and the models obtained with input variables selected by stepwise regression, forward selection and backward elimination methods, which are model-based selection methods. While creating the models, the data set was divided into different training-test percentages and different numbers of neurons were tried in the hidden layer. Accuracy, recall, precision and F1-score metrics were used to compare the classification performances of the models. As a result, the accuracy value for the models with nine input variables selected by the proposed correlation-based filtering method was found to be between 0.93-0.95, which is significantly high. The recall value for our model was obtained between 0.85-0.88 and was sufficient. The precision value for our proposed model was determined to be very high, in the range of 0.98-0.988. The F1-score of the model proposed in this study is between 0.907-0.931, which is high enough. Considering that the proposed model has the lowest number of variables among the compared models, that is, it is the simplest model, and has a good classification performance even with only 10 neurons in the hidden layer, this model can be

used for rapid, lean and efficient classification at low costs, especially compared to model-based models.

Keywords: Input variable selection, Filtering method, Artificial Neural Networks, Classification, Correlation, Hypothesis tests.

1. GİRİŞ

Sınıflandırma, en sık karşılaşılan karar verme faaliyetlerinden biridir. Bir birim veya bireyin, o birimle veya bireyle ilgili bir dizi gözlenen özelliğe dayalı olarak önceden tanımlanmış bir grup veya sınıfa atanması gerektiğinde, bir sınıflandırma problemi ortaya çıkmaktadır. Mühendislik araştırmalarında, farklı endüstrilerde faaliyet gösteren işletmelerde ve tıptaki birçok problem, sınıflandırma problemi olarak ele alınabilir. Örnekler arasında kredi puanlama, kalite kontrol, tıbbi teşhis, kalp atışı, görüntü ve konuşma tanıma sayılabilir (Chuang ve Huang, 2011; Uluskan, 2020; Ramani vd., 2020; Acharya vd., 2017; Ciregan vd., 2012).

Geleneksel istatistiksel sınıflandırma yöntemleri değişkenler arasındaki ilişkilerin matematiksel denklemler biçiminde ifade edilmesine dayanır ve bir dizi varsayım üzerinde çalışır. Örnek olarak, doğrusal regresyon yöntemi, bağımsız ve bağımlı değişkenler arasında doğrusal bir ilişki olması, gözlemlerin birbirinden bağımsız olması ve hatanın normal dağılması gerektiğini varsayar (Uluskan, 2020). İstatistiksel modeller yalnızca temeldeki varsayımlar karşılandığında iyi çalıştığı için bu modellerin başarılı şekilde uygulanabilmesi için kullanıcıların hem veri özellikleri hem de model yetenekleri hakkında iyi bir bilgiye sahip olması gerekir (Zhang, 2000). Öte yandan, veriye dayalı modellemenin dayandığı makine öğrenimi, kural tabanlı programlamaya dayanmadan verilerden öğrenebilen bir algoritmadır. Genel olarak makine öğrenimi algoritmaları, bu varsayımların çoğundan bağımsızdır (Uluskan, 2020).

Özellikle son yıllarda makine öğrenimi teknikleri çok farklı alanlarda kullanılmakta ve karar vericilere daha doğru sonuçlara ulaşmalarında yardımcı olmaktadır. Benzer şekilde sınıflandırma problemleri, makine öğrenimi teknikleri sayesinde kısa sürede uygun modeller oluşturularak çözülebilir ve bu modeller süreç sonunda geleceğe yönelik tahminlerde kullanılabilir. Son yıllarda teknolojinin gelişmesiyle birlikte makine öğrenimi tekniklerinden yapay sinir ağları (YSA) sıklıkla sınıflandırma problemlerinde kullanılmaya başlanmıştır (Ryu vd., 2007).

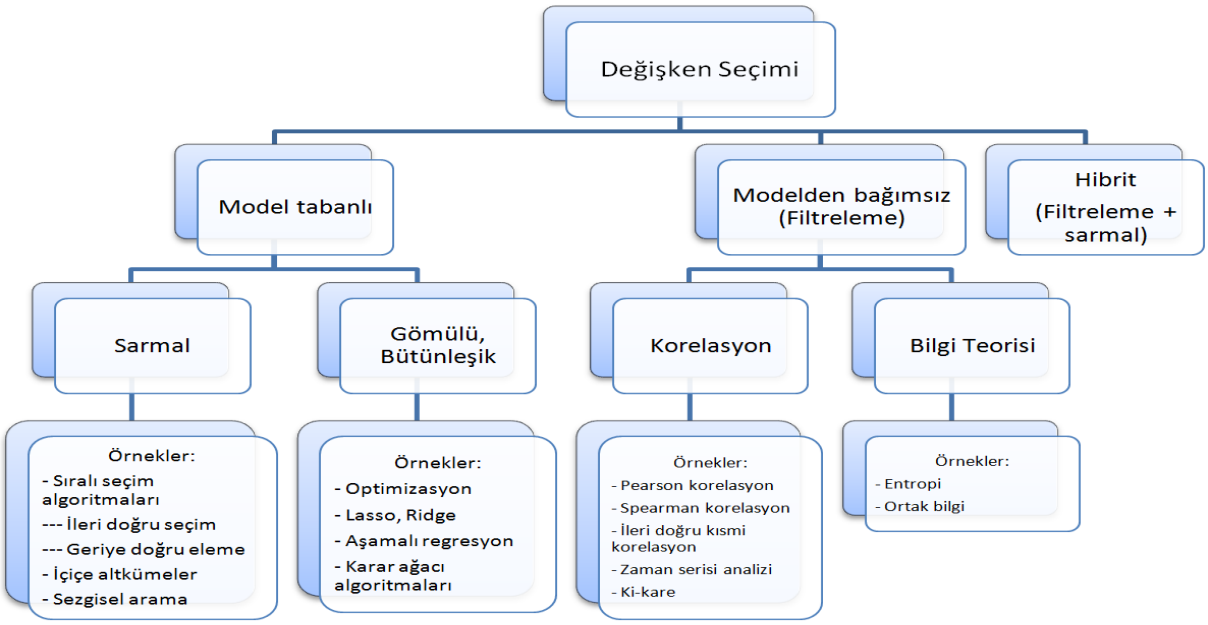
YSA, sınıflandırma için önemli bir araç olarak ortaya çıkmıştır. Son yıllarda gerçekleştirilen çalışmalar, YSA'nın çeşitli geleneksel sınıflandırma yöntemlerine etkili bir alternatif olduğunu ortaya koymuştur. Sinir ağları kendilerini veriye göre ayarlayabildikleri için kendi kendini uyarlayan yöntemlerdir. Ayrıca doğrusal olmayan karmaşık ilişkileri modellemede esnek yapıdadırlar. Son yıllarda YSA ile modelleme yapan kullanıcılar, girdi değişkenlerinin seçiminin gerekli olduğunun giderek daha fazla farkına varmaya başlamış ve bu amaç için farklı yöntemler kullanmışlardır. Girdi değişken seçimi yüksek sayıda potansiyel girdiye sahip modelleri sadeleştirerek kolay anlaşılabilir modeller elde etmek ve kaynakların etkin kullanımı açısından önemlidir. Literatürde girdi değişkenlerinin seçiminde model tabanlı yöntem uygulamalarındaki hesaplama süresi uzunluğu ve iş yükü fazlalığı gibi sebeplerden dolayı düşük maliyetli ve kısa sürede verimli sonuç alınabilecek yöntemlere gereksinim duyulmuştur. Bu durumlarda girdi değişken seçimi için genellikle filtreleme yöntemleri kullanılmıştır. Bu nedenlerle mevcut çalışmada olası girdi sayısı yüksek olan bir sınıflandırma problemi ele alınmış, bu problemde 30 adet olan olası girdi değişken sayısı önerilen korelasyon-hipotez testi tabanlı bir filtreleme yöntemi ile azaltılarak YSA modeli oluşturulmuştur. Önerilen modelin etkinliği, farklı girdi değişken setlerini içeren altı YSA modeli ile karşılaştırılmıştır. Bu altı model tüm girdi değişkenlerini içeren modelle, model tabanlı seçim yöntemlerinden aşamalı regresyonla, ileri doğru seçimle ve geriye doğru eleme yöntemleri ile seçilmiş girdi değişkenleriyle elde edilmiş olan modelleri kapsamaktadır. Performans karşılaştırmasında önerilen model için, tüm girdi değişkenlerini içeren modele göre ve model tabanlı seçim yöntemlerinden aşamalı regresyonla, ileri doğru seçimle ve geriye doğru eleme yöntemleri ile elde edilmiş olan modellere benzer oranlarda yüksek doğruluk, kesinlik, duyarlılık ve F1-skor değerleri elde edilmiştir.

2. BİLİMSEL YAZIN TARAMASI

2.1. Girdi Değişken / Öznitelik Seçimi ve Yöntemleri

Öznitelik olarak da adlandırılan girdi değişkenlerinin seçimi, oluşturulacak modellerin en uygun işlevsel formunun belirlenmesinde temel ve çok önemli bir husustur. Girdi değişkenlerinin seçimi, sınıflandırma, regresyon veya kümeleme gibi problemlerde model oluşturmak için en kullanışlı özelliklerin seçilmesi sürecidir (Solorio-Fernández vd., 2020). Seçim büyük ölçüde, model çıktısının uygun tahmincilerini belirlemek için mevcut veriler arasındaki ilişkilerin keşfedilmesine bağlıdır (Suzuki, 2011). Değişken seçiminde

boyutsallığın azaltılması için ilgisiz ve gereksiz değişkenler atılarak uygun değişkenler seçilmektedir (Remeseiro ve Bolon-Canedo, 2019). Dahası, girdi değişkenlerin seçimi yalnızca verilerin görselleştirilmesini ve anlaşılmasını kolaylaştıracak şekilde boyutsallığını azaltmakla kalmaz, aynı zamanda daha iyi genelleme yeteneğine sahip sade modeller elde edilmesini de sağlar (Solorio-Fernández vd., 2020). Böylece girdi boyutundaki azalma, öğrenme hızını ve model karmaşıklığını azaltarak veya genelleme yeteneğini ve sınıflandırma doğruluğunu artırarak performansı artırabilir. Uygun girdi değişkenlerinin seçimi aynı zamanda ölçüm maliyetini de azaltabilir ve problemin anlaşılmasını kolaylaştırabilir (Remeseiro ve Bolon-Canedo, 2019).



Şekil 1. Girdi değişken seçim yöntemleri

Son yıllarda YSA ile modelleme yapan kullanıcılar, girdi değişkenlerinin seçiminin gerekli olduğunun giderek daha fazla farkına varmaya başlamış ve bu amaç için farklı yöntemler kullanmışlardır. Literatürde değişken seçimi *model tabanlı* ve *modelden bağımsız* yöntemler (filtreleme yöntemleri) olmak üzere temelde ikiye ayrılmaktadır (May vd., 2011; Snieder vd., 2020). Model tabanlı yöntemler *sarmal* (wrapper) ve *gömülü* (embedded) olmak üzere iki gruba ayrılırken benzer şekilde modelden bağımsız filtreleme yöntemleri de *korelasyon* ve *bilgi teorisi* yöntemleri olmak üzere ikiye ayrılmaktadır. Son yıllarda filtreleme ve model tabanlı yöntemlerin birlikte kullanıldığı uygulamalar da *hibrit* yöntemler olarak isimlendirilmektedir (Jović vd., 2015; Solorio-Fernández vd., 2020). Literatür taraması sonunda özetlediğimiz değişken seçim teknikleri Şekil 1'de verilmiştir.

Model tabanlı sarmal ve gömülü yöntemlerin her ikisi de girdi değişken seçimini gerçekleştirmek için bir öğrenme yöntemi gerektirmektedir (Remeseiro ve Bolon-Canedo, 2019). *Sarmal yöntemler*, belirli bir kümeleme algoritmasının sonuçlarını kullanarak girdi değişkenlerden oluşan alt kümeleri değerlendirmektedir. Ancak, sarmal yöntemlerin yüksek hesaplama maliyetine sahip olmaları ve belirli bir kümeleme algoritmasıyla birlikte kullanımlarının sınırlı olması gibi dezavantajları vardır (Solorio-Fernández vd., 2020). Bu nedenle *gömülü yöntemler*, sarmal yöntemlerde oluşturulan farklı alt kümelerin yeniden sınıflandırılması için harcanan hesaplama süresini azaltmayı hedeflemektedir. Gömülü yöntemlerde temel yaklaşım, girdi değişken seçimini eğitim (training) sürecinin bir parçası olarak uygulamaktır (Chandrashekar ve Sahin, 2014). Gömülü yöntemler, girdi değişken seçimini modelleme algoritmasının uygulanması sırasında gerçekleştirmektedir. Dolayısıyla bu yöntemler algoritmanın içine entegre edilmekte, yani gömülmektedir (Jović vd., 2015). Sarmal ve gömülü yöntem uygulamalarındaki iş yükü fazlalığı ve hesaplama süresi uzunluğu gibi nedenlerden dolayı daha maliyet etkin ve daha kısa sürede verimli sonuç alınabilecek yöntemlere gereksinim duyulmuştur. Bu durumlarda girdi değişken seçimi için genellikle uzaklık, bilgi ve korelasyon gibi özellikler kullanılmıştır. Bu yöntemler filtreleme yöntemleri olarak adlandırılmıştır (Sezer ve Çakır, 2022).

Modelden bağımsız filtreleme yöntemleri önceden var olan bir modele dayanmamaktadır (Snieder vd., 2020). Filtreleme yöntemlerinde verinin kendisi aracılığıyla en ilgili girdi değişkenleri seçilmekte; yani girdi değişkenleri, ilgili değişkenlerin aranmasına rehberlik edebilecek herhangi bir kümeleme algoritması kullanılmadan, verinin kendine özgü özelliklerine göre değerlendirilmektedir (Solorio-Fernández vd., 2020). Bu şekilde filtreleme yöntemlerinin odak noktası verilerin genel özellikleri olduğundan bu yöntemler herhangi bir öğrenme yönteminden bağımsızdır. Tümevarım algoritmasından bağımsız olmaları nedeniyle hesaplama açısından maliyetli değildirler ve iyi bir genelleme kapasitesine sahiptirler (Remeseiro ve Bolon-Canedo, 2019). Ayrıca, filtreleme yöntemlerinin temel özellikleri hızları ve ölçeklenebilirlikleridir (Solorio-Fernández vd., 2020).

Son olarak, filtreleme ve sarmal yöntemlerin avantajlarını birleştirmek için *hibrit yöntemler* önerilmiştir. Bu yöntemler uygulanırken ilk önce, boyutsallığı azaltmak ve aday alt kümeler elde etmek için bir filtreleme yöntemi kullanılır. Daha sonra en iyi aday alt kümesini bulmak için bir sarmal yöntem kullanılır (Jović vd., 2015). Hibrit yöntemler, verimlilik (hesaplama çabası) ve etkililik arasında iyi bir denge sağlamaya çalışarak, her iki yaklaşımın niteliklerinden yararlanmaya çalışmaktadır (Solorio-Fernández vd., 2020). Bu şekilde hibrit

yöntemler hem sarmal yöntemlerin yüksek doğruluğuna, hem de filtreleme yöntemlerinin yüksek verimlilik özelliklerine sahip olur (Jović vd., 2015).

2.2. YSA Modellerinde Girdi Değişkenler Seçilirken Dikkat Edilecek Hususlar

Girdi değişkenlerini seçme faaliyetlerinde mevcut olan zorluklar şu şekilde özetlenebilir: (i) çok sayıda olası girdi değişken olması; (ii) olası girdi değişkenleri arasındaki korelasyonlar ve bu korelasyonlar sayesinde ortaya çıkan gereksiz girdi değişkenlerinin bulunması; ve (iii) tahmin gücü çok az olan veya hiç olmayan değişkenlerin bulunmasıdır (Suzuki, 2011).

YSA modelleri, açıklayıcı olmayan yetersiz girdi değişkenleri ile bilgi sağladığı halde diğer girdi değişkenleri ile yüksek korelasyonu olan gereksiz girdi değişkenlerinin dahil edilmesi nedeniyle gerekli olandan daha fazla girdi içerebilmektedir. YSA için en uygun girdi değişkenleri kümesini neyin oluşturduğunun tanımlanması, öncelikle girdi değişkenleri seçiminin model performansı üzerindeki etkisinin dikkate alınmasını gerektirir. Aşağıda bu temel hususlar özetlenmektedir (Suzuki, 2011):

İlgi düzeyi. YSA modellemelerinde sık görülen bir sorun, az sayıda değişkenin seçilmesi veya seçilen girdi değişkenlerinin yeteri kadar açıklayıcı olmamasıdır. Böyle bir durumda, çıktı seçilen girdi değişkenleri tarafından tam olarak açıklanamadığından, sonuçta performansı zayıf bir model elde edilmektedir.

Hesaplama zorluğu. Modele fazla sayıda girdi değişkeni eklemenin bir etkisi, YSA'nın hesaplama yükünü artırmasıdır. Bu durum, ağın eğitim hızının belirlenmesinde önemlidir.

Eğitim zorluğu. Gereksiz girdi değişkenlerinin modelde kullanılmasıyla bir YSA'nın eğitilmesi zorlaşabilmektedir. Böyle modellerde gereksiz parametreler ile hata arasındaki ilişkinin ortaya çıkarılması daha zordur. Gereksiz girdi değişkenleri modele gürültü ekleyerek öğrenme sürecini engellemektedir. Eğitim algoritması, çıktı değişkeni üzerinde hiçbir etkisi olmayan ağırlıkları ayarlamaya çalışarak gereksiz çalışabilir veya gürültü, önemli girdi-çıktı ilişkilerini maskeleyebilmektedir. Sonuç olarak, optimum hatanın belirlenebilmesi için eğitim algoritmasının çok kere yinelenmesi gerekebilir ve bu da modelin hesaplama yükünü artırmaktadır.

Sonuç olarak, istenilen bir girdi değişkeni, diğer girdi değişkenlerinden farklı (yani bağımsız) ve açıklayıcı olmalı, kısaca iyi bir tahminci olmalıdır. O halde optimal girdi değişkeni seti, açıklayıcı olmayan değişkenleri içermeyen, çıktı değişkeninin davranışını

tanımlamak için gereken en az girdi değişkenini içerecektir. Optimum girdi değişkenleri setinin tanımlanması, daha doğru, verimli, uygun maliyetli ve daha kolay yorumlanabilir YSA modelleriyle sonuçlanacaktır.

Bahsedilen zorluklar göz önüne alındığında mevcut çalışmamızda başlıca amaç, yüksek miktardaki olası girdi değişken sayısını, bu değişkenler arasındaki korelasyonlar göz önünde bulundurularak azaltılarak tahmin gücü yüksek YSA modelleri elde etmektir. Bunu yaparken de verimlilik ve harcanan zaman açısından avantajı olan ve fazla iş yükü gerektirmeyen *filtreleme yöntemlerinden korelasyon* kullanılarak girdi değişken seçimi yapılmıştır. Bu şekilde elde edilen sade modellerin etkinliği, tüm girdi değişkenlerini ve girdi değişkenlerin farklı alt kümelerini içeren YSA modellerinin performansları ile karşılaştırılmıştır. Modeller oluşturulurken veri seti farklı eğitim-test yüzdelerine bölünmüş ve gizli katmanda farklı nöron sayıları denenmiştir. Modellerin sınıflandırma performanslarını karşılaştırmak için *duyarlılık* (recall), *kesinlik* (precision) ve *F1-skoru* ölçütleri kullanılmıştır.

2.3. Meme Kanseri Teşhisinde YSA Uygulamaları

Tahmine dayalı modeller, teşhis ve prognostik faaliyetler için çeşitli tıbbi alanlarda kullanılmaktadır. Bu modeller, gerçek vakalardan elde edilen verileri oluşturan 'deneyim' üzerine kuruludur (Dreiseitl ve Ohno-Machado, 2002). Son yıllarda popülerliği artan YSA yöntemi birçok farklı alanda uygulanabilmektedir. Benzer şekilde, literatürde meme kanserini teşhis etmede de YSA uygulamaları görülmektedir.

Örnek olarak Chou (2004), YSA ve çok değişkenli uyarlanabilir regresyon eğrilerini (ÇDURE) bütünleştirerek meme kanseri teşhisi üzerinde çalışmıştır. ÇDURE yöntemiyle elde edilen değişkenler YSA için girdi olarak kullanılmıştır. Sonuç olarak hibrit sistemin geri yayılımlı yapay sinir ağ türünden, diskriminant analizinden ve sadece ÇDURE yöntemiyle elde edilen sonuçlardan daha iyi sonuç verdiği bulunmuştur.

Jeres-Aragonés vd. (2003) YSA ile meme kanserinin tekrar nüksetmesini araştırmışlardır. 85 adet girdi değişkeni medikal uzmanlar tarafından 14'e kadar indirilebilmiştir. Yani derlenen 85 girdi değişkeninden 14'ünün kanseri etkilediği düşünülmüştür. Ardından bir karar ağacı ile 14 girdi değişkeni farklı alt kümelere ayrılıp yapay sinir ağı eğitilmiştir. Bu çalışmayla karar ağacı ile desteklenen YSA'nın farklı bireylerde doğru tedaviyi uygulamaya yardımcı olması amaçlanmıştır.

Zhang vd. (2021) iki farklı gelişmiş YSA olan grafik evrişim ağı (graph convolutional network) ve evrişimsel sinir ağını (convolutional neural networks) birleştirerek yeni bir metot elde etmeye çalışmışlardır. Bu metotta amaç meme mamografilerindeki kanserli olan mamografileri belirlemektir. Oluşturulan yeni YSA 332 resimlik yapay sinir ağında denenmiştir. %96,20±2,90 seviyesinde duyarlılık ve %96,10±1,60 seviyesinde ise doğruluk elde edilmiştir. YSA modeli 5 farklı modelle karşılaştırılmıştır. En iyi sonucu verenin, çalışmada oluşturulan YSA olduğu tespit edilmiştir.

Haddadnia vd. (2012) meme kanseri teşhisi için hastaların termal görüntülerinden alınan nitel ve nicel verileri kullanmışlardır. Bu bilgiler işlendikten sonra teşhis için gerekli en iyi parametreler seçilip teşhisin doğruluğunu arttırmak adına YSA ve genetik algoritmayla oluşturulan model sayesinde iyileştirme sağlanmaya çalışılmıştır. 200 kişiden alınan termal resimlerden 8 adet teşhise yardımcı parametre elde edilmiştir. Bu 8 parametre modellerde girdi değişken olarak kullanılmıştır.

Alshanbari vd.. (2021) evrişimsel sinir ağları ile meme kanserini teşhis etmeye çalışmışlardır. Ağ, 11 katmandan oluşmaktadır. Herkese açık olan BreakHis veri seti çalışmada kullanılmıştır. Sonuç olarak, eğitilen modelle %96 doğruluk oranına ulaşılmıştır.

Diğer taraftan, Irmak vd. (2021) YSA'yı geleneksel makine yöntemleriyle karşılaştırmışlardır. Çalışmada lojistik regresyon, rastgele orman ve k-en yakın komşuluk algoritmaları, destek vektör makineleri, XGboost ve YSA karşılaştırılmıştır. YSA kullanılarak elde edilen %99,36 doğruluk oranının tüm yöntemler arasında en yüksek olduğu tespit edilmiştir.

Bılgıç (2021) evrişimsel sinir ağları ile meme kanserini ve deri kanserini teşhis etmeye çalışmıştır. Daha sonra lojistik regresyon yöntemi ile verileri analiz etmiş ve başarı grafikleri oluşturarak deri kanseri ve meme kanserini karşılaştırmıştır.

Literatürde görüldüğü gibi farklı tip YSA modellerinin teşhiste yardımcı olmada sınıflandırma için kullanıldığı veya geleneksel makine öğrenme yöntemleri ile karşılaştırıldığı çalışmalar yapılmıştır. Ancak literatürde, çok sayıda olası girdi değişkenin olduğu durumlarda teşhis için oluşturulacak modellerde, korelasyon tabanlı bir filtreleme yöntemi yoluyla değişken sayısının azaltılmasını öneren çalışmalara rastlanılmamıştır.

2.4. Girdi Değişken Seçimi Üzerine Yapılmış Çalışmalar

Girdi değişken seçimi üzerine çok farklı alanlarda ve farklı yöntemler kullanılarak çalışmalar gerçekleştirilmiştir. Örnek olarak, Emanet vd. (2021)'nin çalışmasında, literatürde önerilen farklı girdi değişken seçim yöntemleri ve makine öğrenimi teknikleri yardımıyla, girdi değişken seçiminin saldırı tespit sisteminin performansı üzerindeki etkisi incelenmiştir. Seçim için ki-kare testi, Spearman korelasyon katsayısı ve özyinelemeli değişken eleme yöntemleri uygulanmıştır. Elde edilen alt kümeler lojistik regresyon, karar ağacı, çok katmanlı algılayıcı, pasif-agresif ve gradyan artırma gibi makine öğrenimi yöntemleri ile sınıflandırılmış ve performans sonuçları karşılaştırılmıştır.

Sezer ve Çakır, (2022)'ın bankacılık sektöründe sınıflandırma amaçlı değişken alt kümesi seçimi için gerçekleştirdikleri çalışmalarında, sezgisel arama yöntemlerinden ileri/geri yönlü seçim için BestFirst, sıralı arama için RankSearch algoritmaları seçilmiş ve elde edilen yedi farklı değişken alt kümesi için ve tüm değişkenleri içeren veri kümesi için ayrı ayrı bir karar ağacı algoritması ile sınıflandırma yapılarak sekiz farklı karar ağacı oluşturmuştur.

Diğer taraftan Solorio-Fernández vd. (2020) denetimsiz özellik seçim yöntemlerinin bir literatür taramasını sundukları çalışmalarında, bu yöntemlerin temel özelliklerini tanımlayarak bir sınıflandırmasını sunmuş ve yöntemlerin avantaj ve dezavantajlarını belirtmişlerdir. Benzer şekilde Chandrashekar ve Sahin (2014) filtreleme, sarmal ve gömülü girdi değişken seçim teknikleri üzerine bir anket çalışması yapmışlardır.

Snieder vd. (2020) YSA ile nehir taşkınları tahminlemesi üzerine yaptıkları çalışmalarında kısmi korelasyon, kısmi karşılıklı bilgi olmak üzere iki modelden bağımsız girdi değişken seçim yöntemi ile çalışmada geliştirdikleri iki yeni model tabanlı yöntemi karşılaştırmışlardır. Benzer şekilde Fernando vd. (2009)'nin su kaynakları değişkenlerinin modellenmesinde YSA kullandıkları çalışmalarında, kısmi karşılıklı bilgi girdi değişken seçim algoritmasını hesaplama verimliliğini artıracak şekilde değiştirilmiştir. Snelder vd. (2007) çevresel faktörlerin çok değişkenli sınıflandırmaları alanında gerçekleştirdikleri çalışmalarında girdi değişken seçimi için çevresel ve biyolojik matrisler arasındaki korelasyonu maksimuma çıkarma amacına sahip ileri adımlı regresyona benzeyen bir prosedür kullanmışlardır.

Son olarak sağlık alanında Gao vd. (2018) lojistik regresyon, rastgele orman, destek vektör makineleri ve XGboost gibi model tabanlı ve modelden bağımsız yöntemleri Parkinson hastalığı teşhisinde ve klinik sonuçların sınıflandırılması alanında uygulamışlardır.

Sonuç olarak girdi değişken seçimi için literatür incelendiğinde farklı alanlarda model tabanlı ve modelden bağımsız yöntemlerin kullanıldığı ve karşılaştırıldığı çalışmaların bulunduğu görülmektedir. Ancak sağlık alanında çok fazla çalışma karşımıza çıkmamaktadır. Diğer taraftan meme kanseri tahmini ile ilgili özellikli literatür incelendiğinde ilgili alandan girdi değişken seçimi ile ilgili az sayıda çalışma yapıldığı görülmektedir. Bu noktada mevcut çalışmada literatürdeki bu boşluğu doldurmak adına sağlık alanında, özellikli olarak meme kanseri teşhisi konusunda, modelden bağımsız korelasyon tabanlı bir girdi değişken seçim yöntemi önerilmiş ve etkinliği model tabanlı yöntemler dahil farklı modellerle karşılaştırılmıştır.

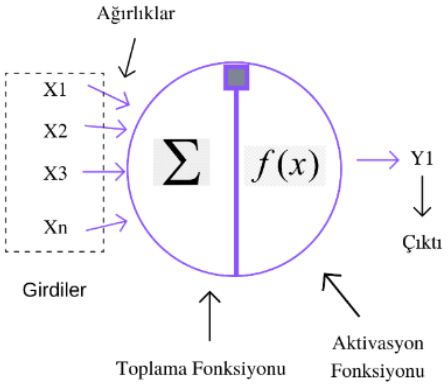
3. YÖNTEM

3.1. Yapay Sinir Ağları

Yapay sinir ağları (YSA) son yıllarda mühendislik, biyoloji, tıp, ekonomi gibi çeşitli alanlarda sınıflandırma, kümeleme, görüntü tanıma ve tahminleme gibi farklı uygulamalarda başarı ile kullanılan popüler teknolojilerden biri haline gelmiştir (Suzuki, 2011, Abiodun vd., 2018; Wu ve Feng, 2018). YSA'nın çok yönlülüğü, yüksek kapasitesinden ve öğrenme fonksiyonundan kaynaklanmaktadır (Suzuki, 2011). Bir makine öğrenimi türü olan YSA, kullanışlılık açısından geleneksel regresyon ve istatistiksel modellere iyi bir alternatiftir (Abiodun vd., 2018). YSA'nın temel özellikleri, karmaşık ve doğrusal olmayan girdi-çıktı ilişkilerini öğrenme, sıralı eğitim prosedürlerini kullanma ve kendilerini verilere uyarlama yeteneğine sahip olmalarıdır (Basu vd., 2010). Sadece sinir ağları olarak da adlandırılabilen YSA, beyindeki biyolojik sinir ağlarının yapısından ve işlevinden ilham alan matematiksel hesaplama modelleridir (Suzuki, 2011; Dongare vd., 2012).

Bir YSA, birbirine sinaptik ağırlıklar yoluyla bağlanan çok sayıda işlem birimlerinden, yani bir dizi yapay nörondan oluşur (Suzuki, 2011; Dongare vd., 2012). Bir sinir ağı genel olarak girdi katmanı, gizli katman ve çıktı katmanı olmak üzere üç katmana ayrılabilir. Dış ortamdan bilgi (veri) almaktan sorumlu girdi katmanı, gizli katmana verileri iletir. Gizli katman verilerin işlenmesinden sorumlu olan nöronlardan oluşmaktadır. Yine nöronlardan oluşan çıktı katmanı ise, son ağ çıktılarının üretilmesinden ve sunulmasından sorumludur (Uluskan, 2020). YSA'lar dışarıdan gelen bilgiyi toplama fonksiyonu ile toplar ve aktivasyon fonksiyonundan geçirerek çıktıyı oluşturur. Toplama fonksiyonu hücreye gelen net girdiyi hesaplayan fonksiyondur. Aktivasyon fonksiyonu ise gelen net girdiyi işleyerek hücrenin bu

girdiye karşı ne çıktı oluşturacağını belirler (Öztemel, 2003). Şekil 2'de basit bir YSA verilmiştir.



Şekil 2. Basit bir yapay sinir ağı hücresi

Geleneksel hesaplama yöntemlerinin aksine, YSA istenen girdi-çıktı ilişkilerini elde edebilmek için 'eğitilir'. Eğitim, yani öğrenme aşamasında, veri örnekleri ağa sunulur ve bir öğrenme algoritması kullanılarak parametreler ayarlanır (Liao ve Wen, 2007). Bu şekilde YSA, ağırlıkları ayarlayarak bir veri setini 'öğrenebilir' (Suzuki, 2011). Ele alınan probleme ilişkin mevcut bilgiye ve kullanıcının amacına göre, kullanılan öğrenme prosedürü 'denetimli', 'denetimsiz' veya her ikisi birden olabilir. Denetimli öğrenme prosedürü, bir çıktıyı gerektirir ve ağa sunulan girdi-çıktı çiftleri ile gerçekleştirilir (Liao ve Wen, 2007). Yani tüm örnekler için girdi ve o girdinin oluşturması gereken çıktılar ağa aktarılır. Ağ, verilen girdilere göre çıktıları oluşturabilmek adına ağırlıkları yeniler. Gerçek çıktılar ile ağın ürettiği çıktılar arasındaki hata hesaplanarak yeni ağırlıklar bu hata oranına göre düzenlenir (Alpaydin, 2020). Denetimsiz öğrenmede ise, ağ, çıktıya gerek duymaz ve bir maliyet fonksiyonuna dayalı olarak öğrenme gerçekleşir (Suzuki, 2011). Denetimsiz öğrenmede ağın eğitimine destek olan herhangi bir çıktı bulunmamaktadır. Sistem sadece girdilerle eğitilir ve buna göre bir çıktı oluşturmaya çalışır (Alpaydin, 2020). Yani denetimsiz öğrenmede ağ, örnek sınıfları arasında doğal olarak var olan ayrımı enbüyükleyebilmek için kendini aşamalı olarak ayarlamaya çalışır (Liao ve Wen, 2007).

YSA gibi öğrenen sistemlerde eğitim, yöntem fark etmeksizin bazı kurallar altında gerçekleştirilir. Bu kurallar çevrimiçi (online) ve çevrimdışı (offline) olarak çalışmaktadır. Çevrimiçi öğrenme kuralları gerçek zamanlı çalışabilmektedir. Bu kurallara göre eğitim gören sistemler fonksiyonların işlemlerini yaparlarken öğrenmeye devam etmektedirler. Çevrimdışı

öğrenme kuralını kullanan sistemler ise kullanılmadan önce eğitilirler. Eğitildikten sonra kullanılmaya başlanan sistemler uygulanırken öğrenme gerçekleştirmezler. Eğer sistemin öğrenmesi gerekli olan yeni bilgiler varsa sistem kullanımdan çıkarılır, çevrimdışı olarak tekrar eğitilir. Eğitimin sonunda sistem tekrar kullanıma açık hale gelmektedir (Goodfellow vd., 2016). Bu çalışmamızda kullanılan YSA'lar denetimli öğrenme ile eğitilmiş çevrimdışı ağlardır. Ayrıca çalışmada oluşturulan YSA'lar çok katmanlı, geri yayılımlı ve ileri beslemelidir.

3.2. Performans Değerlendirme Ölçütleri-Doğruluk, Kesinlik, Duyarlılık ve F1-Skoru

Bu çalışmada, önerilen korelasyon tabanlı filtreleme yöntemiyle girdi değişkenleri seçilmiş olan YSA modelinin etkinliği, tüm girdi değişkenlerini ve girdi değişkenlerin farklı alt kümelerini içeren altı farklı YSA modelinin performansları ile karşılaştırılmıştır. Modellerin sınıflandırma performanslarını karşılaştırmak için doğruluk (accuracy), duyarlılık (recall), kesinlik (precision) ve F1-skoru ölçütleri kullanılmıştır. Tablo 1'de görülen karmaşıklık matrisinin (karışıklık veya hata matrisi de denilmektedir) dört adet elemanı vardır. Bu elemanlar şunlardır:

Doğru Pozitifler (TP- True positive): "Pozitif" olarak doğru şekilde tahmin edilen örneklerin sayısıdır.

Yanlış Pozitifler (FP - False positive): Yanlışlıkla "pozitif" olarak tahmin edilen örneklerin sayısıdır.

Gerçek Negatifler (TN): "Negatif" olarak doğru şekilde tahmin edilen örneklerin sayısıdır.

Yanlış Negatifler (FN): Yanlışlıkla "negatif" olarak tahmin edilen örneklerin sayısıdır.

Tablo 1. Karmaşıklık Matrisi

		Gerçek	
		Pozitif	Negatif
Tahmin	Pozitif	**TP**	**FP**
	Negatif	**FN**	**TN**

Bu çalışmamızda "pozitif"ler "kötü huylu" hücreler olarak "negatif"ler de "iyi huylu" hücreler olarak tanımlanmıştır. O halde doğru sınıflandırılmış pozitif örnekler (TP), gerçekte kötü huylu hücrelerin kötü huylu olarak doğru tahmin edildiği örneklerdir. Doğru

sınıflandırılmış negatif örnekler (TN), gerçekte iyi huylu hücrelerin iyi huylu olarak tahmin edilen örneklerdir. Yanlış sınıflandırılmış pozitif örnek (FP), iyi huylu hücrelere kötü huylu denmesi anlamına gelmektedir. Son olarak, yanlış sınıflandırılmış negatif örnek (FN) kötü huylu olanlara iyi huylu denmesi anlamına gelmektedir.

Karmaşıklık matrisinin bileşenlerini kullanarak, sınıflandırma modellerini değerlendirmek için kullanılan doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skoru gibi çeşitli ölçütler tanımlanabilir.

En kolay anlaşılabilen ölçütlerden biri olan *doğruluk*, Denklem 1'de verildiği gibi toplam doğru tahmin sayısının toplam tahmin sayısına oranıdır. "Yaptığımız tüm tahminlerden ne kadarı doğru çıktı?" sorusuna yanıt vermektedir. Bu çalışmada doğruluk ölçütü doğru yüzde oranı olarak tanımlanmıştır.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Geri çağırma olarak da adlandırılan *duyarlılık* ölçütü gerçek pozitiflerin doğru şekilde belirlenmesinin ölçüsüdür. Yani modelimizin ilgili verileri ne kadar doğru tanımlayabildiğinin bir ölçüsünü vermektedir. Bu nedenle duyarlılık hassasiyet veya gerçek pozitif oran olarak da tanımlanmaktadır. "Doğru olarak tahmin edilmesi gereken tüm veri noktalarından ne kadarını doğru olarak tahmin ettik?" sorusuna yanıt vermektedir. Duyarlılık Denklem 2'de verildiği şekilde hesaplanmaktadır.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{2}$$

Kesinlik, Denklem 3'te verildiği gibi gerçek pozitiflerin tüm pozitiflere oranıdır. Kısaca gerçekte pozitif olanların tüm pozitif olarak tahminlenmişlere oranıdır. "Yaptığımız tüm pozitif tahminlerden ne kadarı doğru çıktı?" sorusuna yanıt vermektedir.

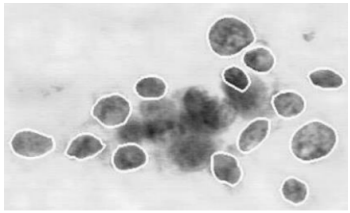
$$Kesinlik = \frac{TP}{TP + FP} \tag{3}$$

Kesinlik ve duyarlılık tanımlarından da görülebileceği gibi bu ölçütler sıkı bir şekilde bağlantılıdır. Kısaca iki ölçüt arasında bir ödünleşme vardır. Bu nedenle kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olan ve Denklem 4'te verilen F1-skoru bu iki ölçütü birleştiren bir ölçüdür. Yapılan analiz açısından kesinlik ve duyarlılık ölçütlerinin ikisi de önemli olduğunda F1-skoru belirlenmelidir.

$$F1\ skoru = \ 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \quad (4)$$

3.3. Veri Seti

Çalışmada University of California Irvine Machine Learning Repository açık erişim internet sitesinden elde edilen ve 569 hastadan derlenmiş olan 30 girdi değişkenli bir veri seti kullanılmıştır (UCI, 2022). Veriler elde edilirken memedeki kistik bölgeden küçük bir hücre topluluğu, iğne aspirasyon biyopsisi yöntemiyle alınmış ve Olympus mikroskobuyla incelenmiştir. Ardından bu görüntüler mikroskobun üzerine takılmış olan bir kamerayla dijital ortama aktarılmıştır. Daha sonra seçilebilen hücrelerin sınırları çizilmiştir. Çizilen sınırlara "snake" adı verilmiştir. Şekil 3'te bu çizilen sınırlar görülebilir.



Şekil 3. Dijitalize edilmiş seçilebilir hücrelerin sınırları, snake

Bilgisayar görü sistemi bu hücrelerin yarıçap (radius), çevre uzunluğu (perimeter), hücrenin alanı (area), yoğunluk (compactness), yumuşaklık (smoothness), içbükeylik (concavity), içbükey noktalar (concavepoints), simetri (symmetry), fraktal boyut (fractialdimension) ve doku (texture) olmak üzere 10 adet özelliğini elde edebilmektedir. Her hücre kümesi için ortalama, maksimum ve standart hata değerleri hesaplanmıştır. Bu şöyle açıklanabilir: Bir hastadan meme hücresi topluluğu alınmıştır. Ardından bu hücre topluluğunun yukarıda verilen 10 adet değişkeni incelenmiştir. 10 değişkenden ilgili değişken için ortalama alınarak ortalama değer (ort) bulunmuştur. Hücre topluluğu içindeki ilgili değişken için en büyük değer, o değişkenin maksimum (maks - en kötü) değeridir. En büyük değere en kötü denilmesinin sebebi ilgili faktörlerin büyüdükçe kanser olasılığının artmasından yani daha kötüye gitmesinden dolayıdır. Daha sonra ise ilgili değişkenin standart hatası (sh) hesaplanmıştır. Böylece toplam 30 adet girdi değişken verisi, 569 hastadan elde edilmiştir. Bu 30 girdi değişkeni mevcut çalışmamızda Tablo 2'deki şekilde isimlendirilmiştir.

Kısaca veri seti 30x569 büyüklüğünde bir matristir. Veri setinde 357 tane iyi huylu (sağlıklı), 212 tane kötü huylu (kanserli) hücre örneği yer almaktadır.

Tablo 2. Girdi değişkenleri

Özellikler	İlgili girdi değişkenleri
Ortalama girdi değişkenleri	X_1 =yarıçap_ort, X_2 =çevre_ort, X_3 =doku_ort, X_4 =alan_ort, X_5 =yoğunluk_ort, X_6 =yumuşaklık_ort, X_7 =içbükeylik_ort, X_8 =içbükey_nkt_ortalama, X_9 =fraktal_boyut_ort, X_{10} =simetri_ort
Standart hata girdi değişkenleri	X_{11} =yarıçap_sh, X_{12} =çevre_sh, X_{13} =doku_sh, X_{14} =alan_sh, X_{15} =yoğunluk_sh, X_{16} =yumuşaklık_sh, X_{17} =içbükeylik_sh, X_{18} =içbükey_nkt_sh, X_{19} =fraktal_boyut_sh, X_{20} =simetri_sh
Maksimum değer girdi değişkenleri	X_{21} =yarıçap_maks, X_{22} =çevre_maks, X_{23} =doku_maks, X_{24} =alan_maks, X_{25} =yoğunluk_maks, X_{26} =yumuşaklık_maks, X_{27} =içbükeylik_maks, X_{28} =içbükey_nkt_maks, X_{29} =fraktal_boyut_maks, X_{30} =simetri_maks

4. ANALİZLER ve BULGULAR

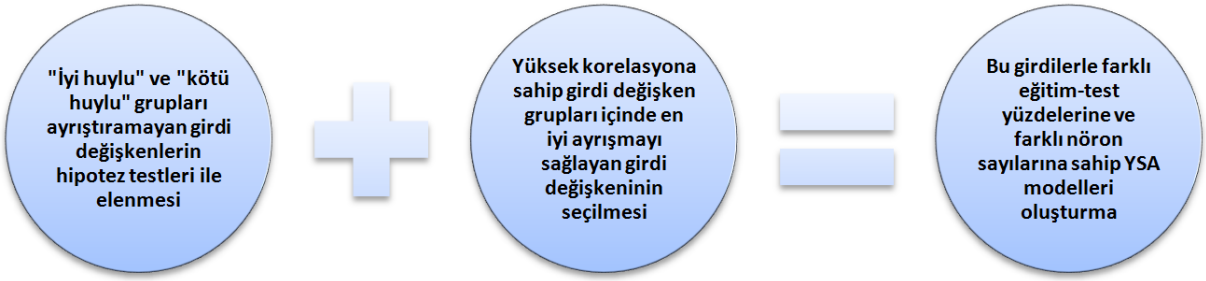
Bu çalışmada başlıca amaç, yüksek miktardaki (30 adet) olası girdi değişken sayısını, bu değişkenler arasındaki korelasyonları göz önünde bulunduran bir yöntem önerisiyle azaltarak maliyet ve zaman açılarından etkin, sınıflandırma performansı yüksek YSA modeli elde etmektir. Bu nedenle mevcut çalışmada korelasyon-hipotez testi tabanlı bir filtreleme yöntemi önerilmiş ve bu yöntemle girdi değişkenleri seçilerek YSA modeli oluşturulmuştur. Bu modelin performansı sarmal ve gömülü yöntemlerle seçilmiş girdi değişkenlerini de içeren altı farklı YSA modeli ile karşılaştırılmıştır. Karşılaştırılan YSA modelleri ile çalışmanın akışı Şekil 4'te verilmiştir.

Şekil 4. Çalışma akışı

İzleyen bölümlerde sırayla, önerilen yöntem ile değişken seçimi anlatılmış ve seçilen dokuz değişken verilmiştir. Sonrasında model tabanlı yöntemlerden aşamalı regresyon, ileri doğru seçim ve geriye doğru eleme yöntemleriyle seçilen değişkenler verilmiştir. Son olarak, önerilen model dahil yedi farklı YSA modelinin performans incelemesi gerçekleştirilmiş ve sonuçlar detaylı olarak tartışılmıştır.

4.1. Önerilen Korelasyon-Hipotez Testi Tabanlı Filtreleme Yöntemiyle Girdi Değişken Seçimi

Önerilen yöntemde ilk olarak, 30 girdi değişken arasından çıktıyı en iyi açıklayanları belirleyebilmek adına, yani sınıflandırmayı en iyi yapanları belirleyebilmek için "iyi huylu" ve "kötü huylu" sınıflarını istatistiksel olarak anlamlı şekilde ayıramayan girdi değişkenleri hipotez testleri yardımıyla elenmiştir. Sonrasında, anlamlı girdi değişkenleri arasındaki korelasyonlar incelenerek en iyi ayrışmayı yapanlar adım adım seçilmiştir. Önerilen yöntemle girdi değişken seçimi Şekil 5'te özetlenmiş ve detayları aşağıda açıklanmıştır.



Şekil 5. Önerilen yöntemle girdi değişken seçimi

4.1.1. "İyi Huylu" ve "Kötü Huylu" Gruplarını Ayrıştıramayan Girdi Değişkenlerin Hipotez Testleri İle Elenmesi

Öncelikli olarak 30 girdi değişkeni arasından, *iyi huylu* ve *kötü huylu* şeklinde etiketlenmiş olan çıktı değişkenini en iyi açıklayan değişkenleri belirleyebilmek için hipotez testleri gerçekleştirilmiştir. Burada iki ortalama farkı için z-testleri, her bir girdi değişken için kötü huylu ve iyi huylu grup ortalamaları arasında fark olup olmadığını belirleyecek şekilde gerçekleştirilmiştir. Hipotez testleri için normallik varsayımı test edilmiş ve normallik varsayımının $p - değeri > \alpha = 0{,}01$ değerleri ile sağlandığı görülmüştür. Bu testlerde kullanılan hipotezler aşağıdaki şekildedir:

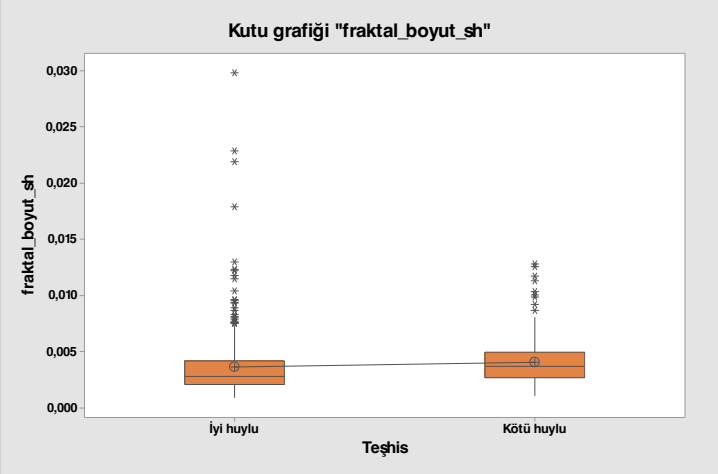
$H_0: \mu_{kötü\ huylu} = \mu_{iyi\ huylu}$

$H_a: \mu_{kötü\ huylu} \neq \mu_{iyi\ huylu}$

Hipotez testleri sonunda Tablo 3'te p-değerleri verilen girdi değişkenlerinin çıktıyı açıklayamadığı, yani kötü huylu ve iyi huylu hücre gruplarını ayrıştıramadığı belirlenmiş ve bu girdi değişkenler elenmiştir. Eleme işlemi yapılırken $\alpha = 0{,}01$ olarak alınmıştır. Burada *fraktal_boyut_ort*, *doku_sh*, *yumuşaklık_sh*, *simetri_sh* ve *fraktal_boyut_sh* değişkenleri için p-değerlerinin $\alpha = 0{,}01$ değerinden büyük olduğu yani sınıflandırma açısından istatistiksel olarak anlamlı olmadıkları belirlenmiştir. Bu girdi değişkenleri arasından *fraktal_boyut_ort* değişkeninin p-değeri küçük olduğundan Şekil 6'da verilen kutu grafiği oluşturularak bu değişkenin iyi bir ayrışma sağlamadığı teyit edilmiştir. Yani, bu değişken temelinde sınıflandırma yapıldığı zaman iyi huylu ve kötü huylu grupları için kutular büyük oranda çakışmaktadır. Bu nedenle *fraktal_boyut_ort* değişkeni dahil *doku_sh*, *yumuşaklık_sh*, *simetri_sh* ve *fraktal_boyut_sh* girdi değişkenleri elenmiştir.

Tablo 3. Anlamlı sınıflandırma yapamayan girdi değişkenleri

Girdi değişken	z-değeri	p-değeri
fraktal_boyut_ort	0,30	0,767
doku_sh	0,21	0,835
yumuşaklık_sh	1,62	0,105
simetri_sh	0,14	0,887
fraktal_boyut_sh	-2,04	0,042



Şekil 6. *fraktal_boyut_sh* değişkeni temelinde sınıflandırma için kutu grafikleri

4.1.2. Yüksek Korelasyona Sahip Girdi Değişken Grupları İçinde En İyi Ayrışmayı Sağlayan Girdi Değişkeninin Seçilmesi

Bu aşamada kalan girdi değişkenleri için Pearson korelasyon katsayıları hesaplanmıştır. Korelasyonları yüksek olan değişkenler için gruplar oluşturulmuş ve ilgili grup içinden en iyi ayrışmayı sağlayan girdi değişken yine hipotez testleri ile yardımıyla seçilmiştir. Birbiri ile korelasyonu yüksek olan girdi değişkenlerinin benzerlik göstermesi sebebiyle bu değişkenlerin bir arada kullanımı yapay sinir ağının eğitimine olumlu bir etki etmeyecektir. Ayrıca fazla sayıda girdi değişkenle yapılan eğitimler bazen ağın sınıflandırmasına olumsuz yansıyabilir (Arı ve Hanbay, 2019).

Korelasyon katsayısı 0,8'den büyük olan değişkenler arasında çok güçlü ilişki bulunmaktadır (LaMorte, 2021). Bu durum ise ciddi çoklu bağlantı (çoklu doğrusal bağlantı) problemi oluşturabilir (Calkins, 2005). Bu nedenle sırayla bir girdi değişkeni ile korelasyonları 0,8 ve üzeri olan girdi değişkenleri grup olarak alınmış, bu grup içinde iyi huylu/kötü huylu ayrışmasını en iyi yapan değişken seçilmiş ve diğerleri elenmiştir. Örnek olarak ilk önce *yarıçap-ort* değişkeni ile 0,8'den yüksek korelasyona sahip girdi değişkenleri belirlenmiştir (Tablo 4). Bu değişkenler *çevre_ort*, *alan_ort*, *yarıçap_maks*, *çevre_maks*, *alan_maks* ve *içbükey_noktalar_ort* değişkenleridir. Bu değişkenler için iyi huylu/kötü huylu grup ortalamaları farkını verecek şekilde z-testleri yapılmıştır. Sınıflandırma açısından anlamsız olan girdi değişkenleri ilk adımda elendiğinden geriye kalan bu anlamlı girdi değişkenleri $(p - değerleri = 0{,}000 < \alpha = 0{,}01)$ arasından z-değeri en büyük olan değişken seçilmiş ve gruptaki diğer değişkenler elenmiştir. Kısaca bu grupta *çevre_maks* girdi değişkeni $(z - değeri = -25{,}33)$ seçilmiştir.

Tablo 4. Grup 1 için korelasyon ve z-değerleri

Grup 1			
Girdi değişkeni	**yarıçap_ort**	**z-değeri**	**p-değeri**
yarıçap_ort	1	-22,21	0,000
çevre_ort	0,998	-22,94	0,000
alan_ort	0,987	-19,64	0,000
yarıçap_maks	0,97	-24,83	0,000
çevre_maks	*0,965*	*-25,33*	*0,000*
alan_maks	0,941	-20,57	0,000
içbükey_noktalar_ort	0,823	-24,84	0,000

Tablo 5. Girdi değişken grupları için korelasyon değerleri ile z-değerleri

Grup 2			
Girdi değişkeni	**doku_ort**	**z-değeri**	**p-değeri**
doku_ort	1	-11,02	0,000
doku_maks	*0,912*	*-12,26*	*0,000*
Grup 3			
Girdi değişkeni	**yoğunluk_ort**	**z-değeri**	**p-değeri**
yoğunluk_ort	1	-15,82	0,000
içbükeylik_ort	0,883	-20,33	0,000
yoğunluk_maks	0,866	-15,16	0,000
içbükeylik_maks	0,816	-19,60	0,000
içbükey_nkt_maks	*0,816*	*-29,12*	*0,000*
Grup 4			
	yumuşaklık_ort	z-değeri	p-değeri
yumuşaklık_ort	1	-9,30	0,000
yumuşaklık_maks	*0,805*	*-10,82*	*0,000*
Grup 5			
	yarıçap_sh	z-değeri	p-değeri
yarıçap_sh	*1*	*-13,30*	*0,000*
çevre_sh	0,973	-12,83	0,000
alan_sh	0,952	-12,16	0,000
Grup 6			
	yoğunluk_sh	z-değeri	p-değeri
yoğunluk_sh	*1*	*-7,08*	*0,000*
içbükeylik_sh	0,801	-6,92	0,000
Grup 7			
	çevre_maks	z-değeri	p-değeri
çevre_maks	1	-25,33	0,000
içbükey_nkt_maks	*0,816*	*-29,12*	*0,000*

Benzer şekilde tüm girdi değişkenleri arasındaki korelasyonlar incelenerek sırayla *doku-ort* değişkeni ile, *yoğunluk-ort* değişkeni ile, *yumuşaklık_ort* değişkeni ile, *yarıçap-sh* değişkeni ile, *yoğunluk_sh* değişkeni ile ve *çevre-maks* değişkeni ile 0,8 ve üzeri korelasyona sahip değişkenler belirlenerek gruplar oluşturulmuş ve her grup için en büyük z-değerine sahip değişken seçilmiş, diğer değişkenler elenmiştir. Tablo 5'te ilgili değişken grupları için korelasyon değerleri ile z-değerleri verilmiştir. Seçilmiş olan girdi değişkenleri Tablo 5'te *italik* olarak işaretlenmiştir.

Sonuç olarak önerilen yöntemle *simetri_ort*, *yarıçap_sh*, *yoğunluk_sh*, *içbükey_nkt_sh*, *doku_maks*, *yumuşaklık_maks*, *içbükey_nkt_maks*, *simetri_maks* ve *fraktal_boyut_maks* olmak üzere 9 adet girdi değişkeni seçilmiştir.

4.2. Model Tabanlı Yöntemlerle Değişken Seçimi

Aşamalı regresyon yöntemi ile değişken seçimi sonunda elde edilmiş olan 13 adet girdi değişkeni Tablo 6'daki ANOVA tablosunda ve ileri doğru seçim yöntemi ile elde edilmiş olan 17 adet girdi değişkeni Tablo 7'deki ANOVA tablosunda verilmiştir. Tablo 7'de bazı değişkenler için p-değerinin $\alpha = 0.05$ değerinden büyük olduğu görülmektedir. Bunun nedeni ileri doğru seçim yönteminde modele giren bir değişkenin tekrar çıkamaması kuralıdır.

Tablo 6. Aşamalı regresyon yöntemi ile değişken seçimi sonunda elde edilmiş olan 13 adet girdi değişkeni

Source	DF	SS	MS	F-Value	P-Value
Regression	13	102,461	7,88160	143,18	0,000
yoğunluk_ort	1	1,837	1,83650	33,36	0,000
içbükey_nkt_ort	1	0,739	0,73864	13,42	0,000
yarıçap_sh	1	0,823	0,82332	14,96	0,000
alan_sh	1	0,239	0,23853	4,33	0,038
yumuşaklık_sh	1	1,195	1,19496	21,71	0,000
içbükeylik_sh	1	0,840	0,83983	15,26	0,000
yarıçap_maks	1	2,813	2,81267	51,10	0,000
doku_maks	1	1,747	1,74723	31,74	0,000
alan_maks	1	1,546	1,54605	28,09	0,000

içbükeylik_maks	1	0,700	0,70012	12,72	0,000
içbükey_nkt_maks	1	0,260	0,26033	4,73	0,030
simetri_maks	1	0,687	0,68728	12,49	0,000
fraktal_boyut_mak	1	0,481	0,48082	8,73	0,003
Error	555	30,551	0,48082		
Total	568	133,012			

Tablo 7. İleri doğru seçim yöntemi ile değişken seçimi sonunda elde edilmiş olan 17 adet girdi değişkeni

Source	**DF**	**SS**	**MS**	**F-Value**	**P-Value**
Regression	17	102,802	6,04716	110,29	0,000
yarıçap_ort	1	0,090	0,08978	1,64	0,201
yoğunluk_ort	1	1,445	1,44546	26,36	0,000
içbükeylik_ort	1	0,142	0,14161	2,58	0,109
içbükey_nkt_ort	1	0,088	0,08784	1,60	0,206
yarıçap_sh	1	0,182	0,18248	3,33	0,069
alan_sh	1	0,040	0,04001	0,73	0,393
yumuşaklık_sh	1	1,050	1,05018	19,15	0,000
yoğunluk_sh	1	0,008	0,00846	0,15	0,695
içbükeylik_sh	1	0,672	0,67192	12,25	0,001
içbükey_nkt_sh	1	0,189	0,18882	3,44	0,064
yarıçap_maks	1	1,460	1,46007	26,63	0,000
doku_maks	1	1,578	1,57841	28,79	0,000
alan_maks	1	1,385	1,38464	25,25	0,000
içbükeylik_maks	1	0,236	0,23628	4,31	0,038
içbükey_nkt_maks	1	0,025	0,02543	0,46	0,496
simetri_maks	1	0,706	0,70575	12,87	0,000
fraktal_boyut_maks	1	0,527	0,52681	9,61	0,002
Error	551	30,211	0,05483		
Total	568	133,012			

Son olarak geriye doğru eleme yöntemi ile değişken seçimi sonunda elde edilmiş olan 13 adet girdi değişkeni Tablo 8'deki ANOVA tablosunda verilmiştir. Bu üç yöntemle elde edilen değişkenlerle YSA modelleri oluşturulup performansları önerilen korelasyon tabanlı yöntemin performansı ile karşılaştırılmıştır.

Tablo 8. Geriye doğru eleme yöntemi ile değişken seçimi sonunda elde edilmiş olan 13 adet girdi değişkeni

Source	DF	SS	MS	F-Value	P-Value
Regression	13	102,608	7,89292	144,08	0,000
yarıçap_ort	1	0,271	0,27104	4,95	0,027
yoğunluk_ort	1	2,033	2,03304	37,11	0,000
içbükey_nkt_ort	1	1,499	1,49871	27,36	0,000
yarıçap_sh	1	0,528	0,52800	9,64	0,002
yumuşaklık_sh	1	1,125	1,12547	20,54	0,000
içbükeylik_sh	1	0,906	0,90597	16,54	0,000
içbükey_nkt_sh	1	0,294	0,29398	5,37	0,021
yarıçap_maks	1	3,542	3,54226	64,66	0,000
doku_maks	1	1,688	1,68843	30,82	0,000
alan_maks	1	3,217	3,21682	58,72	0,000
içbükeylik_maks	1	1,128	1,12755	20,58	0,000
simetri_maks	1	0,724	0,72356	13,21	0,000
fraktal_boyut_maks	1	0,462	0,46180	8,43	0,004
Error	555	30,404	0,05478		
Total	568	133,012			

4.3. YSA Modellerinin Oluşturulup Performanslarının Karşılaştırılması ve Tartışma

İyi huylu/kötü huylu hücre sınıflandırmasında önerilen yöntemle elde edilmiş girdi değişkenleri kullanılarak oluşturulmuş YSA modelinin performansı altı farklı modelle karşılaştırılmıştır. Bu modellerden üç tanesi şu şekildedir: 1) tüm 30 girdi değişkeninin kullanıldığı model, 2) ortalama ve maksimum değerleri içeren 20 adet girdi değişkeninin kullanıldığı model, ve 3) sadece ortalama değerleri içeren 10 adet girdi değişkeninin kullanıldığı model. Diğer üç modeldeki girdi değişkenleri ise bu çalışmada önerilen korelasyon tabanlı filtreleme yönteminin *model tabanlı* yöntemlere göre performansını belirlemek için *sarmal* (wrapped) ve *gömülü* (embedded) yöntemler kullanılarak seçilmiştir.

Böylece son üç model sarmal yöntemlerden *ileri doğru seçim* (forward selection) ve *geriye doğru eleme* (backward elimination) yöntemleri ile, ve gömülü yöntemlerden *aşamalı regresyon* yöntemi ile seçilmiş girdi değişkenleri ile oluşturulmuştur. Bu son üç modelle elde edilmiş olan girdi değişkenleri Bölüm 4.2'de verilmiştir.

Tablo 9. Tüm modeller için doğru sınıflandırma yüzdeleri

Doğru Sınıflandırma Yüzdesi (Doğruluk)				
Girdi sayısı	**Veri seti**	**10 Nöron**	**15 Nöron**	**20 Nöron**
10	**60-20-20**	0,918	0,927	0,930
	70-15-15	0,922	0,930	0,923
	80-10-10	0,926	0,927	0,933
20	**60-20-20**	0,946	0,946	0,938
	70-15-15	0,945	0,948	0,941
	80-10-10	0,946	0,958	0,954
30	**60-20-20**	0,961	0,968	0,958
	70-15-15	0,970	0,963	0,966
	80-10-10	0,958	0,960	0,959
Önerilen yöntem (9 değişken)	**60-20-20**	0,950	0,939	0,936
	70-15-15	0,940	0,930	0,940
	80-10-10	0,940	0,947	0,940
Aşamalı regresyon (13 değişken)	**60-20-20**	0,957	0,962	0,965
	70-15-15	0,965	0,963	0,963
	80-10-10	0,970	0,968	0,963
İleri doğru seçim (17 değişken)	**60-20-20**	0,964	0,952	0,961
	70-15-15	0,966	0,963	0,962
	80-10-10	0,962	0,962	0,970
Geriye doğru eleme (13 değişken)	**60-20-20**	0,955	0,961	0,966
	70-15-15	0,957	0,957	0,962
	80-10-10	0,961	0,954	0,961

YSA modelleri oluşturulurken veri seti farklı eğitim-test yüzdelerine bölünmüş ve gizli katmanda farklı nöron sayıları denenmiştir. Kullanılan eğitim, geçerlilik ve test yüzdeleri 80-10-10, 70-15-15 ve 60-20-20 şeklindedir. Modellerde gizli katmandaki nöron sayısı olarak 10, 15 ve 20 nöron denenmiştir. Modeller 100'er defa çalıştırılmış ve ilgili performans ölçütü için ortalama değerler hesaplanmıştır. Modellerin sınıflandırma performanslarını karşılaştırmak için doğruluk veya doğru sınıflandırma oranı (accuracy), duyarlılık (recall), kesinlik (precision) ve F1-skoru ölçütleri kullanılmıştır. Tüm yedi model için doğru sınıflandırma oranları Tablo 9'da, duyarlılık (recall), kesinlik (precision) değerleri Tablo 10'da ve son olarak F1 skorları Tablo 11'da verilmiştir.

Tablo 9-11 incelendiği zaman genel olarak gizli katmandaki nöron sayısı arttıkça yani 10 nöron yerine 15 veya 20 nöron kullanıldığı zaman model performanslarının iyileştiği görülmektedir. Bu beklenen bir durumdur çünkü veri gizli katmanda daha fazla nöronla tahmin edilmeye çalışılmaktadır. Ancak nöron sayısını arttırmak modeli daha karmaşık bir yapıya büründürmektedir.

Diğer taraftan eğitim seti yüzdesini de arttırmak genel olarak model performanslarını iyileştirmektedir. Özellikle %60 eğitim verisi yerine daha büyük eğitim yüzdelerinde daha iyi sonuçların alındığı görülmektedir.

Doğru sınıflandırma yüzdeleri: Girdi değişkenleri açısından Tablo 9'da verilen doğru sınıflandırma yüzdeleri incelendiği zaman en yüksek sınıflandırma yüzdesinin 0,97 ile üç modele ait olduğu görülmektedir. Bu modellerden ilki 30 girdi değişkenli, veri seti 70-15-15 şeklinde bölünmüş ve gizli katmandaki nöron sayısı 10 olan model, diğerlerinin ise, aşamalı regresyonla elde edilmiş 13 değişkenli, veri seti 80-10-10 şeklinde bölünmüş ve gizli katmandaki nöron sayısı 10 olan model ile ileri doğru seçim yöntemi ile elde edilmiş 17 değişkenli, veri seti 80-10-10 şeklinde bölünmüş ve gizli katmandaki nöron sayısı 20 olan model olduğu görülmektedir. Burada 30 değişkenli model için sınıflandırma performansının yüksek olması girdi değişken sayısının fazla olmasına bağlanabilir. Çünkü modele girdi değişken eklendikçe modelin açıklanabilen oranı artmaktadır. Fakat önemli olan daha az girdi değişkeniyle iyi performanslar yakalayabilmektir. Diğer en iyi modelimiz ileri doğru seçim yöntemi ile elde edilmiş 17 girdi değişkenli, veri seti 80-10-10 şeklinde bölünmüş ve gizli katmandaki nöron sayısı 20 olan model demiştik. Bu modelin de performansının yüksek çıkması yine girdi değişken sayısı yüksekliği ile bağdaştırılabilir. Ayrıca bu modelde gizli

katmandaki nöron sayısı da yüksektir - 20 nöron. Bu nedenle bu modelde hem yüksek sayıda girdi değişken bulunmaktadır hem de model yapı olarak daha karmaşıktır.

En düşük doğru sınıflandırma yüzdesinin de 10 girdi değişkenli 60-20-20 şeklinde ayrılmış 10 nöronlu modele ait olduğu belirlenmiştir. Tablo 9'a bakıldığında genel olarak 10 girdi değişkenli YSA modelinin en düşük değerlere sahip olduğu görülmektedir. Böylece sadece ortalama değerleri içeren 10 adet girdi değişkeninin kullanıldığı modelin doğru yüzde performansının çok yüksek olmadığı söylenebilir. 30 girdi değişkenli, aşamalı regresyonla, ileri doğru seçimle ve geriye doğru eleme yöntemleriyle elde edilmiş modellerin doğru sınıflandırma performansları 0,95-0,97 arasındadır. Diğer taraftan bu çalışmada önerilen korelasyon tabanlı filtreleme yöntemi ile seçilen dokuz girdi değişkenli modelin performansı ise 0,93-0,95 arasında olup azımsanamayacak şekilde iyidir. Önerilen bu modelde girdi değişken sayısı (9 adet) karşılaştırılan diğer tüm modellere göre en düşüktür. Ayrıca önerilen model gizli katmanda sadece 10 nöronla 0,95 oranında doğru yüzde değeri elde edebilmiştir.

Duyarlılık, Kesinlik ve F1-Skor Değerleri: Duyarlılık ölçütü "Doğru olarak tahmin edilmesi gereken tüm veri noktalarından ne kadarını doğru olarak tahmin ettik?" sorusuna yanıt verir demiştik. Kanser gibi ciddi bir hastalıkta duyarlılık önemlidir. Hastanın gerçekten kanserli olduğu ancak modelin bunu kansersiz olarak sınıflandırdığı düşünülürse bu bir sorun oluşturur. Bu sebeple duyarlılık değerinin olabildiğince yüksek tutulması oluşturulacak modeller için önemlidir. Tablo 10'da görüldüğü gibi duyarlılık değerleri 30 değişkenli model, aşamalı regresyon, ileri doğru seçim ve geriye doğru eleme yöntemleri ile elde edilen değişkenlerin kullanıldığı modellerde 0,89 - 0,935 arasında değişmektedir. En iyi model için 0,935 (Aşamalı regresyon ile seçim, 80-10-10 veri seti ve 10 nöron) olduğu görülmüştür. Diğer taraftan bu çalışmada önerilen yöntemle elde edilmiş 9 girdi değişkenli model için duyarlık değerleri 0,845 ile 0,88 arasındadır. 13 girdi değişken kullanan en iyi modelin duyarlılık değerinin 0,935 olduğu düşünüldüğünde ve 30 girdi değişkeniyle eğitilen tam modelin bile duyarlılık değerlerinin 0,901-0,919 aralığında çıktığı göz önünde bulundurulduğunda, sadece 9 değişken ve gizli katmanda 10 nöronla eğitilmiş önerilen YSA modelinin duyarlılık performansının 0,88 olması iyi bir sonuçtur.

Tablo 10. Tüm modeller için duyarlılık ve kesinlik değerleri

Girdi sayısı	Veri seti	10 Nöron		15 Nöron		20 Nöron	
		Kesinlik	**Duyarlılık**	**Kesinlik**	**Duyarlılık**	**Kesinlik**	**Duyarlılık**
10	**60-20-20**	0,966	0,827	0,972	0,832	0,981	0,832
	70-15-15	0,977	0,820	0,973	0,842	0,969	0,825
	80-10-10	0,978	0,824	0,975	0,834	0,967	0,849
20	**60-20-20**	0,986	0,869	0,985	0,870	0,984	0,851
	70-15-15	0,984	0,866	0,971	0,880	0,978	0,862
	80-10-10	0,979	0,866	0,986	0,888	0,985	0,889
30	**60-20-20**	0,988	0,910	0,996	0,913	0,988	0,902
	70-15-15	0,997	0,919	0,992	0,908	0,981	0,927
	80-10-10	0,972	0,912	0,990	0,901	0,992	0,901
Önerilen yöntem (9 değişken)	**60-20-20**	0,988	0,880	0,981	0,858	0,983	0,845
	70-15-15	0,981	0,846	0,979	0,845	0,988	0,856
	80-10-10	0,980	0,863	0,989	0,867	0,982	0,868
Aşamalı regresyon (13 değişken)	**60-20-20**	0,988	0,899	0,989	0,912	0,984	0,923
	70-15-15	0,990	0,908	0,992	0,906	0,988	0,913
	80-10-10	0,986	0,935	0,988	0,925	0,984	0,911
İleri doğru seçim (17 değişken)	**60-20-20**	0,984	0,921	0,982	0,890	0,991	0,905
	70-15-15	0,988	0,915	0,991	0,906	0,991	0,907
	80-10-10	0,993	0,910	0,998	0,908	0,998	0,923
Geriye doğru eleme (13 değişken)	**60-20-20**	0,984	0,897	0,981	0,912	0,983	0,923
	70-15-15	0,988	0,897	0,967	0,913	0,989	0,907
	80-10-10	0,991	0,904	0,985	0,891	0,986	0,908

Benzer şekilde, kanser gibi hastalıklarda kötü huylu şeklinde teşhis konulan hastaların da gerçekten kanserli olması tedavi başarısı açısından önemlidir. Diğer türlü yanlış tedavi uygulamaları yapılabilmektedir. Bu durum göz önüne alındığında teşhis konulurken kesinlik değeri de önem arz etmektedir. Bu noktada yine Tablo 10'da görüldüğü gibi önerilen modelimizin kesinlik değerlerinin 0,98-0,988 ile çok yüksek olduğunu görebilmekteyiz. En yüksek kesinlik değerine sahip modelin ileri doğru seçim yöntemiyle belirlenmiş 17 girdi değişkenli model olduğu ve değerin 0,997 olduğu göz önünde bulundurulduğunda öenrilen modelimizin kesinlik açısından performansının çok yüksek olduğu sonucuna varılabilir.

Tablo 11. Tüm modeller için F1-skorları

F1 skorları				
Girdi değişken sayısı	**Veri seti**	**10 Nöron**	**15 Nöron**	**20 Nöron**
10	**60-20-20**	0,891	0,896	0,900
	70-15-15	0,892	0,903	0,891
	80-10-10	0,895	0,899	0,904
20	**60-20-20**	0,924	0,924	0,913
	70-15-15	0,921	0,923	0,916
	80-10-10	0,919	0,935	0,935
30	**60-20-20**	0,947	0,953	0,943
	70-15-15	0,956	0,948	0,953
	80-10-10	0,941	0,944	0,944
Önerilen yöntem (9 değişken)	**60-20-20**	0,931	0,915	0,909
	70-15-15	0,908	0,907	0,918
	80-10-10	0,922	0,918	0,908
Aşamalı regresyon (13 değişken)	**60-20-20**	0,942	0,949	0,953
	70-15-15	0,947	0,947	0,949
	80-10-10	0,960	0,956	0,946
İleri doğru seçim (17 değişken)	**60-20-20**	0,952	0,934	0,946
	70-15-15	0,950	0,947	0,947
	80-10-10	0,950	0,951	0,959
Geriye doğru eleme (13 değişken)	**60-20-20**	0,939	0,945	0,952
	70-15-15	0,941	0,939	0,946
	80-10-10	0,946	0,936	0,945

Kesinlik ve duyarlılık ölçütlerinin arasında bir ödünleşme olduğu için ve bu iki ölçüt de kanser teşhisinde önemli olduğundan bu iki ölçütü birleştiren F1-skoru aslında bakılması gereken en önemli ölçüdür. Eğitilen farklı modeller için F1-skorları Tablo 11'de verilmiştir. F1-skorları incelendiğinde en yüksek skora (0,96) sahip modelin aşamalı regresyonla seçilen 13 girdi değişkeniyle eğitilmiş model olduğu görülmektedir. Bu çalışmada önerilen yöntemin F1-skoru 0,931 olup yeterince yüksek bir değere sahiptir. 30 girdi değişkenli model, aşamalı regresyon, ileri doğru seçim ve geriye doğru eleme yöntemleriyle seçilmiş girdi değişkenleriyle eğitilmiş modellerde F1-skor aralığı 0,939 - 0,96 şeklindedir. 10 girdi değişkenli model için F1-skor değerlerinin 0,891 ile 0,904 arasında olduğu görülmektedir. Bu durumlar göz önüne alındığında sadece 9 girdi değişkenli modelimizin F1-skoru yönünden iyi bir performans gösterdiği sonucuna ulaşılabilir.

5. SONUÇLAR

Sınıflandırma, uygulamada sıklıkla karşılaşılan karar verme faaliyetlerinden biridir. Son yıllarda YSA modelleri sınıflandırma problemlerinde kullanılmaya başlanmıştır. Diğer taraftan girdi değişkenlerinin seçimi, oluşturulacak modellerin en uygun formunun belirlenmesinde çok önemli bir husustur. Bu nedenle, mevcut çalışmada başlıca amaç, yüksek miktardaki olası girdi değişken sayısını, bu değişkenler arasındaki korelasyonları göz önünde bulundurarak azaltarak sınıflandırma performansı yüksek YSA modelleri elde etmektir.

Kanser gibi hayati önemi olan hastalıkların teşhis aşamasında kullanılacak sınıflandırma modeli için performans ölçütü olarak "Yaptığımız tüm tahminlerden ne kadarı doğru çıktı" sorusuna cevap veren *doğruluk değeri* önemlidir. Bu noktada, önerilen korelasyon tabanlı filtreleme yöntemi ile seçilen dokuz girdi değişkenli modelimiz için doğruluk değeri 0,93-0,95 arasında olup belirgin şekilde iyidir. Diğer taraftan aşamalı regresyonla, ileri doğru seçimle ve geriye doğru eleme yöntemleriyle girdi değişkenleri seçilmiş modeller ile tüm girdi değişkenlerini içeren modelin doğru sınıflandırma performansları 0,95-0,97 arasındadır. Ortalama değerleri içeren 10 adet girdi değişkeninin kullanıldığı model için doğruluk değerleri 0,918-0,933 arasında olup önerilen modelimizin gerisinde performans sergilemiştir. Önerilen modelimizde girdi değişken sayısının (dokuz adet) karşılaştırılan diğer tüm modellere göre en düşük olduğu ve gizli katmanda sadece 10

nöronla 0,95 oranında doğru yüzde değeri elde edebildiği göz önünde bulundurulduğunda önerilen modelin doğruluk performansının yüksek olduğu söylenebilir.

Teşhis aşamasında kullanılacak sınıflandırma modelinin, doğruluk değerinin yanında *kesinlik* ve *duyarlılık değerleri* de önemlidir. Hastanın gerçekten kanserli olduğu ancak modelin bunu kansersiz olarak sınıflandırdığı düşünülürse bu bir sorun oluşturur. Bu sebeple duyarlılık değerinin olabildiğince yüksek tutulması oluşturulacak modeller için önemlidir. Benzer şekilde kanser gibi hastalıklarda kötü huylu şeklinde teşhis konulan hastaların da gerçekten kanserli olması tedavi başarısı açısından önemlidir. Diğer türlü yanlış tedavi uygulamaları yapılabilmektedir. Bu durum göz önüne alındığında teşhis konulurken kesinlik değeri de önem arz etmektedir. "Yaptığımız tüm kötü huylu şeklindeki tahminlerden ne kadarı doğru çıktı?" sorusuna yanıt veren kesinlik değerinin önerilen modelimiz için 0,98-0,988 aralığında ve çok yüksek olduğu belirlenmiştir. Benzer şekilde "Doğru olarak tahmin edilmesi gereken tüm veri noktalarından ne kadarını doğru olarak tahmin ettik?" sorusuna yanıt veren duyarlılık değeri modelimiz için 0,88 bulunmuştur. Karşılaştırılan modeller için elde edilen performanslar incelendiğinde en yüksek 0,935 değerinin elde edilebildiği ve bu değerin 13 girdi değişken kullanan modele ait olduğu düşünüldüğünde ve 30 girdi değişkeniyle eğitilen modelin bile duyarlılık değerlerinin 0,901-0,919 aralığında çıktığı göz önünde bulundurulduğunda, sadece 9 değişken ve gizli katmanda 10 nöronla eğitilmiş önerilen YSA modelimizin duyarlılık performansının 0,88 olması iyi bir sonuçtur.

Son olarak kesinlik ve duyarlılık ölçütleri arasında bir ödünleşme olduğu için, yani genel olarak biri artarken diğeri azaldığı için, bu iki ölçütün de önemli olduğu kanser teşhisi gibi durumlarda iki ölçütü birleştirerek ele alan *F1-skoru* belirlenmelidir. Bu çalışmada önerilen yöntemin F1-skoru 0,907-0,931 olup yeterince yüksek bir değere sahiptir. 30 girdi değişkenli model, aşamalı regresyon, ileri doğru seçim ve geriye doğru eleme yöntemleriyle seçilmiş girdi değişkenleriyle eğitilmiş modellerde F1-skor aralığı 0,939 - 0,96 şeklindedir. 10 girdi değişkenli model için F1-skor değerlerinin 0,891 ile 0,904 arasında olduğu görülmektedir. Bu durumlar göz önüne alındığında sadece 9 girdi değişkenli modelimizin F1-skoru yönünden iyi bir performans gösterdiği sonucuna ulaşılabilir.

Sonuç olarak bu çalışmada ele alınan sınıflandırma problemi için çok sayıdaki olası girdi değişken sayısı düşük maliyetli, hızlı ve kolay anlaşılır bir yöntemle azaltılarak yalın ve etkin bir YSA modeli elde edilmiştir. Literatürde girdi değişkenlerinin seçiminde *model tabanlı* yöntem uygulamalarındaki hesaplama süresi uzunluğu ve iş yükü fazlalığı gibi sebeplerden dolayı daha maliyet etkin ve daha kısa sürede verimli sonuç alınabilecek

yöntemlere gereksinim duyulmuştur. Bu durumlarda girdi değişken seçimi için genellikle filtreleme yöntemleri kullanılmıştır. Benzer şekilde bu çalışmada da korelasyon-hipotez testi tabanlı bir filtreleme yöntemi önerilmiş ve etkinliği farklı girdi değişken setlerini içeren YSA modelleri ile karşılaştırılmıştır. Modeller için performans karşılaştırmasında önerilen model için tüm girdi değişkenleri içeren modele göre ve model tabanlı seçim yöntemlerinden aşamalı regresyonla, ileri doğru seçimle ve geriye doğru eleme yöntemleri ile elde edilmiş olan modellerle benzer oranlarda ve yüksek doğruluk, kesinlik, duyarlılık ve F1-skor değerleri elde edilmiştir. Karşılaştırılan modeller içinde önerilen dokuz-değişkenli modelin değişken sayısının en düşük olduğu, yani en sade model olduğu ve gizli katmanda sadece 10 nöronla bile iyi bir sınıflandırma performansına sahip olduğu göz önüne alındığında bu modelin düşük maliyetle ve hızlı bir şekilde anlaşılır sınıflandırma modelleri elde etmede verimli olacağı belirlenmiştir.

Gelecek çalışmalarda, bu çalışmada önerilen yöntem, farklı alanlardaki sınıflandırma problemleri için farklı veri setleri kullanılarak uygulanabilir ve etkinliği farklı filtreleme veya model tabanlı yöntemlerle karşılaştırılabilir. Bu çalışmada hipotez testleri ile sadece ortalama değerler göz önünde bulundurarak ayrışma durumu incelendiği için gelecek çalışmalarda korelasyon tabanlı farklı filtreleme yöntemleri önerilerek bu yöntemler model tabanlı yöntemlerle karşılaştırılabilir. Son olarak, sınıflandırma için bu çalışmada kullanılan modellere benzer YSA kombinasyonları en iyi parametreler seçilecek şekilde test edilebilir.

ETİK BEYAN

"YSA Sınıflandırma Modellerinde Korelasyon-Hipotez Testi Tabanlı Filtreleme Yoluyla Girdi Seçimi" başlıklı çalışmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş, toplanan veriler üzerinde herhangi bir tahrifat yapılmamış ve bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

KAYNAKÇA

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A. and Arshad, H. (2018), State-of-the-art in artificial neural network applications: A survey, *Heliyon*, *4*(11).

Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adam, M., Gertych, A. and San Tan, R. (2017), A deep convolutional neural network model to classify heartbeats, *Computers in Biology and Medicine*, 89, 389-396.

Alpaydin, E. (2020), *Introduction to Machine Learning*, MIT Press, Cambridge, Massachusetts, ABD.

Alshanbari, E., Alamri, H., Alzahrani, W. and Alghamdi, M. (2021), Breast cancer classification using convolutional neural network, *International Journal of Computer Science and Network Security*, 21(6), 101-106.

Arı, A. and Hanbay, D. (2019), Tumor detection in MR ımages of regional convolutional neural networks, *Journal of the Faculty of Engineering and Architecture of Gazi University*, 34(3), 1395-1408.

Basu, J.K., Bhattacharyya, D. and Kim, T.H. (2010), Use of artificial neural network in pattern recognition, *International Journal of Software Engineering and Its Applications*, 4, 23–34.

Bılgıç, B. (2021), Comparison of breast cancer and skin cancer diagnoses using deep learning method, *29, Signal Processing and Communications Applications Conference (SIU)*, 9-11 Haziran, Istanbul, Türkiye.

Calkins, K.G. (2005), *Correlation Coefficients*, Erişim adresi: https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm

Chandrashekar, G. ve Sahin, F. (2014), A survey on feature selection methods. *Computers & electrical engineering*, *40*(1), 16-28.

Ciregan, D., Meier, U. and Schmidhuber, J. (2012), Multi-column deep neural networks for image classification, *2012 IEEE Conference on computer vision and pattern recognition,* 16-21 Haziran, Providence, RI, ABD, 3642-3649.

Chuang, C. L. and Huang, S. T. (2011), A hybrid neural network approach for credit scoring, *Expert Systems*, 28(2), 185-196.

Chou, S.M., Lee, T.S., Shao, Y.E. and Chen, I.F. (2004), Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regressions lines, *Expert Systems with Applications*, 27(1), 133-142.

Dongare, A.D., Kharde, R.R. and Kachare, A.D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review, *Journal of Biomedical Informatics*, 35(5-6), 352-359.

Emanet, S., Baydoğmuş, G.K. and Demir, Ö. (2021), Öznitelik seçme yöntemlerinin makine öğrenmesi tabanlı saldırı tespit sistemi performansına etkileri, *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, *12*(5), 743-755.

Fernando, T.M.K.G., Maier, H.R. and Dandy, G.C. (2009), Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach, *Journal of Hydrology*, *367*(3-4), 165-176.

Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N.I., Müller, M.L., ... and Dinov, I.D. (2018), Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. *Scientific reports*, *8*(1), 7129.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deeplearning,* MIT Press, Cambridge, Massachusetts, ABD.

Haddadnia, J., Hashemian, M. and Hassanpour, K. (2012), Diagnosis of breast cancer using a combination of genetic algorithm and artificial neural network in medical infrared thermal imaging, *Iranian Journal of Medical Physics*, 9(4), 265-274.

Irmak, M.C., Taş, M.B.H., Turan, S. and Haşiloğlu, A. (2021), Comparative breast cancer detection with artificial neural networks and machine learning methods, *29. Signal Processing and Communications Applications Conference (SIU)*. 9-11 Haziran, Istanbul, Türkiye, 1-4

Jerez-Aragonés, J.M., Gómez-Ruiz, J.A., Ramos-Jiménez, G., Muñoz-Pérez, J. and Alba-Conejo, E. (2003), A combineneural network and decision trees model for prognosis of breast cancer relapse, *Artificial Intelligence in Medicine*, 27(1), 45-63.

LaMorte, W.W. (2021), *Correlation and Regression*, Boston University School of Public Health, Erişim adresi: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-

QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html.

Liao, S.H. and Wen, C.H. (2007), Artificial neural networks classification and clustering of methodologies and applications–literature analysis from 1995 to 2005, *Expert Systems with applications*, *32*(1), 1-11.

May, R., Dandy, G. and Maier, H. (2011), Review of input variable selection methods for artificial neural networks, *Artificial neural networks-methodological advances and biomedical applications*, *10*(1), 19-45.

Öztemel, E. (2003), *Yapay Sinir Ağları*, Papatya Yayıncılık, İstanbul, Türkiye.

Ramani, R., Devi, K.V. and Soundar, K.R. (2020), MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction, *Soft Computing*, 24(21), 16335-16345.

Remeseiro, B. and Bolon-Canedo, V. (2019), A review of feature selection methods in medical applications. *Computers in biology and medicine*, *112*, 103375.

Ryu, Y.U., Chandrasekaran, R. and Jacob, V.S. (2007), Breast cancer prediction using the isotonic separation technique, *European Journal of Operational Research*, 181(2), 842-854.

Sezer, E. and Çakır, Ö. (2022), Sınıflandırma amaçlı değişken alt kümesi seçimi: bir bankacılık uygulaması, *Dicle Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *12*(24), 480-498.

Snelder, T.H., Dey, K.L. and Leathwick, J.R. (2007), A procedure for making optimal selection of input variables for multivariate environmental classifications. *Conservation Biology*, *21*(2), 365-375.

Snieder, E., Shakir, R. and Khan, U.T. (2020), A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. *Journal of Hydrology*, *583*, 124299.

Solorio-Fernández, S., Carrasco-Ochoa, J.A. and Martínez-Trinidad, J.F. (2020), A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*(2), 907-948.

Suzuki, K. (Ed.). (2011), *Artificial neural networks: methodological advances and biomedical applications*. BoD–Books on Demand.

UC Irvine Machine Learning Repository (2022), Erişim adresi: https://archive.ics.uci.edu/datasets.

Uluskan, M. (2020), Artificial neural networks as a quality loss function for six sigma, *Total Quality Management and Business Excellence*, 31(15-16), 1811-1828.

Wu, Y.C. and Feng, J.W. (2018), Development and application of artificial neural network. *Wireless Personal Communications*, *102*, 1645-1656.

Zhang, G.P. (2000), Neural networks for classification: A survey, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4), 451-462.

Zhang, Y.D., Satapathy, S.C., Guttery, D.S., Górriz, J.M. and Wang, S.H. (2021), Improved breast cancer classification through combining graph convolutional network and convolutional neural network, *Information Processing and Management*, 58(2), 102439.

Zou, J., Han, Y. and So, S.S. (2009), Overview of artificial neural networks. *Artificial neural networks: methods and applications*, 14-22.