www.turkishstudies.net/turkishstudies

**Turkish Studies**

eISSN: 1308-2140

INTERNATIONAL BALKAN UNIVERSITY
Sponsored by IBU

***Research Article / Araştırma Makalesi***

# Topluluk Öğrenmesi Yöntemleri ile Göğüs Kanseri Teşhisi*

*Breast Cancer Diagnosis with Ensemble Learning Methods*

Fulya Akcan** - Ahmet Sertbaş***

**Abstract:** Breast cancer is one of the most common types of cancer among women worldwide and has become a major public health problem in today's society. Early diagnosis of breast cancer can significantly increase patients' chances of survival due to enabling patients to reach clinical treatment on time. Machine Learning techniques can be used for developing breast cancer detection tools that can be used as an effective mechanism which physicians can benefit from. In this study, performance comparisons were made according to the accuracy percentages of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF) machine learning methods for breast cancer diagnosis. Bagging, boosting and voting ensemble machine learning methods were applied to increase the accuracy of individual classifiers. Hard voting and soft voting methods were used for the voting ensemble learning method. Adaptive Boosting (Adaboost), Gradient Boosting and XGBoost algorithms were used for the boosting ensemble learning method. The experiments were conducted on Wisconsin Diagnostic Breast Cancer (WDBC) Dataset. Different feature scales of data features negatively affect the modeling of a data set. Therefore, standardization preprocessing has been applied to scale the features. The ensemble learning methods were compared according to precision, recall, f-score, accuracy performance criteria. The highest accuracy percentage was obtained with Soft Voting, Bagging (SVC), Adaboost (SVC) and XGBoost ensemble learning methods.

**Structured Abstract:** Breast cancer is one of the most common types of cancer among women, leading to the majority of new cancer cases and cancer related deaths. Breast tumors are classified as benign and malignant. Malignant tumors increase the mortality rate of patients by proliferating rapidly when left untreated. Early diagnosis of breast cancer can significantly increase patients' chances of survival. Machine learning techniques are used to avoid the limitations of traditional techniques used for breast cancer diagnosis and to assist doctors. Various algorithms and techniques have been used to assist healthcare professionals in the diagnosis of breast cancer, such as Decision Trees (DT), K-Nearest Neighbor (KNN), Naive Bayes (NB),

---

* Makale Prof. Dr. Ahmet Sertbaş danışmanlığında Fulya Akcan tarafından hazırlanan "Kollektif Makine Öğrenmesi Metodları ile Göğüs Kanseri Teşhisi" isimli yüksek lisans tezinden oluşturulmuştur.

** Yüksek Lisans Öğrencisi, İstanbul Üniversitesi- Cerrahpaşa, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Bölümü

*Master's Student, Istanbul University- Cerrahpasa, Institute of Graduate Studies, Department of Computer Engineering*

ORCİD 0000-0003-0197-5347

fulyaakcan0@gmail.com

*** Prof. Dr., İstanbul University- Cerrahpasa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü

*Prof. Dr., Istanbul University- Cerrahpasa, Faculty of Engineering, Department of Computer Engineering*

ORCİD 0000-0001-8166-1211

asertbas@istanbul.edu.tr

**Cite as/ Atıf:** Akcan, F. & Sertbaş, A. (2021). Topluluk öğrenmesi yöntemleri ile göğüs kanseri teşhisi. *Turkish Studies, 16*(2), 511-527. https://dx.doi.org/10.7827/TurkishStudies.

**Received/Geliş:** 02 February/Şubat 2021 *Checked by plagiarism software*

**Accepted/Kabul:** 25 April/Nisan 2021 **Published/Yayın:** 30 April/Nisan 2021

CC BY-NC 4.0

Support Vector Machine (SVM). However, using a single technique does not always guarantee a high accuracy. Because each technique has superiorities and weaknesses regarding classification tasks. Ensemble learning methods are used for increase the strength of individual classifiers and reduce their weaknesses. In this study; bagging, boosting, voting ensemble methods have been applied to increase the power of individual classifiers in breast cancer diagnosis.

Bazazeh and Shubair compared the performance of three most popular machine learning techniques which commonly used for breast cancer detection and diagnosis, namely Support Vector Machine, Random Forest and Bayesian Networks (Bazazeh & Shubair, 2016). Experiments were conducted on the Wisconsin Breast Cancer dataset. Random Forest technique has reached the optimum ROC (Receiver Operator Characteristic) performance (99.9%) compared to the other two techniques. The technique of combining Support Vector Machine, Naive Bayes and J48 Decision Tree with Voting Classifier were used by Kumar and others (Kumar et al., 2017). Experiments were conducted on the Wisconsin Breast Cancer Dataset (WBCD) dataset. 97.13% accuracy was achieved with the proposed ensemble method. Dhanya and others proposed novel model for breast cancer prediction by using existing ensemble techniques along with a combination of supervised machine learning algorithms (Support Vector Machine, Naive Bayes, K-Nearest Neighbor, Logistic Regression) (Dhanya et al., 2020). Variance threshold and f-test feature selection techniques were used to increase the performance of the ensemble method. It has been concluded that the stacking ensemble method combined with the F-test feature selection performed better in predicting breast cancer for Wisconsin, Wisconsin Diagnostic Breast Cancer and MicroRNA datasets.

In this study, the performance comparison of Support Machine Machine, K-Nearest Neighborhood, Naive Bayes, Decision Trees and Random Forest machine learning methods which used in the literature for breast cancer diagnosis was made. Bagging, boosting and voting ensemble machine learning methods were applied to increase the accuracy of individual classifiers and the accuracy values obtained with these methods were compared. Ensemble learning methods were compared according to precision, racall, f-score, accuracy model performance criteria. ROC curves and AUC (Area Under the Receiver Operating Characteristic Curve) scores of bagging and boosting ensemble methods were compared. Bagging, boosting and voting techniques were used for achieve higher accuracy in the ensemble model. Hard voting and soft voting methods were used for the voting ensemble learning method. Adaptive Boosting (Adaboost), Gradient Boosting and XGBoost algorithms were used for the boosting ensemble learning method. Experiments were conducted on the Wisconsin Diagnostic Breast Cancer dataset. The dataset was taken from the University of California Irvine Machine Learning Repository (UCI) (Wolberg et al., 1992). Different feature scales of data features negatively affect the modeling of a data set. Therefore, feature scaling is an important step in modeling algorithms with datasets. StandardScaler function of sklearn library was used for feature scaling. The highest accuracy percentage before principal component analysis was obtained by Gaussian Naive Bayes method (94.74%). The highest accuracy percentage after principal component analysis was obtained by Support Vector Machine method (93.71%). The highest accuracy percentage after standardization preprocessing step was obtained by Support Vector Machine method (96.49%). 96.49% accuracy was obtained with the Random Forest method. 96.49% accuracy was achieved with the Hard Voting ensemble learning method and 97.37% with the Soft Voting ensemble learning method. It has been observed that the highest accuracy percentage obtained by bagging ensemble method was 97.37%. The accuracy percentage is 97.37% with Adaboost (SVC), 96.49% with Adaboost (GaussianNB), and 92.11% with Adaboost (Decision Tree). With Gradient Boosting and XGBoost ensemble methods, respectively; 94.74% and 97.37% accuracy was obtained. The highest percentage of accuracy among all applied methods was achieved by Soft Voting, Bagging (SVC), Adaboost (SVC) and XGBoost ensemble learning methods (97.37%).

This study is aimed to increase the power of individual classifiers in breast cancer diagnosis. A higher accuracy percentage has obtained by ensemble learning methods than individual classifiers. Instead of using Wisconsin Diagnostic Breast Cancer Dataset, researchers can use other datasets for breast cancer prediction to improve the performance of ensemble learning methods.

**Keywords:** Computer Science, Breast Cancer, machine learning, ensemble learning, Wisconsin Breast Cancer dataset

**Öz:** Göğüs kanseri, dünya çapında kadınlar arasında en yaygın kanser türlerinden biridir ve günümüz toplumunda önemli bir halk sağlığı sorunu haline gelmiştir. Göğüs kanserinin erken teşhisi, hastaların klinik

tedaviye zamanında ulaşabilmesini sağlayabileceği için hastaların hayatta kalma şansını önemli ölçüde artırabilmektedir. Makine Öğrenimi teknikleri, göğüs kanserinin erken teşhisi için hekimlerin yararlanabileceği etkili bir mekanizma olarak kullanılabilecek ve hastaların hayatta kalma oranını büyük ölçüde artıracak araçlar geliştirmek için kullanılabilir. Bu çalışmada, göğüs kanseri teşhisi için Destek Vektör Makinesi (SVM), K-En Yakın Komşuluk (KNN), Naive Bayes (NB), Karar Ağacı (DT) ve Rastgele Orman (RF) makine öğrenmesi yöntemlerinin doğruluk yüzdelerine göre performans karşılaştırmaları yapılmıştır. Bireysel sınıflandırıcıların doğruluk oranını arttırmak için bagging, boosting ve voting topluluk makine öğrenmesi yöntemleri uygulanmıştır. Oylama topluluk yöntemi için; çoğunluğa dayalı (hard) oylama ve olasılığa dayalı (soft) oylama yöntemleri kullanılmıştır. Yükseltme topluluk öğrenme yöntemi için; Adaptive Boosting (Adaboost), Gradient Boosting ve XGBoost algoritmaları kullanılmıştır. Kullanılan yöntemlerin performans karşılaştırması, Wisconsin Diagnostic Breast Cancer (WDBC) veri kümesi kullanılarak yapılmıştır. Veri özelliklerin ölçeklerinin farklı olması, bir veri kümesinin modellemesini olumsuz etkilemektedir. Bu nedenle, özelliklerin ölçeklendirilmesi için standardizasyon ön işleme işlemi uygulanmıştır. Topluluk öğrenme yöntemleri; kesinlik, duyarlılık, f-skor ve doğruluk performans kriterlerine göre karşılaştırılmıştır. En yüksek doğruluk yüzdesi; Soft Oylama (Ağırlıklandırılmış), Bagging (SVC), Adaboost (SVC) ve XGBoost topluluk öğrenme yöntemleriyle elde edilmiştir.

**Anahtar Kelimeler:** Bilgisayar Bilimleri, Göğüs Kanseri, makine öğrenmesi, topluluk öğrenmesi, Wisconsin Breast Cancer veri kümesi

## Giriş

Göğüs kanseri, küresel istatistiklere göre yeni kanser vakalarının ve kansere bağlı ölümlerin çoğuna yol açan, kadınlar arasında en yaygın kanser türlerinden biridir. 2018'de, dünya çapında 2.09 milyon yeni vaka tespit edilmiş ve göğüs kanserine bağlı olarak 627.000 ölüm meydana gelmiştir (Meo, 2018). Az gelişmiş ülkelerde erken teşhis ve tedavi tesislerine erişim eksikliği nedeniyle ölüm oranı, gelişmiş ülkere nispeten çok daha yüksektir (International Agency for Research on Cancer, 2013).

Tümörler, iyi huylu(benign) ve kötü huylu(malignant) olarak sınıflandırılmaktadır. İyi huylu tümörlerin, genellikle yavaş büyüdükleri ve vücudun diğer bölgelerine yayılmadıkları bilinmektedir. Kötü huylu(malign) tümörler, tedavi edilmediğinde hızlı ve kontrolsüz bir şekilde çoğalarak tümörün yakınındaki organlara ve dokulara yayılabilmekte ve hastaların ölüm oranını arttırmaktadır. Göğüs kanserinin erken teşhis edilmesiyle, kanser tedaviye daha etkili bir şekilde yanıt vermekte ve bu sayede hastaların hayatta kalma oranı artmaktadır (Caplan, 2014).

Hastalığı teşhis etmek için meme biyopsisi, ultrason, mamografi, termografi ve ince iğne aspirasyon sitolojisi dahil olmak üzere çeşitli yöntemler kullanılmaktadır. Mamografi, göğüs kanseri taraması ve tespiti için günümüzde kullanılan en etkili ve popüler yöntemlerden biridir (Sree vd., 2011). Geleneksel yöntemler bazen teşhis için yeterli olmamakta ve hekimlerin son kararlarını vermeden önce daha fazla test yapmaları gerekmektedir. Literatürde; geleneksel tekniklerin sınırlamalarından kaçınmak, doktorlara ikinci bir görüşle yardımcı olmak ve hastanın hayatına mal olabilecek potansiyel insan hatalarını azaltmak amacıyla makine öğrenmesi teknikleri kullanılmıştır.

Araştırmacılar, sağlık uzmanlarına göğüs kanseri teşhisinde yardımcı olmak için çalışmalarında; Karar Ağaçları (DT), K-En Yakın Komşu (KNN), Naive Bayes (NB), Destek Vektör Makinesi (SVM) gibi çeşitli algoritma ve teknikleri kullanmışlardır. Ancak tek bir tekniğin kullanılması, her zaman yüksek düzeyde doğruluk sağlamamaktadır. Çünkü her bir tekniğin sınıflandırma görevleriyle ilgili birbirlerinden üstün ve zayıf özellikleri vardır. Topluluk öğrenmesi, bireysel sınıflandırıcıların gücünü artırmak ve zayıflıklarını azaltmak için kullanılmaktadır. Topluluk sınıflandırma teknikleri, belirli bir kombinasyon kuralı altında birden fazla sınıflandırıcıyı birleştirme ve daha kısa sürede daha doğru bir teşhis elde etmeye yardımcı olmaktadır.

Bu çalışmada göğüs kanseri teşhisi için; bagging, boosting, voting topluluk yöntemleri uygulanmış ve bu alanda kullanılan temel makine öğrenmesi yöntemleri ile performans karşılaştırılması yapılmıştır. Literatürde yaygın olarak kullanılan klasik makine öğrenmesi yöntemleri karşılaştırılmış ve farklı topluluk öğrenmesi yöntemleri kullanılarak klasik makine öğrenmesi yöntemlerine kıyasla daha yüksek doğruluk oranı elde etmek amaçlanmıştır. Literatürde aynı veri setinin kullanıldığı çalışmalarda uygulanan yöntemlerden farklı olarak; çalışmada birden fazla yükseltme algoritması (Adaboost, Garadient Boosting, XGBoost) ve oylama algoritması (Hard Voting, Soft Voting) uygulanmıştır. Torbalama ve yükseltme topluluk öğrenme yöntemleri; kesinlik, duyarlılık, f-skor ve doğruluk performans kriterlerinin yanı sıra roc eğrileri ve auc skorlarına göre kıyaslanmıştır. Bu çalışmanın, literatürde Wisconsin Diagnostic Breast Cancer veri setinin kullanıldığı çalışmalarla karşılaştırıldığında iyi bir başarı oranı elde ettiği söylenebilmektedir.

**Literatür Taraması**

Aruna ve arkadaşları çalışmalarında; Naive Bayes, SVM-RBF çekirdeği, RBF (Radial Basis Function) sinir ağları, Karar ağaçları J48 ve Simple CART (Classification And Regression Tree) gibi denetimli öğrenme sınıflandırıcılarının performans kriterlerini karşılaştırmışlardır (Aruna vd., 2011). Tüm sınıflandırıcılar; Wisconsin Breast Cancer (WBC), Wisconsin Diagnostic Breast Cancer (WDBC), Breast Tissue verisetleri üzerinde uygulanmıştır. Deneyler WEKA'da yürütülmüştür. SVM-RBF kernel sınıflandırıcısı ile; WBC veri kümesi üzerinde %96.84, WDBC veri kümesi üzerinde %98.06, Breast tissue veri kümesi üzerinde %99 doğruluk oranı ile diğer sınıflandırıcılardan daha iyi performans elde edilmiştir.

Bashir ve arkadaşları çalışmalarında; Naive Bayes, DT-Gini, DT-IG, MBL (Memory Based Learner) ve SVM sınıflandırıcılarının yeni bir kombinasyonunu kullanarak heterojen bir topluluk sınıflandırıcı tanıtmışlardır (Bashir vd., 2015). Deneyler; BCD (Breast Cancer Data), WDBC, WBC ve WPBC (Wisconsin Prognostic Breast Cancer) veri kümeleri üzerinde uygulanmıştır. Veri kümelerini, eğitim ve test seti olarak ayırmak için 10-katlı çapraz doğrulama kullanılmıştır. Önerilen topluluk yöntemi ile; %85.23 doğruluk, %86.18 kesinlik ve %76.68 duyarlılık ortalama oranı elde edilmiştir.

Bazazeh ve Shubair, çalışmalarında göğüs kanseri tespiti ve teşhisi için yaygın olarak kullanılan en popüler makine öğrenimi tekniklerinden üçünü, yani sırasıyla Destek Vektör Makinesi, Rastgele Orman (RF) ve Bayes Ağlarının (BN) performanslarını karşılaştırmışlardır (Bazazeh & Shubair, 2016). Deneyler, Wisconsin Breast Cancer veri kümesi üzerinde yürütülmüştür. Sonuçlar, bayes ağının ortalama duyarlılık (97,1) ve ortalama kesinlik (%97,2) açısından en yüksek performansa sahip olduğunu göstermiştir. Rastgele Orman tekniği ise, diğer iki teknikle karşılaştırıldığında optimum ROC (Receiver Operator Characteristic) performansına (%99.9) ulaşmıştır.

Ahmad ve arkadaşları çalışmalarında, kanserin nüksünü tahmin etmek için Karar Ağaçları, Destek Vektör Makinesi ve Yapay Sinir Ağları (YSA) sınıflandırma teknikleri kullanarak göğüs kanseri verilerini analiz etmiş ve ardından sonuçları karşılaştırmışlardır (Ahmad vd., 2013). Deneyler, Iranian Center for Breast Cancer (ICBC) veri kümesi üzerinde yürütülmüştür. DT, YSA ve SVM ile sırasıyla %93,6, %94,7 ve %95,7 doğruluk elde edilmiştir. SVM sınıflandırma modeli, göğüs kanseri nüksünü en az hata oranı ve en yüksek doğrulukla tahmin etmiştir.

Chaurasia ve Pal çalışmalarında; göğüs kanserini tespit etmek için RepTree (Reduced Error Pruning Tree), RBF Ağı ve Simple Logistic yöntemlerine dayalı bir tanı sistemi önermişlerdir (Chaurasia & Pal, 2014). Test aşamasında, önerilen sistemin performansını değerlendirmek için University Medical Centre (Institute of Oncology, Ljubljana, Yugoslavia) veri tabanına 10 kat çapraz doğrulama yöntemi uygulanmıştır. Önerilen sistem ile, %74.5 doğru sınıflandırma oranı elde edilmiştir.

Punitha ve arkadaşları çalışmalarında, yapay sinir ağında etkili özellik seçimi ve parametre optimizasyonunun paralel olarak işlenmesi için Entegre Yapay Bağışıklık sistemi (Integrated Artificial Immune system- IAIS) ve Yapay Arı Kolonisi (Artificial Bee Colony-ABC) yöntemlerine dayalı göğüs kanseri teşhis yöntemini (IAIS-ABC-CDS) önermişlerdir (Punitha vd., 2021). Wisconsin veri kümesi üzerinde; önerilen IAIS-ABC-CDS yöntemi ile %99.34, yapay sinir ağları ile %99.11 oranında bir ortalama sınıflandırma değeri üretilmiştir.

Dhanya ve arkadaşları, göğüs kanseri tahmini için yeni bir model geliştirmek amacıyla; denetimli makine öğrenmesi algoritmalarının (SVM, NB, KNN, Lojistik Regresyon) bir kombinasyonu ile birlikte mevcut topluluk tekniklerini kullanmışlardır (Dhanya vd., 2020). Topluluk yönteminin performansını artırmak için variance threshold ve f-test özellik seçim teknikleri kullanılmıştır. Topluluk modelinde daha yüksek doğruluk elde etmek için bagging, boosting ve stacking teknikleri kullanılmıştır. F-test özellik seçimi ile birlikte stacking topluluk yönteminin; Wisconsin, WDBC ve MicroRNA verikümeleri için göğüs kanserini tahmin etmede daha iyi performans gösterdiği sonucuna varılmıştır. Stacking topluluk öğrenmesi yöntemi ile; Wisconsin, WDBC ve MicroRNA verikümeleri için sırasıyla %97.14, %97.37, %100 doğruluk elde edilmiştir.

Kumar ve arkadaşları çalışmalarında; SVM, NB ve J48 karar ağacını voting sınıflandırıcı ile birleştirme tekniğini kullanmışlardır (Kumar vd., 2017). Deneyler, Wisconsin Breast Cancer Dataset (WBCD) veri kümesi üzerinde uygulanmıştır. Kullanılan topluluk yöntemi ile, %97.13 doğruluk elde edilmiştir.

Wang ve arkadaşları çalışmalarında; göğüs kanseri teşhisi için WAUCE adında, SVM tabanlı ağırlıklı AUC (Area Under the Receiver Operating Characteristic Curve) topluluk öğrenme modelini önermişlerdir (Wang vd., 2018). Deneyler; Wisconsin Breast Cancer, Wisconsin Diagnostic Breast Cancer ve SEER (Surveillance, Epidemiology, and End Results) Breast Cancer veri kümeleri üzerinde uygulanmıştır. WBC, WDBC ve SEER veri kümeleri üzerinde sırasıyla; %97.10, %97.68, %76.42 doğruluk elde edilmiştir.

Vaka ve arkadaşları, göğüs kanserini tespit etmek için DNNS (Deep Neural Network with Support Value) yöntemini önermişlerdir (Vaka vd., 2020). Deney için, M. G Cancer Hospital & Research Institute 'den elde edilen veri kümesi kullanılmıştır. Önerilen yöntem ile %97.21 doğruluk elde edilmiştir.

Lavanya ve Rani, çalışmalarında hibrit bir yöntem önermişlerdir (Lavanya & Rani, 2012). Önerilen yöntem, özellik seçim yöntemlerinin; veriseti, bagging ve cart decision tree ile birleştirilmesiyle oluşturulmuştur. Çalışmada; Breast Cancer, Breast Cancer Wisconsin (Original) ve Breast Cancer Wisconsin (Diagnostic) verikümeleri kullanılmıştır. Breast Cancer, Breast Cancer Wisconsin (Original) ve Breast Cancer Wisconsin (Diagnostic) verikümeleri için sırasıyla; SVMAttributeEval, PrincipalComponentsAttributeEval, SymmetricUncertAttributesetEval özellik seçim teknikleri kullanılmıştır. Breast Cancer veri kümesi ile %74.47, Breast Cancer Wisconsin (Original) veri kümesi ile %97.85 ve Breast Cancer Wisconsin (Diagnostic) veri kümesi ile %95.96 doğruluk elde edilmiştir.

Assiri ve arkadaşları, çalışmalarında çoğunluk oylama yöntemine dayalı bir topluluk sınıflandırma mekanizması önermişlerdir (Assiri vd., 2020). İlk olarak, son teknoloji makine öğrenimi sınıflandırma algoritmalarının performansı WBCD için değerlendirilmiştir. F3 puanlarına göre en iyi üç sınıflandırıcı seçilmiştir: Lojistik Regresyon, SGD (Stochastic Gradient Descent) Optimizasyonu ile SVM Öğrenimi, MLP (Multilayer Perceptron). Daha sonra; seçilen bu üç sınıflandırıcı, oylama yöntemine dayalı bir topluluk sınıflandırma mekanizması için kullanılmıştır. Çalışma sonunda, önerilen Majority oylama yöntemine dayalı mekanizma ile % 99.42 doğruluk elde edilmiştir.

Rudra ve arkadaşları, göğüs kanseri teşhisinde sınıflandırma işlemi için SVM yöntemini kullanmışlardır. Daha sonra, SVM sınıflandırıcısı ile elde edilen doğruluk oranını arttırmak için Adaboost topluluk yöntemini uygulamışlardır (Rudra vd., 2019). Çalışmada, Coimbra Breast Cancer verikümesi kullanılmıştır. Çalışma sonunda, %92 doğruluk elde edilmiştir.

Gupta ve Garg, çalışmalarında göğüs kanseri teşhisi için K-En Yakın Komşuluk, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, SVM, Adam Gradient Descent Learning ile Derin Öğrenme yöntemlerini kullanmıştır (Gupta & Garg, 2020). Deneyler, Wisconsin Breast Cancer Dataset üzerinde yürütülmüştür. Adam Gradient Descent Learning ile Derin Öğrenme modeli, minimum kayıp ve en yüksek doğruluk oranı (%98.24) elde etmiştir.

Akinnuwesi ve arkadaşları, çalışmalarında göğüs kanserinin risk değerlendirmesi ve erken teşhisi için Temel Bileşen Analizi ve Destek Vektör Makinesi'nin hibrit bir yöntem olarak uygulandığı BC-RAED modelini kullanmışlardır (Akinnuwesi vd., 2020). Verikümesi için, LASUTH Onkoloji Departmanı'ndan seçilen 180 örnek kullanılmıştır. BC-RAED modeli ile; %97.62 doğruluk, % 95.24 duyarlılık ve %100 özgüllük elde edilmiştir.

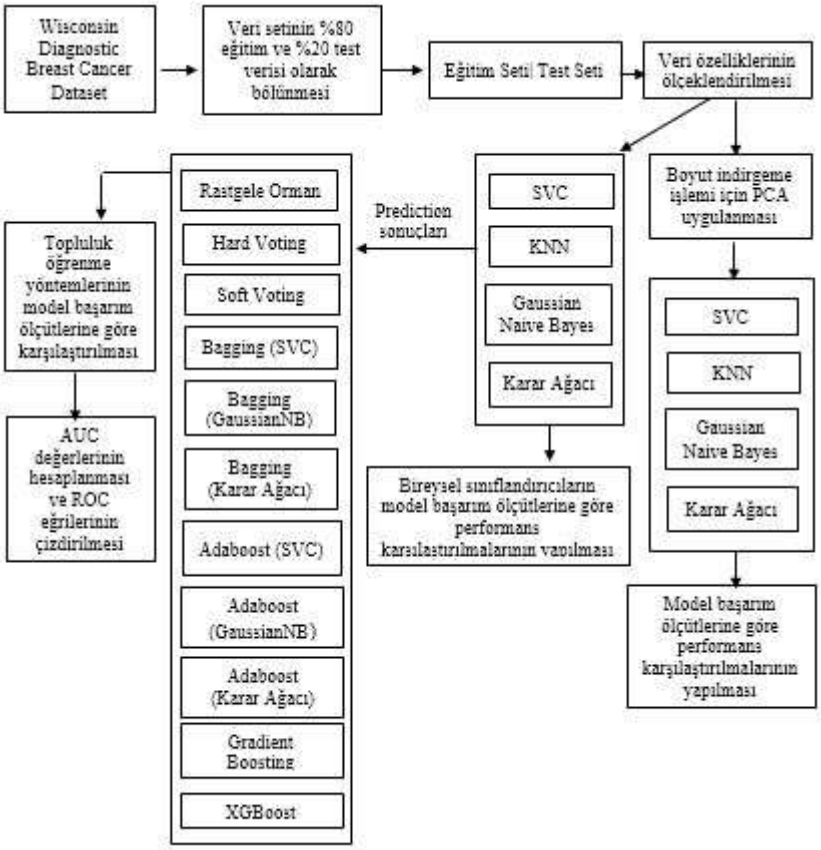
Jabbar, Bayes ağı ve Radyal Tabanlı Fonksiyon ile oluşturulan topluluk modelini kullanarak bir karar destek sistemi oluşturmuştur (Jabbar, 2021). Çalışmada, Wisconsin Breast Cancer verikümesi kullanılmıştır. Önerilen yöntem ile %97 doğruluk elde edilmiştir.

**Yöntem**

Bu çalışmada, göğüs kanseri teşhisi için literatürde yaygın olarak kullanılan Destek Vektör Makinesi, K-En Yakın Komşuluk, Naive Bayes ve Karar Ağaçları makine öğrenmesi yöntemlerinin performanslarını arttırmak için Rastgele Orman, Oylama (Voting), Torbalama (Bagging) ve Yükseltme (Boosting) topluluk öğrenme (ensemble methods) yöntemleri uygulanmış ve yöntemlerin karşılaştırmalı performans analizleri yapılmıştır. Önerilen sistem temel olarak beş adımdan oluşmaktadır:

(i) Wisconsin Diagnostic Breast Cancer veri kümesinin sisteme girdi olarak verilmesi,
(ii) Wisconsin Diagnostic Breast Cancer veri kümesinin %80 eğitim ve %20 test verisi olmak üzere bölünmesi,
(iii) Eğitim ve test verilerinin özellik ölçeklendirme işlemine tabi tutulması,
(iv) SVC, KNN, Gaussian Naive Bayes, Karar Ağacı bireysel sınıflandırıcılarına hem PCA uygulanmadan hem de PCA uygulandıktan sonra kesinlik, duyarlılık, f-skor ve doğruluk model başarım ölçütlerine göre performans karşılaştırılmalarının yapılması,
(v) Bireysel sınıflandırıcıların doğruluk oranını arttırmak için Rastgele Orman, Bagging, Boosting, Hard Voting, Soft Voting topluluk makine öğrenmesi yöntemleri uygulanması ve kesinlik, duyarlılık, f-skor, doğruluk model başarım ölçütlerine göre performans karşılaştırılmalarının yapılması.

Bu çalışma için seçilen SVC, K-En Yakın Komşuluk, Gaussian Naive Bayes ve Karar Ağacı modellerinden oluşan bir sınıflandırıcı grubu üzerinde topluluk yöntemlerinden Hard Voting, Soft Voting; klasik sınıflandırıcı grubun üç üyesi (SVC, Gaussian Naive Bayes, Karar Ağacı) ile Bagging ve Adaptive Boosting, ayrıca Rastgele Orman, Gradient Boosting ve XGBoost algoritmaları uygulanmıştır (Şekil 1). Burada Boosting topluluk yöntemlerinde (Adaptive Boosting, Gradient Boosting, XGBoost), öğrenme oranı = 0,1 olarak belirlenmiştir. Çalışmada, literatürdeki kanser teşhisi çalışmalarında yaygın olarak kullanılan Scikit-Learn kütüphanesi kullanılmıştır.




**Şekil 1.** Önerilen sistemin çalışma mekanizması

### Veri Ön İşleme

Veri özelliklerin (Feature) ölçeklerinin farklı olması, bir veri kümesinin modellemesini olumsuz etkilemektedir. Bu nedenle özelliklerin ölçeklendirilmesi, algoritmaların veri kümeleriyle modellenmesinde önemli bir adımdır.

Özellik Ölçeklendirme, verilerde bulunan bağımsız özellikleri sabit bir aralıkta standartlaştırmak için kullanılan bir tekniktir. Ölçeklendirme işlemi için, sklearn kütüphanesinin StandardScaler fonksiyonu kullanılmıştır. StandardScaler fonksiyonu ile ölçeklendirme işleminde; x özellik, μ ortalama, σ ortalamadan standart sapma olup standardizasyon aşağıdaki şekilde hesaplanmaktadır:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

### Boyut İndirgeme

Veri setinin boyutunu en aza indirmek için Temel bileşen analizi (PCA) kullanılmıştır. Temel bileşen analizinin temel amacı, çok değişkenli bir veri kümesinin boyutluluğunu azaltmak ve verideki temel özellikleri bularak daha az sayıda değişkenle daha büyük veri setini özetlemeyi sağlamaktır. PCA, bu amaca ulaşmak için ana bileşenler adı verilen yeni değişkenleri hesaplamaktadır. Bu değişkenler, orijinal değişkenlerin doğrusal kombinasyonları olarak elde edilmektedir. Bu çalışmada, temel bileşen analizi hem eğitim hem de test verisine uygulanmıştır.

**Sınıflandırma Yöntemleri**

**Destek Vektör Makinesi**

Destek Vektör Makinesi, kanser tanı alanında yaygın olarak uygulanan denetimli makine öğrenmesi sınıflandırma tekniklerinden biridir. Tüm sınıflardan destek vektörleri olarak bilinen kritik örnekleri seçip bu destek vektörlerini kullanarak sınıfları olabildiğince geniş bir şekilde bölen doğrusal bir işlev oluşturarak sınıfları ayırır. Veri kümesini en iyi şekilde iki sınıfa ayıran bir hiper düzlem bulma fikrine dayanmaktadır (Williams, 2011). Çalışmada, sklearn kütüphanesinin Destek Vektör Makinesi için geliştirilmiş SVC sınıfı kullanılmıştır. SVC; kritik örnek noktaları ve hiper düzlem arasındaki mesafeyi maksimize ederek, farklı sınıfları ayırmak için en iyi hiper düzlemi bulmaya çalışmaktadır.

**K-En Yakın Komşu**

K-En Yakın Komşu algoritması, yeni veri noktalarının değerlerini tahmin etmek için "özellik benzerliği" kullanır. KNN, eğitim setindeki veri noktalarıyla olan benzerlik ölçüsüne bağlı olarak yeni veri noktalarını sınıflandırır. KNN modeli geliştirilirken en yakın komşu sayısı, k ve veri noktaları arasındaki mesafe olmak üzere iki tasarım parametresi bulunmaktadır. Veri noktaları arasındaki mesafe; Öklid, Manhattan veya Minkowski uzaklık bağıntıları ile hesaplanabilir (Saçlı vd., 2019).

**Naive Bayes**

Naive Bayes sınıflandırıcısı, olasılıksal bir sınıflandırıcıdır ve her bir özelliğin hedef sınıfa bağımsız ve eşit bir katkı yaptığını varsaymaktadır. Bu yöntem, Bayes teoremini ve koşullu bağımsızlık varsayımını kullanarak yeni veriler için tahminler yapmak için kullanılabilir (Mitchell, 2005). Tahmin ediciler sürekli bir değer aldığında ve ayrık olmadığında, bu değerlerin bir gauss dağılımından örneklendiğini varsayılmaktadır.

**Karar Ağaçları**

Karar ağaçları; kök, iç karar düğümleri, dallar ve yapraklardan oluşan ağaç yapısı şeklinde bir sınıflandırma algoritmasıdır. Veri kümesini en iyi ayıran özellik kök olarak adlandırılıp ağacın en üstünde yer almaktadır. Öngörülen sınıf değeri olan bir yaprak düğümüne ulaşıncaya kadar özellik değerlerini ağacın diğer iç düğümleriyle karşılaştırma işlemi yapılmaktadır. ID3, C4.5 gibi karar ağacı algoritmalarında, her nitelik için bilgi kazancı ölçülerek bölme kriteri belirlenmektedir (Quinlan, 1993).

**Topluluk Yöntemleri**

**Rastgele Orman**

Rastgele Orman, daha iyi bir tahmin performansı elde etmek için birkaç karar ağacının birden çok modelini kullanan bir topluluk sınıflandırıcıdır. Yeni bir örneği sınıflandırmak için, her karar ağacı girdi verileri için bir sınıflandırma sağlar; rastgele orman sınıflandırmaları toplar ve sonuç olarak en çok oylanan tahmini seçer. RF sınıflandırıcı; aşırı öğrenme(overfitting) sorunu olmaksızın, doğruluk açısından diğer sınıflandırma yöntemlerinin çoğundan daha iyi performans gösterme eğilimindedir (Misra & Li, 2019). Random Forest yönteminin çalışma mekanizması, Şekil 2'de verilmiştir.

Adım 1: Eğitim setinden rastgele n örnek seçilir.

Adım 2: n örneği temel alan bir karar ağacı oluşturulur.

Adım 3: Adım 1 ve Adım 2, k kere tekrar edilir, where k = ağaç sayısı

Adım 4: Sınıf etiketini oy çokluğuyla atamak için; her ağacın tahmini toplanır.

**Şekil 2.** Random Forest modeli çalışma mekanizması

**Oylama (Voting)**

Oylama Sınıflandırıcısı, birden çok model oluşturan ve daha sonra bunları daha iyi sonuçlar elde etmek için birleştiren bir topluluk yöntemidir. Oylama Topluluğu yöntemi için; çoğunluğa dayalı (hard) oylama ve olasılığa dayalı (soft) oylama olmak üzere başlıca iki yaklaşım bulunmaktadır. Olasılığa dayalı oylamada; her bir sınıflandırıcı, belirli bir veri noktasının belirli bir hedef sınıfa ait olduğuna dair bir olasılık değeri sağlamaktadır. Tahminler, sınıflandırıcının önemi ile ağırlıklandırılmakta ve toplanmaktadır. En fazla ağırlıklı olasılık toplamına sahip olan hedef etiket, oylamayı kazanır. Çoğunluğa dayalı oylamada; her model, her bir test örneği için oylama yapmakta ve oyların yarısından fazlasını alan, son çıktı tahmini olmaktadır. Tahminlerin hiçbiri oyların yarısından fazlasını alamazsa, topluluk yaklaşımının bu durum için istikrarlı bir tahmin yapamayacağı varsayabilir. Ayrı ayrı modeller oluşturmak ve bunların her biri için doğruluğu bulmak yerine, çok sayıda modelden oluşan bir grup üzerinde eğitim veren ve her çıktı sınıfı için toplam oylama çoğunluğuna göre çıktıyı tahmin eden tek bir model oluşturulmaktadır. SVC, KNN, Gaussian Naive Bayes ve Karar Ağacı modellerinden oluşan bir grup üzerinde oylama sınıflandırıcısı uygulanmıştır.

Çoğunluğa dayalı oylamada; sınıf etiketi Y, her bir Cj sınıflandırıcısının çoğunluk oyu ile tahmin edilmektedir:

$$Y = mode\{C_1(x), C_2(x), \dots, C_m(x)\} \qquad [j = 1,2,3,\dots,m] \tag{2}$$

Olasılığa dayalı oylamada; sınıf etiketi, her bir i sınıflandırıcısının Pij tahmini olasılıklarına bağlı olarak tahmin edilmektedir. Wj, j. sınıflandırıcıyla dağıtılabilen ağırlıktır.

$$Y = argmax_i \sum_{j=1}^{m} W_j P_{ij} \qquad [j = 1, 2,3, \dots, m; i = 1, 2, 3, \dots,n] \tag{3}$$

**Torbalama (Bagging)**

Torbalama sınıflandırıcısı, son bir tahmin oluşturmak için rastgele oluşturulmuş eğitim setlerinin sınıflandırmalarını birleştirerek çalışmaktadır. Torbalama, sınırlı veriye sahip olunduğunda etkili bir yöntemdir ve örnekleri kullanarak birçok örnek üzerinde skorları bir araya getirerek bir tahmin elde edilmektedir. Karar ağacına dayalı torbalama topluluğu sınıflandırıcı, birden çok karar ağacı kullanmaktadır. Burada her bir karar ağacı zayıf öğrenen iken, tüm karar ağaçları birlikte alındığında güçlü öğrenendir. Yeni bir örneğin sınıflandırılması gerektiğinde, tüm ağaçlar bir sınıfa oy verir. Yeni örneğin sınıfı için nihai tahmin, maksimum oyu olan sınıf tarafından kazanılmaktadır. Bu çalışmada; SVC, Gaussian Naive Bayes, Karar Ağacı algoritmalarıyla ayrı ayrı torbalama sınıflandırıcıları oluşturulmuştur.

$$c_i = argmax_i\{\frac{1}{N} \sum_{T=1}^{N} h_T(x, c_i)\} \qquad [T = 1, \dots, N] \tag{4}$$

Burada $h_T$ tahmin modelini, T bootstrap eğitim alt kümesini temsil etmektedir. Ortalama sınıf puanı; tüm $h_T$ (x, $c_i$) modellerinde hesaplanan, en yüksek ortalama puanla tahmin edilen sınıfı almaktadır.

**Yükseltme (Boosting)**

Yükseltme yöntemi, sınıflandırma ve regresyon problemlerini çözmek için kullanılan bir tekniktir; bireysel performansı zayıf kabul edilebilen bir dizi modeli birleştirir, bireysel tahminler bir araya getirildiğinde yüksek doğrulukta bir model elde edilir. Yükseltme yöntemi zayıf öğrenicileri sırayla eğitir, her bir öğrenici bir öncekini düzeltmeye çalışmaktadır. Veri setini tahmin ederek başlanılır ve her gözleme eşit ağırlık verilir. Sonraki yinelemelerde, önceki gözlemde yanlış tahmin edilen gözlemlere daha fazla ağırlık verilerek önceki yinelemelerin hatalarından öğrenme işlemi yapılmaktadır. Yinelemeli bir süreç olarak, model sayısında veya doğrulukta bir sınıra ulaşılana kadar öğrenci eklemeye devam edilir.

$$H_m(x) = \sum_{m=1}^{M} y_m \; h_m(x) \qquad [m=1,\ldots,M] \qquad (5)$$

Burada $y_m$, m. hedef değerini temsil etmektedir. $h_m(x)$ temel öğreniciyi temel etmektedir ve her aşamada kayıp fonksiyonunu en aza indirmek için seçilmektedir.

Bu çalışmada Adaptive Boosting (AdaBoost), Gradient Boosting (GBM) ve XGBoost (XGBM) yükseltme algoritmaları kullanılmıştır. Gradient Boosting algoritmasında da AdaBoost algoritmasında olduğu gibi, bir topluluğa sırayla öğreniciler eklenir, her bir öğrenici bir öncekini düzeltmeye çalışır. Bununla birlikte, AdaBoost gibi her yinelemede her yanlış sınıflandırılmış gözlem için ağırlıkları değiştirmek yerine, Gradient Boosting yöntemi yeni tahmin ediciyi önceki tahmin edici tarafından yapılan hataları göz önüne alarak oluşturur. Kayıp fonksiyonunun (loss function) negatif gradyanı ile maksimum olarak ilişkilendirilebilecek yeni öğreniciler oluşturulmaktadır. XGBoost, gradient boosting algoritmasının hız ve performans açısından daha gelişmiş bir uygulamasıdır. Paralel işleme, XGBoost'un önemli iyileştirmelerinden biridir. Çalışmada; SVC, Gaussian Naive Bayes, Karar Ağacı algoritmalarıyla ayrı ayrı yükseltme sınıflandırıcıları oluşturulmuştur.

**Materyal ve Değerlendirme Ölçütleri**

**Veri Seti**

Bu çalışmada, Wisconsin Diagnostic Breast Cancer veri kümesi kullanılmıştır. Veri seti, California Üniversitesi'nin Irvine Makine Öğrenimi Havuzundan (UCI) alınmıştır (Wolberg vd., 1992). Wisconsin Diagnostic Breast Cancer veri kümesi; 569 örnek ve 32 özellikten (id, tanı, 30 gerçek-değerli özellik) oluşmaktadır. Özellikler, bir göğüs kitlesinin ince iğne aspirasyonunun (FNA) dijitalleştirilmiş bir görüntüsünden hesaplanmıştır. Veri kümesindeki toplam 569 göğüs kanseri verisinin 212'si kötü huylu (malignant), 357'si iyi huyludur (benign).

Özellik bilgileri:

1. Örnekler için kullanılan sıra numarası (ID)
2. Kanser verisinin iyi huylu veya kötü huylu olduğunu belirten tanı sınıfı (M = malignant, B = benign)
3. Her bir tümör hücresi çekirdeği üzerinden; kitle görüntülerinin sayısallaştırılması ile, 10 gerçek-değerli özellik ölçümlenmiştir. Ölçümlenen gerçek değerli özellikler şunlardır:
    - Yarıçap (Çekirdek merkezinden çevredeki noktalara olan mesafelerin ortalaması)
    - Doku (Gri ölçek değerlerinin standart sapması)
    - Çevre
    - Alan
    - Pürüzsüzlük (Yarıçap uzunluklarında yerel değişim)

- Kompaktlık (çevre² / alan- 1.0)
- İçbükeylik
- İçbükey Noktalar
- Simetri
- Fraktal Boyut

Bu özelliklerin her biri için ortalama, standart hata ve en kötü değer (ilgili özelliğin aldığı en büyük üç değerin ortalaması) hesaplanmış ve sonuçta 30 adet tanımlayıcı özellik elde edilmiştir.

**Model Başarım Ölçütleri**

Sınıflandırıcı doğruluğu (accuracy), sınıflandırıcının vakaları ne kadar iyi tahmin edebildiğinin bir ölçüsüdür. Kesinlik (Precision), yapılan doğru pozitif tahminlerin sayısını ölçen bir ölçüdür. Duyarlılık (Recall), yapılabilecek tüm olumlu tahminlerden yapılan doğru pozitif tahminlerin sayısını ölçen bir metriktir. F-Skor, hem kesinlik hem de duyarlılığa aynı ağırlığı vererek tek bir ölçü halinde birleştirmenin bir yolunu sağlamaktadır. ROC grafiğinin altındaki alan (AUC), sınıflandırıcının performansını yansıtmaktadır. Grafiğin altındaki alanı, grafiğin toplam alanına bölerek elde edilmektedir.

Gerçek pozitif (TP), yanlış pozitif (FP), gerçek negatif (TN) ve yanlış negatif (FN) sayıları göz önüne alındığında, aşağıdaki performans değerlendirme ölçümleri hesaplanmıştır:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (9)$$

**Bulgular**

Göğüs kanseri teşhisi için literatürde yaygın olarak kullanılan makine öğrenmesi yöntemlerinin doğruluk yüzdelerine göre performans karşılaştırmaları yapılmıştır. SVC, KNN, Gaussian Naive Bayes ve Karar Ağacı bireysel sınıflandırıcılarının standardizasyon + temel bileşen analizi (PCA) öncesi ve sonrası doğruluk yüzdeleri Tablo 1'de verilmiştir.

**Tablo 1:** Sınıflandırıcılara Göre Standardizasyon+Temel Bileşen Analizi (PCA) Öncesi ve Sonrası Doğruluk

| Sınıflandırıcı | Standardizasyon ve PCA Öncesi | Standardizasyon ve PCA Sonrası |
|---|---|---|
| SVC | %93.86 | %93.71 |
| KNN | %93.86 | %90.21 |
| Gaussian Naive Bayes | %94.74 | %91.61 |
| Karar Ağacı | %92.98 | %86.01 |

Tablo 1'e göre standardizasyon + temel bileşen analizi öncesi en yüksek doğruluk, Gaussian Naive Bayes yöntemiyle %94.74 oranında elde edirken; standardizasyon + temel bileşen analizi sonrasında en yüksek doğruluk, SVC yöntemiyle %93.71 oranında elde edilmiştir. SVC, KNN, Gaussian Naive Bayes ve Karar Ağacı sınıflandırıcılarının; standardizasyon ön işleme adımı sonrası model başarım ölçüt değerleri Tablo 2'de karşılaştırılmıştır.

**Tablo 2:** Sınıflandırıcılara Göre Standardizasyon Ön İşleme İşlemi Sonrası Model Başarım Ölçütleri

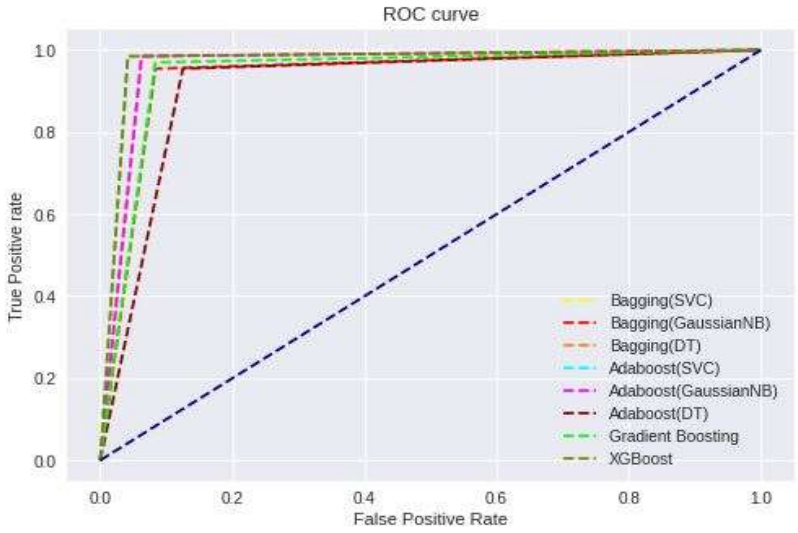
| Sınıflandırıcı | Kesinlik | Duyarlılık | F-Skor | Doğruluk |
|---|---|---|---|---|
| SVC | 0.96 | 0.98 | 0.96 | %96.49 |
| KNN | 0.94 | 0.97 | 0.95 | %94.74 |
| Gaussian Naive Bayes | 0.94 | 0.95 | 0.94 | %93.86 |
| Karar Ağacı | 0.93 | 0.95 | 0.93 | %92.98 |

Tablo 2'ye göre; standardizasyon ön işleme adımı sonrası en yüksek doğruluk, SVC sınıflandırıcısıyla elde edilmiştir (%96.49). Tablo 1 ve Tablo 2'de verilen doğruluk yüzdeleri birlikte değerlendirildiğinde; temel bileşen analizi uygulamanın, sınıflandırma doğruluk yüzdesini azalttığı gözlemlenmiştir. Bu durumun, temel bileşen analizinin sınıflandırma için önemli olan bazı bilgileri kaybettiğinden kaynaklandığı düşünülmüştür. Topluluk öğrenme yöntemlerinin; kesinlik, duyarlılık, f-skor, doğruluk model başarım ölçütlerine göre karşılaştırılması Tablo 3'te yer almaktadır.

**Tablo 3:** Topluluk Öğrenme Yöntemlerinin Model Başarım Ölçütlerine Göre Karşılaştırılması

| Topluluk Yöntemleri | Kesinlik | Duyarlılık | F-Skor | Doğruluk |
|---|---|---|---|---|
| Rastgele Orman | 0.96 | 0.98 | 0.96 | %96.49 |
| Hard Voting | 0.96 | 0.98 | 0.96 | %96.49 |
| Soft Voting | 0.97 | 0.98 | 0.97 | %97.37 |
| Bagging (SVC) | 0.97 | 0.98 | 0.97 | %97.37 |
| Bagging (GaussianNB) | 0.94 | 0.95 | 0.94 | %93.86 |
| Bagging (Karar Ağacı) | 0.96 | 0.98 | 0.96 | %96.49 |
| Adaboost (SVC) | 0.97 | 0.98 | 0.97 | %97.37 |
| Adaboost (GaussianNB) | 0.96 | 0.98 | 0.96 | %96.49 |
| Adaboost (Karar Ağacı) | 0.91 | 0.95 | 0.92 | %92.11 |
| Gradient Boosting | 0.94 | 0.97 | 0.95 | %94.74 |
| XGBoost | 0.97 | 0.98 | 0.97 | %97.37 |

Tablo 3'e göre topluluk öğrenme yöntemleri arasında en yüksek doğruluk yüzdesi; Soft Voting, Bagging (SVC), Adaboost (SVC) ve XGBoost yöntemleri ile elde edilmiştir (%97.37). Torbalama ve Yükseltme topluluk öğrenme yöntemlerinin ROC eğrileri, Şekil 3'te verilmiştir.




**Şekil 3.** Torbalama ve Yükseltme topluluk öğrenme yöntemlerinin ROC eğrileri

Torbalama ve Yükseltme topluluk öğrenme yöntemlerinin AUC skorları, Tablo 4'te verilmiştir.

**Tablo 4:** Torbalama ve Yükseltme Topluluk Öğrenme Yöntemlerinin AUC Skorları

| Topluluk Yöntemleri | AUC |
|---|---|
| Bagging (SVC) | %97.16 |
| Bagging (GaussianNB) | %93.56 |
| Bagging (Karar Ağacı) | %96.12 |
| Adaboost (SVC) | %97.16 |
| Adaboost (GaussianNB) | %96.12 |
| Adaboost (Karar Ağacı) | %91.48 |
| Gradient Boosting | %94.32 |
| XGBoost | %97.16 |

Tablo 4'e göre en yüksek AUC değeri; Bagging (SVC), Adaboost (SVC) ve XGBoost yöntemleri ile elde edilmiştir (%97.16).

**Sonuç**

Makine öğrenimi teknikleri tıp alanında yaygın olarak kullanılmaktadır ve hekimlere mevcut verileri analiz etmede yardımcı olan yararlı bir teşhis aracı olarak hizmet etmektedir.

Bu çalışmada; göğüs kanseri teşhisi için voting, bagging ve boosting topluluk öğrenme yöntemleri uygulanmıştır. Ayrıca, göğüs kanseri teşhisinde yaygın olarak kullanılan makine öğrenmesi yöntemleri incelenmiş ve doğruluk yüzdelerine göre birbirleriyle karşılaştırılmıştır. Araştırılan tekniklerin performans karşılaştırması Wisconsin Diagnostic Breast Cancer veri kümesi kullanılarak yapılmıştır.

Temel bileşen analizi öncesi en yüksek doğruluk yüzdesi, Gaussian Naive Bayes yöntemi ile elde edilmiştir (%94.74). Temel bileşen analizi sonrası en yüksek doğruluk yüzdesi, SVC yöntemi ile elde edilmiştir (%93.71). Standardizasyon ön işleme işlemi sonrası, bireysel sınıflandırıcılar arasında en yüksek doğruluk yüzdesi SVC yöntemi ile elde edilmiştir (%96.49).

Rastgele Orman yöntemi ile %96.49 doğruluk elde edilmiştir. Hard Oylama topluluk öğrenme yöntemi ile %96.49, Soft Oylama topluluk öğrenme yöntemi ile %97.37 doğruluk elde edilmiştir (Tablo 3). Torbalama topluluk yöntemi ile edilen en yüksek doğruluk yüzdesinin %97.37 olduğu gözlenmiştir (Tablo 3). Adaboost (SVC) ile %97.37, Adaboost (GaussianNB) ile %96.49, Adaboost (Karar Ağacı) ile %92.11 doğruluk elde edilmiştir. Gradient Boosting ve XGBoost topluluk yöntemleri ile sırasıyla; %94.74 ve %97.37 doğruluk elde edilmiştir. Uygulanan tüm yöntemler arasında en yüksek doğruluk yüzdesi; Soft Voting, Bagging (SVC), Adaboost (SVC) ve XGBoost topluluk öğrenme yöntemleri ile elde edilmiştir (%97.37).

Literatürde Wisconsin Diagnostic Breast Cancer veri kümesi ile birçok makine öğrenme algoritmaları kullanılarak yüksek doğruluk değerleri elde edilmesine rağmen iyileştirilmiş algoritmaların geliştirilmesi hala gereklidir. Çalışmanın literatürde yapılan benzer çalışmalar ile (genellikle benzer çalışmalarda sadece doğruluk değerleri verildiğinden) doğruluk model başarım ölçütüne göre karşılaştırılması Tablo 5'te verilmiştir.

**Tablo 5:** Çalışmanın WDBC Veri Setinin Kullanıldığı Literatürdeki Benzer Çalışmalar ile Doğruluk Başarım Ölçütüne Göre Karşılaştırılması

| Araştırmacılar | Sınıflandırma Yöntemi | Veri Seti | Doğruluk |
|---|---|---|---|
| Lavanya & Rani (2012) | Cart karar ağacını bagging sınıflandırıcı ile birleştirme tekniği | Breast Cancer Veri Kümesi | %74.47 |
| | | Breast Cancer Wisconsin (Original) | %97.85 |
| | | **Breast Cancer Wisconsin (Diagnostic)** | **%95.96** |
| Bashir vd. (2015) | Naive Bayes, DT-Gini, DT-IG, MBL SVM sınıflandırıcılarının yeni bir kombinasyonu kullanılarak oluşturulan heterojen bir topluluk sınıflandırıcı | Breast Cancer Data, **Wisconsin Diagnostic Breast Cancer**, Wisconsin Breast Cancer ve Wisconsin Prognostic Breast Cancer Veri Setleri | **%85.23** |
| Wang vd. (2018) | WAUCE adında, SVM tabanlı ağırlıklı AUC topluluk öğrenme modeli | Wisconsin Breast Cancer | %97.10 |
| | | **Wisconsin Diagnostic Breast Cancer** | **%97.68** |
| | | SEER Breast Cancer | %76.42 |
| Dhanya vd. (2020) | Bagging (Naive Bayes) | Wisconsin Veri Seti | Wisconsin Diagnostic Breast Cancer Veri Seti ile; |
| | Bagging (SVM linear) | **Wisconsin Diagnostic Breast Cancer** | Naive Bayes: %92.98 |
| | Bagging (SVM rbf) | MicroRNA | SVM linear: %95.61 |
| | Boosting (Naive Bayes) | | SVM rbf: %58.77 |
| | Boosting (SVM linear) | | Bagging (Naive Bayes): %92.11 |
| | Boosting (SVM rbf) | | |
| | Stacking | | |

| | | | | |
|---|---|---|---|---|
| | | | | Bagging (SVM linear): %95.61<br>Bagging (SVM rbf): %58.77<br>Boosting (Naive Bayes): %94.74<br>Boosting (SVM linear): %95.61<br>Boosting (SVM rbf): %58.77<br>Stacking: %97.37 |
| Bu çalışmada | SVC | **Wisconsin Diagnostic Breast Cancer** | | %96.49 |
| | KNN | | | %94.74 |
| | Gaussian Naive Bayes | | | %93.86 |
| | Karar Ağacı | | | %92.98 |
| | Rastgele Orman | | | %96.49 |
| | Hard Voting | | | %96.49 |
| | Soft Voting | | | %97.37 |
| | Bagging (SVC) | | | %97.37 |
| | Bagging (GaussianNB) | | | %93.86 |
| | Bagging (Karar Ağacı) | | | %96.49 |
| | Adaboost (SVC) | | | %97.37 |
| | Adaboost (GaussianNB) | | | %96.49 |
| | Adaboost (Karar Ağacı) | | | %92.11 |
| | Gradient Boosting | | | %94.74 |
| | XGBoost | | | %97.37 |

Tablo 5'e göre, WDBC veri setinin kullanıldığı literatürdeki çalışmalardan farklı olarak; çalışmada birden fazla yükseltme algoritması (Adaboost, Garadient Boosting, XGBoost) ve oylama algoritması (Hard Voting, Soft Voting) uygulanmıştır. Literatürde yapılan çalışmalarda da gözlendiği gibi, topluluk öğrenmesi yöntemlerinin bireysel sınıflandırıcıların doğruluk oranını arttırdığı görülmüştür (Dhanya vd., 2020; Assiri vd., 2020). Bu çalışmada, Wisconsin Diagnostic Breast Cancer veri seti kullanılmış ve bireysel sınıflandırıcılar arasında en yüksek doğruluk oranı SVC sınıflandırıcısı ile elde edilmiştir (%96.49). Bu çalışmada seçilen modellerden SVC tabanlı Bagging ve Adaboost topluluk öğrenme yöntemleri, GaussianNB tabanlı Bagging ve Adaboost topluluk öğrenme yöntemlerine göre daha yüksek doğruluklu bir sınıflandırma performansı göstermişlerdir. Çalışma sonunda en yüksek doğruluk; Soft Voting, Bagging (SVC), Adaboost (SVC) ve XGBoost topluluk öğrenme yöntemleri ile elde edilmiştir (%97.37). Tablo 5'te yer alan, WDBC veri setinin kullanıldığı literatürdeki benzer çalışmaların sınıflandırma doğrulukları %58.77 ile %97.68 arasında değişim göstermektedir. Çalışmada kullanılan Soft Voting, Bagging (SVC), Adaboost (SVC) ve XGBoost topluluk öğrenme yöntemleri; elde edilen %97.37'lik doğruluk ile literatürde WDBC veri setinin kullanıldığı çalışmalar arasında iyi bir başarı oranı elde etmiştir.

Sonuç olarak; sağlık uzmanlarına ikinci bir görüş sunarak yardımcı olmak amacıyla, bu tür makine öğrenmesi yöntemleriyle göğüs kanseri teşhisi çalışmalarının geliştirilmesi sağlık alanına önemli bir katkı sağlayacaktır.

**Kaynakça**

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M. & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3. http://dx.doi.org/10.4172/2157-7420.1000124

Akinnuwesi, B. A., Macaulay, B. O. & Aribisala, B. S. (2020). Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. *Informatics in Medicine Unlocked*, *21*, 100459.

Aruna, S., Rajagopalan, S. P. & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology*, *2*, 37-45.

Assiri, A. S., Nazir, S. & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, *6*(6), 39.

Bashir, S., Qamar, U. & Khan, F. H. (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. *Quality & Quantity*, *49*(5), 2061-2076.

Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 1-4.

Caplan, L. (2014). Delay in breast cancer: implications for stage at diagnosis and survival. *Frontiers in public health*, *2*, 87. https://doi.org/10.3389/fpubh.2014.00087

Chaurasia, V. & Pal, S. (2014). Data mining techniques: To predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing IJCSMC*, *3*(1), 10-22.

Dhanya, R. Paul, I. R., Akula, S. S., Sivakumar, M., & Nair, J. J. (2020). F-test feature selection in Stacking ensemble model for breast cancer prediction. *Procedia Computer Science*, *171*, 1561-1570. https://doi.org/10.1016/j.procs.2020.04.167

Gupta, P. & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science*, *171*, 593-601.

International Agency for Research on Cancer. (2013). Latest world cancer statistics Global cancer burden rises to 14.1 million new cases in 2012: Marked increase in breast cancers must be addressed. *World Health Organization*, *12*.

Jabbar, M. A. (2021). Breast Cancer Data Classification Using Ensemble Machine Learning. *Engineering and Applied Science Research*, *48*(1), 65-72.

Kumar, U., Nikhil, M.S., & Sumangali, K. (2017). Prediction of breast cancer using voting classifier technique. In *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, 108-114.

Lavanya, D. & Rani, K. U. (2012). Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, *2*(1), 17-24.

Meo, A. S. (2018). Blood Groups and Breast Cancer. *Pakistan journal of medical sciences*, *34*(6), 1589. https://doi.org/10.12669/pjms.346.16824

Mitchell, T. M. (2005). Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. *Draft book chapter*.

Misra, S. & Li, H. (2019). Noninvasive fracture characterization based on the classification of sonic wave travel times. *Machine Learning for Subsurface Characterization*, 243-287.

Punitha, S., Al-Turjman, F. & Stephan, T. (2021). An automated breast cancer diagnosis using feature selection and parameter optimization in ANN. *Computers & Electrical Engineering*, *90*, 106958. https://doi.org/10.1016/j.compeleceng.2020.106958

Rudra, S., Uddin, M. & Alam, M. M. (2019). Forecasting of breast cancer and diabetes using ensemble learning. *International Journal of Computer Communication and Informatics*, *1*(1), 1-5.

Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, *38*(48), 49.

Saçlı, B., Aydınalp, C., Cansız, G., Joof, S., Yilmaz, T., Çayören, M., Önal, B. & Akduman, I. (2019). Microwave dielectric property based classification of renal calculi: Application of a kNN algorithm. *Computers in biology and medicine*, *112*, 103366. https://doi.org/10.1016/j.compbiomed.2019.103366

Sree, S. V., Ng, E. Y., Acharya, R. U., & Faust, O. (2011). Breast imaging: A survey. *World journal of clinical oncology*, *2*(4), 171–178. https://doi.org/10.5306/wjco.v2.i4.171

Vaka, A. R., Soni, B. & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*, *6*(4), 320-324. https://doi.org/10.1016/j.icte.2020.04.009

Wang, H., Zheng, B., Yoon, S. W. & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, *267*(2), 687-699. https://doi.org/10.1016/j.ejor.2017.12.001

Williams, G. (2011). Descriptive and predictive analytics. In *Data Mining with Rattle and R* (ss. 171-177). Springer. https://doi.org/10.1007/978-1-4419-9890-3_8

Wolberg, W. H., Street, W. N. & Mangasarian, O. L. (1992). Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository*.