
GAZİ ÜNİVERSİTESİ
Gazi
1926

YAPILANDIRILMIŞ VERİ KÜMELERİNİN GAN İLE ARTIRILMASININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ ÜZERİNDEKİ ETKİSİNİN ANALİZİ

Amel Sulaimna Mandan SALIHI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR BİLİMLERİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

ŞUBAT 2024

Amel Sulaimna Mandan SALIHI tarafından hazırlanan "YAPILANDIRILMIŞ VERİ KÜMELERİNİN GAN İLE ARTIRILMASININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ ÜZERİNDEKİ ETKİSİNİN ANALİZİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi bilgisayar bilimleri Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Oktay YILDIZ

Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ..……

Başkan: Doç. Dr. Ömer Özgür TANRIÖVER

Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Doç. Dr. Ümit ATİLA

Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 08/02/2024

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

…….……

Prof. Dr. Aslıhan TÜFEKCİ

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Amel Sulaimna Mandan SALIHI

08/02/2024

YAPILANDIRILMIŞ VERİ KÜMELERİNİN GAN İLE ARTIRILMASININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ ÜZERİNDEKİ ETKİSİNİN ANALİZİ

(Yüksek Lisans Tezi)

Amel Sulaiman Mandan SALIHI

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

ŞUBAT 2024

ÖZET

Makine öğrenmesi yöntemleri genellikle belirli bir amacı gerçekleştirmek üzere önceden sunulan veri kümelerini kullanır. Ancak, yetersiz veya temsil özelliği zayıf veri kümeleri modelin öğrenme başarısını olumsuz yönde etkiler. Yetersiz veya temsil edilmemiş örnek problemini aşmak için veri artırma yöntemleri sıklıkla kullanılmaktadır. Veri artırımı, sentetik veri oluşturarak bir veri kümesinin boyutunu ve çeşitliliğini artırmak için kullanılan bir metodolojidir. Özellikle orijinal veri kümesi yetersiz veya dengesizlik gösteriyorsa kullanılır. Bu tez çalışmasında genele açık iki farklı veri kümesi üzerinde Çekişmeli Üretici Ağ (GAN) ile veri artırımı gerçekleştirilmiş; elde edilen yeni veri kümesinin tahmin modellerindeki başarıma etkisi incelenmiştir.

Deneysel çalışmada kullanılan ilk veri kümesi reklam tıklama tahmini amacıyla araştırmacılar tarafından sıklıkla kullanılan Müşteri Reklam Tıklama Tahmini veri kümesidir. Diğeri sağlık alanında kalp hastalığı tanısından kullanılan Koroner Arter Hastalığı veri kümesidir.Deneysel çalışmalarda, Çekişmeli Üretici Ağlar her iki problem için de ek eğitim örnekleri üretmek için kullanılmıştır. Daha sonra Çekişmeli Üretici Ağ tabanlı veri artırımının varlığı veya yokluğu dikkate alınarak birkaç makine öğrenimi yönteminin doğruluğu değerlendirilmiştir. Ayrıca, koroner arter hastalığı tahmini için daha başarılı olan özellikleri çıkarmak için özellik seçimi yaklaşımları kullanılmıştır. Bulgulara göre, algoritmaların çoğunluğu, Çekişmeli Üretici Ağ kullanımı ile veri artırımı sonrasında doğruluk, hassasiyet ve özgüllükte belirgin bir iyileşme göstermiştir. Bu sonuçlar, oluşturulan verinin, pozitif ve negatif olaylar arasındaki farkı doğru bir şekilde ayırt etme kapasitesini etkili bir şekilde artırdığını göstermektedir. Sonuçlara göre, Çekişmeli Üretici Ağ tabanlı veri artırımının tüm modeller için çeşitli derecelerde iyileşmeye neden olduğu bulunmuştur. Ayrıca, özellik seçimi stratejilerinin kullanımının koroner arter hastalığı tahmininin etkililiğini artırdığı gösterilmiştir. Bu çalışmada Çekişmeli Üretici Ağların veri artırma yöntemi olarak çeşitli alanlarda kullanılabileceği gösterilmiştir.

Bilim Kodu	:	92429
Anahtar Kelimeler	:	Veri Artırımı, Çekişmeli Üretici Ağ , Makine Öğrenimi
Sayfa Adedi	:	80
Danışman	:	Doç.Dr. Oktay YILDIZ

ANALYSİS OF THE IMPACT OF AUGMENTİNG STRUCTURED DATASETS WİTH GAN ON MACHİNE LEARNİNG METHODS

(M. Sc. Thesis)

Amel Sulaiman Mandan SALIHI

GAZİ UNIVERSITY

INSTITUTE OF INFORMATICS

FEBRUARY 2024

ABSTRACT

Machine learning algorithms commonly leverage present datasets to accomplish a certain objective. Insufficient datasets might have a detrimental effect on the model's learning performance. Data augmentation approaches are commonly used to address the issue of inadequate instances. Data augmentation is a technique employed to enhance the amount and variety of a dataset by creating artificial data. It is especially beneficial when the initial dataset is poor or shows an imbalance. This thesis explores the application of Generative Adversarial Networks (GAN) for data augmentation. Two publically accessible datasets were used for this purpose. The main objective was to examine how the supplemented dataset affects the performance of prediction models. The first set of data employed in the experimental investigation is the Customer Ad Click Prediction dataset, which is widely utilized by academics for the specific aim of ad click prediction. The second dataset is the Coronary Artery Disease dataset, which is utilized in the health field for the purpose of diagnosing heart disease. Generative Adversarial Networks were used in experimental research to create additional training examples for both issues. Afterward, the precision of various machine learning techniques was evaluated by considering whether GAN-based data augmentation was present or not. Moreover, feature selection methodologies were employed to identify characteristics that demonstrate higher efficacy in predicting coronary artery disease. The findings indicate that most algorithms had a substantial enhancement in accuracy, precision, and specificity following the implementation of GAN-based data augmentation. The results demonstrate that the obtained data significantly improves the ability to differentiate between positive and negative instances. All models improved on different scales with GAN-based data augmentation. Feature selection algorithms also enhance coronary artery disease prediction. This study shows how Generative Adversarial Networks may enhance data across domains.

Science Code	:	92429
Key Words	:	Data Augmentation, Generative Adversarial Networks, Machine Learning
Page Number	:	80
Supervisor	:	Assist. Prof. Dr. Oktay YILDIZ

TEŞEKKÜR

Tez çalışmalarım boyunca bana yol gösteren, bilgi ve birikimlerini aktaran, her zaman yanımda olan ve desteklerini esirgemeyen danışman hocam Doç. Dr. Oktay Yıldız'a sonsuz teşekkür ve saygılarımı sunarım.

Hayatımın her aşamasında beni destekleyen, her zaman yanımda olan aileme, özellikle de bu süreçte büyük fedakarlıklar yapan kocama ve çocuğuma en içten dileklerimle teşekkür ederim.

Tez çalışmam boyunca yardımlarını hiçbir zaman esirgemeyen, değerli çalışma arkadaşlarıma da teşekkürü bir borç bilirim.

Bu çalışmanın gerçekleşmesinde emeği geçen herkese şükranlarımı sunarım.

İÇİNDEKİLER

Sayfa

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

ŞEKİLLERİN LİSTESİ viii

SİMGELER VE KISALTMALAR x

1.GİRİŞ 1

2.LİTERATÜR TARAMASI 3

3.GENEL KAVRAMLAR 13

3.1 Makine Öğrenmesi 13

3.2. Veri Artırma 18

3.3. Çekişmeli Üretici Ağlar (Generative Adversarial Networks - GAN) 24

4. DENEYSEL ÇALIŞMA VE SONUÇLAR 27

4.1 Dengeli Bir Veri Kümesiyle Deneysel Tasarım 29

4.1.1 Reklam veri kümesi 29

4.1.2 Veri artırma olmaksızın elde edilen sonuçlar 30

4.1.3 Veri Artırma ile elde edilen sonuçlar 34

4.1.4 Deneysel sonuçların irdelenmesi 39

4.2 Dengesiz Veri Kümesi için Veri Artırma 41

4.2.1 Z-Alizadeh sani veri kümesi 41

4.2.2 Veri artırımı ve özellik seçimi olmaksızın elde edilen sonuçlar 43

4.2.3 Veri Artırımı ve özellik seçimi ile elde edilen sonuçlar 48

4.2.4 Deneysel sonuçların irdelenmesi 53

5.SONUÇ VE DEĞERLENDİRME 57

KAYNAKLAR 59

ÖZGEÇMİŞ 64

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 1. Çekişmeli Üretici Ağların Mimarisi 24
Şekil 2. Araştırma sürecinin adımları 28
Şekil 3. Veri Kümesi Örneği..... 29
Şekil 4. Rastgele Orman Karışıklık Matrisi..... 32
Şekil 5. Gradient Boosting Karışıklık Matrisi 32
Şekil 6. Lojistik Regresyon Karışıklık Matrisi 32
Şekil 7. XGBoost Karışıklık Matrisi..... 33
Şekil 8. Karar Ağacı Karışıklık Matrisi 33
Şekil 9. K-En Yakın Komşular Karışıklık Matrisi..... 33
Şekil 10. Doğruluk, Özgüllük. ve Veri Artırımı olmayan Modeller için Hassasiyet..... 34
Şekil 11. Rastgele Orman Karışıklık Matrisi..... 36
Şekil 12. Gradient Boosting Karışıklık Matrisi 36
Şekil 13. Lojistik Regresyon Karışıklık Matrisi 37
Şekil 14. XGBoost Karışıklık Matrisi..... 37
Şekil 15. Karar Ağacı Karışıklık Matrisi 37
Şekil 16. K-En Yakın Komşular Karışıklık Matrisi..... 38
Şekil 17. Doğruluk, Özgüllük. ve Veri Büyütmeli Modeller için Hassasiyet..... 38
Şekil 18. Artırılmış verinin görüntüsü 39
Şekil 19. Veri Artırımı Olmadan ve Veri Artırımı ile Doğruluk Karşılaştırması 40
Şekil 20. Veri Artırımı Olmadan ve Veri Artırımı ile Özgüllüğün Karşılaştırılması 40
Şekil 21. Veri Artırımı Olmadan ve Veri Artırımı ile Hassasiyet Karşılaştırması 41
Şekil 22. Z-Alizadeh sani veri kümesinin özellik kategorisi 42
Şekil 23. Rastgele Orman Karışıklık Matrisi..... 45
Şekil 24. Gradient Boosting Karışıklık Matrisi 45
Şekil 25. Lojistik Regresyon Karışıklık Matrisi 46
Şekil 26. XGBoost Karışıklık Matrisi..... 46
Şekil 27. Karar Ağacı Karışıklık Matrisi 46
Şekil 28. K-En Yakın Komşular Karışıklık Matrisi..... 47
Şekil 29. Destek Vektör Makinesi Karışıklık Matrisi..... 47
Şekil 30. Veri Artırımı ve Özellik Seçimi Olmayan Modeller için Doğruluk, Özgüllük ve Duyarlılık 47

Şekil 31. Rastgele Orman Karışıklık Matrisi 50
Şekil 32. Gradient Boosting Karışıklık Matrisi 50
Şekil 33. Lojistik Regresyon Karışıklık Matrisi 51
Şekil 34. XGBoost Karışıklık Matrisi..... 51
Şekil 35. Karar Ağacı Karışıklık Matrisi 51
Şekil 36. K-En Yakın Komşular Karışıklık Matrisi..... 52
Şekil 37. Destek Vektör Makinesi Karışıklık Matrisi..... 52
Şekil 38. Veri Artırımı ve Özellikli Modellerde Doğruluk, Özgüllük ve Duyarlılık 52
Şekil 39. Veri Artırımı ve Özellik Seçimiyle Doğruluk Karşılaştırması 54
Şekil 40. Veri Artırımı ve Özellik Seçimiyle ve Bunlar Olmadan Özgüllük Karşılaştırılması 55
Şekil 41. Veri Artırma ve Özellik Seçimi olmadan ve Veri Artırma ve Özellik Seçimi ile duyarlılık karşılaştırması..... 55

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
XGBoost	eXtrem Gardient Bossting
NLP	Natural Lnguage Processing
ID3	Iterative Dichotomiser 3
KNN	K-Nearest Neighbors
DVM	Destek Vektor Makinesi
RBF	Radial Basis Function
GAN	Genrative Adverial Networks
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
UCI	University of California Irvine
BMI	Body Mass Index
DM	Diabetes Mellitus
FBS	Fasting Blood Sugar
CRP	C-Reactive Protein
HTN	Hypertension
FH	Family History
ÇÜA	Çekişmeli Üretici Ağ
CNN	Convolutional Neural Network
DNN	Deep Neural Network
BCI	Brain Computer Interface
FR	Random Forest
GB	Gargient Boosting
LR	Logistic Regression
BP	Blood Pressure
PR	Pulse Rate
CAD	Coronary Artery Disease

Kısaltmalar	Açıklamalar
DA	Data Augmentation
ECG	Electrocardiogram
ST	ST Segment
LVH	Left Ventricular Hypertrophy
LBBB	Left Bundle Branch Block
RBBB	Right Bundle Branch Block
HDL-C	High-Density Lipoprotein Cholesterol
LDL-C	Low-Density Lipoprotein Cholesterol
BUN	Blood Urea Nitrogen
ESR	Erythrocyte Sedimentation Rate
HB	Hemoglobin
K	Potassium
Na	Sodium
WBC	White Blood Cell Count
PLT	Platelet Count
EF-TTE	Ejection Fraction by Transthoracic Echocardiography
RWMA	Regional Wall Motion Abnormality
VHD	Valvular Heart Disease
LAD	Left Anterior Descending
LCX	Left Circumflex
RCA	Right Coronary Artery

1. GİRİŞ

Makine öğrenimi modellerinin etkinliği, kullanılan eğitim veri kümesinin kalitesinden önemli ölçüde etkilenir. Dengesiz ve sınırlı veri kümeleri, yanlış etiketlenmiş hedef değişkenler, anomali veri noktaları ve yüksek boyutlu girdi uzayları gibi zorluklar, bir makine öğrenimi modelinin başarımını önemli ölçüde etkileyebilir. Bunların dışında sınıflandırıcıların performansı, eğitim veri kümesinin boyutundan da etkilenir. Ayrıca derin öğrenme modelleri, geniş veri kümelerinde eğitilmiş olsalar bile, tanıdık olmayan veri kümeleriyle karşılaştıklarında belirgin bir tutarsızlık gösterirler. Ancak derin öğrenme modelleri, eğitim veri kümesine hızlı bir şekilde uyum sağlayabilir, etiket ve/veya tam piksel gürültüsü varlığında bile yüksek genelleme potansiyeline sahiptir. Bu modellerin performansını iyileştirmek için düzenleme teknikleri kullanılmış olsa da eğitim veri kümesindeki etiket gürültüsünü düzeltmekte sıklıkla problemler yaşanır[1]. Belirli bir makine öğrenimi modelinin kullanılmasından bağımsız olarak, modelin, yetersiz eğitim veri kümesi kalitesi ve miktarı nedeniyle, tanıdık olmayan verilerle karşılaştığında performansı genellikle düşer.

Aşırı uydurma sorunu, eğitim veri kümesinin altta yatan popülasyonu yeterince yansıtmadığında veya model parametreleri açısından aşırı karmaşık olduğunda öne çıkar. Aşırı uydurma sorununu ele almak için, düzenleme yöntemleri olarak adlandırılan bir dizi teknik kullanılır. Uygun bir düzenleme yönteminin seçimi, belirli bir kullanım durumuna bağlıdır. Çünkü bazı yöntemler yalnızca belirli sınıflandırıcılar, veri türleri veya alanlar için uygulanabilir. Öte yandan, diğer yöntemler, sınıflandırma sorunundan bağımsız olarak veri düzeyinde kullanılabilir. Veri artırma teknikleri, sentetik gözlemlerin üretilmesi yoluyla eğitim veri kümesindeki verilerin ölçeğini ve çeşitliliğini artırmak için etkili bir yöntem sağlar. Bu teknikler, veri düzeyinde çalıştıkları için çeşitli problem türleri ve sınıflandırıcılar için düzenleme yöntemleri olarak yaygın bir şekilde kullanılır. Veri artırma, çeşitli durumlarda ve problemlerde makine öğrenimi modeli doğruluğunu ve ölçeklenebilirliğini artırmak için yaygın olarak kullanılır[2].

Tezin Amacı

Bu araştırmanın amacı, farklı alanlarda tahminsel analizlerde kullanılmak üzere veri artırma yaklaşımlarının potansiyelini keşfetmektir. Temel amaç, sınırlı veri koşulları altında tahmin modelinin verimliliğini artırmaktır. Bu problem sentetik örnekler oluşturarak ve bunları eğitim setine dahil ederek başarılabilir. Bu çalışmada, mevcut kaynakları kullanarak yeni veri kümeleri oluşturma sürecini otomatikleştirerek makine öğrenimi zorluklarını başarıyla çözmek için faydalı bir teknik sunulmuştur. Özellikle sınırlı veriler kullanılarak gerçekleştirilen tahmin modellerinde, yüksek doğruluk için veri artırma önemli olabilir.

Kısıtlar

Sağlık, finans ve görüntü tanıma gibi alanlarda makine öğreniminin yaygın olarak kullanılması, eğitim veri kümesi kalitesi ve miktarı sorunlarını ortaya çıkarmaktadır. Dengesiz ve dengeli veri kümeleri, model performansını ve genelleştirilebilirliğini etkileyen sorunlar ortaya çıkabilir. Bu çalışma, veri artırmanın dengeli ve dengesiz veri kümeleriyle nasıl başa çıkabileceğini araştırmaktadır. Veriler sınırlı ve sınıflar arasında eşit dağılmamış olduğunda, makine öğrenimi modellerinin daha iyi performans göstermesi gerekmektedir. Dengeli veri kümelerinin doğruluğu, hassasiyeti, hatırlama oranı ve sınıflandırma performansı, veri takviyesi yoluyla iyileştirilebilir. Bu çalışma, tahmin dengesini koruyan sınıflandırıcılar için en iyi veri artırma yöntemlerini belirlemeyi amaçlamaktadır. Dengesiz veri kümeleri bağlamında, bu çalışma, azınlık sınıfları için sentetik örnekler oluşturarak sınıf dengesizliği ele almak için veri artırma tekniklerinin potansiyelini araştırmaktadır. Geri çağırma, doğruluk ve F1 puanı gibi değerlendirme metrikleri, veri artırmanın dengesiz senaryolarda model performansı üzerindeki etkisini değerlendirmek için kullanılır. Bu tez, her iki veri kümesi türü için de veri artırmanın faydalarını ve dezavantajlarını incelemektedir. Bu, kapsamlı deneyler ve çalışmalar yoluyla başarılacaktır. Bu çalışma, pratik durumlarda makine öğrenimi modeli performansını artırmak için veri artırma yaklaşımlarını optimize etmeye yardımcı olacaktır

Tez Organizasyonu

Birinci bölüm tez konusunu, problemi ve amaçları tanıtır. İkinci bölüm mevcut çalışmaları değerlendirir. Üçüncü bölüm tanımları, teorik çerçeveyi ve varsayımları ele alır. Dördüncü bölüm tasarımı, veri toplama, analitik metodolojileri ve istatistiksel analizi içerir. Beşinci bölüm sonuçların analizini, özetini ve sonuçları kapsar. Her bölüm farklı bir amaca hizmet eder ve çalışmanın genel yapısına ve tutarlılığına katkıda bulunur.

2. LİTERATÜR TARAMASI

Makine öğrenimi alanı, çeşitli teknoloji alanlarında ve gerçek dünya zorluklarında uygulamalarının hızlı bir şekilde artmasına neden olan önemli bir büyüme yaşamaktadır. Makine öğrenimi modellerinin başarısı, eğitim amaçlı kullanılan verilerin yeterli mevcudiyeti, hassasiyeti ve heterojenliğine bağlıdır. Veriler, insan girdisinden veya sensör okumalarından kaynaklanabilen sembolik ve sayısal öznitelikler şeklinde temsil edilebilir. Bu öznitelikler geniş bir karmaşıklık yelpazesine sahip olabilir ve farklı güvenilirlik seviyelerine sahip olabilir. Makine öğrenimi algoritmalarının gerçek dünya zorluklarına yönelik optimizasyonu, çeşitli ve yüksek kaliteli verilerin dahil edilmesine ağırlık verir. Son çalışmalar, uygun veri artırma yöntemlerinin uygulanmasıyla bu sınırlamaların azaltılabileceğini göstermiştir. Bu teknikler, mevcut eğitim verilerini artırarak veya çeşitli yaklaşımlar kullanarak yeni örnekler oluşturarak ek eğitim örnekleri oluşturmayı içerir. Çok sayıda çalışma ve halka açık veri kümesi, makine öğrenimi algoritmalarının birçok alanda hassasiyetini artırmak için yapılmıştır. Literatürde araştırmacılar çeşitli veri artırma yöntemlerini kullanmışlardır:

Liu ve arkadaşlarının 2023 yılında yapmış oldukları çalışmada [3], sınırlı bir örneklem büyüklüğü kullanarak pulmoner nodüllerin tanımlanması ve sınıflandırılmasının iyileştirilmesine yönelik yeni bir metodoloji sunulmaktadır. Bu çalışmada önerilen yöntem, 3D piksel düzeyi istatistik algoritmasını içeren bir veri artırma tekniğinin kullanılmasını içermektedir. Bu algoritma, sentetik pulmoner nodüllerin sağlıklı akciğer verileriyle birleştirilerek yeni örnekler oluşturur. Ayrıca, piksel verilerinin pulmoner nodül örneklerindeki anlaşılmasını artırmak için gizli değişkenler içeren yeni bir gömme mekanizması önerilmektedir. Bu tekniklerin değerlendirmesi, pulmoner nodüllerin bilgisayarlı tomografi (BT) görüntülerinden oluşan bir veri kümesinde gerçekleştirilir. Sonuçlar, veri artırma yönteminin Çekişmeli üretici ağ (ÇÜA) çerçevesine kıyasla üstünlüğünü göstermektedir. Ayrıca, gömme mekanizmasının nodül sınıflandırması için doğruluğu ve direnci artırmadaki etkinliği kanıtlanmıştır. Özetle, önerilen veri artırma tekniği ve gömme mekanizması, özellikle mevcut verilerin sınırlı olduğu senaryolarda, pulmoner nodüllerin hassas ve dirençli tanımlanması ve sınıflandırılması için önemli geliştirmeler sunmaktadır.

S. Afzal ve arkadaşları [4], Alzheimer hastalığı aşamalarının tespitinde sınıf dengesizliği sorununu hafifletmek için veri artırma kullanan bir çerçeve sunmaktadır. Çerçeve, veri artırma ve transfer öğrenimi olmak üzere iki temel aşamadan oluşur. Veri artırma sürecinde, döndürme, çeviri, ölçeklendirme ve gürültü ekleme gibi bir dizi teknik, eğitim veri kümesini yapay olarak artırmak için kullanılır. Transfer öğrenimi süreci, artırılmış görüntüler üzerinde sınıflandırma yapmak için önceden var olan bir konvolüsyonel sinir ağı (CNN) modelinin kullanılmasını içerir. Çerçevenin performansı, Alzheimer hastalığı ile ilgili manyetik rezonans görüntüleri içeren halka açık bir derleme olan OASIS veri kümesi kullanılarak değerlendirilir. Bulgular, önerilen metodolojinin sınıf dengesizliği sorununu ele almadaki etkinliğini göstermektedir ve Alzheimer hastalığı aşamalarının tespiti için %95,11'lik dikkate değer bir doğruluk oranı ile sonuçlanmaktadır. Genel olarak, veri artırma tabanlı bir çerçevenin kullanımı, Alzheimer aşamalarının tespitinde sınıf dengesizliği sorununu ele almak için potansiyel olarak etkili bir yaklaşım sunar. Bu yöntem, uygulamasının kolaylığı ve tıbbi görüntülerin sınıflandırılmasıyla ilgili diğer görevlere uyarlama potansiyeli ile karakterize edilir.

P. Cao ve çalışma arkadaşları[5], derin sinir ağları (DNN'ler) kullanarak atriyal fibrilasyon (AF) tespiti için yeni bir veri artırma yöntemi önermektedir. Önerilen veri artırma yöntemi, Pan-Tompkins QRS dedektörü tarafından belirlenen karakteristik noktalara göre EKG bölümlerinin birleştirilmesine dayanmaktadır. Önerilen yöntem, MIT-BIH Arritmia Veritabanı'nda sıkı bir şekilde değerlendirilmiş ve sonuçlar, AF tespiti için DNN'lerin performansını etkili bir şekilde artırma yeteneğini göstermektedir. Bu yenilikçi yaklaşım, derin sinir ağlarının kullanımıyla AF tespitinin doğruluğunu ve verimliliğini artırmada umut verici bir potansiyele sahiptir.

L. Nanni ve çalışma arkadaşları[6], veri artırma tekniklerinin hayvan sesi sınıflandırma modellerinin performansını artırmadaki uygulamalarını araştırıyor. Veri artırma, mevcut veri noktalarına dönüş, çeviri, ölçeklendirme ve gürültü ekleme gibi çeşitli dönüşümler uygulayarak eğitim veri setinin boyutunu artırmak için kullanılan bir tekniktir. Bu makale, özellikle hayvan sesi sınıflandırma için tasarlanmış veri artırma tekniklerinin kapsamlı bir incelemesini sunmaktadır. Ele alınan teknikler arasında zaman gerdirme, pitch kaydırma, gürültü ekleme ve karışık ses artırma yer almaktadır. Veri artırmanın etkinliğini değerlendirmek için birden fazla deney yapılmıştır ve bulgular, hayvan seslerini sınıflandıran modeller üzerinde kayda değer ve faydalı bir etki göstermiştir. Genel olarak,

veri artırma, eğitim veri setini genişleterek aşırı uyumu azaltmaya ve genel genelleştirme performansını iyileştirmeye hizmet ederek, hayvan sesi sınıflandırma modellerinin etkinliğini artırmak için güçlü bir teknik olarak tanımlanmıştır.

A. Ramires ve X. Serra[7], ses efektleriyle karşılaştığında enstrüman sınıflandırma modellerinin direncini artırmak için veri artırma tekniklerinin kullanımını inceler. Bu makale, enstrüman sınıflandırması bağlamında kullanılan çeşitli veri artırma tekniklerinin bir genel bakışını sunar. Bu teknikler arasında, ses kliplerinin süresini ayarlamayı içeren zaman uzatma, pitch değiştirme, çevresel gürültüyü simüle eden gürültü ekleme ve yankı odasının etkilerini çoğaltan yankı bulunur. Veri artırmanın etkililiği deneysel analiz yoluyla değerlendirilir ve enstrüman sınıflandırma modellerinin ses efektlerine karşı direncinde önemli bir artış gösterir. Özetle, veri artırma, enstrüman sınıflandırma modellerinin ses efektlerine karşı sağlamlığını artırmak için etkili bir teknik olarak hizmet verir. Eğitim veri kümesini yapay olarak genişleterek, aşırı uyum sorunlarını etkili bir şekilde ele alır ve genel genelleme performansını iyileştirir.

A. Chatziagapi ve arkadaşları [8], konuşma duygu tanıma (KDT) alanında Çekişmeli üretici ağ (ÇÜA) kullanımını veri artırma yöntemi aracılığıyla inceliyor. ÇÜA'lar, yapay öğrenme modellerinin bir sınıfı olarak, mevcut örnekleri yakından taklit eden sentetik veriler üreterek eğitim veri kümelerini zenginleştirmede büyük potansiyel göstermektedir. Makale, özel olarak konuşma duygu tanıma için tasarlanmış, ÇÜA'lara dayalı bir veri artırma yöntemini sunuyor. Bu yöntemin etkinliği, IEMOCAP veri kümesi üzerinde yapılan bir değerlendirme ile değerlendirilmektedir. Bulgular, KDT modelinin performansında kayda değer bir artış olduğunu göstermektedir. Makalede vurgulanan ana noktalar, KDT bağlamında gerçekçi sentetik veri üretmek için ÇÜA'ların kullanımını içerir. Ayrıca, ÇÜA'lara dayalı bir KDT veri artırma tekniği sunmaktadır. Önerilen yöntem, IEMOCAP veri seti üzerinde etkili bir şekilde uygulanmış ve KDT modelinin performansını artırmıştır. ÇÜA'lara dayalı bir veri artırma tekniği kullanımı, KDT modellerini geliştirmek için basit bir teknik olmasına rağmen oldukça umut vericidir. Ayrıca, bu yöntem başka KDT veri kümelerine kolayca uygulanabilir.

D. Refai ve arkadaşları [9], Arapça metin sınıflandırma modellerinin etkinliğini artırmak için veri büyütme tekniklerinin uygulanmasını araştırıyor. Veri büyütme, önceden var olan veri noktalarına eşanlamlı değiştirme, kelime sırası karıştırma ve gürültü ekleme gibi çeşitli

dönüşümler uygulayarak, eğitim veri setinin hacmini yapay olarak genişletmek için kullanılan bir metodolojidir. Makale, transformatörlerin ve benzerlik ölçümlerinin entegrasyonunu içeren, Arapça metin sınıflandırması için veriyi büyütmeye yönelik yeni bir yaklaşım sunmaktadır. Sinirsel ağların belirli bir kategorisi olan transformatörler, doğal dil işleme ile ilgili görevlerde olağanüstü performans gösterir. Metin içindeki, özellikle geniş mesafeler boyunca uzanan karmaşık ilişkileri ve bağımlılıkları öğrenme ve anlama yeteneğine sahiptirler. Sonuç olarak, transformatörler veri büyütme amacına çok uygundur. Önerilen yöntem, genişletilmiş metnin anlamsal benzerliğini orijinal metinle koruyabilmek için benzerlik ölçümlerini kullanır. Yöntem, dört Arapça metin sınıflandırma veri seti üzerinde değerlendirilmiş ve sonuçlar Arapça metin sınıflandırma modellerinin performansında kayda değer bir iyileşme olduğunu göstermiştir. Makale, Arapça metin sınıflandırmasını geliştirmek için veri büyütme tekniklerinin kullanımını, transformatörleri ve benzerlik ölçümlerini kullanan yeni bir yaklaşımın ortaya çıkışını ve dört farklı Arapça metin sınıflandırma veri seti üzerinde gözlenen cesaret verici sonuçları vurgulamaktadır. Bu veri büyütme tekniğinin kullanımı, Arapça metin sınıflandırma modellerinin etkinliğini artırmada umut vadetmekte ve aynı alan içindeki diğer veri setlerine kolayca uygulanabilir.

P. Rana ve arkadaşları [10], kan hücresi görüntü sınıflandırma modellerinin etkinliğini artırmak için veri büyütme tekniklerinin kullanımını araştırıyor. Makale, veri büyütme tekniği kullanarak kan hücresi görüntülerinin sınıflandırmasını geliştirmek için yeni bir yaklaşım sunuyor. Bu teknik, Wasserstein diverjans ÇÜA, ve yeni önerilen doğrusal olmayan mixup yönteminin kullanımını birleştiriyor. Wasserstein diverjans yaratıcı Çekişmeli üretici ağ (ÇÜA), gerçek görüntülere çok benzeyen sentetik görüntüler üretebilir. Öte yandan, mixup tekniği, iki farklı görüntüyü birleştirerek yeni bir bileşik görüntü oluşturmayı içerir. Ek olarak, mixup'ın yeni bir varyantı olan doğrusal olmayan mixup, görüntü birleştirme sürecini gerçekleştirmek için doğrusal olmayan bir fonksiyon kullanır. Ayrıca, bu sınıflardaki yeterli temsili garanti altına almak için azınlık sınıflarına öncelik veren bir örnekleme stratejisi önerilmektedir. Değerlendirme konusu olan yöntem, halka açık erişilebilir iki kan hücresi görüntü veri seti kullanılarak değerlendirilmiş olup, kan hücresi görüntü sınıflandırma modelinin performansında önemli bir gelişme göstermektedir. Makale, birkaç önemli bulguyu vurgulamaktadır. Birincisi, kan hücresi görüntü sınıflandırmasını geliştirmede veri büyütmenin önemini vurgulamaktadır. İkincisi, Çekişmeli üretici ağ (ÇÜA) ve mixup varyantlarını birleştiren yeni bir yaklaşım ortaya koymaktadır. Sonuncusu ise, azınlık sınıfına öncelik veren bir örnekleme stratejisinin

uygulanmasını tartışmaktadır. Bu özel veri büyütme tekniğinin kullanımı, kan hücresi görüntü sınıflandırma modellerinin performansını artırmada potansiyel göstermekte olup, diğer kan hücresi görüntü veri setlerinde kolayca uygulanabilir.

Q. Zhu ve çalışma arkadaşları[11], veri büyütme teknikleri ve ikili dikkat mekanizmalarını kullanan yeni bir ince işlenmiş görüntü sınıflandırma yaklaşımı sunmaktadır. Önerilen yöntem, üç ana bileşeni kullanarak ince işlenmiş görüntü sınıflandırma zorluğunun üstesinden gelmektedir: veri büyütme modülü, ikili dikkat ağı ve etkileşim modülü. Veri büyütme modülü, döndürme, çevirme, ölçekleme ve gürültü ekleme gibi çeşitli dönüşümler yoluyla ek veri noktaları oluşturarak eğitim veri setini geliştiren oldukça etkili bir tekniktir. Görüntü analizi bağlamında, ikili dikkat ağı, görüntünün önemli özellik bölgelerine öncelik verecek şekilde tasarlanmıştır. Bu, iki farklı dikkat mekanizmasının kullanımıyla elde edilir: biri küresel özellikleri yakalamak için, diğeri yerel özellikleri yakalamak için. Etkileşim modülü, daha ayırıcı bir görüntü temsili elde etmek için küresel ve yerel özellikleri birleştirmede kritik bir işlev üstlenir. Etkililiğini değerlendirmek için, önerilen metodoloji ince işlenmiş görüntü sınıflandırması ile ilgili üç veri seti üzerinde değerlendirilmiştir. Bu değerlendirme, modelin performansında kayda değer gelişmeler olduğunu ortaya koymaktadır. Söz konusu metodoloji, ince işlenmiş görüntü sınıflandırma modellerinin performansını iyileştirme potansiyeli göstermektedir. Basit uygulaması, çeşitli diğer ince işlenmiş görüntü sınıflandırma çalışmalarına kolaylıkla uygulanmasını kolaylaştırır.

R. Gulakala ve arkadaşları [12], COVID-19 tespiti için yeni bir veri artırma tekniği sunmaktadır ve bunun için bir Çekişmeli üretici ağ (ÇÜA) kullanmaktadır. Önerilen yaklaşım, COVID-19 tarafından etkilenmiş akciğerleri tasvir eden görüntüler oluşturmak için bir koşullu Çekişmeli üretici ağ (ÇÜA) kullanmaktadır. Bu yöntem, üretilen görüntülerin enfekte olmuş akciğerlerin durumunu doğru bir şekilde temsil etmesi şartıyla çalışmaktadır. Çekişmeli üretici ağ (ÇÜA), COVID-19 ile enfekte olmuş bireylerden alınan gerçek akciğer görüntülerini içeren bir veri kümesi kullanılarak eğitilmiştir. Bu eğitim, ÇÜA'ın gerçek enfekte akciğerlere yakından benzeyen sentetik görüntüler üretebilmesine olanak tanır. Sentetik görüntüler daha sonra COVID-19 tespiti modeli için temel alan bir konvolüsyonel sinir ağı (CNN) için eğitim veri setini geliştirmek için kullanılmaktadır. Bulgular, COVID-19 tespiti için kullanılan konvolüsyonel sinir ağı (CNN) etkinliğinde dikkate değer bir iyileşme göstermektedir. Makale, bir koşullu ÇÜA'ın COVID-19'u tasvir eden gerçek akciğer görüntüleri üretmek için kullanılması, bu sentetik görüntülerin eğitim

veri setini geliştirmek için entegrasyonu ve sonrasında COVID-19 tespit modelinin etkinliğinde önemli bir artış gibi birkaç dikkat çekici bulguyu sunmaktadır. İncelenen yöntem, COVID-19 tespit modellerini geliştirmede potansiyel göstermekte olup, basit uygulanabilirliği nedeniyle diğer tıbbi görüntü sınıflandırma görevlerinde de kolayca kullanılabilir.

S. Yu ve arkadaşları[13], Rahim ağzı hücrelerinin sınıflandırılması için veri artırma yöntemi sunmaktadır ve bu yöntem bir Çekişmeli üretici ağ (ÇÜA) temellidir. Önerilen metodoloji, anormal rahim ağzı hücrelerinin görüntülerni oluşturmak için bir koşullu Çekişmeli üretici ağ (ÇÜA) kullanmaktadır. Bu süreç, üretilen görüntülerin anormal rahim ağzı hücrelerini doğru bir şekilde temsil etmesine bağlıdır. Çekişmeli üretici ağ (ÇÜA), gerçek anormal rahim ağzı hücrelerini içeren bir veri kümesi kullanılarak eğitilmiştir, bu da ona bu gerçek örnekleri yakından taklit eden sentetik görüntüler üretebilme yeteneği kazandırmıştır. Sentetik görüntüler daha sonra bir konvolüsyonel sinir ağı (CNN) temelli hücre sınıflandırma modeli için eğitim veri setini geliştirmek için kullanılmaktadır. Bulgular, Konvolüsyonel Sinir Ağlarına (CNN) dayalı rahim ağzı hücre sınıflandırma modelinin etkinliğinde dikkate değer bir iyileşme göstermektedir. İncelenen yöntem, rahim ağzı hücrelerini sınıflandırmak için modelleri geliştirmede potansiyele sahiptir ve diğer tıbbi görüntü sınıflandırma görevlerinde de kolayca kullanılabilir.

S. Lamba ve arkadaşları[14], pirinç hastalığı tespitinin doğruluğunu artırmak amacıyla veri kümesini geliştirmek için bir teknik sunmaktadır. Bu, Çekişmeli Üretici Ağ (ÇÜA'lar) kullanılarak gerçekleştirilmektedir. Bu çalışmada kullanılan yöntem, belirli bir duruma, yani belirli bir hastalığa yakalanmış pirinç bitkilerini tasvir eden görüntüler oluşturmak için bir koşullu Çekişmeli Üretici Ağ (ÇÜA) kullanımını içermektedir. ÇÜA, belirli bir hastalığı sergileyen gerçek pirinç bitkilerini içeren bir veri kümesi kullanılarak eğitilir. Bu eğitim süreci, ÇÜA'ın gerçek örnekleri yakından taklit eden enfekte pirinç bitkilerinin sentetik görüntülerini üretebilmesine olanak tanır. Sentetik görüntüler daha sonra bir destek vektör makinesi temelli pirinç hastalığı tespit modeli için eğitim veri setini geliştirmek için kullanılır. Bulgular, pirinç hastalığı tespit modelinde kullanılan destek vektör makinesinin etkinliğinde dikkate değer bir iyileşme göstermektedir. Önerilen yöntem, pirinç hastalığı tespit modellerini geliştirmede potansiyele sahiptir ve basit uygulanabilirliği nedeniyle diğer bitki hastalığı tespit görevlerinde kolayca kullanılabilir.

S. Shao ve arkadaşları[15], makine arıza teşhisini artırmak için Çekişmeli Üretici Ağ (ÇÜA) kullanarak veri artırma yöntemi üzerine yeni bir yaklaşım sunmaktadır. Bu yaklaşım, belirli bir duruma, özellikle arıza gösteren makine bileşenlerini tasvir eden görüntüler üretmek için bir koşullu Çekişmeli Üretici Ağ (ÇÜA) kullanmaktadır. ÇÜA, bozuk bir durumda olan gerçek makine bileşenlerini içeren bir veri kümesi kullanılarak eğitilir. Bu eğitim, ÇÜA'ın orijinal veri kümesinde bulunan aynı arızaları doğru bir şekilde tasvir eden makine bileşenlerinin sentetik görüntülerini üretebilmesine olanak tanır. Oluşturulan sentetik görüntüler daha sonra bir makine arıza teşhis modeli için eğitim veri setini geliştirmek için kullanılır. Bulgular, makine arıza teşhis modelinin etkinliğinde dikkate değer bir iyileşme göstermektedir. Önerilen yöntem, makine arıza teşhis modellerini geliştirmede potansiyele sahiptir.

F. Fahimi ve arkadaşları[16], beyin-bilgisayar arayüzü (brain-computer interface BCI) sistemlerini veri artırma için Çekişmeli Üretici Ağ (ÇÜA) kullanarak geliştirmek için yeni bir yaklaşım sunmaktadır. Bu çalışmada kullanılan yaklaşım, belirli bir koşula uygun görüntüler üretebilen bir koşullu Çekişmeli Üretici Ağ (ÇÜA) temellidir. Bu durum, belirli bir kullanıcı niyeti ile ilişkilendirilmiş beyin sinyalleri ile ilgilidir. ÇÜA, belirli bir kullanıcı niyeti ile ilişkilendirilmiş gerçek beyin sinyallerini içeren bir veri kümesi kullanılarak eğitilir. Bu eğitim süreci, ÇÜA'ın gerçek beyin sinyallerini yakından taklit eden, aynı kullanıcı niyetini sürdüren sentetik beyin sinyali görüntülerini üretebilmesine olanak tanır. Ardından, oluşturulan sentetik görüntüler, bir Beyin-Bilgisayar Arayüzü (BCI) sisteminin eğitim veri setini geliştirmek için kullanılır. Bulgular, beyin-bilgisayar arayüzü (BCI) sisteminin performansında dikkate değer bir gelişme göstermektedir.

S. Jain ve arkadaşları[17], yüzey kusuru tespiti ve sınıflandırmasındaki derin öğrenme modellerinin performansını artırmak amacıyla sentetik veri artırma yoluyla veri miktarını ve kalitesini artırmak için yeni bir strateji sunmaktadır. İncelenen yöntem, bir Çekişmeli Üretici Ağ (ÇÜA) kullanımına dayanmaktadır. Çekişmeli Üretici Ağ (ÇÜA), gerçek yüzey kusuru örneklerini içeren bir veri kümesi kullanılarak eğitilir. Sonuç olarak, GAN, gerçek dünya örneklerine yakından benzeyen sentetik yüzey kusuru görüntüleri üretme yeteneğine sahiptir. Yapay olarak üretilen görüntüler daha sonra yüzey kusurlarını tespit etmek ve sınıflandırmak amacıyla tasarlanmış derin öğrenme modeli için eğitim veri setini geliştirmek için kullanılır. Bulgular, bu çalışmada önerilen yöntemin, yüzey kusurlarını tespit etme ve

sınıflandırmaya yönelik görevde derin öğrenme modelinin performansında dikkate değer bir artış sağladığını göstermektedir.

Gu, R. ve arkadaşları [18], geleneksel öneri sistemlerinde yetersiz veri nedeniyle yüksek derecede kişiselleştirilmiş önerilerin sunulmasındaki zorluğu ele aldılar. Bu engeli aşmak için, makale cihazda öğrenmeyi bulutla koordine edilmiş veri artırma ile birleştiren yenilikçi bir yaklaşım önermektedir. Cihazda öğrenme, kullanıcıların kişisel verilerini doğrudan cihazlarında kullanarak kişiselleştirilmiş öneri modellerini eğitmeyi mümkün kılar, bulutla koordine edilmiş veri artırma ise bu veriyi gizliliği koruyarak buluttan ek bilgilerle zenginleştirir. Bu yeni strateji, son derece özelleştirilmiş öneriler sunmayı amaçlayarak kullanıcıların benzersiz tercihlerini ve bağlamını yansıtmaktadır. Bu yaklaşımın potansiyel faydaları arasında, geleneksel yöntemlere kıyasla geliştirilmiş öneri doğruluğu ve uygunluğu, artan kullanıcı memnuniyeti ve artırılmış gizlilik koruması bulunmaktadır. Makale, muhtemelen algoritmalar, veri artırma teknikleri ve gizliliği koruma mekanizmaları gibi teknik yönleri ele almaktadır ve önerilen yöntemi mevcut yaklaşımlarla karşılaştırmak için deneysel değerlendirmeler içerebilir.

Chen, L.ve arkadaşları [19], kişiselleştirilmiş öneri sistemlerindeki adil önerilerin iyileştirilmesini, veri artırma yoluyla araştırıyorlar. Geleneksel sistemler, sınırlı kullanıcı verisi nedeniyle yüksek derecede kişiselleştirilmiş öneriler sunmada zorluklarla karşılaşır, bu da yenilikçi çözümlerin gerekliliğini ortaya çıkarır. Önerilen yaklaşım, bu sorunu ele almak için cihazda öğrenmeyi bulutla koordine edilmiş veri artırma ile birleştiriyor. Cihazda öğrenmeyi kullanarak, kullanıcı verisi yerel olarak kullanılarak kişiselleştirilmiş öneri modelleri eğitilir, bulutla koordine edilmiş veri artırma ise bu veriyi buluttan ek bilgilerle zenginleştirir ve gizliliği korur. Bu yeni strateji olan aşırı model kişiselleştirmesi, her kullanıcının benzersiz tercihlerini ve bağlamını yansıtan son derece özelleştirilmiş öneriler sunmayı amaçlamaktadır, bu da öneri doğruluğunu, uygunluğunu ve kullanıcı memnuniyetini artırırken gizliliği korur. Makale, önerilen yöntemi mevcut yaklaşımlarla karşılaştırmak için algoritmalar ve deneysel değerlendirmeler gibi teknik yönleri tartışıyor. Bu çalışma, öneri sistemlerinde adillik üzerine süregelen tartışmaya katkı sağlayarak, veri artırma yoluyla öneri adilliğini iyileştirmede umut verici sonuçlar sunan esnek bir çerçeve sunuyor.

Çoğunluğu etiketlenmiş örnek sayısı sınırlı olan az sayıda sınıfa sahip dengesiz veri üzerinde sınıflandırma yapma görevi, az atış (few-shot) öğrenme açısından zorluklar oluşturmaktadır[20]. Bu senaryo, tıbbi tanı ve dolandırıcılık tespiti alanlarında kullanılmaktadır. Makale, az sayıda sınıflandırmanın dengesizliği ile başa çıkmak için özel olarak tasarlanmış bir veri artırma tekniği sunmaktadır. Metodoloji, iki temel aşamadan oluşmaktadır. Veri örnekleme için SMOTE-ENN tekniği kullanılmaktadır. Azınlık sınıfını aşırı örnekleme ve çoğunluk sınıfına yakın veri noktalarını kaldırma süreci, yeni veri noktalarının oluşturulmasına yol açmaktadır. Ayrıca, veri noktaları döndürme, çevirme, ölçeklendirme ve gürültü ekleme gibi çeşitli dönüşümlere tabi tutulmaktadır. Bu metodolojinin kullanımı, artırılmış veri seti içinde daha fazla çeşitlilik getirerek sınıflandırıcıya ek karmaşıklık katmaktadır. Önerilen teknik, geniş çapta tanınan doğal görüntüler koleksiyonu olan CIFAR-100 veri seti üzerinde gerçekleştirilen deneylerle değerlendirilmiştir. Bu çalışmada önerilen veri artırma stratejisi, verilen veri setinde az sayıda sınıflandırma modellerinin performansını önemli ölçüde artırmaktadır.

S. Sundaram ve arkadaşları[21], göğüs X-ışınlarının sınıflandırılmasında veri artırma yöntemi olarak Çekişmeli Üretici Ağ (ÇÜA) kullanımını incelemektedir. Önerilen yöntem, iki temel adımdan oluşan Çekişmeli Üretici Ağ (ÇÜA) temellidir ve yeni bir veri artırma tekniği sunmaktadır. İlk olarak, gerçek göğüs X-ışını görüntülerini içeren bir veri kümesini kullanarak ÇÜA eğitilir. Bu eğitim sürecinin temel amacı, gerçek X-ışını görüntülerine yakından benzeyen sentetik göğüs X-ışını görüntüleri üretmektir. Ardından, eğitimden geçmiş olan ÇÜA, eğitim veri setini geliştirmek için kullanılır ve yeni sentetik göğüs X-ışını görüntüleri üreterek bunları mevcut eğitim verisine entegre eder. Önerilen metodolojinin etkinliği, kamuoyuna açık olan göğüs X-ışını görüntülerini içeren CheXpert veri seti kullanılarak değerlendirilir. Bulgular, Çekişmeli Üretici Ağ (ÇÜA) temelli önerilen veri artırma tekniğinin CheXpert veri setine uygulandığında göğüs X-ışını sınıflandırma modellerinin performansını etkili bir şekilde artırdığını göstermektedir.

Y. Li ve çalışma arkadaşları[22], çözünmüş gaz analizi bağlamında dengesiz örnek veri setleri sorununu ele almaktadır. Farklı çözünmüş gaz türleri için değişen örnek boyutlarının mevcudiyetinin, sınıflandırıcılarda önyargılara ve azınlık sınıflarının doğru bir şekilde tespitinde zorluklara neden olabileceğine dikkat çekmektedir. Bu sorunu çözmek için, makale, Çekişmeli Üretici Ağ (ÇÜA) kullanarak bir veri artırma tekniği önermektedir. Çekişmeli Üretici Ağ (ÇÜA), gerçek çözünmüş gaz görüntülerini içeren bir veri kümesi

kullanılarak eğitilir. Amacı, az temsil edilen sınıflar için orijinal görüntülere yakından benzeyen sentetik veri üretmektir. Önerilen yöntemi değerlendirmek için UCI Çözünmüş Gaz veri seti kullanılmıştır ve bulgular, çözünmüş gaz analizi bağlamında dengesiz örnek veri setleri ile başa çıkarken sınıflandırıcı performansında önemli bir iyileşme gösterdiğini göstermektedir.

J. Hemmerich ve arkadaşları[23], moleküler bağlamda veri artırma yöntemi olarak konformasyonel aşırı örnekleme (COVER) kullanımını incelemektedir. COVER yöntemi, atom döndürme, çeviri ve ölçeklendirme manipülasyonları aracılığıyla yeni moleküler konformasyonlar üretme sürecini içermektedir. Bu yaklaşımın amacı, ilaç keşfi ve toksisite tahmini gibi görevlerde makine öğrenme modellerinin etkinliğini artırmak için kapsamlı ve çeşitli bir eğitim veri seti oluşturmaktır. Makale, COVER'a dayalı moleküller için bir veri artırma yaklaşımını sunmaktadır ve bu yaklaşım iki temel adımdan oluşmaktadır. İlk olarak, yöntem, eğitim veri setinde bulunan moleküller kümesinden yeni moleküler konformasyonlar üretmek için COVER algoritmasını kullanır. Ayrıca, bu yeni konformasyonların eğitim veri setine eklenmesi, veri setinin boyutunu artırmak ve çeşitliliğini artırmak amacıyla hizmet eder.
Araştırma hedefini gerçekleştirmek için, başlangıçta, Çekişmeli Üretici Ağlar (ÇÜA) kullanarak veri artırma işlemini gerçekleştireceğiz. Bu, tüketici reklam tıklama tahmini veri kümesi ve koroner arter hastalığı veri kümesi için ek eğitim örnekleri üretmeyi içerir. Makine öğrenimi algoritmalarının doğruluğunu GAN tabanlı veri artırma ile değerlendireceğiz ve koroner arter hastalığını tahmin etmek için daha etkili özelliklerin keşfi için özellik seçimi yaklaşımlarını kullanacağız. Son olarak, GAN'lar kullanılarak veri artırma uygulandıktan sonra makine öğrenimi modellerinin performansındaki iyileşmeyi değerlendireceğiz.

3. GENEL KAVRAMLAR

3.1 Makine Öğrenmesi

Makine öğrenimi, yapay zeka alanında özel bir bilim dalı olup, algoritmaların ve istatistik modellerin geliştirilmesine odaklanmaktadır. Bu araçlar, bilgiyi verilerden elde ederek belirli bir görevde bilgisayar sistemlerinin performansını artırmaya yöneliktir ve açık programlamaya gerek duymadan gerçekleştirilir. Makine öğreniminin amacı, bilgisayarların bilgiyi otonom bir şekilde edinmesini ve geliştirmesini sağlamak ve bu da insan müdahalesine olan ihtiyacı ortadan kaldırmaktadır.

Makine öğreniminin türleri:

- Denetimli Öğrenme: Etiketlenmiş eğitim verisi içerir. Böylece bilgisayar, eğitim verisi içinde bulunan bilgileri kullanarak yeni verileri tahmin etme veya sınıflandırma yeteneği kazanır.
- Denetimsiz Öğrenme: Bu özel paradigmada, bilgisayara herhangi bir etiketleme olmadan bir veri kümesi sunulur ve bilgisayar, veri içindeki desenleri ve yapıları tespit etmek ve tanımak için bir öğrenme süreci geçirir.
- Yarı Denetimli Öğrenme: Bilgisayar sistemi, hem etiketlenmiş hem de etiketlenmemiş veri kombinasyonu ile donatılmıştır, bu da bilgisayarın hem veri türlerinden bilgi kullanarak bilgi edinmesine hem de yeni verileri tahmin etmesine veya sınıflandırmasına olanak tanır.
- Pekiştirmeli Öğrenme: Bu, bir ajanın eylemlerine bağlı olarak ödüller veya cezalar alarak kararlar verme yeteneğini kazandığı bir makine öğrenme yaklaşımıdır. Bu yaklaşımın temel amacı, belirli bir süre zarfında ödüllerin birikimini optimize etmektir, bu da özellikle robotik ve oyun gibi alanlarda uygulanabilir kılar.

Makine öğrenimi, bilgisayar görüsü, doğal dil işleme, konuşma tanıma, robotik, sağlık, finans ve birçok diğer alanı içeren farklı alanlarda geniş bir uygulama yelpazesine sahiptir. Teknolojinin gelişimi, geniş veri setlerini analiz edebilen, tahminler üretebilen ve karmaşık görevleri otomatikleştirebilen akıllı sistemlerin ortaya çıkmasını kolaylaştırmıştır[24]. Yaygın olarak kullanılan algoritmalar şunlardır;

- Random Forest (RF)

  Çoklu ve bireysel ağaçları bir araya getirerek hem sınıflandırma hem de regresyon görevlerini gerçekleştirebilen bir denetimli öğrenme yöntemidir. Algoritma, her biri bir sınıfı tahmin etmek için özellikleri kullanan ağaçlardan oluşan bir orman oluşturur. Nihai tahmin, tüm ağaçlarda alınan oyların en yüksek olduğu sınıfın seçilmesiyle belirlenir. Önceki çalışmalara göre[25], bir RF sınıflandırıcısına daha fazla ağaç eklemenin performansını artırdığını göstermektedir. Algoritma, çoklu karar ağaçlarının kombinasyonunu içeren bir topluluk model oluşturmak için matematiksel bir formül kullanır. Formül, ormandaki ağaçların miktarına bağlı olarak değişkenlik gösterir. Aşağıdaki formül, RF sınıflandırıcısını matematiksel olarak temsil etmektedir:

$$n_{ij} = W_I C_j - W_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)} \qquad (1)$$

  Burada:
  ni sub(j), j düğümünün önemini belirtir
  w sub(j), j düğümüne ulaşan örneklerin ağırlıklı sayısını ifade eder
  C sub(j), j düğümünün safsızlık değerini temsil eder
  Sol(j), j düğümündeki bölünmeden sonra elde edilen sol alt düğümü belirtir.
  Sağ(j), j düğümündeki bölünmeden sonra elde edilen sağ alt düğümü temsil eder.

- Gradient Boosting (GB)

  Regresyon ve sınıflandırma amaçları için yaygın olarak kullanılan bir enseble öğrenme yöntemidir. Gradient iniş optimizasyonu kullanarak daha az hata ile çoklu modeller eğitmek mümkündür. Temel modeller, az sayıda düğüm ve yaprak içeren zayıf karar ağaçlarıdır. Algoritmanın başlıca parametreleri ağaç sayısı, öğrenme hızı, maksimum derinlik ve minimum düğüm bölme verisidir. Gradient Boosting, çeşitli giriş kaynaklarını ele alabilir ve yüksek tahmin doğruluğu sağlar, bu nedenle makine öğreniminde kullanışlıdır. Ancak Gradient Boosting modelleri eğitim verilerine aşırı uyum sağlayabilir ve bilinmeyen veriler üzerinde kötü performans sergileyebilir. Doğru bir şekilde ayarlanmadığında, Gradient Boosting model eğitimi hesaplama açısından pahalı olabilir[26]. Gradient Boosting'in matematiksel formülasyonunu oluşturan iki ana parça vardır. Hedef fonksiyon ve tahmin fonksiyonu, Gradient Boosting'de düşünülmesi gereken unsurlardır. Hedef fonksiyonu, bir kayıp fonksiyonu ve düzenleme terimi içerir ve eğitimin amacı model parametrelerini maksimize etmektir.

$$obj = \Sigma L(y, \hat{y}) + \Omega(f) \quad (2)$$

Burada obj amaç fonksiyonu, y gerçek etiket, ŷ tahmin edilen etiket, L kayıp fonksiyonu, f karar ağacı ve Ω düzenleme terimidir. Gradient Boosting'deki tahmin işlevi, nihai tahmini yapmak için birden fazla karar ağacının tahminlerini birleştirir ve şu şekilde yazılabilir:

$$\hat{y} = \Sigma f(x) \quad (3)$$

Burada ŷ tahmin edilen etiket, f bireysel karar ağacı ve x ise girdi verileridir[27].

- Lojistik Regresyon (LR)

Genellikle ikili sınıflandırma görevlerinde kullanılan istatistiksel bir modeldir. Temel amacı, verilen girdi özellikleri temelinde, bir girdinin iki ayrı sınıftan birine ait olma olasılığını tahmin etmektir. Herhangi bir bağımsız değişken için eğitim verilerinin gözlemlerini artırmak için katsayılar kullanılır. Lojistik fonksiyonlardaki olasılıkların hesaplanması, doğrusal regresyonun bulgularına dayanır. Modelin optimize edilmesi için çözücü yöntemi ve düzenleme parametresi kullanılır. Finans, pazarlama, sağlık ve sosyal bilimler gibi alanlarda Lojistik Regresyon'un popülerliği, basitliği, yorumlanabilirliği ve büyük veri kümeleri ile başa çıkma yeteneğine dayanabilir. Lojistik regresyon formülü aşağıdaki gibidir:

$$p(z) = \frac{1}{1+e^{-z}} \quad (4)$$

Denklem, 'z' ile gösterilen giriş özelliklerinin doğrusal bir kombinasyonunu kullanır, 'p' pozitif sınıf olasılığını ve 'e' 2.71828 sayısını temsil eder. z, giriş özelliklerinin doğrusal bir kombinasyonudur:

$$z = b0 + b1x1 + b2x2 + ... + bnxn \quad (5)$$

Burada, b0 kesişme katsayısıdır, b1'den bn'ye kadar olan katsayılar ise x1'den xn'ye kadar olan girdi özelliklerinin katsayılarıdır.[28].

- eXtreme Gradient Boosting (XGBoost)

Gradient Boosting tekniğinin bir optimizasyonudur. Algoritma, makine öğrenimi alanında geniş bir şekilde kullanılmaktadır. Sınıflandırma ve regresyon uygulanabilir. Tianqi Chen tarafından 2014 yılında oluşturulan algoritma, yapısal verilerle etkili bir şekilde başa çıkma yeteneği nedeniyle büyük ilgi görmüştür[29]. XGBoost, aşırı uyum sorununu hafifletmek için karar ağaçları ve düzenleme yöntemlerini kullanan sofistike bir gradyan artırma algoritmasıdır. Yazılım, çok sayıda özelliği içeren büyük veri

kümelerini etkili bir şekilde işleme yeteneğine sahiptir, bu da onu regresyon ve sınıflandırma gibi çeşitli görevler için uygun hale getirir. XGBoost, karar ağaçlarını otomatik olarak budama ve özel kayıp fonksiyonları ile değerlendirme metriklerini içerme yeteneğine sahiptir. Algoritmanın dikkat çekici hızı, hassasiyeti ve uyarlanabilirliği, öneri sistemleri, sahtekarlık tespiti, görüntü sınıflandırma ve doğal dil işleme (NLP) gibi çeşitli alanlarda yaygın bir şekilde uygulanmasına yol açmıştır. XGBoost matematiksel ifadesi iki bölümden oluşur. XGBoost, eğitim sırasında model parametrelerini optimize eden Hedef Fonksiyon ile başlar. Kayıp ve düzenleme terimleri matematiksel ifadeyi oluşturur.

$$Obj = L(y, \hat{y}) + \Omega(\hat{W}) \quad (6)$$

XGBoost modeli matematiksel denklemi, gerçek etiketi (y), öngörülen etiketi (ŷ), kayıp işlevini (L), model parametreleri kümesini (Ŵ) ve düzenleme terimini (Ω) kullanan bir nesnel işleve (obj) sahiptir. Tahmin fonksiyonu, son tahmini yapmak için birden fazla karar ağacını birleştirir ve şu şekilde yazılabilir:

$$\hat{y} = \sum f(x, T) \quad (7)$$

Tahmin edilen etiket nerede, bireysel karar ağacıdır, x girdi verisidir ve Bu ağaçtaki bölme kuralları kümesidir[27].

- Karar Ağacı

Hem kategorik hem de sayısal verilere uygun esnek bir sınıflandırma tekniğidir. Bu varlıkların yapıları, ağaç benzeri bir yapıya benzer, bu da sezgisel ve basit bir anlayışı mümkün kılar. Karar ağaçları, verileri hiyerarşik bir formatta düzenleme yetenekleri nedeniyle karar verme sürecini kolaylaştırmak için tıbbi alanında sıkça kullanılır. Ağaç, belirli koşullar tarafından belirlenen farklı yolları veya seçenekleri görsel bir temsil olarak görev yapar. Süreç, birincil bir düğümle başlar ve ağaç benzeri bir yapı sergileyerek çeşitli ardışık yanıtlara genişler. Karar ağaçlarının basitliği ve etkinliği, onları tıbbi veri kümelerinde geniş çapta tanınan ve kullanılan yapmaktadır[30]. ID3 algoritması, sınıf ve öznitelik entropilerini türetir, her öznitelik için bilgi kazancını 9, 10 ve 11 numaralı denklemleri kullanarak hesaplar ve en yüksek bilgi kazancına sahip kök düğümü seçer. Bu özellik en bilgi dolu olarak kabul edilir. Yukarıdaki yöntemi tüm özellikler ağaç yapısında olana kadar tekrarlayın[31].

$$Info = -\sum_{i=1}^{m} Pi * log_2(Pi) \quad (8)$$

$$Infon_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * info(D_j) \quad (9)$$

$$Gain(A) = info\ (D) - \ info_A(D) \qquad (10)$$

- K-En Yakın Komşular (KNN)

Bilgisayarlar tarafından verileri sınıflandırmada kullanılan bir makine öğrenimi tekniğidir. Bunu, yeni verileri daha önce gözlemlediği benzer verilerle eşleştirerek gerçekleştirir. Bu yaklaşım, sınıflandırmayla ilgili görevlerdeki etkinliği ve verimliliği ile bilinir. Algoritma, benzerlik gösteren örneklerin bir araya gelme eğiliminde olduğu temel prensibe dayanır, bu da yakın konumda bulunan örneklerin daha büyük bir olasılıkla birbirine bağlı olduğu anlamına gelir. Parametre k'ya atanan sayısal değer, belirli bir veri noktası için dikkate alınan komşu veri noktalarının miktarını belirler. Algoritma, yeni edinilen verinin mevcut veri kümesine olan yakınlığını ölçmek için mesafe fonksiyonlarını hesaplar. Yeni verinin sınıf atanması, k en yakın komşularının öznitelik değerlerine dayanarak belirlenir. Sonuç olarak, veri noktası uygun bir etiketle belirlenir[32].[33]'e göre, KNN'de kullanılan mesafe metriği değişebilir; ancak genellikle Öklidyen mesafe ve Manhattan mesafesi kullanılan mesafe ölçümlerindendir. Diyelim ki y1, y2,..., yn özelliklere sahip bir eğitim veri noktamız xj ve x1, x2,..., xn özelliklere sahip bir test veri noktamız xi olsun. Bu iki nokta arasındaki Öklidyen mesafesini hesaplamak için kullanılan formül şöyledir:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (11)$$

Ve bu iki nokta arasındaki Manhattan mesafesi şu şekilde hesaplanır:

$$d = \sum_{i=1}^{n} |x_i - \ y_i| \qquad (12)$$

- Destek Vektör Makinesi (DVM)

Denetimli makine öğrenimi alanında yaygın olarak kullanılan bir tekniktir. Kullanılan metodoloji, veriyi daha yüksek boyutlu bir uzayda sınıflandırmak için bir hiperdüzlem kullanır ve marjların optimize edilmesi yoluyla bir doğrusal sınıflandırıcı olarak işlev görür[34]. Marj, hiperdüzlem ile her sınıftan en yakın destek vektörler arasındaki mesafedir. İki belirli sınıf için hiperdüzlem ideal bir ayıracıdır[35]. Destek Vektör Makinesi ilk olarak doğrusal bir sınıflandırıcı olarak geliştirilmiştir. Ancak o zamandan beri, daha yüksek boyutlu bir uzayda çalışan çekirdek tekniğini kullanarak doğrusal olmayan problemleri ele almak için geliştirmeler yaşamıştır. Bu araştırmada, giriş uzayını özellik uzayına eşlemek için RBF çekirdeği kullanılmaktadır[36]. Marjı maksimize etmek için destek vektörlerini kullanarak ideal hiperdüzlem belirlenir. Denklem 14'teki w ve b'nin en iyi değerleri.

$$f(x) = sign(w.x + b) \qquad (13)$$

DVM'Yİ eğitmek için optimizasyon problemi, nesnel işlevi en aza indirmeyi içerir:

$$minimize: \frac{1}{2||w||^2} + \ C \sum max(0, 1 - \ y_i(w.x_i + b)) \qquad (14)$$

Burada:

$||w||^2$, ağırlık vektörünün kare Öklid normunu temsil eder (düzenleme için kullanılır).

C, marjın maksimize edilmesi ile sınıflandırma hatalarının en aza indirilmesi arasındaki dengeyi kontrol eden bir hiperparametredir.

∑max (0,1-yi (w·xi + b)), yanlış sınıflandırılmış örnekleri cezalandıran menteşe kaybını temsil eder.

3.2. Veri Artırma

Bir Makine Öğrenimi çözümü, doğruluk beklentilerini karşılayamadığında, genellikle modeli suçlamak doğal bir tepkidir. Bu nedenle, performansı artırmak amacıyla kullanıcılar genellikle daha karmaşık yöntemlerle değiştirmeye yönlenebilirler. Ancak, genellikle kalitenin modelde değil, modele sağlanan veri olduğu bilinmelidir. Veriyi kasıtlı olarak değiştirerek eğitim verisi eklemek için güçlü bir makine öğrenimi yöntemi olan veri artırma, alan uzmanlığı veya Derin Öğrenme kullanılarak yapılabilir. Bu da güncellenmiş örnekler oluşturmak için otomatik olarak yapılabilir. Örneğin, giriş verilerinin çeşitliliği olmayabilir. Veri artırma yaklaşımları, makine öğrenimi modellerine yapay değişiklikler getirerek, modellerin geniş bir koşul yelpazesine karşı direncini ve esnekliğini artırarak çalışırlar[37]. En yüksek performans seviyesine ulaşmak için, veri artırma tekniklerini dahil etmek önemlidir. Veri artırma, denetimli öğrenimde temel gerçek etiketlemeleri sağlayarak önemli bir rol oynar. Bu süreç, modellerin etkili bir şekilde desenleri tanımlamasına ve kesin tahminler yapmasına olanak tanır. Bu prosedür, modellerin güçlü genelleme yeteneklerine sahip olmasını sağlar, bu da aşırı uyuma karşı olasılığı azaltır ve daha önce karşılaşılmamış verilerde performanslarını artırır. Ek veri ekleyerek, veri kümesinin çeşitliliğini ve direncini artırarak çeşitli düzenlemeler aracılığıyla yeni eğitim örnekleri oluştururuz. Sonuç olarak, artırılmış verinin eklenmesi, veri kümesinin genel boyutunu artırır ve bu da aşırı uyuma karşı sorunu azaltır ve modellerin etkili bir şekilde genelleme yeteneklerini artırır. Bu, genellikle küçük ölçekte olan veya dağılımlarında dengesizlik gösteren orijinal veri kümeleri ile uğraşılırken özellikle avantajlıdır[38]. Ayrıca, veri artırma tekniklerinin kullanımı, modelin

daha geniş bir varyasyon ve koşul yelpazesi ile karşılaşmasına olanak tanır, bu da gerçek dünya senaryolarını simüle eder ve nihayetinde modelin genel performansını artırır[39].

Veri artırma, makine öğrenimi modellerinin etkinliğini büyük ölçüde artıran birçok avantaj sunmaktadır. Başlangıçta, bu yaklaşım bize eğitim verilerinin yetersizliği kısıtlamasını yapay olarak artırarak ele alma imkanı sağlar. Bu, başlangıç veri setinin sınırlı boyutta veya gerçek hayat durumlarını doğru bir şekilde yansıtmak için gerekli çeşitliliği taşımadığı durumlarda özellikle önemlidir. Ayrıca, Veri Artırmanın kullanımı, modelin üretim aşamasındaki performansını önemli ölçüde artırır. Geniş bir eğitim verisi yelpazesi ekleyerek model, farklı senaryolarda etkili bir şekilde gezinme ve güçlü genelleme yetenekleri sergileme yeteneğini artırır, böylece gerçek dünya bağlamında pratikte daha etkili olur[37].

Veri Artırmanın önemli avantajlarından biri, ek manuel etiketleme veya ön işleme gerektirmeksizin yeni veri üretme kapasitesidir. Bu uygulama, önemli zaman ve kaynakların korunmasına neden olur, bu da eğitim sürecinin verimliliğini artırır. Ayrıca, Veri Artırmanın kullanımı, makine öğrenimi alanında karşılaşılan yaygın bir engel olan aşırı uyumu azaltma konusunda erken bir strateji olarak hizmet eder. Veri setini genişleterek ve çeşitli varyasyonları dahil ederek, modelin eğitim verilerini ezberleme eğilimi azalır ve böylece önemli desenlere odaklanabilir. Sonuç olarak, model, yeni, görülmemiş verilerle karşılaştığında gelişmiş genelleme yetenekleri sergiler. Ayrıca, bu metodoloji, modelimizin gerçek dünya durumlarına uygulandığında etkili olmasında kritik bir rol oynayan belirli özellikleri vurgulamamıza olanak tanır. Veri Artırma, çeşitli makine öğrenimi zorluklarını ele alır. Modern makine öğrenimi iş akışları bunu gerektirir çünkü eğitim verisini iyileştirir, model performansını artırır ve manuel çaba olmadan stabiliteyi destekler[38].

- Veri Artırmanın Avantajları

Veri artırmanın kullanımı, daha geniş bir giriş veri yelpazesiyle makine öğrenimi modellerini eğitme bağlamında bir dizi avantaj sunmaktadır. İlk olarak, artırılmış veri kullanımı, modele çeşitli senaryoları etkili bir şekilde ele alma yeteneği kazandırarak modellerin hassasiyetini artırır[46]. Ayrıca, veri artırma tekniklerinin kullanımı, aşırı uyumu azaltmaya yardımcı olur; bu, modelin eğitim veri setinde tatmin edici performans sergilemesine rağmen yeni, görülmemiş verilerde zayıf performans gösteren yaygın bir

endişe olarak karakterizedir. Modelin artırılmış örneklerle karşılaşması sayesinde, daha önce karşılaşmadığı örneklerle daha etkili bir şekilde genelleme yapabilme yeteneği kazanır, bu da genel performansını artırır[47]. Ayrıca, veri artırmanın kullanımının, modelin genelleme kapasitesini artırmada avantajlı olduğu ve bu da yeni ve bilinmeyen bilgiler üzerinde daha kesin tahminler yapmasına izin verdiği bulunmuştur. Veri artırmanın dikkate değer bir avantajı, önemli bir eğitim verisi hacmini edinme ve etiketleme ile ilişkili zorlukları azaltma kapasitesidir. Mevcut veriyi geliştirerek, veri kümesini etkili bir şekilde genişletmek mümkündür, bu da modele bilgi edinme için daha fazla örnek sağlar ve ek veri toplama maliyetleriyle ilgili önemli masraflardan kaçınmaya yardımcı olur. Bu olay, model geliştirme sürecini hızlandırır ve gerekli kaynakları azaltır[48]. Ayrıca, veri artırma, bazı sınıfların eksik olduğu dengesiz veri setleriyle çalışırken bir kurtarıcıdır. Azınlık sınıfının örneklerine artırma yöntemlerini uygulayarak, sınıf dağılımını dengeleyebilir ve modelin tüm sınıfları tanıma ve işleme konusunda daha iyi olmasını sağlayabiliriz. Bu genellikle kategorilendirme zorluklarında güvenilir sonuçlar elde etmek için esastır[49].

- Veri Artırmanın Dezavantajları

Veri artırmanın birçok avantajı olmasına rağmen, belirli sınırlamaların varlığını kabul etmek önemlidir. Dikkate değer bir sorun, veri artırma sürecinin önemli ölçüde zaman ve hesaplama kaynakları yatırımını gerektirebilmesinden kaynaklanır. Birçok artırma tekniği, giriş parametrelerinin dikkatlice ayarlanmasını gerektirir, bu da daha uzun eğitim süreleri ve artan kaynak talepleri ile sonuçlanabilir. Başka bir sınırlama, veri artırma süreci sırasında ortaya çıkabilecek bilgi kaybı olasılığıdır. Kesme veya yakınlaştırma gibi işlemler sırasında veriden önemli ayrıntıların kazara çıkarılması, modelin belirli görevlerdeki performansını olumsuz etkileyebilir[50]. Ayrıca, yeterli denetim olmadan veri artırmanın uygulanması, gerçeklikten sapma eğiliminde olan veri oluşturulmasına neden olabilir. Sentetik veriler, gerçek dünyada gözlemlenen dağılığı doğru bir şekilde yansıtmayabilir, bu da modele aldatıcı desenlerden öğrenme ve ardından pratik durumlarda performansını etkileme olasılığına yol açabilir. Dahası, başlangıç veri kümesi önyargılı olduğunda, veri artırma sürecinin bu önyargıların etkisini artırma potansiyeli bulunmaktadır. Eğer artırma süreci önceden var olan önyargıları güçlendirmeye hizmet ediyorsa, modelin karar verirken adil olmayan bir davranış sergileme olasılığı vardır. Bu sınırlamaları azaltmak için artırma yöntemlerini ve parametreleri dikkatlice seçmek önemlidir. Geliştirilmiş verinin model performansını nasıl etkilediğini belirlemek için kapsamlı çalışmalara ihtiyaç vardır. Ayrıca,

başlangıç veri kümesindeki herhangi bir önyargıyı tanımak ve dikkatlice artırma sırasında bunları artırmaktan kaçınmak önemlidir. Bu kısıtlamaları dikkate alarak ve veri artırma yöntemlerini dikkatli bir şekilde uygulayarak, makine öğrenimi modeli üzerindeki olumsuz etkileri en aza indirirken avantajları en üst düzeye çıkarabiliriz[51].

- Veri Artırma Uygulama Alanları

Veri büyütme uygulamaları, veri büyütme tekniklerinin kullanılmasının makine öğrenimi modellerinin etkinliğini ve hassasiyetini artırabileceği belirli örnekleri ifade eder. Daha geniş ve daha çeşitli bir eğitim verisi kümesi oluşturmak, bu hedefin gerçekleştirilebileceği araçtır. Veri büyütme stratejilerinin başarıyla uygulandığı birçok alandan bazıları şunlardır:

- Bilgisayar görüntüleme, resim sınıflandırma, nesne tespiti, yüz tanıma, semantik bölümlendirme ve resim oluşturma gibi çeşitli görevleri içeren bir alanı kapsar. Bu görevlerin performansını artırmak için veri artırma tekniklerinin faydalı olduğu bulunmuştur. Veri artırma, resimlerin yönlendirmelerini, boyutlarını, dönmelerini, renklerini ve gürültü eklemelerini içeren diğer değişikliklerle resimlerin ek varyasyonlarını üretmek için kullanılan bir tekniktir[52].

- Doğal Dil İşleme, metin sınıflandırma, duygu analizi, soru cevaplama, metin özetleme ve metin oluşturma gibi çeşitli görevleri içeren bir alandır. Bu NLP görevlerinin performansını artırmak için veri artırma tekniklerinin faydalı olduğu bulunmuştur. Veri artırma, kelime dağarcığı, cümle yapısı, dil bilgisi, yazım ve diğer dilbilgisel yönler açısından çeşitlilik ekleyerek ek metin örnekleri üretmek için kullanılan bir tekniktir[53].

- Konuşma işleme, konuşma tanıma, konuşmacı tanımlama, konuşma sentezi ve konuşma duygu tanıma gibi bir dizi uygulamayı içeren bir alanı kapsar. Veri artırma tekniklerinin kullanımının, söz konusu uygulamaların performansını artırmada avantajlı sonuçlar doğurduğu gözlemlenmiştir. Veri artırma teknikleri, gürültü, ton, hız, ses düzeyi ve diğer ilgili özelliklerde çeşitlilik gösteren ek ses örnekleri üretmek için kullanılabilir[54].

- ❖ Tıbbi görüntü veri artırma, görüntü analizi, teşhis, segmentasyon, tespit ve oluşturma konularında yardımcı olabilir. Veri artırma, çeşitli kontrastlar, parlaklıklar, dönmeler, deformasyonlar ve benzeri özelliklere sahip yeni tıbbi görüntüler oluşturabilir[55].

- ➢ Veri Artırma Teknikleri

Veri artırma teknikleri, önceden var olan veri noktalarının değiştirilmiş kopyalarını oluşturarak bir veri kümesinin hacmini ve değişkenliğini yapay olarak geliştirmek için kullanılan bir dizi metodolojiyi ifade eder. Kullanımlarının potansiyel faydalarından biri, makine öğrenimi modellerinin, özellikle de derin öğrenme tekniklerini kullananların, aşırı uyum sorununu hafifleterek ve iyileştirilmiş genelleme yeteneklerini teşvik ederek performansını ve hassasiyetini artırma potansiyelidir[39]. Veri büyütme teknikleri, ses, metin, resim ve video dahil olmak üzere çeşitli veri yöntemlerinde kullanılabilir. Veri arttırma tekniklerine örnek Görüntü, Metin, Ses için verilebilir.

Görüntü veri artırma teknikleri, bir makine öğrenimi modelini eğitmek için yeni ve çeşitli örnekler oluşturmak amacıyla görüntüleri değiştirme yöntemleridir. Bu teknikler, aşırı uydurmayı önleyerek ve genelleştirmeyi artırarak görüntü sınıflandırma, bölütleme, tespit ve üretim modellerinin performansını ve doğruluğunu artırmaya yardımcı olabilir[40]. Birkaç görüntü veri artırma tekniği kullanılabilir, bunlar arasında boyutlandırma ve ölçeklendirme bulunmaktadır. Boyutlandırma, görüntünün boyutlarını belirli bir boyuta değiştirme işlemidir, ölçeklendirme ise piksel değerlerini belirli bir aralığa uydurmak, örneğin [0, 1] veya [-1, 1] gibi bir aralığa ayarlamak anlamına gelir. Kırpma, bir görüntünün belirli bir bölümünü yeni görüntü olarak kullanma eylemini ifade eder. Çevirme, bir görüntüyü yatay veya dikey olarak tersine çevirme sürecidir. Dolgu, bir görüntüye sınırlar veya kenarlar eklemenin bir yöntemidir. Döndürme, bir görüntünün yönetimini belirli bir açıyla değiştirme işlemidir. Rastgele afin dönüşüm, görüntüye birleştirme, döndürme, ölçekleme ve kaydırma işlemlerinin kombinasyonunu uygulama işlemidir. Gauss bulanıklığı işlemi, bir görüntüye gürültüyü azaltma ve mevcut detay seviyesini azaltma amacıyla bir bulanıklık filtresi uygulama işlemidir. Renk uzayı dönüşümleri, bir görüntünün çeşitli görsel özelliklerini değiştirme işlemidir, örneğin parlaklık, kontrast, renk tonu, doygunluk veya gamma gibi. Rastgele silme, bir görüntünün belirli bölümlerini rastgele bir şekilde seçici bir şekilde silme işlemidir. Birden çok görüntüyü birleştirme ve karıştırma işlemi, farklı görsel öğelerin birleştirilmesini içerir[41].

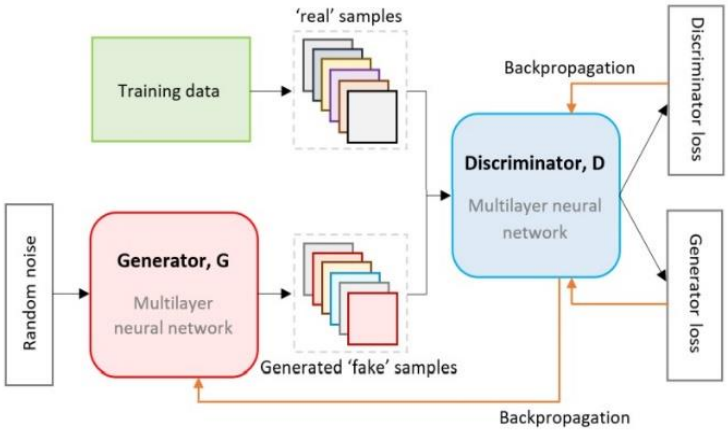
Metin verisi artırma teknikleri, bir makine öğrenimi modelini eğitmek amacıyla metin verisini çeşitlendirmek için kullanılan bir dizi yöntemi ifade eder. Düzenleme tekniklerinin kullanılmasının potansiyel bir faydası, doğal dil işleme (NLP) modellerinde performans ve doğruluğun potansiyel olarak artırılabilmesidir. Bu teknikler, uygulandığında aşırı uyum sorununu etkili bir şekilde azaltabilir ve metin sınıflandırma, soru cevaplama, özetleme ve üretim modelleri de dahil olmak üzere çeşitli NLP görevlerinde geliştirilmiş genelleme sağlayabilir[42]. Metin verisi artırma tekniklerine örnek olarak şunlar sayılabilir: Kelime veya cümle karıştırma, belirli bir metin içinde bir kelimenin veya cümlenin yerini rastgele değiştirme eylemini ifade eder. Kelime değiştirme, kelimelerin sinonimleri, antonimleri, hiperonimleri, hiponimleri veya yanlış yazılmış varyasyonları ile değiştirilmesini içerir. Sözdizimi ağacı manipülasyonu, aynı kelimeleri kullanarak ancak farklı dilbilgisi yapıları kullanarak bir cümleyi tekrar ifade etme işlemini içerir. Rastgele kelime ekleme tekniği, belirli bir metin içinde rastgele konumlara kelimeler eklemenin eylemini içerir. Diğer yandan, rastgele kelime silme, belirli bir metin içinde rastgele konumlardan kelimeleri çıkarma işlemine karşılık gelir. Geri çeviri işlemi, belirli bir metni farklı bir dile çevirme ve ardından onu orijinal dile geri çevirme işlemini içerir. Maskeleme dil modelleme tekniği, belirli bir metin içinde belirli kelimeleri bilerek gizleme ve daha sonra önceden eğitilmiş bir dil modelini kullanarak bu maskeleme kelimeleri tahmin etme işlemini içerir[43].

Ses verisi artırma teknikleri, bir makine öğrenimi modelini eğitmek amacıyla ses verisini çeşitlendirmek için kullanılan bir dizi yöntemi ifade eder. Düzenleme tekniklerinin kullanılmasının, konuşma tanıma, konuşmacı tanıma, ses sınıflandırma ve ses üretim modelleri de dahil olmak üzere ses işleme modellerinde performans ve hassasiyetin artırılmasına katkıda bulunabileceği belirtilmiştir. Bu, aşırı uyumun önlenmesi ve genelleme yeteneklerinin teşvik edilmesi yoluyla gerçekleştirilir[44]. Yaygın olarak kullanılan ses verisi artırma tekniklerinden biri, ses verisine Gaussian veya rastgele gürültü eklemeyi içeren gürültü enjeksiyonudur. Bu teknik, modelin direncini artırmayı amaçlar. Önerilen işlevsellik, sesin çalma konumunu rastgele bir saniye sayısını sola veya sağa kaydırarak ayarlama yeteneğini içerir. Hızı değiştirmek, zaman serisini sabit bir faktöre göre ayarlamayı içerir, bunu uzatma veya sıkıştırma şeklinde gerçekleştirebilir. Diğer yandan, pitch'i değiştirmek, sesin tonuna rastgele varyasyonlar eklemeyi içerirken hızı sabit tutmayı içerir. Arka plan gürültüsü karıştırma işlemi, farklı ses kaynaklarını içeren sesleri karıştırma ve entegrasyonu içerir. Frekans maskeleme işlemi, ses spektrumu içinde belirli frekans

bantlarının rastgele silinmesini içerir. Yankı işlemi kavramı, çeşitli akustik ortamları çoğaltmak için oda darbe tepkisi kullanımını içerir[45].

3.3. Çekişmeli Üretici Ağlar (Generative Adversarial Networks - GAN)

Makine öğrenimi ve bilgisayarla görme, mevcut olanlardan yeni veri örnekleri oluşturarak eğitim veri kümelerini genişletmek ve geliştirmek için veri artırmayı kullanır. Ham veri kümesine kırpma, çevirme, döndürme ve gürültü ekleme yöntemin bir parçasıdır. Veri büyütmenin temel amacı, bir modelin genelleme yeteneğini geliştirmek için bir veri kümesinin miktarını ve çeşitliliğini genişletmektir. Modeli daha geniş bir örnek dizisine tabi tutarak, aşırı uyuma karşı daha az savunmasız hale gelir ve bilinmeyen verileri işlemede daha iyi hale gelir. Bu yöntem, modelin yetkin öğrenme için veri eksikliğiyle karşılaşabileceği ve sonuç olarak aşırı uyum eğilimini gösterebileceği kısıtlı veri kümelerini içeren senaryolarda özellikle faydalıdır. Eldeki zorluk, algoritmanın eğitimini iyileştirmek için ek verilerin sağlanmasını içeren veri artırmanın benimsenmesiyle etkin bir şekilde ele alınmaktadır[37]. Sinir ağları, beyni modelleyen makine öğrenimi algoritmalarıdır. Birbirine bağlı düğümler olan nöronlar, katmanlar halinde sinir ağları oluşturur. Her katmanın nöronları, ağırlıklı bağlantılar yoluyla bir sonraki katmana bağlanır. Bir sinir ağı, bağlantı ağırlıklarını yinelemeli olarak ayarlayarak girdileri ve çıktıları ilişkilendirmeyi öğrenir. Buna geri yayılım denir. Geri yayılım algoritması, bağlantı ağırlıklarını ayarlayarak beklenen çıktı ile gözlemlenen çıktı arasındaki değişimi en aza indirir. Sinir ağları konuşma, resim tanımlama, nesne algılama, doğal dil anlayışı ve genetikte etkilidir. Araştırmalar, karmaşık ve çok boyutlu veri kümesi problemlerini çözmedeki etkinliğini göstermektedir[56].



Şekil 1. Çekişmeli Üretici Ağların Mimarisi

Çekişmeli Üretici Ağlar (ÇÜA), mevcut veriyi taklit eden yeni veri üretebilen makine öğrenimi algoritmalarıdır. ÇÜA'lar, bir üretici model ve bir ayırt edici modelden oluşur. Üretici model, yeni veri üretme sorumluluğuna sahiptir. Bu süreç, rastgele bir girişle başlamayı ve daha sonra girişi iteratif olarak değiştirmeyi içerir, böylece mevcut veriyi yakından taklit eden bir çıkış üretir. Ayırt edici model, gerçek veri ile sentezlenmiş veri arasında ayırt etme sorumluluğuna sahiptir. Bu, gerçek veriyi sentezlenmiş veriden ayırt etmek için gerekli olan özellikleri tanımayı içeren bir süreçle elde edilir. İki model, bir minimax oyununda eğitilir. Üretici modelin amacı, ayırt edici modelin onu gerçek veriden ayırt edemediği kadar gerçekçi veri üretmektir. Diğer taraftan, ayırt edici model, gerçek ve sentezlenmiş veri arasındaki ayrım yeteneğini geliştirmeye çalışır. Her iki model de eğitim yoluyla görevlerini geliştirir. Ayırt edici model, gerçek veriyi oluşturulan veriden ayırt etme konusunda daha iyileşirken, üretici model, gerçek veriyi taklit eden veri üretme konusunda daha iyileşir[57]. ÇÜA eğitim hedefi şu şekildedir:

$$V_{GAN}(D, G) = E_{x \sim pdata(x)}[\log(D(x))] + E_{z \sim pz(z)}[\log(1 - D(G(z)))], \qquad (15)$$

Burada:
Ex∼pdata(x): Beklenen veri dağılımı. Log (D (x)) işlevinin tüm gerçek veri noktalarında ortalaması alınır.
log(D(x)): Ayırıcı ağ için kayıp işlevi. Gerçek bir veri noktası için ayırıcı çıktısının logaritmasıdır.
Ez∼pz(z): Gürültü dağılımı üzerindeki beklenti. log (1-D (G (z))) fonksiyonunun tüm gürültü noktaları üzerindeki ortalaması.
log(1-D(G(z))): Jeneratör ağı için kayıp fonksiyonu. Oluşturulan bir veri noktası için ayırıcı çıktısının tamamlayıcısının logaritmasıdır.

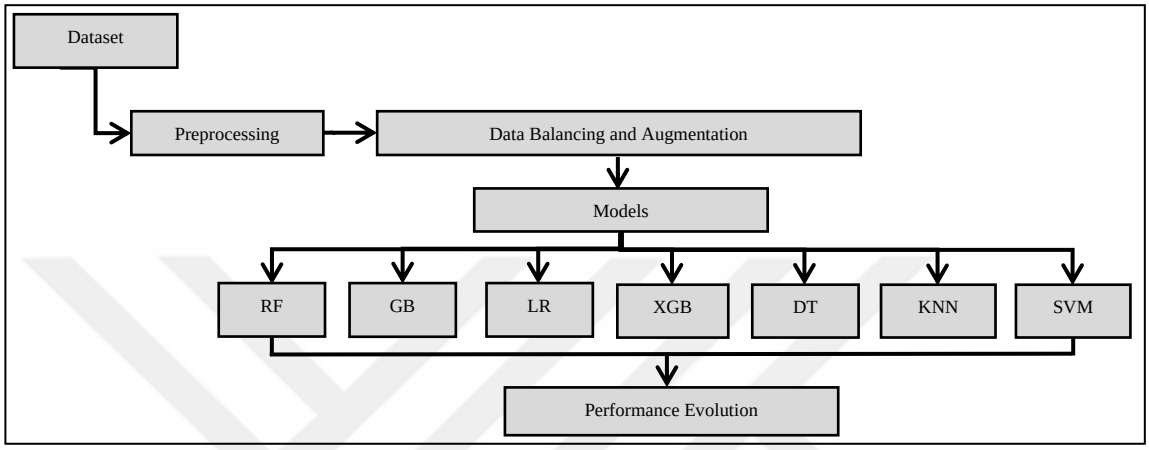
Ayrımcı model, eğitim verilerine doğrudan erişimi olmayan üretken modele geri bildirim sağlar. Ayrımcı model ise gerçeği sahte örneklerden ayırt etmek için eğitim verilerini kullanır. Ayrımcı model, sentetik verileri tespit etme yeteneğini artırmaya çalışırken, üretken model onu aldatmaya çalışır. Rekabet, her iki modelin performansını artırır. ÇÜA'lar gerçeğe yakın görüntüler, metinler ve diğer veri türleri oluşturur. Uygulamaları resim sentezi, metin oluşturma ve veri geliştirmeyi içerir[58].

4. DENEYSEL ÇALIŞMA VE SONUÇLAR

Bu çalışmada Çekişmeli Üretici Ağlar kullanılarak yetersiz veya temsil özelliği zayıf veri kümelerinin artırmak ve böylece makine öğrenimi algoritmalarının başarısını iyileştirmek amaçlanmıştır. ÇÜA tarafından oluşturulan örnekler, orijinal veri kümesinin temel dağılımını etkili bir şekilde yakalamış ve daha önce keşfedilmemiş kalıpları ortaya çıkarmaya yardımcı olan yeni veri noktaları türetebilmiştir. Makine öğrenimi ve bilgisayarla görme, mevcut olanlardan yeni veri örnekleri oluşturarak eğitim veri kümelerini genişletmek ve geliştirmek için veri artırmayı kullanır. Ham veri kümesine kırpma, çevirme, döndürme ve gürültü ekleme yöntemin bir parçasıdır. Veri büyütmenin temel amacı, bir modelin genelleme yeteneğini geliştirmek için bir veri kümesinin miktarını ve çeşitliliğini genişletmektir. Modeli daha geniş bir örnek dizisine tabi tutarak, aşırı uyuma karşı daha az savunmasız hale gelir ve bilinmeyen verileri işlemede daha usta hale gelir. Bu metodoloji, modelin yetkin öğrenme için veri eksikliğiyle karşılaşabileceği ve sonuç olarak aşırı uyum eğilimini gösterebileceği kısıtlı veri kümelerini içeren senaryolarda özellikle faydalıdır. Eldeki zorluk, algoritmanın eğitimini iyileştirmek için ek verilerin sağlanmasını içeren veri artırmanın benimsenmesiyle etkin bir şekilde ele alınmaktadır[37].

Bu çalışmanın temel amacı, veri büyütme tekniklerini kullanma bağlamında, makine öğrenimi algoritmalarının yetersiz veri nedeniyle bozulmalarına karşın iyileştirilmesidir. Bu amaca ulaşmak için, çekişmeli üretici ağlar kullanımı, ilk veri kümesinin dağılımına yüksek derecede benzerliğe sahip ek örnekler oluşturmaya yönelik bir yaklaşım olarak kullanıldı, çekişmeli üretici ağ yeni veri noktalarının tanıtılmasına izin verdi. Bu daha sonra, daha önce makine öğrenimi modelleri tarafından gözlemlenmeyen kalıpların tanımlanmasını ve tanınmasını kolaylaştırdı. Şekil 2, tezdeki deneyleri yürütmek için kullanılan yaklaşımı göstermektedir. Metinsel özellikler analizden çıkarılarak sadece sayısal özelliklere odaklanıldı. Bulgularımız, modelin performansını veri büyütme tekniklerinin uygulanmasıyla ve uygulanmadan karşılaştırırken doğrulukta kayda değer bir artış olduğunu göstermektedir. Bulgular, ÇÜA kullanılarak veri büyütmenin modellerin doğruluğunda iyileşme gösterdiğini göstermektedir.

Deneysel çalışmada, veri önişlemde scikit-learn PowerTransformer kullanılmıştır. Özellik varyanslarını stabilize etmek ve normal olmayan dağılımları ele almak için Yeo-Johnson yaklaşımı tercih edilmiştir. Veri, analizi iyileştirmek ve makine öğrenimi algoritmalarını daha esnek hale getirmek için daha Gauss benzeri ve aykırı değerlere karşı daha esnek hale getirilmiştir.



Şekil 2. Araştırma sürecinin adımları

Deneysel çalışmada kullanılan performans değerlendirme metrikleri Tablo 1 de verildiği gibidir.

Tablo 1. Performans Değerlendirme Metrikleri

Precision	TP / (TP + FP)
Sensitivity (recall)	TP / (TP + FN)
Specificity	TN / (TN + FP)
F1-Score	2 * (precision * sensitivity) / (precision + sensitivity)
Accuracy	(TP + TN) / (TP + TN + FP + FN)

Burada:

TP = doğru tahmin edilen pozitif örnekleri
FP = yanlış tahmin edilen pozitif örnekleri,
TN = doğru tahmin edilen negatif örnekleri,
FN = yanlış tahmin edilen negatif örnekleri temsil eder.

4.1 Dengeli Bir Veri Kümesiyle Deneysel Tasarım

4.1.1 Reklam veri kümesi

Reklam veri kümesi Kaggle ‘da paylaşılan, gerçek dünya kaynaklarından elde edilen 1000 örnekten oluşan kapsamlı bir koleksiyon olup, her biri 8 benzersiz özellik içermektedir. " Daily Time Spent on Site " 32.6 dakikadan 91.43 dakikaya kadar herhangi bir yerde değişebilir; ortalama 65 dakika ve bu süre içinde standart sapma 15.85 dakikadır. Kullanıcıların yaşı 19 ile 61 arasında değişir ve ortalama yaş 36'dır ve bu yaş aralığında standart sapma 8.79'dur. Ortalama olarak, kullanıcılar 36 yaşındadır. " Area Income " aralığı 13.996,50 dolardan 79.484,80 dolara kadar uzanır, ortalama 55.000 dolar ve bu gelir aralığında standart sapma 13.414 dolardır. Ortalama 180 dakika ve standart sapma 43,90 dakika ile " Daily Internet Usage " 104,78 dakikadan 269,96 dakikaya kadar değişir. Genel olarak, ortalama 180 dakikadır. Veri setine göre, yaklaşık %48,1'i erkek olan cinsiyet dağılımı benzerdir. Örneğe göre, katılımcıların yüzde elli "Reklama Tıkladı", bu da katılımın eşit bir derecede olduğunu gösterir.

Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
68.95	35	61833.9	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	3/27/2016 0:53	0
80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	4/4/2016 1:39	0
69.47	26	59785.94	236.5	Organic bottom-line service-desk	Davidton	0	San Marino	3/13/2016 20:35	0
74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	1/10/2016 2:31	0
68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	6/3/2016 3:36	0
59.99	23	59761.56	226.74	Sharable client-driven software	Jamieberg	1	Norway	5/19/2016 14:30	0
88.91	33	53852.85	208.36	Enhanced dedicated support	Brandonstad	0	Myanmar	1/28/2016 20:59	0
66	48	24593.33	131.76	Reactive local challenge	Port Jefferybury	1	Australia	3/7/2016 1:40	1
74.53	30	68862	221.51	Configurable coherent function	West Colin	1	Grenada	4/18/2016 9:33	0
69.88	20	55642.32	183.82	Mandatory homogeneous architecture	Ramirezton	1	Ghana	7/11/2016 1:42	0
47.64	49	45632.51	122.02	Centralized neutral neural-net	West Brandonton	0	Qatar	3/16/2016 20:19	1
83.07	37	62491.01	230.87	Team-oriented grid-enabled Local Area Network	East Theresashire	1	Burundi	5/8/2016 8:10	0
69.57	48	51636.92	113.12	Centralized content-based focus group	West Katiefurt	1	Egypt	6/3/2016 1:14	1
79.52	24	51739.63	214.23	Synergistic fresh-thinking array	North Tara	0	Bosnia and Herzegovina	4/20/2016 21:49	0
42.95	33	30976	143.56	Grass-roots coherent extranet	West William	0	Barbados	3/24/2016 9:31	1
63.45	23	52182.23	140.64	Persistent demand-driven interface	New Travistown	1	Spain	3/9/2016 3:41	1
55.39	37	23936.86	129.41	Customizable multi-tasking website	West Dylanberg	0	Palestinian Territory	1/30/2016 19:20	1
82.03	41	71511.08	187.53	Intuitive dynamic attitude	Pruittmouth	0	Afghanistan	5/2/2016 7:00	0
54.7	36	31087.54	118.39	Grass-roots solution-oriented conglomeration	Jessicastad	1	British Indian Ocean Territory (Chagos Archipelago)	2/13/2016 7:53	1
74.58	40	23821.72	135.51	Advanced 24/7 productivity	Millertown	1	Russian Federation	2/27/2016 4:43	1
77.22	30	64802.33	224.44	Object-based reciprocal knowledgebase	Port Jacqueline	1	Cameroon	1/5/2016 7:52	0
84.59	35	60015.57	226.54	Streamlined non-volatile analyzer	Lake Nicole	1	Cameroon	3/18/2016 13:22	0
41.49	52	32635.7	164.83	Mandatory disintermediate utilization	South John	0	Burundi	5/20/2016 8:49	1
87.29	36	61628.72	209.93	Future-proofed methodical protocol	Pamelamouth	1	Korea	3/23/2016 9:43	0
41.39	41	68962.32	167.22	Exclusive neutral parallelism	Harperborough	0	Tokelau	6/13/2016 17:27	1
78.74	28	64828	204.79	Public-key foreground groupware	Port Danielleberg	1	Monaco	5/27/2016 15:25	0
48.53	28	38067.08	134.14	Ameliorated client-driven forecast	West Jeremyside	1	Tuvalu	2/8/2016 10:46	1
51.95	52	58295.82	129.23	Monitored systematic hierarchy	South Cathyfurt	0	Greece	7/19/2016 8:32	1
70.2	34	32708.94	119.2	Open-architected impactful productivity	Palmerside	0	British Virgin Islands	4/14/2016 5:08	1
76.02	22	46179.97	209.82	Business-focused value-added definition	West Guybury	0	Bouvet Island (Bouvetoya)	1/27/2016 12:38	0
67.64	35	51473.28	267.01	Programmable asymmetric data-warehouse	Phelpschester	1	Peru	7/2/2016 20:23	0
86.41	28	45593.93	207.48	Digitized static capability	Lake Melindamouth	1	Aruba	3/1/2016 22:13	0

Şekil 3. Veri Kümesi Örneği

4.1.2 Veri artırma olmaksızın elde edilen sonuçlar

Aşağıdaki bölüm, kullanıcı davranışı tahmini için uygulanan altı makine öğrenimi modelinin elde ettiği sonuçların analizini sunmaktadır. Bu araştırmanın ana odak noktası, veri artırmanın model performansı üzerindeki potansiyel etkisini değerlendirmektir. Araştırma, Rastgele Orman, Gradyan Artırma, Lojistik Regresyon, XGBoost, Karar Ağacı ve K-En Yakın Komşular gibi çeşitli denetimli makine öğrenimi tekniklerini içermektedir. Sonuçlar aşağıda özetlenmiştir.

Tablo 2'deki bulgulara dayanarak, rastgele orman modelinin kapsamlı bir doğruluk oranına sahip olduğu gözlemlenebilir, bu oran %97'dir. Ek olarak, modelin doğru pozitif örnekleri doğru bir şekilde tanıma yeteneğini gösteren bir hassasiyet değeri olan %98,05 ve doğru negatif örnekleri doğru bir şekilde tanıma kapasitesini gösteren bir özgünlük değeri olan %96,89 ve f1-skor %96,99 elde edilmiştir. Şekil 4'te sunulan karışıklık matrisi, Rastgele Orman modelinin 292 örneği doğru bir şekilde tanımladığını göstermektedir. Özellikle, 141 örneği doğru pozitif olarak sınıflandırdı ve 151 örneği doğru negatif olarak sınıflandırdı. Ancak, model üç yanlış negatif örneği ve beş yanlış pozitif örneği sergilemiştir. Rastgele Orman yöntemi, varsayılan ayarlar kullanılarak uygulanmıştır.

Gradient Boosting modeli, toplam 300 veri içerisinde 138 tıklama ve 152 tıklama olmama durumu doğru bir şekilde tanımlayarak kullanıcı davranışını doğru bir şekilde tahmin etti, bu durum Şekil 5'te gösterilmiştir. Sınıflandırmanın tıklamaları tıklama olmama durumu olarak yanlış tanıdığı iki durum ve tıklama olmama durumlarını yanlışlıkla tıklama olarak kategorize ettiği sekiz durum bulunmaktadır. GB modelinin özgünlüğü %94.52 ve hassasiyeti %98,70'dir, bu da tıklama olmama durumlarını ve tıklama durumlarını doğru bir şekilde tanımlama yeteneğini göstermektedir.

Lojistik Regresyon modeli, toplam 300 örnek içinden 290'ı doğru bir şekilde tahmin ederek iyi bir performans sergilemiştir. Modelin negatif bir sonuç tahmin ettiği 6 durum ve pozitif bir sonuç tahmin ettiği 4 durum, Şekil 6'te görülen karışıklık matrisinde belirtilmiştir. Model, kabul edilebilir bir hassasiyet %96,10 ve özgünlük %97,26 seviyesi gösterdi, bu da tıklama ve tıklama olmama durumları arasında etkili bir şekilde ayırt etme yeteneğini göstermektedir. Lojistik Regresyon model modelinin %96,66 doğruluk oranına sahip olduğu

görülmektedir. Genel olarak, LR modeli durumları ilgili sınıflara etkili bir şekilde sınıflandırmıştır.

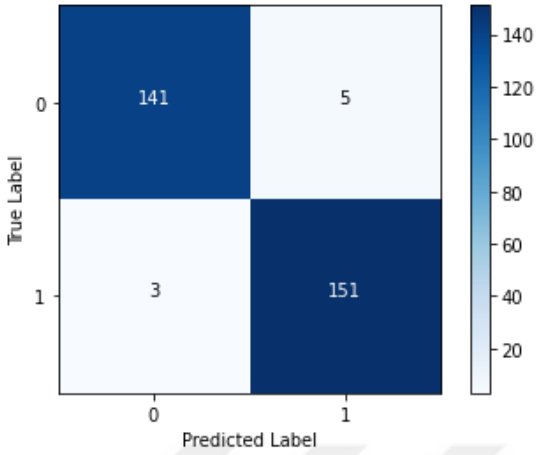
XGBoost modeli, tüm veri seti içinde 288'i doğru bir şekilde kategorize etti. Model, 139 gerçek pozitif (TP) tahmini ve 149 gerçek negatif (TN) tahmini yaptı. Şekil 7, bunu bir karışıklık matrisi şeklinde göstermektedir. Tablo 2, modelin tıklama ve tıklama olmama durumlarını başarıyla ayırt ettiğini, doğruluk %96,66 ve hassasiyet %96,40 elde edildiği görülmüştür.

Karar ağacı modeli, toplam örneklerin 284'ünü doğru bir şekilde tanımlamış, 135 TP ve 149 TN tahmini yapmış, ancak aynı zamanda Şekil 8'deki karışıklık matrisine göre 11 FP ve 5 FN tahmin edilmiştir. Modelin duyarlılığı %96,10 idi, bu da tıklama durumlarının %96'sını tanıma yeteneğini gösterirken, özgünlüğü %92,46 idi, bu da tıklama olmama durumlarının %92'sini tanıma yeteneğini gösterir. Bu sonuçlar, model için varsayılan parametreleri kullanarak elde edilmiştir.

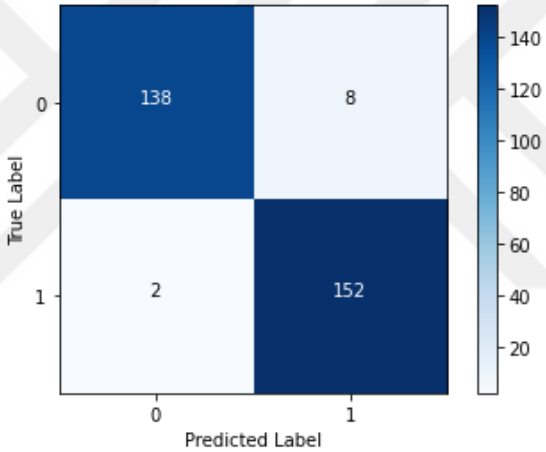
K-En Yakın Komşular modelinin sonuçlarına dayanarak, karışıklık matrisi Şekil 9'de görüldüğü üzere toplam örneklerin 222'sini doğru bir şekilde sınıflandırdı. Gerçek negatif durumlar için özgünlük %77,65 idi ve 146'dan 113'ü doğru bir şekilde tanımlandı. Gerçek pozitif durumlar için duyarlılık %70,53 idi ve 154'ten 109'u doğru bir şekilde tanımlandı. K-En Yakın Komşular sınıflandırıcısı %74,99 doğruluk ve f1-skoru %70,99 elde etti. Bu sonuçlar, k değerini 3'e ayarlayarak elde edilmiştir.

Tablo 2. Veri Artırımı olmaksızın elde edilen sonuçlar (%)

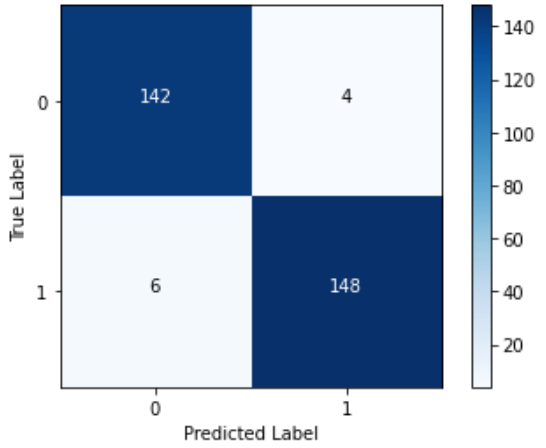
Model	Acc.	Spec.	Sens.	F1-Score
Random Forest	97,00	96,89	98,05	96,99
Gradient Boosting	96,66	94,52	98,70	96,65
Logistic Regression	96,66	97,26	96,10	96,66
XGBoost	96,66	95,89	96,40	96,66
Decision Tree	94,33	92,46	96,10	94,32
K-Nearest Neighbors	74,99	77,65	70,53	70,99



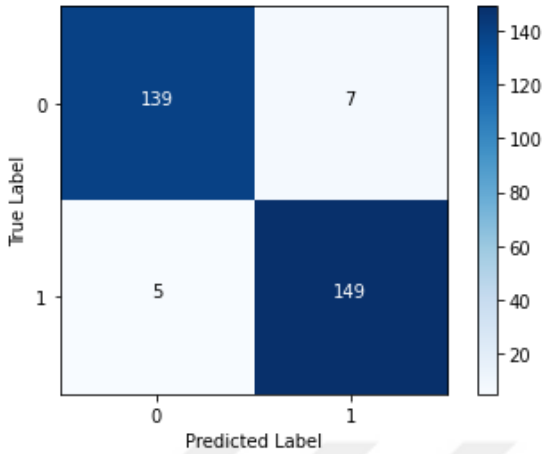
Şekil 4. Rastgele Orman Karışıklık Matrisi



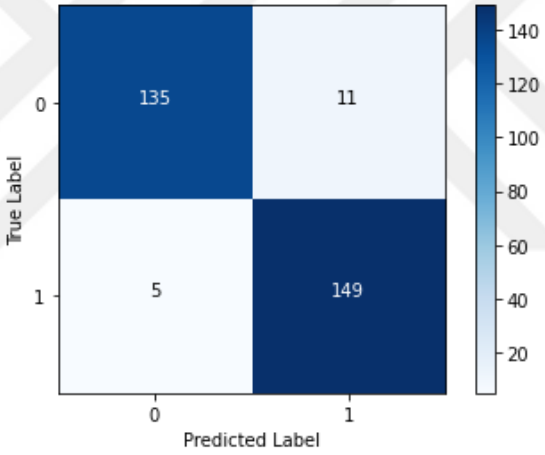
Şekil 5. Gradient Boosting Karışıklık Matrisi



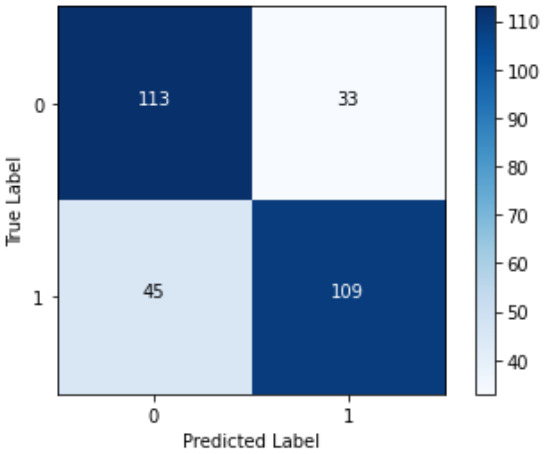
Şekil 6. Lojistik Regresyon Karışıklık Matrisi



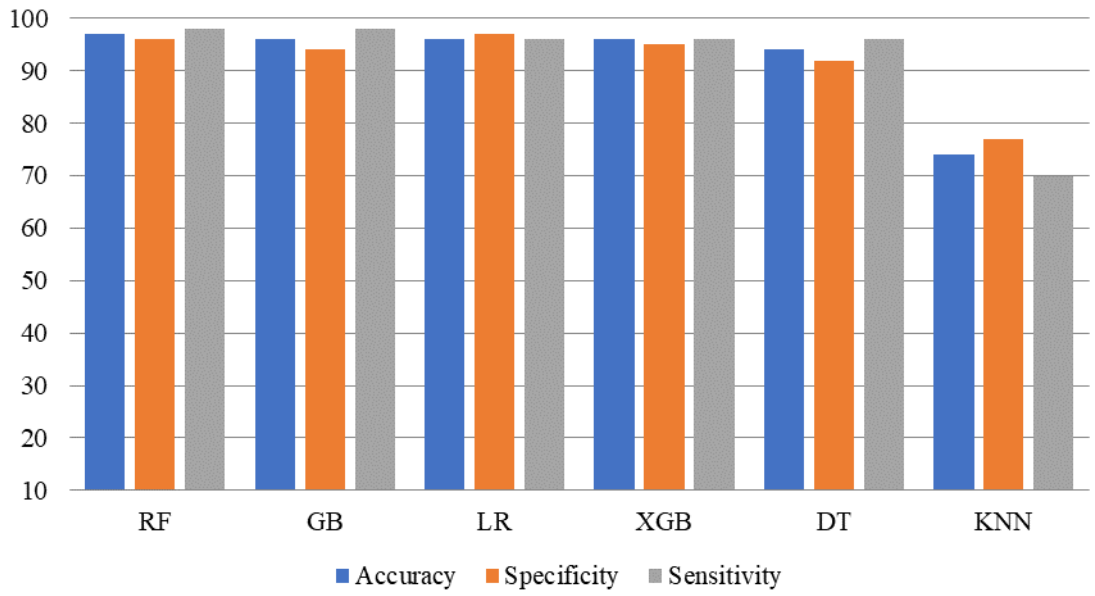
Şekil 7. XGBoost Karışıklık Matrisi



Şekil 8. Karar Ağacı Karışıklık Matrisi



Şekil 9. K-En Yakın Komşular Karışıklık Matrisi



Şekil 10. Doğruluk, Özgüllük. ve Veri Artırımı olmayan Modeller için Hassasiyet

4.1.3 Veri Artırma ile elde edilen sonuçlar

Mevcut veri kümesine GAN uyguladıktan sonra; Toplam 1500 örnekten, Rastgele Orman modeli, Şekil 11'deki karışıklık matrisinde gösterildiği gibi, 759 non-click durumunu ve 727 click olayını doğru bir şekilde tanımlanmıştır. Rastgele Orman modeli için doğruluk, özgünlük %99,13 ve duyarlılık %98,64 elde edilmiştir.

Gradient Boosting modelinin yüksek bir doğruluk skoru olan 98,93'a sahip olması, test verilerinin %98,93'unu etkili bir şekilde tanımladığını göstermektedir. Şekil 12'de gösterilen karışıklık matrisine dayanarak, modelin 759 tıklama olmayını ve 726 tıklama olayını yüksek doğrulukta tahmin ettiği görülebilir. Ayrıca, Tablo 3'te görüldüğü gibi, modelin duyarlılığı %98,50 ve özgünlüğü %98,84'tur, bu da pozitif click durumlarının %98'ini ve negatif non-click durumlarının %98'unu güvenilir bir şekilde tanımlama yeteneğini göstermektedir. Model, her iki sınıf için %98,93'luk yüksek doğruluk ile örnekleri doğru bir şekilde sınıflandırmada güçlü bir performans sergilemiştir. ÇÜA kullanılarak genişletilmiş veri kümesi üzerinde eğitilen GB modelinin doğruluğu ve güvenilirliği, click ve non-click durumlarını doğru bir şekilde tespit etme yeteneğinde açıkça görülmektedir.

Lojistik Regresyon modeli, Şekil 13'teki karışıklık matrisinde gösterildiği gibi, toplam örneklerden 710 tıklama ve 748 tıklama olmayanı doğru bir şekilde tanımlayarak %97,26

doğruluk elde etmiştir. Ayrıca, Tablo 3'te görüldüğü gibi, modelin özgüllüğü %,98,16 ve duyarlılığı %96,33 olup, olumlu tıklama örneklerinin %96'sini ve olumsuz tıklama dışı örneklerin %98'sini güvenilir bir şekilde tanımlama yeteneğini göstermektedir.

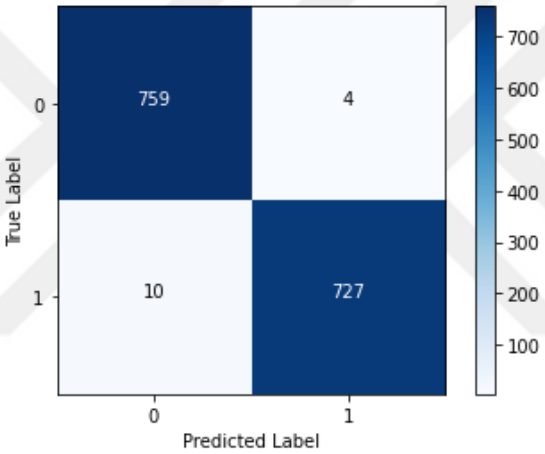
Tablo 3'te gösterildiği gibi. XGBoost modeli, ÇÜA kullanılarak zenginleştirilmiş veri kümesi üzerinde test edildikten sonra %98,93 doğruluk elde etti. Bu, modelin örneklerin %98'ini doğru bir şekilde tespit ettiğini gösterir. Şekil 14'te görülen karışıklık matrisi, modelin sırasıyla 728 click olayını ve 756 non-click olayını başarıyla tanıdığını göstermektedir. Modelin duyarlılık ve özgünlüğü sırasıyla %98,64 ve 99,21 idi, bu da modelin click durumlarının %98'ini ve non-click durumlarının %99'unu doğru bir şekilde tespit ettiğini göstermektedir.

Şekil 15'te gösterilen karışıklık matrisine göre, Karar Ağacı algoritması olayları doğru bir şekilde sınıflandırmada yüksek bir doğruluk seviyesi sergiledi. Özellikle, kullanıcının tıklamadığı 758 örneği ve tıkladığı 722 örneği başarıyla tanıdı. 758 doğru pozitif tahmin ve 722 doğru negatif tahmin yapma yeteneği, bu doğruluğa ulaşmasına yardımcı oldu. Model, %98,66 doğruluk oranına ulaştı ve örneklerin %98'ini doğru bir şekilde tespit ettiğini gösterdi. Bulgular, modelin tıklama durumlarının %97,10'unu ve tıklama olmayan durumlarının %99,21'inin doğru bir şekilde tanımladığını. Bu sonuçlar, model için varsayılan parametreleri kullanarak elde edilmiştir.

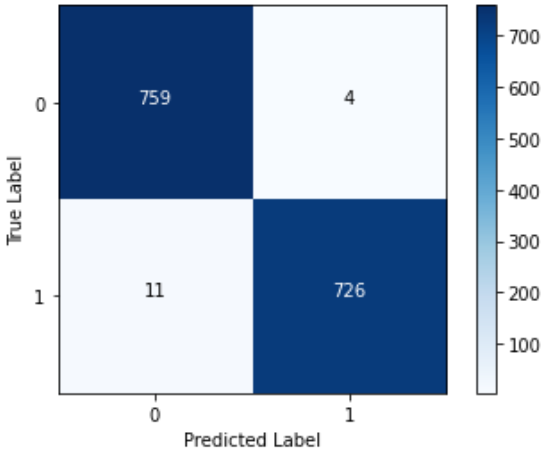
K-En Yakın Komşular modelinin doğruluğu %94,73 idi, bu da tüm örneklerin %94'ünü doğru bir şekilde tanıdığını gösterir. Şekil 16'da gösterildiği gibi, 694 tıklama durumunu ve 730 tıklama olmayan durumu tanımada etkili oldu, bunların sırasıyla 730 doğru pozitif ve 694 doğru negatif tahmin yaptı. Modelin hassasiyeti tıklama ile ilgili senaryolar için %94,21 özgünlüğü tıklama olmayan senaryolar için %95,23 ve f1-skoru 93,73 idi. Bu sonuçlar, k parametresinin 3 olarak ayarlanmasıyla elde edildi, bu Tablo 3'te gösterilmiştir.

Tablo 3. Veri Artırma ile elde edilen sonuçlar (%)

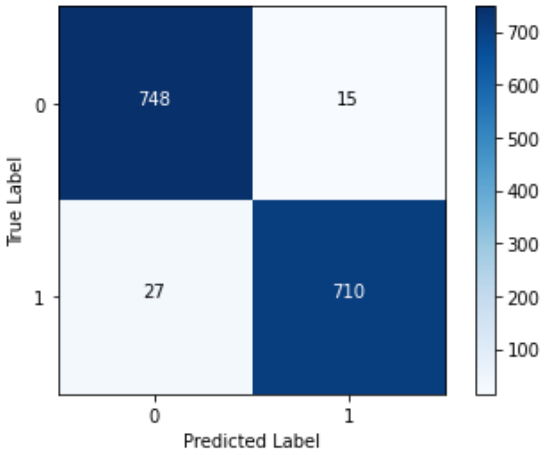
Model	Acc.	Spec.	Sens.	F1-Score
Random Forest	99,13	99,60	98,64	99,13
Gradient Boosting	98,93	98,84	98,50	98,93
Logistic Regression	97,26	98,16	96,33	97,26
XGBoost	98,93	99,21	98,64	98,93
Decision Tree	98,66	99,21	97,10	98,66
K-Nearest Neighbors	94,73	95,23	94,21	93,73



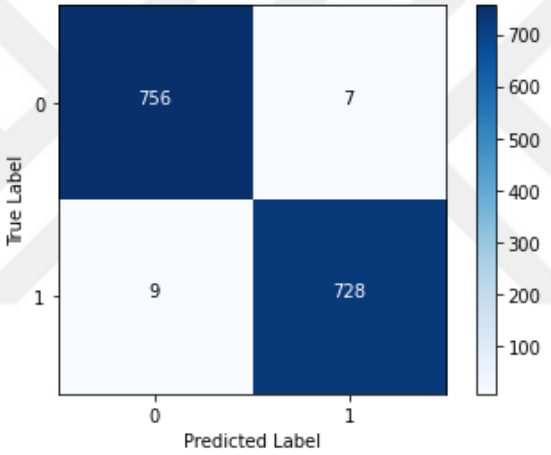
Şekil 11. Rastgele Orman Karışıklık Matrisi



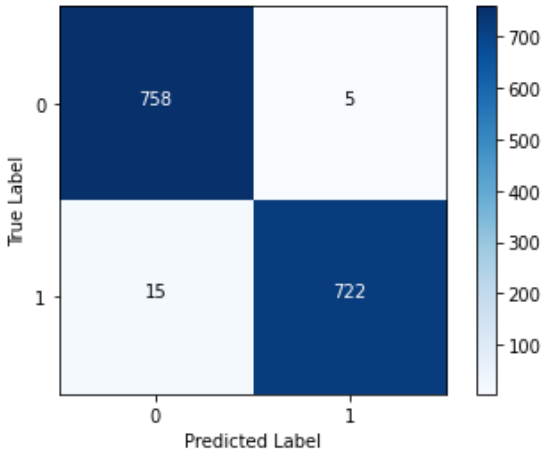
Şekil 12. Gradient Boosting Karışıklık Matrisi



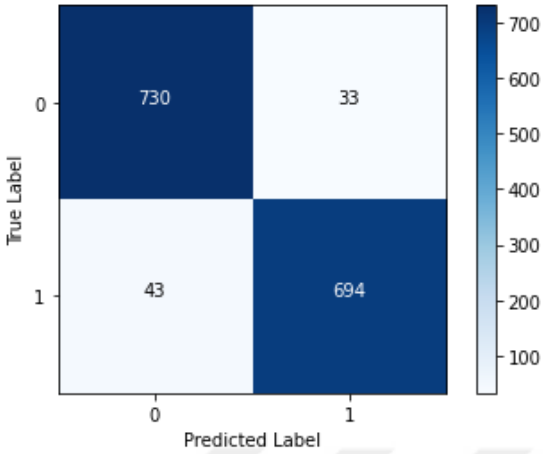
Şekil 13. Lojistik Regresyon Karışıklık Matrisi



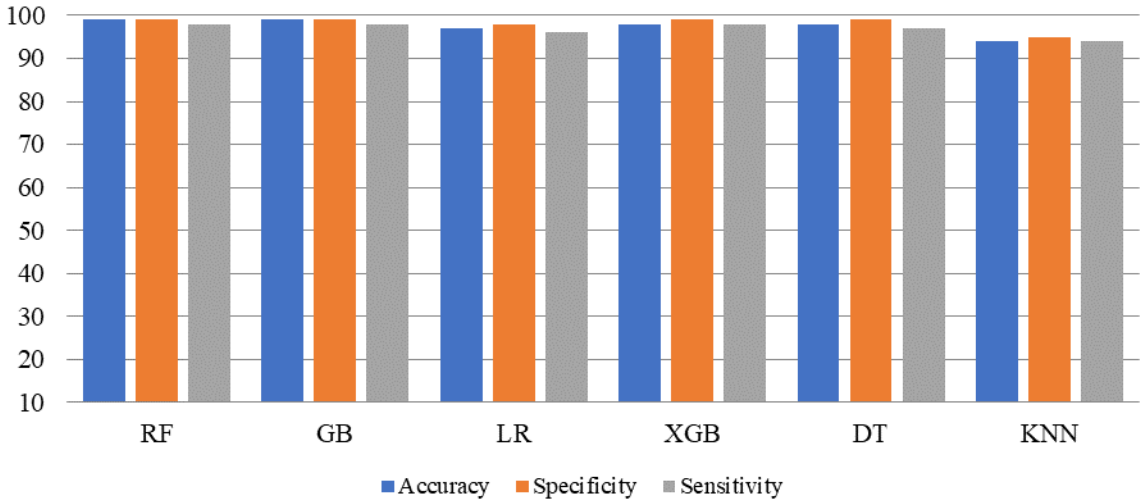
Şekil 14. XGBoost Karışıklık Matrisi



Şekil 15. Karar Ağacı Karışıklık Matrisi



Şekil 16. K-En Yakın Komşular Karışıklık Matrisi



Şekil 17. Doğruluk, Özgüllük. ve Veri Büyütmeli Modeller için Hassasiyet

Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Male	Clicked on Ad
81.98	40	65461.92	229.22	0	0
66.01	23	34127.21	151.95	0	1
61.57	53	35253.98	125.94	1	1
53.3	34	44893.71	111.94	0	1
34.87	40	59621.02	200.23	0	1
43.6	38	20856.54	170.49	0	1
77.88	37	55353.41	254.57	0	0
75.83	27	67516.07	200.59	0	0
49.95	39	68737.75	136.59	0	1
60.94	41	76893.84	154.97	0	1
89.15	42	59886.58	171.07	0	0
78.7	30	53441.69	133.99	0	1
57.35	29	41356.31	119.84	0	1
34.86	38	49942.66	154.75	0	1
70.68	31	74430.08	199.08	0	0
76.06	23	58633.63	201.04	0	0
66.67	33	72707.87	228.03	1	0
46.77	32	31092.93	136.4	1	1
62.42	38	74445.18	143.94	0	1
78.32	28	49309.14	239.52	0	1
37.32	50	56735.14	199.25	1	1
40.42	45	40183.75	133.9	1	1
76.77	36	58348.41	123.51	0	1
65.65	30	72209.99	158.05	0	0

Şekil 18. Artırılmış verinin görüntüsü

4.1.4 Deneysel sonuçların irdelenmesi

ÇÜA'ın veri artırma olarak uygulanması, tüm modellerin genel performansında önemli artışlara yol açmıştır. Bu da Şekiller 19–21'e gösterildiği gibi ve sırasıyla Tablolar 2 ve 3'te bildirilen doğruluk sonuçlarıyla görülmektedir. Tablo 2'de görüldüğü gibi, ÇÜA'ın uygulanmasından önce, K-En Yakın Komşular modelinin doğruluk puanı %74 ile en düşükken, Rastgele Orman modeli en yüksek doğruluk puanına sahip %97 idi. Öte yandan, Tablo 3, ÇÜA'ın veri artırma için kullanılmasının tüm modellere faydalı olduğunu göstermektedir. Rastgele Orman modeli %99 doğruluk puanı elde etti, bu da öğrendikleri daha karmaşık ve geniş desenlerin, performanslarını artırdığını göstermektedir. XGBoost , Karar Ağacı ve Gradient Boosting modelleri doğrulukları için %98 ‘dir.

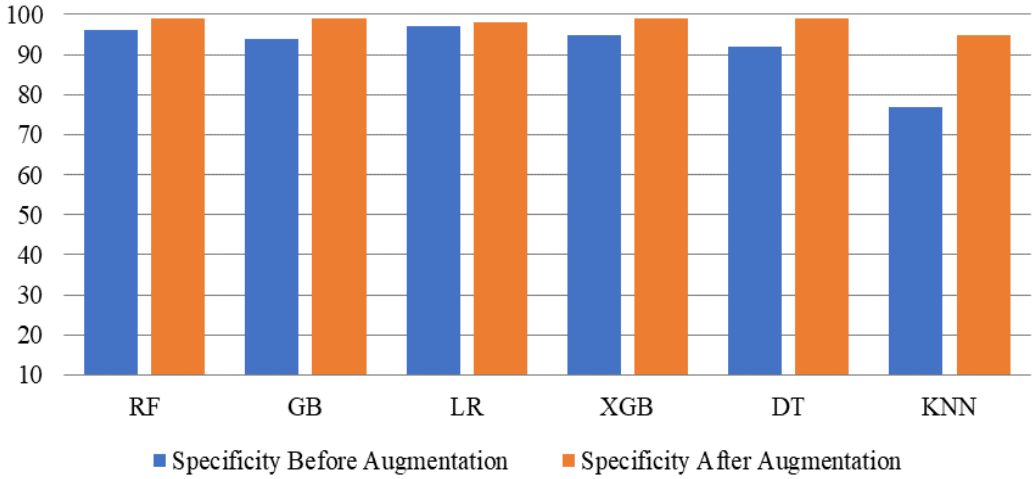
K-En Yakın Komşular modeli, doğruluk açısından en önemli artışı gösterdi. Veri artırma amacıyla ÇÜA'ların kullanılması, çoğu model için artan duyarlılık ve özgüllük üretmiştir. Bu sonuç, üretilen verilerin modellerin pozitif ve negatif durumları doğru bir şekilde ayırt etme kapasitesini güçlendirdiğini göstermektedir. ÇÜA tabanlı veri artırmanın eklenmesi,

genel olarak model performansını önemli ölçüde artırmıştır, ancak modelden modele göre değişen derecelerde. Genel olarak, bu eğilim tutarlıdır.

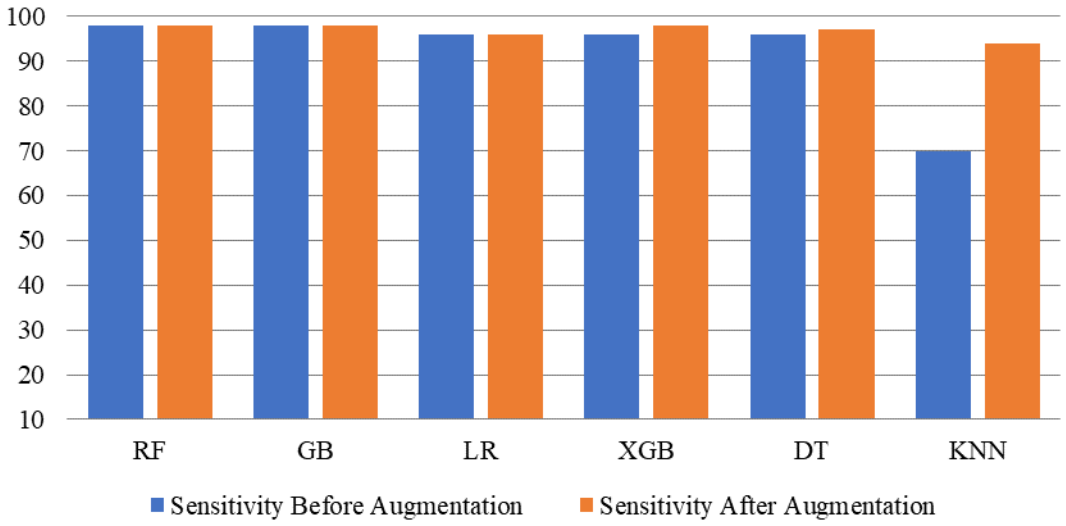
Bulgular, ÇÜA kullanılarak yeni sentetik veri oluşturmanın, modellerin gerçek ve yanlış olayları ayırt etme kapasitesini önemli ölçüde artırdığını göstermektedir. Bulgular bunu kanıtlamıştır. Örneğin, Rastgele Orman ve Gradient Boosting modellerinin doğruluk oranı sırasıyla daha önce %97 ve %96 seviyelerindeyken, her ikisi de %99 ve %98'e yükselmiştir. Diğer modellerle karşılaştırıldığında, K-En Yakın Komşular modeli en önemli iyileştirmeye sahiptir ve doğruluk kazancı diğer modellerden %20 daha yüksektir. Daha büyük veri setinin etkisiyle, modellerin çoğunluğu duyarlılık ve özgüllüklerinde artış görmüştür.



Şekil 19. Veri Artırımı Olmadan ve Veri Artırımı ile Doğruluk Karşılaştırması



Şekil 20. Veri Artırımı Olmadan ve Veri Artırımı ile Özgüllüğün Karşılaştırılması



Şekil 21. Veri Artırımı Olmadan ve Veri Artırımı ile Hassasiyet Karşılaştırması

4.2 Dengesiz Veri Kümesi için Veri Artırma

4.2.1 Z-Alizadeh sani veri kümesi

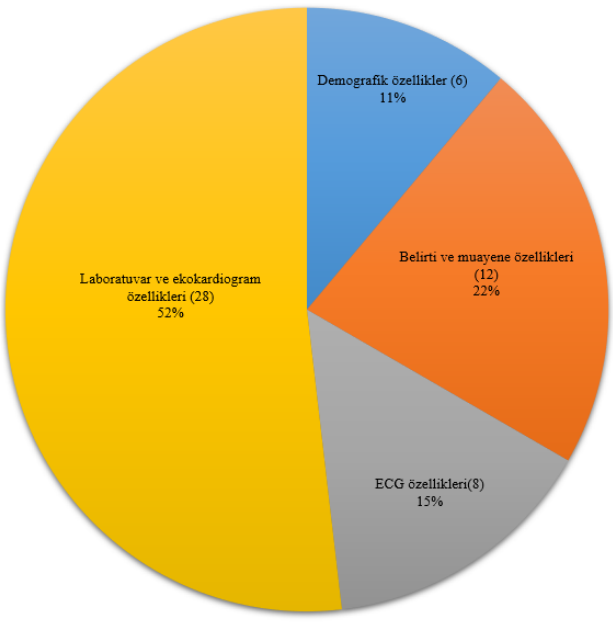
Bu çalışmada UCI veri deposunda bulunan Z-Alizadeh Sani veri kümesi kullanılmıştır. Kapsamlı veri setinde 303 örnek bulunmaktadır ve her biri 54 farklı özelliğe sahiptir. Veri seti, 54 farklı özelliği içerir. Veri seti, sağlık sorunu olmayan bireyleri temsil eden değer 0'ı, kalp hastalığı teşhisi konmuş olanları temsil eden değer 1'i içeren ikili etiketlerden oluşur. Veri seti, yaş, ağırlık ve cinsiyet gibi demografik bilgileri içerir. Yaş aralığı 30 ila 86 yıl arasındadır. Ağırlık değerleri, 48 ila 120 arasında değişen vücut ağırlıklarını kapsar. Veri seti, cinsiyet bilgilerini içerir; toplamın %58'ini oluşturan 176 erkek ve %42'sini oluşturan 127 kadın bulunmaktadır. Mevcut tıp literatürüne dayanarak, bu özellikleri bireylerde kardiyovasküler hastalıkların olası göstergeleri olarak düşünmek makuldür. Şekil 22'de gösterildiği gibi, Özellikler dört ayrı kategoriye ayrılır:

- Demografik özellikler (6 özellik): Bunlar yaş, ağırlık, boy, cinsiyet, BMI (vücut kitle indeksi) ve DM (diyabet) içerir ve bir hastanın diyabet geçmişini ifade eder.
- Belirti ve muayene özellikleri (12 özellik): Bunlar açlık kan şekeri (FBS), CRP (C-reaktif protein), HTN (hipertansiyon geçmişi olan bir hastanın tıbbi geçmişi) içerir. "Mevcut sigara içen" terimi, sigara içme eylemiyle şu anda ilgili olanları ifade eder; "eski sigara içen" terimi, daha önce sigara içme geçmişi olan ancak bu alışkanlığı

sona erdirmiş olan kişileri tanımlar; "FH" aile geçmişi anlamına gelir ve kan bağı olan kişilerde kardiyovasküler hastalığın varlığını ifade etmek için kullanılır; obezite, BP (kan basıncı), PR (kalp atış hızı), ödem, tipik göğüs ağrısı ve nefes darlığı.

- ECG özellikleri (8 özellik): Bunlar EKG (elektrokardiyogram) sonucu, Q dalga, ST yükselmesi, ST depresyonu, T dalga inversiyonu, LVH (sol ventriküler hipertrofi), LBBB (sol demet dal bloğu) ve RBBB (sağ demet dal bloğu) içerir.
- Laboratuvar ve ekokardiogram özellikleri (28 özellik): Bunlar HDL-C (yüksek yoğunluklu lipoprotein kolesterol), LDL-C (düşük yoğunluklu lipoprotein kolesterol), BUN (kan üre azotu), ESR (eritrosit sedimantasyon hızı), HB (hemoglobin), K (potasyum), Na (sodyum), WBC (beyaz kan hücresi), Lenf, Nöt, PLT (trombosit), EF-TTE (transtorasik ekokardiyografi ile ejeksiyon fraksiyonu), Bölge RWMA (bölgesel duvar hareketi anormalliği), VHD (kapak hastalığı), LAD (sol ön inen arter), LCX (sol sirkumfleks arter), RCA (sağ koroner arter), Cath (anjiyografi sonucu) ve üç ana arterin daralma derecesi ile ilgili diğer 10 özellik içerir.

Her hasta, koroner arter hastalığı (CAD) veya Normal olmak üzere iki gruptan birine kategorize edilebilir. Koroner arter hastalığı (CAD), hastaların koroner arterlerindeki çap daralma derecesi %50'ye ulaştığında veya aştığında teşhis edilir. Alternatif olarak, belirli bir hastalık için kriterleri karşılamayan bireyler, tipik veya standart bir duruma sahip olarak kategorize edilir[60].



Şekil 22. Z-Alizadeh sani veri kümesinin özellik kategorisi

4.2.2 Veri artırımı ve özellik seçimi olmaksızın elde edilen sonuçlar

Bu bölüm, koroner arter hastalığı oluşumunu öngörmek amacıyla yedi farklı makine öğrenimi modelinin benimsenmesinden elde edilen bulguların değerlendirmesini sağlar. Bu çalışmanın temel amacı, Çekişmeli üretici ağlar kullanarak veri seti dengeleme ve artırma amaçlı etkinliğini değerlendirmektir. Ayrıca, çalışma, özellik seçimi tekniklerini içererek model performansında potansiyel bir iyileşmeyi araştırmayı amaçlamaktadır. Deney, Rastgele Orman, Gradient Boosting, Lojistik Regresyon, XGBoost, Karar Ağacı, K-En Yakın Komşular, Destek Vektör Makinesi ve Çekişmeli üretici ağlar gibi birçok denetimli makine öğrenimi yönteminin kullanımını ve 5 katlı çapraz doğrulama yaklaşımıyla içermektedir. Bulgular bir sonraki bölümde sunulmaktadır.

Tablo 4'te sunulan bulgulara göre, Rastgele Orman modelinin doğruluğu %87,80'dir. Rastgele Orman modeli, verilen veri kümesi için sınıf etiketlerini doğru bir şekilde tahmin etme konusunda iyi bir doğruluk düzeyi göstermektedir. Şekil 23'teki karışıklık matrisine göre, model, 1 sınıf etiketine sahip 87 örnekten 56'sını ve 0 sınıf etiketine sahip 216 örnekten 208'ini doğru bir şekilde tahmin etmiştir.

Gradient Boosting modelinin doğruluğu %86,79 olarak gözlendi, bu Rastgele Orman modelinin doğruluğuna kıyasla biraz daha düşüktü. Bununla birlikte, özgürlük ve duyarlılık puanları sırasıyla %92,44 ve %72,36 değerlerini göstermektedir. Bu, Gradient Boosting modelinin veri setinin sınıf etiketlerini etkili bir şekilde tahmin ettiği anlamına gelir. Şekil 24'ta görüldüğü gibi, karışıklık matrisi, Gradient Boosting modelinin 0 sınıf etiketine sahip 216 örnekten 200'ünü ve 1 sınıf etiketine sahip 87 örnekten 63'ünü doğru bir şekilde tahmin ettiğini göstermektedir.

Lojistik Regresyon modelinin doğruluğu %77,88'dir, bu Rastgele Orman ve Gradient Boosting modellerine kıyasla karşılaştırmalı olarak daha düşüktür. Lojistik regresyon modeli, veri setinin sınıf etiketlerini doğru bir şekilde tahmin etme konusunda yüksek bir doğruluk düzeyi sergiler. Şekil 25'teki karışıklık matrisine göre lojistik regresyon modeli, 0 sınıf etiketine sahip 216 örnekten 193'ünü ve 1 sınıf etiketine sahip 87 örnekten 43'ünü doğru bir şekilde tanımladı.

XGBoost modeli, Rastgele Orman ve Gradient Boosting modellerine kıyasla daha yüksek bir doğruluk seviyesi sergiler, yüzde 85,13'e ulaşır. özgürlük ve duyarlılık puanları sırasıyla yüzde 91,59 ve yüzde 72,11'dir. XGB modelinin veriye sınıf etiketleri atama konusunda kabul edilebilir bir tahmin performansı sergilediğini göstermektedir. XGBoost modelinin karışıklık matrisi Şekil 26'dekine göre, 0 sınıf etiketine ait 216 örnekten 197'sini doğru bir şekilde kategorize ettiğini ve 1 sınıf etiketine ait 87 örnekten 63'ünü doğru bir şekilde sınıflandırdığını göstermektedir.

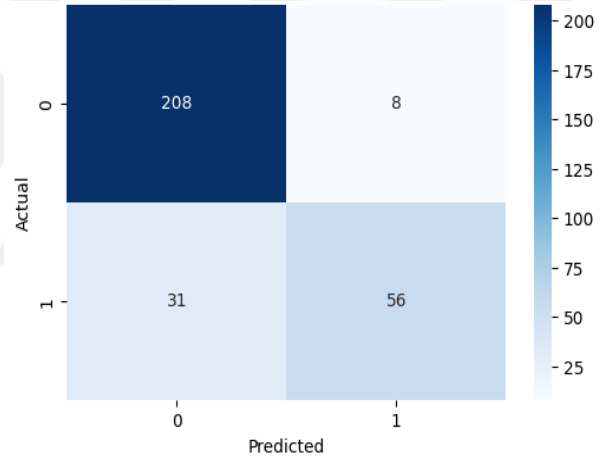
Karar Ağacı modeli, Rastgele Orman, Gradient Boosting ve XGB modellerine kıyasla yüzde 77,55'lik bir doğruluk oranına sahiptir. Özğürlük ve hassasiyet puanları sırasıyla yüzde 84,18'lik ve yüzde 60,62'dir. Karar ağacı modeli, verilen verinin sınıf etiketlerini doğru bir şekilde tahmin etme konusunda yüksek bir doğruluk seviyesine sahiptir. Şekil 27'da gösterilen karışıklık matrisine dayanarak, modelin 0 sınıf etiketine ait 216 örneğin 183'ünü doğru bir şekilde kategorize ettiği ve 1 sınıf etiketine ait 87 örneğin 53'ünü doğru bir şekilde sınıflandırdığı görülmektedir.

K-en yakın komşular modelinin doğruluk oranı %66 olarak görülmektedir ki bu değerlendirilen modellerin en düşüğüdür. K-en yakın komşular modelinin sınıf etiketlerini tahmin etme konusundaki doğruluğu, diğer modellere kıyasla daha düşüktür. Şekil 28'de gösterilen karışıklık matrisine göre, K-en yakın komşular modeli, 0 sınıf etiketine ait 216 örnekten 180'ini doğru bir şekilde tanımladı ve 1 sınıf etiketine ait 87 örnekten 20'sini doğru bir şekilde tanımladı.

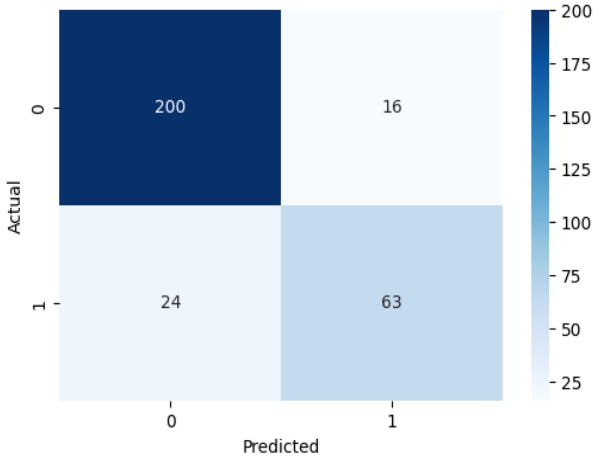
Destek Vektör Makinesi, bir sınıflandırma doğruluğu yüzde 71,28 sergileyerek dikkat çekici bir özgünlük seviyesi olan yüzde 100'ü gösterdi. Destek Vektör Makinesi modeli, veri için sınıf etiketlerini doğru bir şekilde tahmin etme konusunda suboptimal bir performans sergilemektedir. Şekil 29'de gösterilen Destek Vektör Makinesi için karışıklık matrisi, modelin 0 sınıf etiketine ait olan 216 örneği doğru bir şekilde tanımladığını, ancak 1 sınıf etiketine ait olan 87 örneği yanlışlıkla etiketlediğini göstermektedir. Bulgulara göre, Rastgele Orman, Gradient Boosting, XGBoost ve Karar Ağacı modellerinin veri setinin sınıf etiketlerini doğru bir şekilde sınıflandırdığı görüldü. K-en yakın komşular ve Destek Vektör Makinesi modellerinin etkinliği oldukça düşüktür.

Tablo 4. Veri Artırımı ve Özellik Seçimi olmaksızın elde edilen Sonuçlar

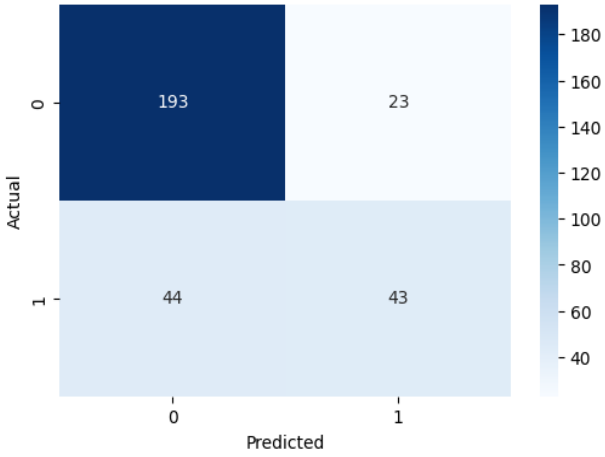
Model	Acc.	Spec.	Sens.	F1-Score
Random Forest	87,80	96,44	64,36	81,36
Gradient Boosting	86,79	92,44	72,36	83,40
Logistic Regression	77,88	89,35	49,42	70,70
XGBoost	85,13	91,59	72,11	82,44
Decision Tree	77,55	84,18	60,62	72,20
K-Nearest Neighbors	66,00	83,33	22,98	52,86
Support Vector Machine	71,28	100	0	41,61



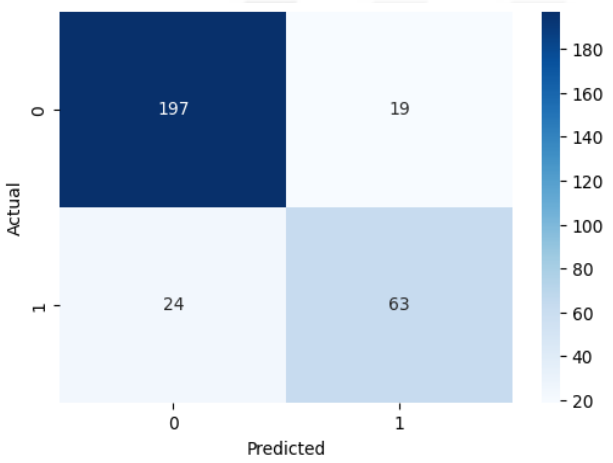
Şekil 23. Rastgele Orman Karışıklık Matrisi



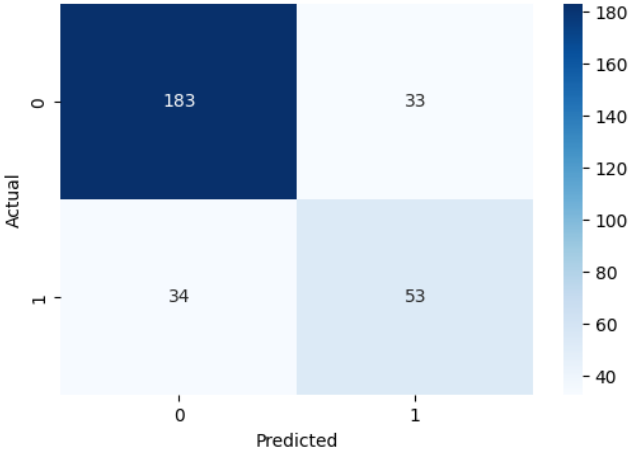
Şekil 24. Gradient Boosting Karışıklık Matrisi



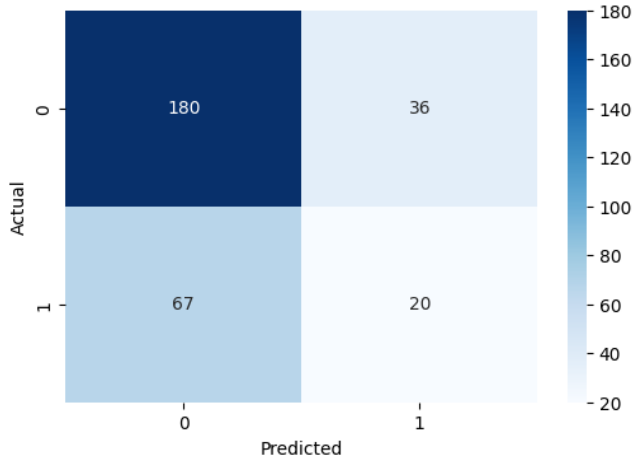
Şekil 25. Lojistik Regresyon Karışıklık Matrisi



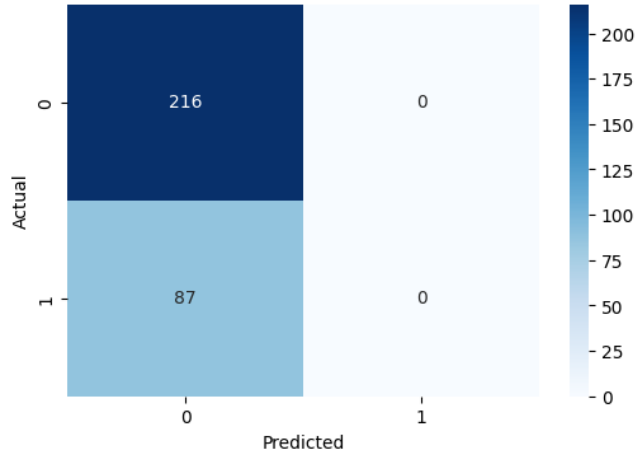
Şekil 26. XGBoost Karışıklık Matrisi



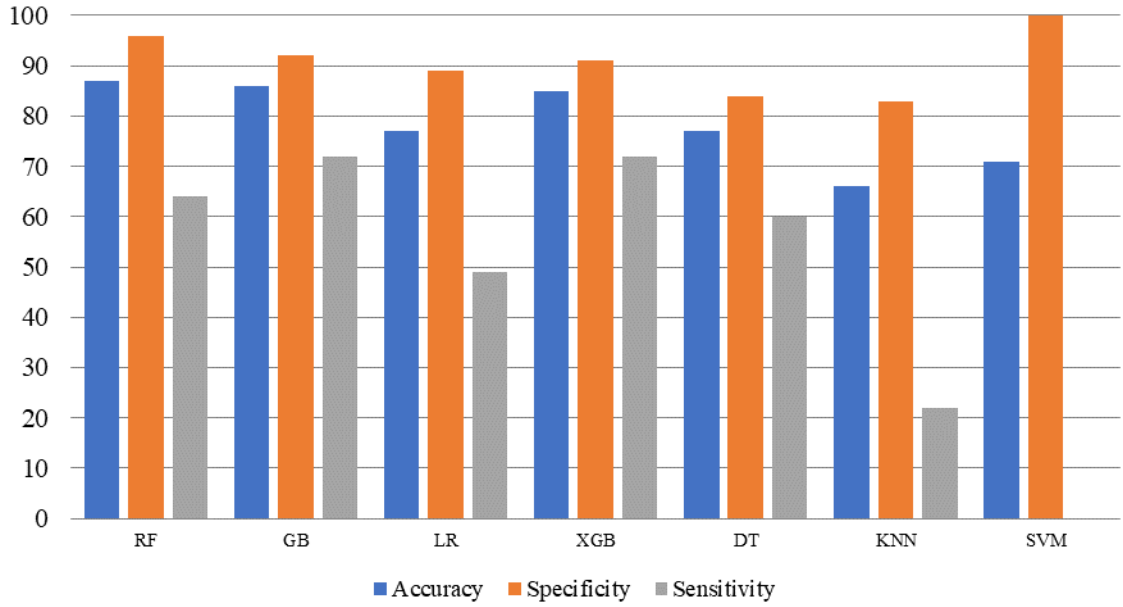
Şekil 27. Karar Ağacı Karışıklık Matrisi



Şekil 28. K-En Yakın Komşular Karışıklık Matrisi



Şekil 29. Destek Vektör Makinesi Karışıklık Matrisi



Şekil 30. Veri Artırımı ve Özellik Seçimi Olmayan Modeller için Doğruluk, Özgüllük ve Duyarlılık

4.2.3 Veri Artırımı ve özellik seçimi ile elde edilen sonuçlar

Veri artırma uygulandıktan ve Öznitelik Seçimi için Destek Vektör Makineleri kullanıldıktan sonra, algoritma tahmini için 54 özelliğin 21'inin önemli olduğu belirlendi. Tablo 5'te gösterilen bulgulara göre, Rastgele Orman modelinin yüzde 97,35'e varan en iyi doğruluk seviyesini sergilediği görülmektedir. Özgürlük, duyarlılık ve f1 skor ölçütleri %97,30 , %97,40 ve %97,34 değerlerini göstermektedir. Bu, Rastgele Orman modelinin veri setinin sınıf etiketlerini etkili bir şekilde tahmin ettiği anlamına gelir. Şekil 31'de gösterilen karışıklık matrisi, Rastgele Orman modelinin performansını göstermektedir. Bu, 0 sınıf etiketine sahip toplam 1000 örnekten Rastgele Orman modelinin 970 örneği doğru bir şekilde tahmin ettiğini göstermektedir. Benzer şekilde, 1 sınıf etiketine sahip örneklerin toplamından Rastgele Orman modeli doğru bir şekilde 978 örneği tanımlamıştır.

Gradient Boosting modelinin hassasiyeti yüzde 97,35'tir. Özgürlük yüzde 96,92 ve duyarlılık yüzde 97,81 değerlerini göstermektedir. Yukarıdaki bulgulara dayanarak, Gradient Boosting modelinin veri sınıf etiketlerini tahmin etmede yüksek bir doğruluk seviyesine sahip olduğu gözlemlenebilir. Şekil 32'de sunulan karışıklık matrisine dayanarak, Gradient Boosting modeli toplam 1000 örnekten 0 sınıf etiketi olarak doğru bir şekilde tanımlanan 969 örneği başarıyla tahmin etti. Ayrıca, 1000 örnekten 1 sınıf etiketi olarak doğru bir şekilde tahmin ettiği 978 durumu göstermektedir.

Ayrıca, Rastgele Orman ve Gradient Boosting modelleri ile karşılaştırıldığında, lojistik regresyon modelinin daha yüksek bir doğruluk oranına sahip olduğu görülmektedir, yüzde 97,61 Hassasiyet ve özgürlük 97,21 puanları, veri setinin sınıf etiketlerini tahmin etmede yüksek bir doğruluk seviyesini göstermektedir, yüzde 97,41. Bu, lojistik regresyon modelinin veri setinin sınıf etiketlerini tahmin etmede yüksek bir doğruluk seviyesine sahip olduğunu göstermektedir. Şekil 33'teki karışıklık matrisine göre, lojistik regresyon modeli 1000 örneğin 972'sini doğru bir şekilde 0 sınıf etiketi olarak sınıflandırmış ve 24 örneği yanlış bir şekilde 1 sınıf etiketi olarak tanımıştır.

Ayrıca, XGBoost'un bir sınıflandırma doğruluğu yüzde 97,34'tir. Hassasiyet %97,35 ve özgürlük metrikleri yüzde 97,25 değerine sahiptir. Bu, XGB modelinin veri setinin sınıf etiketlerini doğru bir şekilde tahmin ettiği anlamına gelir. XGBoost modeli, Şekil 34'de sunulan karışıklık matrisinde görüldüğü gibi, 1000 örnekten 975'ini 1 sınıf etiketi için ve

1000 örnekten 973'ünü 0 sınıf etiketi için doğru bir şekilde tanımlayarak tahminlerinde yüksek bir doğruluk seviyesi sergilemiştir.

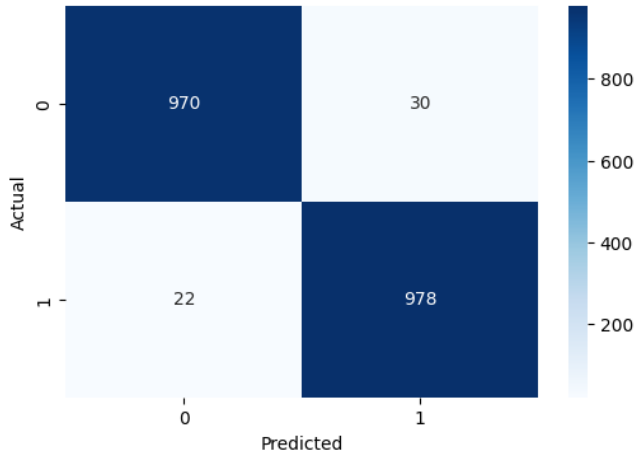
Rastgele Orman, Gradient Boosting, XGBoost ve lojistik regresyon modellerine karşı olarak, karar ağacı modelinin daha düşük bir doğruluk oranı yüzde 95,91'dir. Bulgular, karar ağacı modelinin veri setinin sınıf etiketlerini tahmin etme konusunda yüksek bir doğruluk seviyesine sahip olduğunu göstermektedir. Şekil 35'de sunulan karışıklık matrisine dayanarak, karar ağacı modeli, toplam 1000 durumun 0 sınıf etiketi olarak doğru bir şekilde sınıflandırma doğruluğu yüzde 95 sergilemiştir. Bununla birlikte, 37 durumu 1 sınıf etiketi olarak yanlış sınıflandırmıştır.

K-en yakın komşular modelinde özellik seçiminin kullanılması, yüzde 96,85'lik bir doğruluk oranına neden olmuştur. Ayrıca, sırasıyla yüzde 97,31 ve yüzde 96,41'lık özgürlük ve duyarlılık oranlarına sahiptir. Özellik seçimi yaklaşımı, sınıf etiketi 0'a sahip örnekleri sınıflandırmada K-en yakın komşular modelinin tahmin performansını önemli ölçüde artırmıştır. Ancak, modelin sınıf etiketi 1'e ait örnekleri doğru bir şekilde tahmin etme kapasitesinde önemli bir iyileşme olmamıştır. Şekil 36, K-en yakın komşular modelinin sınıf etiketi 0'a ait 973 örneği doğru bir şekilde sınıflandırdığını, ancak 27 örneği yanlış sınıflandırdığını gösteren bir karışıklık matrisini göstermektedir.

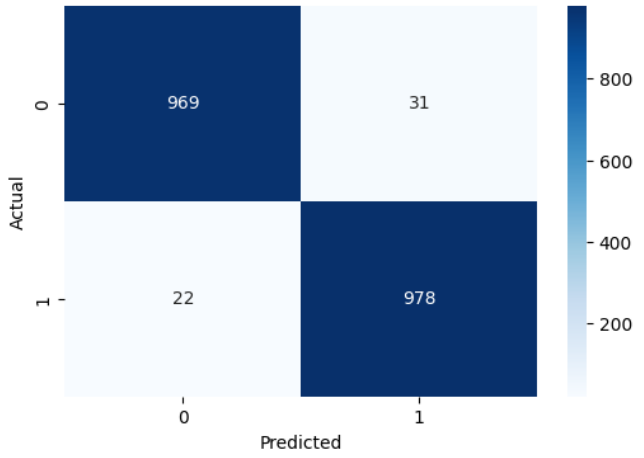
Özellik seçimi için Destek Vektör Makinesi modeli, DVM kullanımından sonra %95,41'lik bir doğruluk oranına ulaştı. Model, etiket 1'e ait örnekleri sınıflandırmada geliştirilmiş performans potansiyeli sergilemektedir. Şekil 37'deki karışıklık matrisine göre, sınıf etiketi 0'a ait toplam 1000 örneğin 995'ini doğru bir şekilde sınıflandırdı. Sınıf etiketi 1, %91,31'lik bir hassasiyet puanına sahiptir. Model f1 skoru %95,39 değerine sahiptir. Bu da gerçek pozitif durumları doğru bir şekilde tespit etme konusundaki yüksek doğruluğu göstermiştir.

Tablo 5. Veri Artırımı ve Özellik Seçimi ile elde edilen sonuçlar

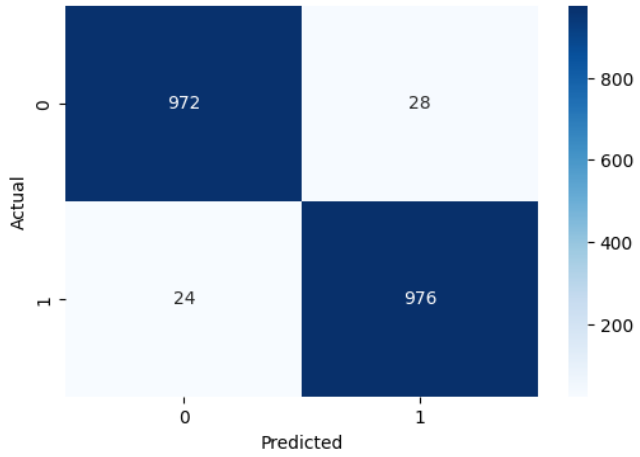
Model	Acc.	Spec.	Sens.	F1-Score
Random Forest	97,35	97,30	97,40	97,34
Gradient Boosting	97,35	96,92	97,81	97,34
Logistic Regression	97,41	97,21	97,61	97,39
XGBoost	97,34	97,25	97,35	97,20
Decision Tree	95,91	95,51	96,31	95,89
K-Nearest Neighbors	96,85	97,31	96,41	96,84
Support Vector Machine	95,41	99,51	91,31	95,39



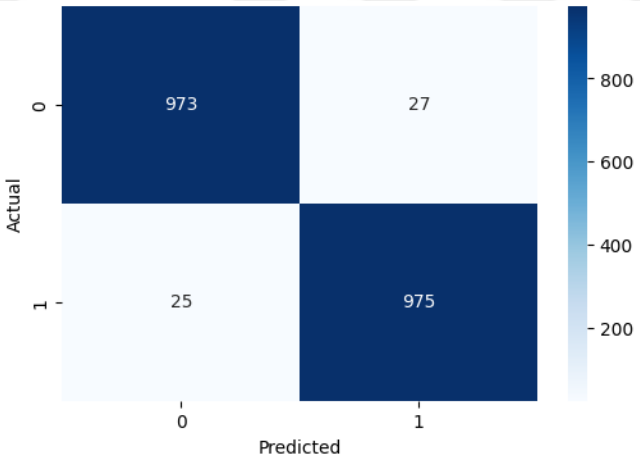
Şekil 31. Rastgele Orman Karışıklık Matrisi



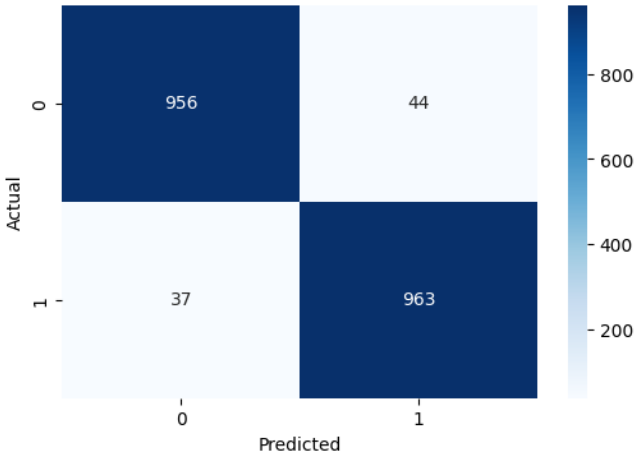
Şekil 32. Gradient Boosting Karışıklık Matrisi



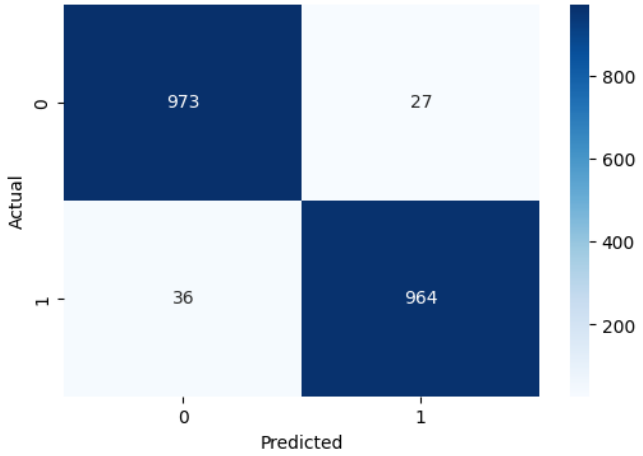
Şekil 33. Lojistik Regresyon Karışıklık Matrisi



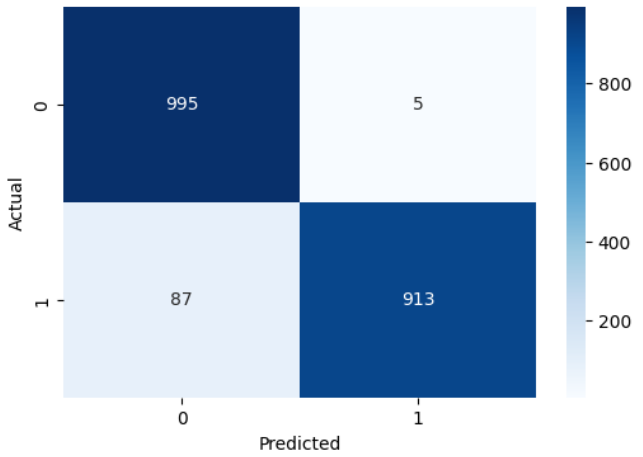
Şekil 34. XGBoost Karışıklık Matrisi



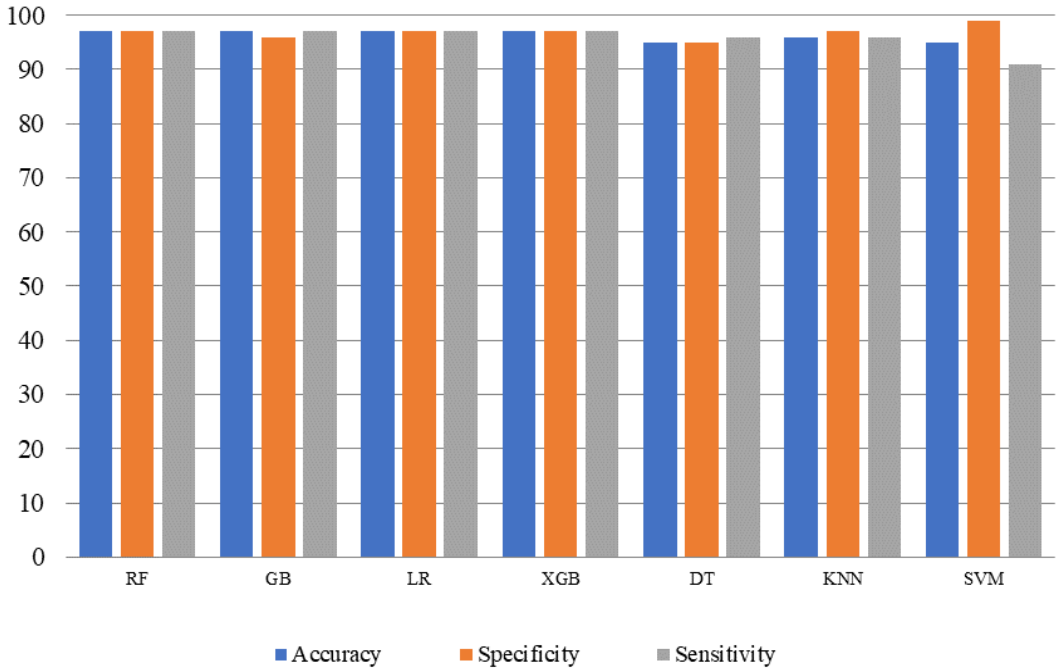
Şekil 35. Karar Ağacı Karışıklık Matrisi



Şekil 36. K-En Yakın Komşular Karışıklık Matrisi



Şekil 37. Destek Vektör Makinesi Karışıklık Matrisi



Şekil 38. Veri Artırımı ve Özellikli Modellerde Doğruluk, Özgüllük ve Duyarlılık

4.2.4 Deneysel sonuçların irdelenmesi

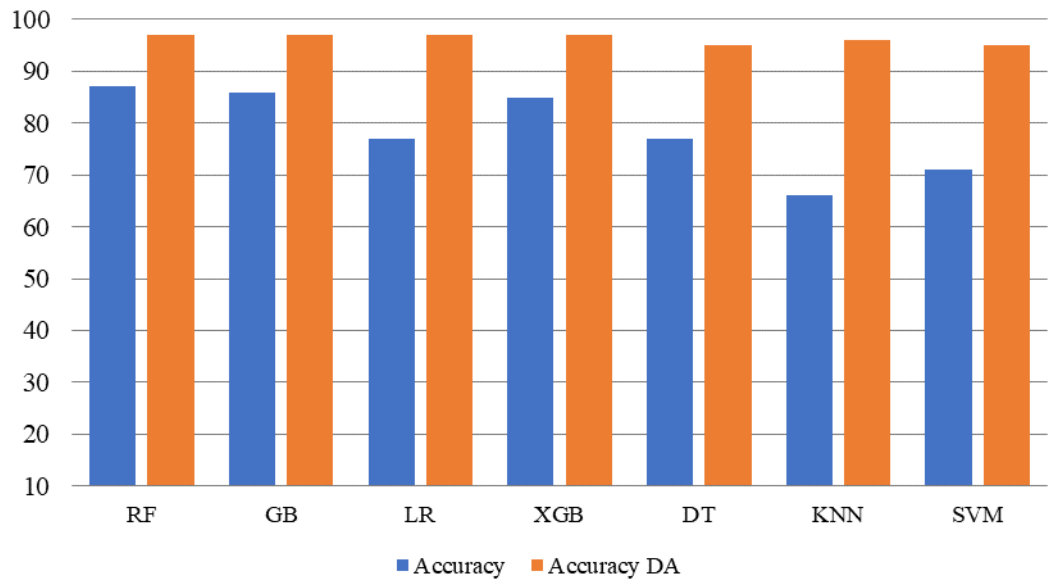
Alizadeh Koroner Arter veri kümesi, daha fazla CAD vakasına (217) sahipken normal vakalara (87) kıyasla bir dengesizlik sergilemektedir. Çekişmeli Üretici Ağ (GAN) algoritmasını kullanarak örnekler oluşturduk. Bu örnekler, ardından Farklı makine öğrenimi modellerini eğitmek için kullanıldı, bunlar arasında Rastgele Orman, Gradient Boosting, Lojistik Regresyon, XGBoost, Karar Ağaçları, K-En Yakın Komşular ve Destek Vektör Makineleri bulunmaktadır. En önemli özelliği belirlemek için Destek Vektör Makineyi bir özellik seçimi tekniği olarak kullandık. Bu bölüm, hem veri artırma hem de özellik seçimi tekniklerinin kullanılmasını dikkate alarak hesaplamalı tahmin sonuçlarının bir analizini sunmaktadır.

Bölüm 4.2.2'deki veri artırma ve özellik seçimi içermeyen tahmin sonuçlarının analizi ve bu teknikleri kullanan Bölüm 4.2.3, ilginç bulgular sunmaktadır. Bölüm 4.2.2'de modeller farklı hassasiyet seviyelerini gösterdi. İlk Rastgele Orman modeli %87'lik bir doğruluk oranına sahipti, ardından gelen modeller ise karşılaştırmalı olarak daha düşük doğruluk oranları elde etti. Gradient Boosting modeli, Lojistik Regresyon, XGBoost, Karar Ağaçı, K-En Yakın Komşular ve Destek Vektör Makine dahil modeller, sırasıyla doğruluk seviyelerine göre sıralandı. Ancak, modeller arasında doğruluk açısından belirgin bir fark olmadığı gözlenmiştir.

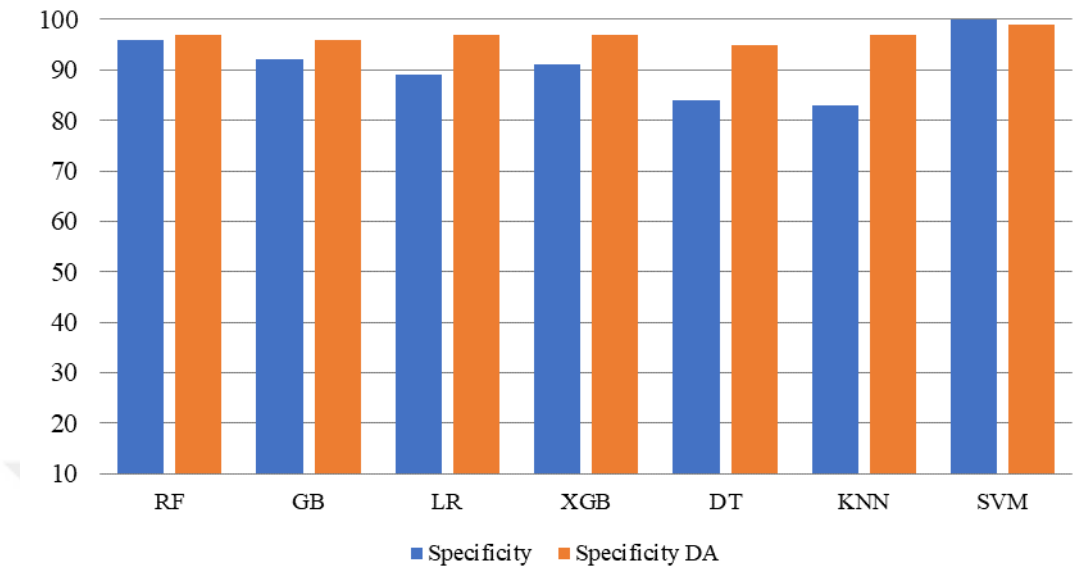
Bölüm 4.2.3'de, modeller, Bölüm 4.2.2'de tartışılan tahminlere göre artan bir doğruluk gösterdi. Rastgele Orman modeli, dikkate değer bir %97'lik bir doğruluk oranı elde ederek olağanüstü bir performans sergiledi. Ayrıca, değerli hassasiyet ve özgürlük puanları gösterdi. Bölüm 4.2.3'de, Gradient Boosting, Lojistik Regresyon, XGBoost ve Karar Ağaçı modelleri artan doğruluk göstererek sırasıyla %97, %97, %97 ve %95 doğruluk oranlarına ulaşmıştır.

K-En Yakın Komşular modeli, başlangıçtaki %66'lık doğruluktan %96'ya yükselen %30'luk bir artışla önemli bir gelişme gösterdi. Bu artış, modelin sınıflandırma performansını önemli ölçüde iyileştirdi. Destek Vektör Makine modelinin doğruluk puanı, Destek Vektör Makinenin özellik seçimi yöntemi olarak kullanıldığında %71'lik önceki skora kıyasla %95'e yükseldi. Ayrıca, Bölüm 4.2.3, doğruluk, hassasiyet açısından önemli gelişmeleri gösterdi. Örneğin, Rastgele Orman modeli, doğruluk, hassasiyet açısından diğer modellere kıyasla üstün performans sergiledi. Rastgele Orman modelinin sağlamlığı ve esnekliği, geniş bir

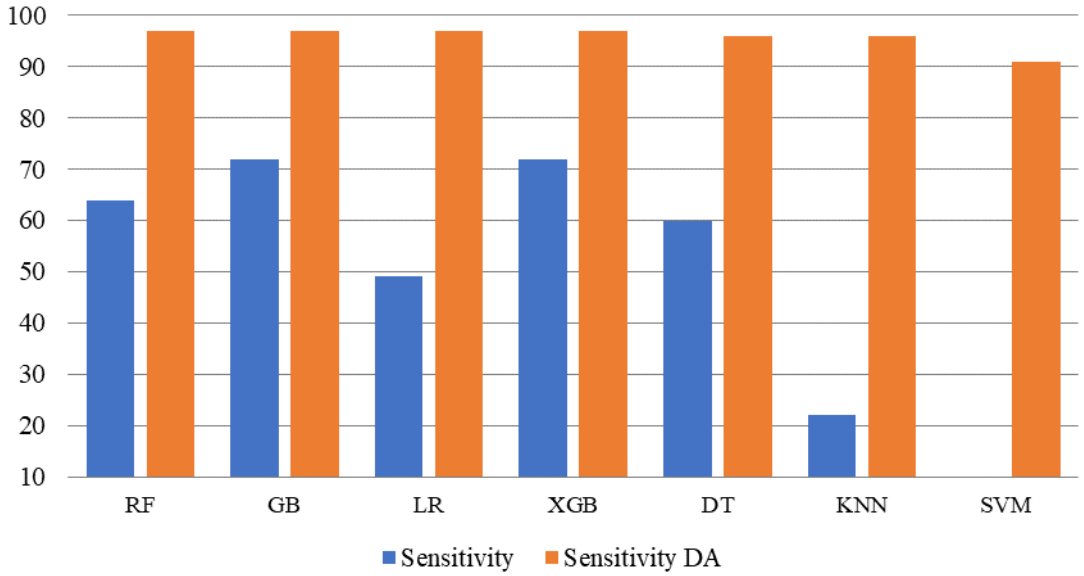
görev yelpazesine uygun bir şekilde kullanılmasına olanak tanır. Veri artırma ve özellik seçimi tekniklerinin uygulanmasının ardından, K-En Yakın Komşular modelinin hassasiyetinde bir iyileşme gözlendi. Veri artırma ve özellik seçimi, K-En Yakın Komşular gibi parametrik olmayan modellerin performansını potansiyel olarak artırabilecekleri bulunmuştur, bu da Şekil 39-31'de gösterildi. Deneysel 2 kanıt, makine öğrenimi modellerinin veri artırımı ve özellik seçimi faydalarından yararlanabileceğini göstermektedir. Gradient Boosting, Karar Ağacı, K-En Yakın Komşular ve destek vektör makine modelleri Rastgele Orman modelinden daha az doğruydu. Veri artırımı ve özellik seçimi sonrasında K-En Yakın Komşular modelinin hassasiyeti önemli ölçüde iyileşti. Bu, K-En Yakın Komşular ve diğer parametrik olmayan modellerin veri artırımı ve özellik seçimi faydalarından yararlanabileceği açıktır. Sonuçlar, model performansının veri artırımı ve özellik seçimi yöntemlerinin kullanılmasıyla büyük ölçüde gelişmiş olduğunu göstermektedir. Bu tekniklerin uygulanması, hem normal hem de koroner arter vakalarının daha hassas sınıflandırılmasını sağlamıştır. çekişmeli üretici ağ ve özellik seçimi yöntemleri gibi belirli tekniklerin kullanımı modelin performansı üzerinde gözle görülür bir etkiye sahip olmuştur.



Şekil 39. Veri Artırımı ve Özellik Seçimiyle Doğruluk Karşılaştırması



Şekil 40. Veri Artırımı ve Özellik Seçimiyle ve Bunlar Olmadan Özgüllük Karşılaştırılması



Şekil 41. Veri Artırma ve Özellik Seçimi olmadan ve Veri Artırma ve Özellik Seçimi ile duyarlılık karşılaştırması

5. SONUÇ VE DEĞERLENDİRME

İlk deney, veri artırımı amacıyla çekişmeli üretici ağ entegrasyonunun tüm modellerin genel performansında önemli bir iyileşme sağladığını göstermiştir. Çekişmeli üretici ağ uygulamasından önce, K-En Yakın Komşu modeli %74'lük bir doğruluk seviyesiyle en düşük doğruluk seviyesini sergiledi. Buna karşılık, Rastgele Orman modeli %97'lik bir doğruluk seviyesiyle en yüksek doğruluk seviyesini gösterdi. Çekişmeli üretici ağ artırımının uygulanmasını takiben, Rastgele Orman ve Gradient Boosting modellerinin her ikisi de %99'luk dikkate değer bir doğruluk oranı sergiledi. Bu sonuç, öğrenme süreci sırasında karmaşık desenleri anlama yeteneğinin arttığını göstermektedir. XGBoost ve Karar Ağacı modelleri, %98 doğruluk oranı elde ederek yüksek bir doğruluk seviyesi sergilediler. Ayrıca, K-En Yakın Komşu modeli, modeller arasında en önemli iyileşmeyi sergiledi. ÇÜA kullanarak veri artırımının uygulanması, hassasiyet ve özgüllükte iyileşme sağladı ve modelin pozitif ve negatif örnekler arasında doğru bir şekilde ayrım yapma yeteneğinin arttığını gösterdi. Genel olarak, çekişmeli üretici ağ, makine öğrenimi modellerinin performansını artırmada tutarlı bir şekilde etkili olduklarını göstererek, bir dizi uygulama alanında etkililiklerini kanıtlamışlardır.

İkinci deneyin bulguları, Alizadeh Koroner arter veri kümesinde dengesizlik gösterdi ve koroner arter hastalığı ile ilgili vakaların normal vakalara kıyasla daha yüksek bir yaygınlığı olduğunu gösterdi. Veri artırımı ve özellik seçimi teknikleri bağlamında çekişmeli üretici ağ kullanımı, model doğruluğunda önemli bir artış göstermiştir. Artırım ve seçim teknikleri olmadan, modeller farklı doğruluk seviyeleri sergiledi ve başlangıçtaki Rastgele Orman modeli %87'lik bir doğruluk oranı elde etti. Buna karşılık, teknik uygulamasını takiben, Rastgele Orman modeli %97'lik olağanüstü bir performans sergiledi. Ayrıca, hem hassasiyet hem de özgürlük puanlarında önemli gelişmeler gözlemlendi. Gradient Boosting, Lojistik Regresyon, XGBoost ve Karar Ağaçları gibi birkaç diğer model, doğruluğu artırdı. K-En Yakın Komşu modeli, hassasiyet ölçisinde önemli bir iyileşme sergiledi ve %22'den %96'ya yükseldi. Bu, modelde veri artırım tekniklerinin ve ilgili özelliklerin seçilmesinin faydalı etkilerini vurgulamaktadır. Rastgele Orman modeli, diğer modellerle karşılaştırıldığında doğruluk, hassasiyet ve özgürlük açısından sürekli olarak olağanüstü sonuçlar sergiledi. Çalışmanın bulguları, bu tekniklerin K-En Yakın Komşu gibi parametrik olmayan modellerin performansını önemli ölçüde artırabileceğini, özellikle normal ve koroner arter

vakalarının doğru sınıflandırılmasında belirgin bir potansiyele sahip olduğunu göstermektedir. Bu, bu tekniklerin model performansını artırmada önemli bir potansiyeli olduğunu göstermektedir.

Deneysel sonuçlar, Makine Öğrenimi modellerinin performansını artırmak için veri artırımı ve özellik seçimi teknikleri bağlamında çekişmeli üretici ağ etkililiği için ilginç kanıtlar sağlamaktadır.

İlk deneyde, yeni veri oluşturmak için ÇÜA kullanımı, özellikle pozitif ve negatif örnekler arasındaki farklılıkları belirleme yeteneğini artırdı. Örneğin, hem Rastgele Orman hem de Gradient Boosting modellerinin doğruluğunda bir iyileşme gözlendi ve her ikisi de %99'luk yüksek bir seviyeye ulaştı. Daha önce, Rastgele Orman modelinin doğruluğu %97 iken, Gradient Boosting modelinin doğruluğu %96 idi. K-En Yakın Komşular modeli, diğer modellere kıyasla %20'lik bir doğruluk artışı elde ederek önemli bir iyileşme sergiledi. Geliştirilen veri kümesinin bir sonucu olarak birçok modelde hassasiyet ve özgüllükte gelişmeler gözlenmektedir.

İkinci deneyde, veri artırım teknikleri kullanıldığında modeller daha iyi doğruluk gösterdi. Rastgele Orman modeli olağanüstü bir performans sergileyerek %97'lik yüksek bir doğruluk oranı elde etti. Ayrıca, önemli hassasiyet ve özğürlük puanları sergiledi. Gradient Boosting, Lojistik Regresyon, XGBoost ve Karar Ağacı modelleri de sırasıyla %97, %97, %97 ve %95 doğruluk oranları elde ederek doğruluğu artırdılar. K-En Yakın Komşular modeli, doğrulukta önemli bir iyileşme sergileyerek, %66'lık başlangıç puanından %96'lık bir doğruluk puanı elde etti. Destek Vektör Makinesi'nin bir özellik seçimi yöntemi olarak kullanılması, Destek Vektör Makine modelinin hassasiyet puanında bir iyileşmeye neden oldu ve 91 değerine ulaştı. Rastgele Orman modeli, daha yüksek performansı sürekli olarak vurgulayarak, çeşitli görevlerdeki güvenilirliğini ve uyum sağlayabilirliğini vurguladı. Sonuçlar, çekişmeli üretici ağ ve özellik seçimi stratejileri gibi belirli metodolojilerin uygulanmasının modellerin etkililiği üzerinde önemli bir etkiye sahip olabileceğini göstermektedir. Ayrıca, bu metodolojilerin tıbbi veri kümelerinde, örneğin Alizadeh Koroner arter veri kümesinde daha doğru kategorizasyonlara neden olabileceği sonucuna varılmıştır.

KAYNAKLAR

1. Afzal, S., Maqsood, M., Nazir, F., Khan, U.,(2019). A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. *IEEE access* 7, 115528-115539
2. Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights, 1*(1), 104.
3. Alomar, K., Aysel, H. I., & Cai, X. (2023). Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *Journal of Imaging 2023, Vol. 9, Page 46*.
4. Bakır, H., Çayır, A. N., & Navruz, T. S. (2023). A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification. *Multimedia Tools and Applications*.
5. Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: a statistical viewpoint. Acta numerica, 30, 87-201.
6. Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, *55*(7).
7. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967.
8. Cao, P., Li, X., Mao, K., Lu, F., Ning, G., Fang, L., & Pan, Q. (2020). A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation. Biomedical Signal Processing and Control, 56, 101675.
9. Chao, X., & Zhang, L. (2023). Few-shot imbalanced classification based on data augmentation. *Multimedia Systems*, *29*(5), 2843–2851.
10. Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., ... & Narayanan, S. (2019, September). Data Augmentation Using GANs for Speech Emotion Recognition. In Interspeech (pp. 171-175).
11. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
12. Chengchuang, L., Chun, S., Gansen, Z., & al, et. (2021). Review of Image Data Augmentation in Computer Vision. *Journal of Frontiers of Computer Science & Technology*, *15*(4), 583.
13. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., Chlap, P., Phd, M., Vandenberg, ; N, Dowling, M. ; J., & Holloway, ; L. (2021). A review of

medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563.

14. Dhanani, J., & Rana, K. (2018). Logistic regression with stochastic gradient ascent to estimate click through rate. *Lecture Notes in Networks and Systems*, *9*, 319–326.
15. Dulhare, U. N., & Hussam Ali, M. (2023). Underwater human detection using faster R-CNN with data augmentation. *Materials Today: Proceedings*, *80*, 1940–1945.
16. Fahimi, F., Dosen, S., Ang, K. K., Mrachacz-Kersting, N., & Guan, C. (2020). Generative adversarial networks-based data augmentation for brain–computer interface. IEEE transactions on neural networks and learning systems, 32(9), 4039-4051.
17. Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988.
18. Gu, R., Niu, C., Yan, Y., Wu, F., Tang, S., Jia, R., ... & Chen, G. (2022). On-Device Learning with Cloud-Coordinated Data Augmentation for Extreme Model Personalization in Recommender Systems. arXiv preprint arXiv:2201.10382.
19. Chen, L., Wu, L., Zhang, K., Hong, R., Lian, D., Zhang, Z., ... & Wang, M. (2023). Improving Recommendation Fairness via Data Augmentation. arXiv preprint arXiv:2302.06333.
20. Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*.
21. Gulakala, R., Markert, B., Reports, M. S., (2022), Generative adversarial network based data augmentation for CNN based detection of Covid-19. *scientificreports*.
22. Hemmerich, J., Asilar, E., & Ecker, G. F. (2020). COVER: Conformational oversampling as data augmentation for molecules. *Journal of Cheminformatics*, *12*(1).
23. Hüttenrauch, B. (2016). Limitations of data augmentation and outlook. *Targeting Using Augmented Data in Database Marketing*, 279–290.
24. Jain, S., Seth, G., Paruthi, A., Soni, U., & Kumar, G. (2022). Synthetic data augmentation for surface defect detection and classification using deep learning. *Journal of Intelligent Manufacturing*, *33*(4), 1007–1020.
25. Jha, R., Lovering, C., & Pavlick, E. (2020). *Does Data Augmentation Improve Generalization in NLP.*, *arXiv*. URL: http://arxiv.org/abs/2004.15012,Son Erişim Tarihi: 5/11/2023.

26. Lamba, S., Baliyan, A.,V. K.(2022). Generative Adversarial Networks based Data Augmentation for Paddy Disease Detection using Support Vector Machine. *10th International Conference on Reliability, Infocom Technologies and Optimization*.
27. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning for AI. *Communications of the ACM*, *64*(7), 58-65.
28. Li, S., Zhang, K., Chen, Q., Wang, S., & Zhang, S. (2020). Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and Genetic Algorithm. *IEEE Access*, *8*, 139512–139528.
29. Li, Y., Xu, Y., Li, X., Li, R.,(2022). Addressing imbalance of sample datasets in dissolved gas analysis by data augmentation: Generative adversarial networks. *Wiley Online Library*, *Transmission & Distribution,*, *16*(22), 4494–4504.
30. Little, C., Elliot, M., Allmendinger, R., & Samani, S. S. (2021). Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. https://arxiv.org/abs/2112.01925v1
31. Liu, Y., Hou, Y.-J., Qin, C.-X., Li, X.-H., Du, Q.-M., Li, S.-J., Wang, B., & Zhou, C.-C. (2023). A Data Augmentation Method and the Embedding Mechanism for Detection and Classification of Pulmonary Nodules on Small Samples. http://arxiv.org/abs/2303.12801
32. Manoharan, A., Begam, K. M., Aparow, V. R., & Sooriamoorthy, D. (2022). Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review. *Journal of Energy Storage*, *55*, 105384.
33. Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: a review. *Procedia Manufacturing*, *35*, 698–703.
34. Modaresi, F., & Araghinejad, S. (2014). A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. *Water Resources Management*, *28*(12), 4095–4111.
35. Moreno-Barea, F. J., Strazzera, F., Jerez, J. M., Urda, D., & Franco, L. (2019). Forward Noise Adjustment Scheme for Data Augmentation. *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, 728–734.
36. Mumuni, A., Array, F. M.(2022). Data augmentation: A comprehensive survey of modern approaches. *Array*.
37. Nanni, L., Maguolo, G., Informatics,(2020). Data augmentation approaches for improving animal audio classification. *Ecological Informatics*.

38. Ramires, A., & Serra, X. (2018). Data augmentation for instrument classification robust to audio effects. *Proceedings of the International Conference on Digital Audio Effects, DAFx*, 226–231.
39. Rana, P., Sowmya, A., Meijering, E., reports, Y.(2022). Data augmentation with improved regularisation and sampling for imbalanced blood cell image classification. *Scientific Reports.*
40. Refai, D., Abo-Soud, S., & Abdel-Rahman, M. (2022). Data Augmentation using Transformers and Similarity Measures for Improving Arabic Text Classification.
41. Saini, M., Computing, S. S.(2020). Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Applied Soft Computing*.
42. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1).
43. Shao, S., Wang, P., Industry, R.(2019). Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*.
44. Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, *8*(1).
45. Snider, E., Hernandez-Torres, S., Diagnostics, R. H.-, & 2023, undefined. (2023). Using ultrasound image augmentation and ensemble predictions to prevent machine-learning model overfitting. *Mdpi. Diagnostics*. 417.
46. Song, K., Huang, Q., Zhang, F. en, & Lu, J. (2021). Coarse-to-fine: A dual-view attention network for click-through rate prediction. *Knowledge-Based Systems*, *216*.
47. Strelcenia, E., Extraction, S. P.-M. L. and K.(2023). A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud Detection. *Machine Learning and Knowledge Extraction.*
48. Sundaram, S.,(2021).Gan-based data augmentation for chest x-ray classification. *Arxiv.*
49. Taha Jijo, B., & Mohsin Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28.
50. Turkmenoglu, B. K., & Yildiz, O. (2021). Predicting the survival of heart failure patients in unbalanced data sets Dengesiz veri setlerinde kalp yetmezliği hastalarinin sağkalim tespiti. *29th IEEE Conference on Signal Processing and Communications Applications, Proceedings*.

51. Woodman, R. J., & Mangoni, A. A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. *Aging Clinical and Experimental Research*, 1–35.
52. Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition*, *137*.
53. Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X., & Zhu, T. (2017). Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework. IEEE 2nd International Conference on Big Data Analysis.
54. Xu, Y., Noy, A., Lin, M., Qian, Q., Li, H., & Jin, R. (2020). *WeMix: How to Better Utilize Data Augmentation*. http://arxiv.org/abs/2010.01267.
55. Yu, S., Zhang, S., Wang, B., Dun, H., Xu, L., Huang, X., Shi, E., & Feng, X. (2021). Generative adversarial network based data augmentation to improve cervical cell classification model. *Mathematical Biosciences and Engineering*, *18*(2), 1740–1752.
56. Zhang, J., Ma, C., Zhong, C., Zhao, P., & Mu, X. (2022). Multi-scale and multi-channel neural network for click-through rate prediction. *Neurocomputing*, *480*, 157–168.
57. Zhang, R., Zhou, B., Lu, C., in, M. M.-M. P.,(2022). The performance research of the data augmentation method for image Classification. *Mathematical Problems in Engineering.*
58. Zhu, Q., Kuang, W., Image, Z. L.-J. of V. C.(2022). Dual attention interactive fine-grained classification network based on data augmentation. *Journal of Visual Communication and Image Representation*.
59. Zipes, D. P. (2018). Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine (11th Edition). USA: ELSEVIR.
60. Zou, D., Wang, Z., Zhang, L., Zou, J., Li, Q., Chen, Y., & Sheng, W. (2021). Deep Field Relation Neural Network for click-through rate prediction. *Information Sciences*, *577*, 128–139.



GAZİLİ OLMAK AYRICALIKTIR.