Avrupa Bilim ve Teknoloji Dergisi
Sayı 19, S. 263-271, Ağustos 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 19, pp. 263-271, August 2020

Research Article

Korelasyon Temelli Özellik Seçimi, Genetik Arama ve Rastgele Ormanlar Tekniklerine Dayanan Yeni Bir Rahim Ağzı Kanseri Teşhis Yöntemi

Can Eyüpoğlu[1*]

[1] Milli Savunma Üniversitesi, Hava Harp Okulu, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-6133-8617)

(İlk Geliş Tarihi 22 Nisan 2020 ve Kabul Tarihi 23 Mayıs 2020)

(**DOI:** 10.31590/ejosat.725305)

ATIF/REFERENCE: Eyüpoğlu, C. (2020). Korelasyon Temelli Özellik Seçimi, Genetik Arama ve Rastgele Ormanlar Tekniklerine Dayanan Yeni Bir Rahim Ağzı Kanseri Teşhis Yöntemi. *Avrupa Bilim ve Teknoloji Dergisi*, (19), 263-271.

Öz

Rahim ağzı kanseri kadınlarda en sık görülen kanser türlerinden biridir. Bu kanser türü nedeniyle gerçekleşecek ölümlerin sayısını azaltmanın yolu erken teşhistir. Hastalığı erken teşhis ederken doktorlara yardımcı olmak için makine öğrenmesi ve veri madenciliği teknikleri kullanılmaktadır. Bu çalışmada rahim ağzı kanseri teşhisi için korelasyon temelli özellik seçimi (correlation-based feature selection-CFS), genetik algoritma (genetic algorithm-GA) ve rastgele ormanlar (random forests-RF) tekniklerinden yararlanan yeni bir yöntem önerilmiştir. Veri önişleme, özellik seçimi ve sınıflandırma olmak üzere üç aşamadan oluşan yöntemin performansı; sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü metrikleri kullanılarak test edilmiştir. Ardından performans sonuçları klasik makine öğrenmesi teknikleri ve literatürde var olan çalışmalarla karşılaştırılmıştır. Deneysel sonuçlardan önerilen yöntemin etkili olduğu ve rahim ağzı kanserini erken teşhis etmede doktorlar tarafından yardımcı bir araç olarak kullanılabileceği görülmektedir.

Anahtar Kelimeler: Rahim Ağzı Kanseri Teşhisi, Korelasyon Temelli Özellik Seçimi, Genetik Arama, Rastgele Ormanlar, Makine Öğrenmesi, Veri Madenciliği.

A New Cervical Cancer Diagnosis Method Based on Correlation-based Feature Selection, Genetic Search and Random Forests Techniques

Abstract

Cervical cancer is one of the most common types of cancer in women. The way to reduce the number of deaths due to this type of cancer is early diagnosis. Machine learning and data mining techniques are used to assist doctors while early diagnosing the disease. In this study, a new method exploiting correlation-based feature selection (CFS), genetic algorithm (GA) and random forests (RF) techniques is proposed for the diagnosis of cervical cancer. The performance of the proposed method consisting of three stages: data preprocessing, feature selection and classification has been tested using classification accuracy, precision, recall, and F-measure metrics. In the sequel, the performance results are compared with the conventional machine learning techniques and the existing studies in the literature. It can be seen from the experimental results that the proposed method is effective and can be used as an auxiliary tool by doctors in diagnosing cervical cancer early.

Keywords: Cervical Cancer Diagnosis, Correlation-based Feature Selection, Genetic Search, Random Forests, Machine Learning, Data Mining.

[*] Sorumlu Yazar: Milli Savunma Üniversitesi, Hava Harp Okulu, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, ORCID: 0000-0002-6133-8617, caneyupoglu@gmail.com, ceyupoglu@hho.edu.tr

1. Giriş

Rahim ağzı kanseri, 2018 yılında tahmini 570.000 vaka ile kadınlarda dördüncü en sık görülen kanser türüdür ve tüm kadın kanserlerinin %7,5'ini temsil etmektedir. 2018 yılında yaklaşık 311.000 kadın rahim ağzı kanseri sebebiyle ölmüştür. Bu ölümlerin %85'inden fazlası düşük ve orta gelirli ülkelerde meydana gelmiştir (World Health Organization, 2019).

Rahim ağzı kanseri sebebiyle gerçekleşen ölümlerin sayısını azaltmanın yolu erken teşhistir. Özellikle tarama programlarının bulunmadığı ülkelerde bu kanser türünü erken teşhis etmek ve etkili tedaviye başlamak hayatta kalma olasılığını önemli ölçüde artırmaktadır. Mevcut şartlarda bu hastalık ilerleyene kadar ya da tedaviye erişilinceye kadar genellikle tanılanmamaktadır. Bu durum da rahim ağzı kanserinde yüksek ölüm oranını neden olmaktadır. Rahim ağzı kanserinin belirtilerinin anlaşılması ve tespit edilmesi sayesinde hastalara erken teşhis koyulabilmektedir (World Health Organization, 2020).

Bilgisayar biliminde, rahim ağzı kanserini teşhis ederken doktorlara erken tanıda yardımcı olmak amacıyla veri madenciliği ve makine öğrenmesi teknikleri kullanılmaktadır. 2017 yılında rahim ağzı kanseri tanısı için demografik bilgilerden, alışkanlıklardan ve tıbbi kayıtlardan oluşan yeni bir veri seti (Fernandes vd., 2017a; Fernandes vd., 2017b) oluşturulmuştur. Literatürde şimdiye kadar bu veri setini kullanan bazı çalışmalar yapılmıştır ve bu çalışmalarda çeşitli teknikler, modeller ve sistemler önerilmiştir. Wu ve Zhou (2017) tarafından yapılan çalışmada rahim ağzı kanseri teşhisi için özyinelemeli özellik eleme (recursive feature elimination-RFE), temel bileşen analizi (principal component analysis-PCA) ve destek vektör makinesi (support vector machine-SVM) temelli iki yöntem önerilmiştir. İlk olarak hastalık tanısı için kullanılan veri setinden RFE ve PCA ile özellik çıkarma işlemi yapılmıştır. Ardından çıkarılan özellikler SVM ile sınıflandırılmıştır. Deneysel sonuçlar, özellik çıkarma için PCA'nın kullanıldığı yöntemin rahim ağzı kanseri teşhisinde daha iyi sonuçlar verdiğini göstermektedir.

Abdoh vd. (2018) ise kanser tanısı için oversampling, RFE, PCA ve rastgele ormanlar (random forests-RF) tabanlı bir teknik ileri sürmüştür. Simülasyon sonuçları, önerilen tekniğin rahim ağzı kanseri teşhisinde kullanılabileceğini göstermektedir. Rayavarapu ve Krishna (2018), rahim ağzı kanseri tahmini için bir derin sinir ağı (deep neural network-DNN) kullanmıştır. Deng vd. (2018) tarafından yapılan çalışmada ise XGBoost (eXtreme Gradient Boosting), SVM ve RF tekniklerinin rahim ağzı kanseri sınıflandırmadaki performansı incelenmiştir. Çalışmanın sonucunda XGBoost ve RF tekniklerinin SVM'den daha iyi performansa sahip olduğu görülmüştür.

Sawhney vd. (2018) kanser teşhisi için ateş böceği algoritması (firefly algorithm-FA) temelli bir özellik seçimi metodu önermiştir. Rahim ağzı kanseri veri setinde FA ile özellik seçimi yapıldıktan sonra RF kullanılarak sınıflandırma yapılmıştır. Son zamanlarda yapılan bir çalışmada ise Adem vd. (2019) rahim ağzı kanseri teşhisi için derin öğrenmeye dayalı bir sınıflandırma modeli geliştirmiştir. Test sonuçları, önerilen modelin sınıflandırma doğruluğu performansının literatürdeki bazı çalışmalardan daha iyi olduğunu göstermektedir.

Bu makalede ise literatürde var olan çalışmalardan farklı olarak rahim ağzı kanseri teşhisi için korelasyon temelli özellik seçimi (correlation-based feature selection-CFS), genetik algoritma (genetic algorithm-GA) ve RF tekniklerine dayanan yeni bir yöntem önerilmiştir. Önerilen yöntemin performansı; sınıflandırma doğruluğu, kesinlik (precision), duyarlılık (recall veya sensitivity) ve F-ölçütü (F-measure veya F_1 score) metrikleri kullanılarak incelenmiş, klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile karşılaştırılmıştır.

Çalışmanın diğer bölümlerinin organizasyonu şu şekildedir. Materyal ve Metot bölümünde, bu çalışmada rahim ağzı kanseri teşhisi için kullanılan veri seti ve önerilen yöntem tanıtılmaktadır. Araştırma Sonuçları ve Tartışma bölümünde önerilen yöntem, performans metrikleri kullanılarak klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile kıyaslanmaktadır. Sonuç bölümünde ise çalışmanın genel sonuçlarından bahsedilmektedir.

2. Materyal ve Metot

Bu bölümde ilk olarak çalışmada önerilen yöntemin performansının değerlendirilmesi için kullanılan veri seti açıklanmaktadır. Ardından veri önişleme, özellik seçimi ve sınıflandırma aşamalarından oluşan önerilen yöntem ayrıntılı olarak anlatılmaktadır.

2.1. Kullanılan Veri Seti

Bu çalışmada önerilen yöntem, Rahim Ağzı Kanseri Veri Seti (Cervical Cancer Data Set) (Fernandes vd., 2017a; Fernandes vd., 2017b) üzerinde test edilmiştir. Bu veri seti Kaliforniya Üniversitesi – Irvine Makine Öğrenmesi Deposunda (University of California – Irvine Machine Learning Repository) açık erişimli olarak bulunmaktadır.

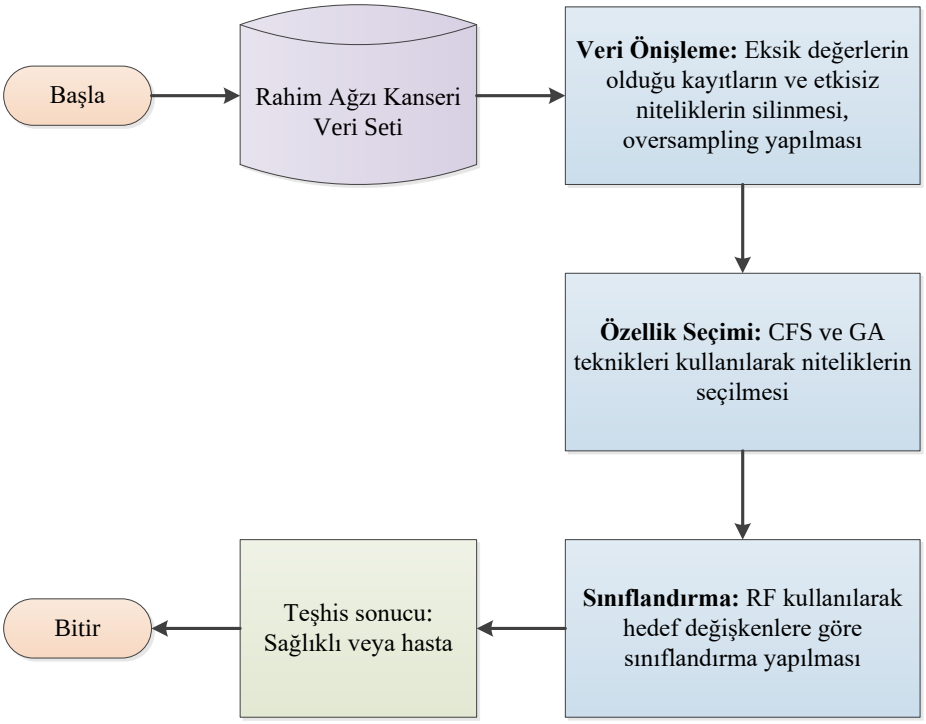
Veri seti, toplam 858 hastaya ait demografik bilgiler, alışkanlıklar ve tıbbi kayıtlardan oluşmaktadır. Veri setindeki nitelik sayısı 36'dır. Bunlardan 32'si rahim ağzı kanseri risk faktörleri ile ilgilidir, 4'ü ise sınıflandırma için etiket olarak kullanılan hedef değişkenleridir. Bu hedef değişkenleri rahim ağzı kanseri tanısında kullanılan Hinselmann, Schiller, Sitoloji ve Biyopsi adlı teşhis araçları ve tekniklerdir. Veri setinde bulunan nitelikler ve bu niteliklere ait bilgiler Tablo 1'de özetlenmektedir.

Tablo 1. Rahim Ağzı Kanseri Veri Setinin Özellikleri

Nitelik Numarası	Nitelik Adı (Risk Faktörleri)	Eksik Değer Sayısı	Farklı Değer Sayısı	Minimum Değer	Maksimum Değer
1	Yaş	0	44	13	84
2	Cinsel partner sayısı	26	12	1	28
3	İlk cinsel ilişki (yaş)	7	21	10	32
4	Gebelik sayısı	56	11	0	11
5	Sigara kullanımı	13	2	0	1
6	Sigara kullanımı (yıl)	13	30	0	37
7	Sigara kullanımı (paket/yıl)	13	62	0	37
8	Hormonal kontraseptifler	108	2	0	1
9	Hormonal kontraseptifler (yıl)	108	40	0	30
10	IUD (intrauterine device-rahim içi cihaz kullanımı)	117	2	0	1
11	IUD (yıl)	117	26	0	19
12	STDs (sexually transmitted diseases-cinsel yolla bulaşan hastalıklar)	105	2	0	1
13	STDs (sayı)	105	5	0	4
14	STDs: Kondilomatoz	105	2	0	1
15	STDs: Servikal kondilomatoz	105	1	0	0
16	STDs: Vajinal kondilomatoz	105	2	0	1
17	STDs: Vulvo-perineal kondilomatoz	105	2	0	1
18	STDs: Sifiliz	105	2	0	1
19	STDs: Pelvik inflamatuar hastalık	105	2	0	1
20	STDs: Genital herpes	105	2	0	1
21	STDs: Molluscum contagiosum	105	2	0	1
22	STDs: AIDS (acquired immune deficiency syndrome)	105	1	0	0
23	STDs: HIV (human immunodeficiency virus)	105	2	0	1
24	STDs: Hepatit B	105	2	0	1
25	STDs: HPV (human papillomavirus)	105	2	0	1
26	STDs: Tanı sayısı	0	4	0	3
27	STDs: İlk tanıdan bu yana geçen süre	787	18	1	22
28	STDs: Son tanıdan bu yana geçen süre	787	18	1	22
29	Dx (diagnosis-teşhis): Kanser	0	2	0	1
30	Dx: CIN (cervical intraepithelial neoplasia)	0	2	0	1
31	Dx: HPV	0	2	0	1
32	Dx	0	2	0	1
	Hedef Değişken Adı (Sınıflandırma Etiketi)	**Sağlıklı Kişi Sayısı**		**Hasta Kişi Sayısı**	
33	Hinselmann	823		35	
34	Schiller	784		74	
35	Sitoloji	814		44	
36	Biyopsi	803		55	

2.2. Önerilen Yöntem

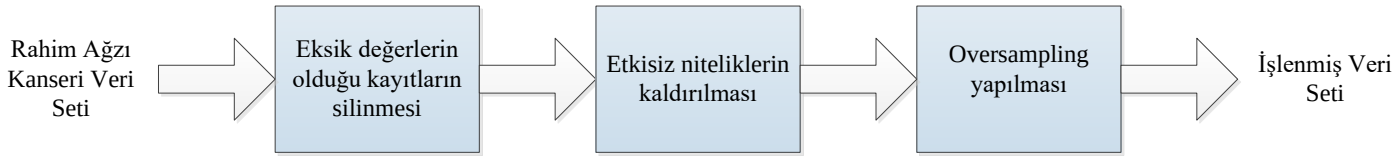
Bu çalışmada önerilen yöntem, Şekil 1'deki akış diyagramında da görüldüğü üzere veri önişleme, özellik seçimi ve sınıflandırma aşamalarından oluşmaktadır. Her bir adımda gerçekleştirilen işlemler alt bölümlerde ayrıntılı olarak anlatılmaktadır.



Şekil 1. Önerilen Yöntemin Akış Diyagramı

2.2.1. Veri Önişleme

Bu çalışmada rahim ağzı kanseri teşhisi başarısını artırmak için uygulanan veri önişleme süreci Şekil 2'de gösterilmektedir. Çalışmada kullanılan veri setinde Tablo 1'de görüldüğü üzere eksik değerlerin olduğu birçok kayıt vardır ve bu kayıtlar hastalık teşhisi için uygulanan sınıflandırma işlemini olumsuz etkilemektedir. Bu nedenle ilk olarak veri setinden eksik değerlerin olduğu kayıtlar silinmiştir. Bu işlem sonucunda 668 hastaya ayıt kayıtlar kalmıştır.



Şekil 2. Veri Önişleme Sürecinin Blok Diyagramı

Yine Tablo 1'e bakıldığında "STDs: İlk tanıdan bu yana geçen süre" ve "STDs: Son tanıdan bu yana geçen süre" adlı 27 ve 28. niteliklerin neredeyse tüm kayıtlarda eksik olduğu görülmektedir. Bu sebeple bu iki nitelik veri setinden silinmiştir ve ardından veri setinde risk faktörleri ile ilgili 30 nitelik kalmıştır. Bu işlemlerden sonra hedef değişkenlerinin özellikleri Tablo 2'de gösterilmektedir. Tablodan görüldüğü üzere sağlıklı ve rahim ağzı kanseri hastası etiketlerinin veri setindeki oranı çok dengesizdir. Sınıflandırma performansını kötü etkileyen bu problemin üstesinden gelmek için etiket sayıları aynı olacak şekilde oversampling yapılmıştır.

Tablo 2. Önişleme Sonrası Hedef Değişkenlerin Özellikleri

Hedef Değişken	**Sağlıklı Kişi Sayısı**	**Hasta Kişi Sayısı**
Hinselmann	638	30
Schiller	605	63
Sitoloji	629	39
Biyopsi	623	45

2.2.2. Özellik Seçimi

Önerilen yöntemin özellik seçimi aşamasında CFS tekniği (Hall, 1999) kullanılmıştır. CFS, her bir özelliğin bireysel tahmin yeteneğini aralarındaki fazlalık (redundancy) derecesi ile birlikte dikkate alarak bir nitelik alt kümesinin değerini ölçer. Düşük karşılıklı korelasyona sahipken sınıfla yüksek derecede korelasyona sahip özelliklerin alt kümeleri tercih edilir. Bu teknik ile ilgili daha ayrıntılı bilgi için (Hall, 1999) kaynağına başvurulabilir. CFS tekniğinde arama metodu olarak ise GA (Goldberg, 1989) kullanılmıştır. GA'lar doğal seleksiyon ve genetik ilkelerine dayanan arama yöntemleridir (Bremermann, 1958; Fraser, 1957; Holland, 1975; Sastry vd. 2005). Bu yöntemler, bir arama sorununun karar değişkenlerini belirli kardinalitelerin sınırlı uzunluktaki alfabe dizelerine kodlar. Arama problemine aday çözümler olan dizelere kromozomlar, alfabelere genler ve genlerin değerlerine ise aleller denir. GA'lar geleneksel optimizasyon tekniklerinin aksine parametrelerin kendisinden ziyade parametrelerin kodlanmasıyla çalışır. İyi çözümler geliştirmek, doğal seleksiyonu gerçeklemek ve iyi çözümleri kötü çözümlerden ayırmak için bir ölçüt gereklidir. Bu ölçüt, matematiksel model veya bilgisayar simülasyonu olan nesnel bir fonksiyon ya da insanların daha kötü olanlara göre daha iyi çözümler seçtiği öznel bir işlev olabilir. Esas itibarıyla uygunluk ölçütü, daha sonra iyi çözümlerin gelişimini yönlendirmek için GA tarafından kullanılacak bir aday çözümün göreceli uygunluğunu belirlemelidir. GA'lardaki bir başka önemli konsept nüfus kavramıdır ve klasik arama yöntemlerinden farklı olarak aday çözümlerin popülasyonuna dayanır. Genellikle kullanıcı tanımlı bir parametre olan popülasyon büyüklüğü, genetik algoritmaların ölçeklenebilirliğini ve performansını etkileyen önemli faktörlerden biridir (Sastry vd. 2005). Bu çalışmada CFS ve GA tekniklerlerinin beraber kullanılması sonucu seçilen nitelikler Tablo 3'te gösterilmektedir.

Tablo 3. CFS ve GA Kullanılarak Seçilen Nitelikler

Hedef Değişken	Seçilen Niteliklerin Numarası	Seçilen Niteliklerin Adları
Hinselmann	1, 2, 3, 4, 5, 9, 11, 12, 14, 23, 29	Yaş, Cinsel partner sayısı, İlk cinsel ilişki (yaş), Gebelik sayısı, Sigara kullanımı, Hormonal kontraseptifler (yıl), IUD (yıl), STDs, STDs: Kondilomatoz, STDs: HIV, Dx: Kanser
Schiller	2, 3, 4, 8, 9, 10, 17, 29, 30	Cinsel partner sayısı, İlk cinsel ilişki (yaş), Gebelik sayısı, Hormonal kontraseptifler, Hormonal kontraseptifler (yıl), IUD, STDs: Vulvo-perineal kondilomatoz, Dx: Kanser, Dx: CIN
Sitoloji	2, 3, 4, 7, 8, 12, 29	Cinsel partner sayısı, İlk cinsel ilişki (yaş), Gebelik sayısı, Sigara kullanımı (paket/yıl), Hormonal kontraseptifler, STDs, Dx: Kanser
Biyopsi	1, 2, 3, 4, 5, 8, 11, 12, 29	Yaş, Cinsel partner sayısı, İlk cinsel ilişki (yaş), Gebelik sayısı, Sigara kullanımı, Hormonal kontraseptifler, IUD (yıl), STDs, Dx: Kanser

2.2.3. Sınıflandırma

Önerilen yöntemin sınıflandırma aşamasında RF tekniğinden (Breiman, 2001) yararlanılmıştır. RF, birçok farklı alanda yaygın olarak kullanılan bir gözetimli öğrenme tekniğidir ve güçlü bir öğrenici oluşturmak için zayıf öğreniciler grubunu kullanma ilkesiyle çalışmaktadır. Bu teknikte her ağaç bağımsız bir karar ağacı oluşturmak için veri setinin bir alt kümesini rastgele olarak seçer. Seçilen rastgele alt küme, her ağaç budama olmadan bir yaprak düğüme ulaşana kadar kök düğümden bir alt düğüme tekrarlı olarak bölünür. Her ağaç, özelliklerin ve hedef değişkenin sınıflandırmasını bağımsız olarak yapar ve son ağaç sınıfı için oy kullanır. Son olarak elde edilen çoğunluk ağaç oylamasına dayanarak nihai genel sınıflandırmaya karar verilir (Abdoh vd., 2018).

3. Araştırma Sonuçları ve Tartışma

Bu bölümde sırasıyla performans ölçümü için kullanılan metriklere, önerilen yöntemin performansının klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile karşılaştırılmasına yer verilmektedir. Bu çalışmada önerilen yöntem Intel Core i7 8565U işlemci (1.80 GHz) ve 8 GB RAM ile Windows 10 Pro 64-bit işletim sisteminde çalışan Weka 3.8.4'te gerçeklenmiştir. Tüm testler veri bölümleme için 10-kat çapraz geçerleme tekniği kullanılarak gerçekleştirilmiştir.

3.1. Performans Ölçümü için Kullanılan Metrikler

Bu makalede önerilen yöntemin performansı; sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü metrikleri kullanılarak test edilmiştir. Bu metrikler şu şekilde hesaplanır (Eyupoglu vd., 2018; Yavuz vd., 2017; Yavuz ve Eyüpoğlu, 2019):

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

$$Kesinlik = \frac{TP}{TP + FP} \tag{2}$$

$$Duyarlılık = \frac{TP}{TP + FN} \tag{3}$$

$$F\text{–}ölçütü = \frac{2 \times Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \tag{4}$$

Bu formüllerdeki *TP* (true positive-gerçek pozitif), *TN* (true negative-gerçek negatif), *FP* (false positive-yanlış pozitif) ve *FN* (false negative-yanlış negatif) sayıları karışıklık matrisinde (confusion matrix) yer alan 4 farklı sınıflandırma sonucunu temsil etmektedir.

3.2. Önerilen Yöntemin Klasik Makine Öğrenmesi Teknikleri ile Karşılaştırılması

Önerilen yöntemin performansı; her bir hedef değişkeni için sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü metrikleri kullanılarak klasik makine öğrenmesi teknikleri ile kıyaslanmıştır. Kıyaslama için kullanılan teknikleri şunlardır: radyal tabanlı fonksiyon ağı (radial basis function network-RBFN) (Frank, 2014), lojistik regresyon (logistic regression-LR) (Le Cessie ve Van Houwelingen, 1992), AdaBoostM1 (Freund ve Schapire, 1996), Naive Bayes (NB) (John ve Langley, 1995), OneR (Holte, 1993), lojistik model ağaçları (logistic model trees-LMT) (Landwehr vd., 2005), SVM (Keerthi vd., 2001), C4.5 (Quinlan, 1993), K* (Cleary ve Trigg, 1995) ve voted perceptron (VP) (Freund ve Schapire, 1999). Tablo 4-7'de sırasıyla Hinselmann, Schiller, Sitoloji ve Biyopsi hedef değişkenleri için önerilen yöntemin performansı klasik makine öğrenmesi teknikleri ile karşılaştırılmaktadır.

Tablo 4. Önerilen Yöntemin Performansının Hinselmann Hedef Değişkeni için Klasik Makine Öğrenmesi Teknikleri ile Karşılaştırılması

Yöntem	**Performans Metriği**			
	Doğruluk	**Kesinlik**	**Duyarlılık**	**F-ölçütü**
RBFN	%78,61	0,755	0,848	0,799
LR	%66,22	0,665	0,654	0,659
AdaBoostM1	%77,66	0,812	0,719	0,763
NB	%71,00	0,740	0,647	0,691
OneR	%88,71	0,825	0,983	0,897
LMT	%92,16	0,937	0,904	0,920
SVM	%61,76	0,582	0,835	0,686
C4.5	%93,34	0,930	0,937	0,934
K*	%88,87	0,899	0,876	0,887
VP	%60,11	0,600	0,607	0,603
Önerilen Yöntem	%97,41	0,959	0,991	0,975

Tablo 5. Önerilen Yöntemin Performansının Schiller Hedef Değişkeni için Klasik Makine Öğrenmesi Teknikleri ile Karşılaştırılması

Yöntem	**Performans Metriği**			
	Doğruluk	**Kesinlik**	**Duyarlılık**	**F-ölçütü**
RBFN	%71,57	0,685	0,800	0,738
LR	%63,06	0,602	0,774	0,677
AdaBoostM1	%72,56	0,738	0,699	0,718
NB	%63,47	0,598	0,820	0,692
OneR	%81,49	0,758	0,926	0,833
LMT	%89,17	0,882	0,904	0,893
SVM	%62,98	0,590	0,850	0,696
C4.5	%90,00	0,882	0,924	0,902
K*	%81,41	0,785	0,864	0,823
VP	%52,89	0,523	0,648	0,579
Önerilen Yöntem	%92,56	0,906	0,950	0,927

Tablo 6. Önerilen Yöntemin Performansının Sitoloji Hedef Değişkeni için Klasik Makine Öğrenmesi Teknikleri ile Karşılaştırılması

Yöntem	Performans Metriği			
	Doğruluk	**Kesinlik**	**Duyarlılık**	**F-ölçütü**
RBFN	%82,35	0,744	0,987	0,848
LR	%67,81	0,667	0,712	0,689
AdaBoostM1	%81,40	0,830	0,790	0,809
NB	%56,28	0,716	0,208	0,323
OneR	%87,36	0,798	1,000	0,888
LMT	%92,13	0,908	0,938	0,923
SVM	%61,13	0,592	0,719	0,649
C4.5	%92,61	0,896	0,963	0,929
K*	%83,39	0,814	0,865	0,839
VP	%60,81	0,622	0,552	0,585
Önerilen Yöntem	%93,88	0,926	0,954	0,940

Tablo 7. Önerilen Yöntemin Performansının Biyopsi Hedef Değişkeni için Klasik Makine Öğrenmesi Teknikleri ile Karşılaştırılması

Yöntem	Performans Metriği			
	Doğruluk	**Kesinlik**	**Duyarlılık**	**F-ölçütü**
RBFN	%82,18	0,763	0,933	0,840
LR	%66,45	0,621	0,844	0,716
AdaBoostM1	%76,81	0,764	0,775	0,770
NB	%66,29	0,625	0,817	0,708
OneR	%84,67	0,796	0,933	0,859
LMT	%92,30	0,919	0,928	0,923
SVM	%63,32	0,589	0,884	0,707
C4.5	%92,46	0,917	0,934	0,925
K*	%87,24	0,850	0,904	0,876
VP	%56,18	0,555	0,620	0,586
Önerilen Yöntem	%95,91	0,948	0,971	0,960

Önerilen yöntemin Hinselmann hedef değişkeni için olan doğruluk, kesinlik, duyarlılık ve F-ölçütü performansı sırasıyla %97,41, 0,959, 0,991 ve 0,975'tir. Tablo 4'te görüldüğü üzere önerilen yöntem tüm metriklerde klasik makine öğrenmesi tekniklerinden daha iyi performans göstermiştir. Schiller hedef değişkeni için elde edilen performans sonuçları sınıflandırma doğruluğu için %92,56, kesinlik için 0,906, duyarlılık için 0,950 ve F-ölçütü için 0,927'dir. Önerilen yöntemin performansı, tüm metriklerde klasik tekniklerden daha iyidir. Önerilen yöntem, Sitoloji hedef değişkeninde ulaştığı %93,88'lik doğruluk, 0,926'lık kesinlik ve 0,940'lık F-ölçütü başarısıyla klasik tekniklerden daha iyi performans göstermiştir. Duyarlılık değişkenindeki performansı ise RBFN, OneR ve C4.5 tekniklerine yakındır. Son olarak önerilen yöntemin Biyopsi hedef değişkeninde elde ettiği sırasıyla %95,91, 0,948, 0,971 ve 0,960'lık performansı diğer tüm tekniklerden daha iyidir.

3.3. Önerilen Yöntemin Literatürdeki Çalışmalar ile Karşılaştırılması

Bu çalışmada önerilen yöntemin sınıflandırma doğruluğu performansı Hinselmann, Schiller, Sitoloji ve Biyopsi hedef değişkenleri için literatürdeki diğer çalışmalarla (Abdoh vd., 2018; Adem vd., 2019; Deng vd., 2018; Rayavarapu ve Krishna, 2018; Wu ve Zhou, 2017) kıyaslanmıştır. Tablo 8'de diğer çalışmalarda kullanılan yöntemler ve her bir hedef değişkeni için elde edilen doğruluk oranları verilmektedir.

Tablo 8. Önerilen Yöntemin Sınıflandırma Doğruluğu Açısından Literatürdeki Çalışmalarla Karşılaştırılması

Çalışma	**Yöntem**	**Hedef Değişken**			
		Hinselmann	**Schiller**	**Sitoloji**	**Biyopsi**
Wu ve Zhou (2017)	PCA – SVM	%93,79	%90,18	%92,46	%94,03
Abdoh vd. (2018)	RFE – RF	%95,88	%92,91	%95,89	%95,87
Rayavarapu ve Krishna (2018)	DNN	-	-	%90,00	%95,00
Deng vd. (2018)	XGBoost	%96,34	%95,59	%96,30	%95,59
Adem vd. (2019)	Derin öğrenme	%95,50	%96,70	%96,60	%97,30
Önerilen Yöntem	CFS – GA – RF	%97,41	%92,56	%93,88	%95,91

Tablo 8'den görüldüğü üzere, önerilen yöntemin Hinselmann hedef değişkeni için olan %97,41'lik sınıflandırma doğruluğu performansı literatürdeki diğer çalışmalardan daha iyidir. Schiller değişkeni için Wu ve Zhou (2017) tarafından yapılan çalışmadan, Sitoloji için Wu ve Zhou (2017) ve Rayavarapu ve Krishna (2018) çalışmalarından, Biyopsi için ise Adem vd. (2019) çalışması haricindeki tüm çalışmalardan daha iyi performans göstermiştir.

4. Sonuç

Kadınlarda yaygın olarak görülen kanser türlerinden biri olan rahim ağzı kanseri sebebiyle yaşanan ölümlerin oranını azaltmak için erken teşhis çok büyük bir öneme sahiptir. Bilgisayar bilimindeki makine öğrenmesi ve veri madenciliği teknikleri bu kanser türünü erken teşhis ederken doktorlara yardımcı olmak amacıyla kullanılmaktadır. Bu çalışmada CFS, GA ve RF tekniklerine dayanan yeni bir rahim ağzı kanseri teşhis yöntemi önerilmiştir. Önerilen yöntem; veri önişleme, özellik seçimi ve sınıflandırma olmak üzere üç aşamada gerçekleşmektedir. Performans değerlendirmesi için sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü metrikleri kullanılmıştır. Elde edilen performans sonuçları klasik makine öğrenmesi tekniklerinin ve literatürdeki çalışmaların performanslarıyla kıyaslanmıştır. Deneysel sonuçlar önerilen yöntemin başarılı olduğunu ve bu kanser türüne erken tanı koyarken doktorlara destek olabileceğini göstermiştir.

Kaynakça

Abdoh, S. F., Rizka, M. A., & Maghraby, F. A. (2018). Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access*, *6*, 59475-59485.

Adem, K., Kiliçarslan, S., & Cömert, O. (2019). Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications*, *115*, 557-564.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Bremermann, H. J. (1958). *The evolution of intelligence: The nervous system as a model of its environment*. University of Washington, Department of Mathematics.

Cleary, J. G., & Trigg, L. E. (1995, July). K*: An instance-based learner using an entropic distance measure. In *12th International Conference on Machine Learning* (pp. 108-114).

Deng, X., Luo, Y., & Wang, C. (2018). Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 631-635). IEEE.

Eyupoglu, C., Aydin, M. A., Zaim, A. H., & Sertbas, A. (2018). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, *20*(5), 373.

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017a). *Cervical cancer (Risk Factors) Data Set* [Data file]. Available from http://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017b). Transfer learning with partial observability applied to cervical cancer screening. In *Iberian Conference on Pattern Recognition and Image Analysis* (pp. 243-250). Springer, Cham.

Frank, E. (2014). Fully supervised training of Gaussian radial basis function networks in WEKA. Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Fraser, A. S. (1957). Simulation of Genetic Systems by Automatic Digital Computers II. Effects of Linkage on Rates of Advance Under Selection. *Australian Journal of Biological Sciences*, *10*(4), 492-500.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *13th International Conference on Machine Learning* (pp. 148-156).

Freund, Y., & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, *37*(3), 277-296.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. PhD thesis. University of Waikato, Hamilton, New Zealand.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*(1), 63-90.

John, G. H., & Langley, P. (1995, August). Estimating continuous distributions in Bayesian classifiers. In *10th Conference on Uncertainty in Artificial Intelligence (UAI'95)* (pp. 338-345).

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation*, *13*(3), 637-649.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*(1-2), 161-205.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(1), 191-201.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rayavarapu, K., & Krishna, K. K. (2018, March). Prediction of Cervical Cancer using Voting and DNN Classifiers. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-5). IEEE.

Sastry, K., Goldberg, D., & Kendall, G. (2005). Genetic algorithms. In *Search methodologies* (pp. 97-125). Springer, Boston, MA.

Sawhney, R., Mathur, P., & Shankar, R. (2018, May). A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In *International Conference on Computational Science and Its Applications* (pp. 438-449). Springer, Cham.

World Health Organization. (2019). *Human papillomavirus (HPV) and cervical cancer*. Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer

World Health Organization. (2020). *Cervical cancer*. Retrieved from https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

Wu, W., & Zhou, H. (2017). Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access*, *5*, 25189-25195.

Yavuz, E., & Eyüpoğlu, C. (2019). Meme Kanseri Teşhisi İçin Yeni Bir Skor Füzyon Yaklaşımı. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *7*(3), 1045-1060.

Yavuz, E., Eyupoglu, C., Sanver, U., & Yazici, R. (2017). An ensemble of neural networks for breast cancer diagnosis. In *2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 538-543). IEEE.