T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Doktora Tezi

İLAÇ – HEDEF ETKİLEŞİMİNDE (DTI) PROTEİNLERİN ETKİLİ ÖZELLİKLERİNİN BELİRLENMESİNDE YAPAY ZEKA ALGORİTMALARI TABANLI BİR ÇERÇEVE GELİŞTİRME

Sara NAGHİB ZADEH

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Prof. Dr. Zümrüt ECEVİT SATI

II. DANIŞMAN
Doç. Dr. Ali GHANBARİ

Ağustos, 2025

İSTANBUL

Bu çalışma, 15.08.2025 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı, Enformatik Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. **Zümrüt ECEVİT SATI**(Danışman)
İstanbul Üniversitesi
Fakülte

Prof. Dr. **SEVINÇ GÜLSEÇEN**
İstanbul Üniversite
Fakülte

Doç. Dr. YALÇIN ÖZKAN
İstinye Üniversite
Fakülte

Dr. Öğr. Üyesi GÖKÇE KARAHAN ADALI
Beykent Üniversite
Fakülte

Prof. Dr. ÇİĞDEM EROL
İstanbul Üniversite
Fakülte

İntihal Programı Beyanı

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Proje Destekleri

Bu tez, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Yürütücü Sekreterliğinin ..... numaralı projesi ile desteklenmiştir.

Bu tez, ... numaralı ..... projesi ile desteklenmiştir

Tezden Üretilmiş Yayınların Künye Bilgileri

Naghib Zadeh, S., Ecevit Satı, Z., & Ghanbari Sorkhi, A. (2025). Improving Drug-Target Interaction Prediction Using SVM-SMOTE: A Solution for Imbalanced Datasets. International Journal of Management Information Systems and Computer Science, 9(1), 10-28. https://doi.org/10.33461/uybisbbd.1661593

Naghib Zadeh, S., & Ecevit Satı, Z. (2025). A BIBLIOMETRIC EVALUATION OF SMART AGRICULTURE RESEARCH. International Journal of 3D Printing Technologies and Digital Industry, 9(2), 194-206. https://doi.org/10.46519/ij3dptdi.1648618

ÖNSÖZ

Bilimsel bilgiye katkıda bulunma amacıyla çıktığım bu doktora yolculuğu, yalnızca bireysel çaba ve araştırmalarla değil, aynı zamanda birçok değerli insanın desteğiyle şekillenmiş; sabır, kararlılık ve inançla örülmüş bir süreç olmuştur. Bu tez, hem bilimsel bir çalışmanın sonucu hem de bu yolculuğun tanığı olarak kalacaktır.

Bu süreçte her zaman bilgi ve deneyimiyle bana yol gösteren, akademik gelişimimde büyük katkıları olan değerli danışmanım Prof. Dr. Zümrüt Ecevit Satı'ya en içten teşekkürlerimi sunuyorum. Kendisinin rehberliği, teşvik edici tutumu ve yönlendirmeleri, bu çalışmanın ortaya çıkmasında belirleyici olmuştur. Aynı şekilde, değerli ikinci danışmanım Doç. Dr. Ali Ghanbari'ye; bilimsel bakış açımı geliştirmeme katkı sağlayan kıymetli yorumları, sabrı ve desteği için sonsuz teşekkür ederim. Bölüm başkanımız Prof. Dr. Sevinç Gülseçen'e, sağladığı akademik ortam ve destekleyici tutumu için içtenlikle teşekkür ederim. Onun liderliği altında böyle bir çalışma gerçekleştirmek benim için bir onur olmuştur.

Hayatım boyunca en büyük destekçilerim olan, her daim yanımda olduklarını hissettiren sevgili annem ve babama olan minnettarlığımı kelimelerle ifade etmek güçtür. Onların sonsuz sevgisi, duaları ve karşılıksız fedakârlıkları, bu yolculuğun en sağlam temelini oluşturmuştur. En zor anlarımda bana güç verdiler, inançlarını hiç kaybetmeden daima arkamda durdular. Bu başarı, onların sevgisi ve desteği olmadan mümkün olmazdı. Bu tezi, her şeyden önce onların ellerine ve yüreklerine ithaf ediyorum.

Son olarak, bilimsel araştırmaların insanlığın ortak geleceği için umut ışığı olduğuna yürekten inanıyor; bu yolda atılan her adımın, bir hastalığın daha tedavisinin bulunmasına vesile olmasını diliyorum. Dünyada hiçbir hastalığın çaresiz kalmadığı bir gelecek umuduyla, bilimin insanlık yararına hizmet etmeye devam etmesini temenni ediyorum.

Ağustos 2025 SARA NAGHIB ZADEH

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **ix**
TABLO LİSTESİ ..... **xi**
SİMGE VE KISALTMA LİSTESİ ..... **xiv**
ÖZET ..... **xxii**
SUMMARY ..... **xxv**
1. GİRİŞ **1**
1.1. ÇALIŞMANIN AMACI 3
1.2. ARAŞTIRMA HEDEFLERİNİN GERÇEKLEŞTİRİLMESİNE YÖNELİK ÖNERİLEN ADIMLAR 4
1.3. ÇALIŞMANIN VARSAYIMLARI 4
2. GENEL KISIMLAR **6**
2.1 GİRİŞ 6
2.2 BİYOLOJİK KAVRAMLAR 7
2.2.1 Gen 7
2.2.3 Protein 9
2.2.4 İlaç Ve Kimyasal Yapısı 10
2.3 İLAÇ GELİŞTİRME AŞAMALARI 11
2.4. PROTEİN ÖZELLİKLERİNİN TAHMİNİ VE UYGULAMALARI 11
2.5 DTI TAHMİN MODELLERİ 12
2.5.1 Kantitatif yapı-aktivite ilişkisi modeli (Quantitative Structure–Activity Relationship (QSAR)) 13
2.5.2 Proteokimimetrik Model (Proteochemometric Modeling (PCM)) 13
2.5.3 Moleküler yerleştirme (Molecular docking) 14
2.6 TEMEL HESAPLAMALI YAKLAŞIMLARLA DTI TAHMİNİ 15
2.6.1 Yerleştirme Simülasyonları 15
2.6.2 Makine Öğrenimi Yöntemleri 16
2.7 DTI TAHMİNİNDE KULLANILAN MAKİNE ÖĞRENMESİ YÖNTEMLERİ 16
2.7.1.Benzerlik/Mesafe Tabanlı Yöntemler 16

2.7.2.Özellik Tabanlı Yöntemler 17
2.7.3 Matris Ayrıştırma Yöntemleri 17
2.7.4 Ağ Tabanlı Yöntemler 18
2.7.5 Derin Öğrenme Yöntemleri 18
2.7.6 Kombine Yöntemler 19
2.8 DTI ÜZERINE YAPILAN ÇALIŞMALARIN İNCELENMESI 19
3. MALZEME VE YÖNTEM 23
3.1 ARAŞTIRMANIN MODELİ 23
3.2 İLAÇ-HEDEF VERI KÜMELERI 26
3.3 ÖZELLIK ÇIKARMA YÖNTEMLERI 27
3.3.1 Veri Çeşitliliğinin Önemi Ve Özelliklerin Alt Kümeleri 28
3.3.2 İlaçlardan Özellikler Çıkarma 30
3.3.3 Proteinlerden Özellikler Çıkarma 31
3.4 TEMEL BİLEŞENLER ANALİZİ (PRINCIPAL COMPONENTS ANALYSİS (PCA)) 39
3.5 VERI DENGELEME TEKNIĞI 40
3.5.1 Siyam Sinir Ağları (Siamese Neural Networks (SNN)) 42
3.5.2 Sentetik Azınlık Aşırı Örnekleme Tekniği – Tomek Bağlantıları(Synthetic Minority Over-sampling Technique – Tomek Links(SMOTE-Tomek)) 52
3.6 DEĞERLENDİRİLEN VERİ DENGELEME YÖNTEMLERİ VE ÖRNEKLEME TEKNİKLERİ 53
3.6.1 Kümeleme Merkezi Algoritması(Cluster Centroid Algorithm) 54
3.6.2 Uyarlanabilir Yoğunlaştırılmış En Yakın Komşu (Adaptive Condensed Nearest Neighbor (Ada-CNN)) 54
3.6.3 Düzenlenmiş En Yakın Komşu (Edited Nearest Neighbour (ENN)) 55
3.6.4 Tekrarlamalı Düzenlenmiş En Yakın Komşu (Repeated Edited Nearest Neighbour (RENN)) 55
3.6.5 Komşuluk Temizleme Kuralı (Neighbourhood Cleaning Rule (NCL)) 56
3.6.6 Tek Taraflı Seçim(One-Sided Selection (ONESS)) 57
3.6.7 Tüm k-en yakın komşular (All k-nearest neighbors (All-KNN)) 57
3.6.8.RandomOverSampler 58
3.6.9 RandomUnderSampler 58
3.6.10 Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique(SMOTE)) 58
3.6.11 Uyarlanabilir Sentetik Örnekleme (Adaptive Synthetic Sampling(ADASYN)) 59

3.6.13 SMOTE ve Düzenlenmiş En Yakın Komşular (SMOTE Edited Nearest Neighbours (SMOTEENN)) 60
3.6.14 Sınırda Sentetik Azınlık Aşırı Örnekleme Tekniği (Borderline Synthetic Minority Over-sampling Technique (Borderline-SMOTE)) 60
3.6.15 SVM-SMOTE 61
3.6.16 Tomek Bağlantıları (Tomek Links) 62
3.6.17. SMOTE Nominal ve Sürekli Değişkenler için (SMOTE for Nominal and Continuous (SMOTE-NC)) 62
3.6.18 Nominal Veriler için Sentetik Azınlık Aşırı Örnekleme TekniğiSMOTEN (Synthetic Minority Over-sampling Technique for Nominal Data(SMOTEN)) 63
3.7 DEĞERLENDİRİLEN MAKİNE ÖĞRENMESİ ALGORİTMALARI 63
3.7.1 K-En Yakın Komşu (k-Nearest Neighbors (K-NN)) Algoritması 64
3.7.2 Yapay Sinir Ağları 65
3.7.3 Naive Bayes Algoritması 67
3.7.4 Destek Vektör Makinesi (Support Vector Machine (SVM)) 68
3.7.5 Karar Ağacı Algoritması 69
3.7.6 Kareli Ayrımcı Analiz (Quadratic Discriminant Analysis(QDA)) 71
3.7.7 Doğrusal Ayırıcı Analiz (Linear Discriminant Analysis (LDA)) 72
3.7.8 Kolay Topluluk Sınıflandırıcısı (Easy Ensemble Classifier (EEC)) 73
3.7.9 Dengeli Torba Sınıflandırıcı (Balanced Bagging Classifier (BBC)) 74
3.7.10 Uyarlanabilir Güçlendirme (Adaptive Boosting (AdaBoost)) 75
3.7.11 Rastgele Azaltma ve Güçlendirme (RandomUndersampling ve Boosting (RUSBoost)) 76
3.8 DEĞERLENDİRME ÖLÇÜTLERİ 77
4. BULGULAR 79
4.1 ÇEŞİTLİ DTI VERİ SETLERİ ÜZERİNDE AKILLI TAHMİN MODELİNİN PERFORMANS ANALİZİ 83
4.2 NÜKLEER RESEPTÖR VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI 85
4.2.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması 85
4.2.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması 90
4.2.3 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması 93
4.2.4 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması 98
4.3 İYON KANALLARI VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI 103
4.3.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması 103
4.3.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması 107

4.3.3 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması..... 111
4.3.4 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması... 115
4.4 G PROTEİNİNE BAĞLI RESEPTÖRLER VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI..... 118
4.4.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması 119
4.4.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması 123
4.4.3 CKSAAGP_GDPC_GTPC Özellikli Veri Setinde Performans Karşılaştırması 128
4.4.4 PSEAAC_PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması 132
4.5 ENZİM VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI..... 137
4.5.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması 137
4.5.2 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması... 142
4.5.3 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması 145
4.5.4 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması..... 150
4.6 ÖNERİLEN MODELİN PERFORMANSININ ROC VE RECALL EĞRİLERİ KULLANILARAK ANALİZ EDİLMESİ..... 154
4.6.1 İyon Kanalları Veriseti: Önerilen Modelin ROC & Recall Analizi 154
4.6.2 G Proteinine Bağlı Reseptörler Veriseti: Önerilen Modelin ROC & Recall Analizi..... 160
4.6.3 Nükleer Reseptör Veriseti: Önerilen Modelin ROC & Recall Analizi 165
4.6.4 Enzim Veriseti: Önerilen Modelin ROC & Recall Analizi 170
4.7 MODELİN FARKLI VERİ KÜMELERİNDE GENELLENEBİLİRLİĞİNİN VE KARARLILIĞININ DEĞERLENDİRİLMESİ 175
5. TARTIŞMA VE SONUÇ 178
KAYNAKLAR 183
EKLER..... Error! Bookmark not defined.
ÖZGEÇMİŞ 212

ŞEKİL LİSTESİ

Sayfa No

Şekil 3. 1: İlaç-hedef etkileşimlerini tahmin etmek için önerilen modelin süreç akışı (yazar tarafından). 26

Şekil 3. 2: Altın standart veri kümesinin özeti 27

Şekil 3. 3: İlaç–protein etkileşim verilerinde özellik çıkarma yöntemlerinin sınıflandırılması 28

Şekil 3. 4: Özellik Çıkarımı ve Özellik Sayısının Özeti. 29

Şekil 3. 5: İlaç ve hedef özelliklerinden türetilen dört özellik alt kümesi. 29

Şekil 3. 6: DTI tahmininde veri dengesizliğine çözüm tekniği 42

Şekil 3. 7: Siyam sinir ağı mimarisi (Utkin ve diğ., 2021) 44

Şekil 3. 8: Veriye Benzerlik Hesaplama için Siamese Ağının İç Yapısı (yazar tarafından). 51

Şekil 3. 9: k-en yakın komşu yönteminin gösterimi 65

Şekil 3. 10: Perceptron şeması 66

Şekil 3. 11: SVM'deki hiper düzlemler (Otten, 2024) 69

Şekil 3. 12: Örnek bir karar ağacı (Wang ve Saddik, 2023) 71

Şekil 4. 1: DTI tahmininde yapay zeka algoritmalarına dayalı önerilen çerçevenin genel yapısı (yazar tarafından). 79

Şekil 4. 2: Önerilen İlaç–Hedef Etkileşimi Tahmin Modelinin Akış Diyagramı (yazar tarafından). 80

Şekil 4. 3: Önerilen modelin farklı dengeleme ve makine öğrenimi yöntemleriyle karşılaştırma yapısı (yazar tarafından). 83

Şekil 4. 4: IC Veriseti AAC DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 155

Şekil 4. 5: IC Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 157

Şekil 4. 6: IC Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 158

Şekil 4. 7: IC Veriseti PSEAAC-PSEPSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 160

Şekil 4. 8: GPCR Veriseti AAC-DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 162

Şekil 4. 9: GPCR Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 163

Şekil 4. 10: GPCR Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 164

Şekil 4. 11: GPCR Veriseti PSEAAC-PSEPSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 165

Şekil 4. 12: NR Veriseti AAC-DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 166

Şekil 4. 13: NR Veriseti PseAAC-PsePSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 168

Şekil 4. 14: NR Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 169

Şekil 4. 15: NR Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 170

Şekil 4. 16: EN Veriseti AAC-DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 171

Şekil 4. 17: EN Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 172

Şekil 4. 18: EN Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 174

Şekil 4. 19:EN Veriseti PSEAAC-PSEPSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması 175

TABLO LİSTESİ

Sayfa No

Tablo 4. 1:Tez Çalışmasında Kullanılan Donanım ve Yazılım Bileşenleri 81

Tablo 4. 2: Akıllı DTI Tahmin Modelinin Farklı Veri Setleri Üzerindeki Performans Sonuçları 84

Tablo 4. 3: NR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 88

Tablo 4. 4: NR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 89

Tablo 4. 5: NR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 92

Tablo 4. 6 : NR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 93

Tablo 4. 7: NR veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı 96

Tablo 4. 8: NR veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı 97

Tablo 4. 9: NR veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı 101

Tablo 4. 10: NR veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı 102

Tablo 4. 11: IC veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 106

Tablo 4. 12: IC veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 107

Tablo 4. 13: IC veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 110

Tablo 4. 14: IC veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 111

Tablo 4. 15: IC veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı 114

Tablo 4. 16: IC veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı 115

Tablo 4. 17: IC veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı 117

Tablo 4. 18: IC veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı 118

Tablo 4. 19: GPCR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 122

Tablo 4. 20: GPCR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 123

Tablo 4. 21: GPCR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 126

Tablo 4. 22: GPCR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 127

Tablo 4. 23: GPCR veri kümesi üzerinde CKSAAGP_GDPC_GTPC özellikleriyle eğitilen Siyam modellerinin performansı 130

Tablo 4. 24 :GPCR veri kümesi üzerinde CKSAAGP_GDPC_GTPC özellikleriyle eğitilen Siyam modellerinin performansı 131

Tablo 4. 25: GPCR veri kümesi üzerinde PSEAAC_PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı 135

Tablo 4. 26: GPCR veri kümesi üzerinde PSEAAC_PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı 136

Tablo 4. 27: EN veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 140

Tablo 4. 28: EN veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı 141

Tablo 4. 29: EN veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı 144

Tablo 4. 30: EN veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı 145

Tablo 4. 31: EN veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 148

Tablo 4. 32: EN veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı 149

Tablo 4. 33: EN veri kümesi üzerinde pseAAC-psePSSM özellikleriyle eğitilen Siyam modellerinin performansı 152

Tablo 4. 34: EN veri kümesi üzerinde pseAAC-psePSSM özellikleriyle eğitilen Siyam modellerinin performansı 153

Tablo 4. 35: EN , IC , GPCR ve NR Veri Kümelerinde Siamese ve SMOTE-Tomek ile Farklı Sınıflandırıcıların (5 Katlı Çapraz Doğrulama) Performansı 177

Tablo 5. 1: Önerilen Modelin Dört Veri Kümesinde Mevcut Yöntemlerle Karşılaştırılması 179

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar	Açıklama
DTI	: Drug – Target Interaction
EN	: Enzyme
IC	: Ion Channel
GPCR	: G Protein-Coupled Receptor
NR	: Nuclear Receptor
KNN	: K Nearest Neighbors
SVM	: Support Vector Machine
DT	: Decision Tree
RF	: Random Forest
AB	: AdaBoost
NB	: Naive Bayes
QDA	: Quadratic Discriminant Analysis
LDA	: Linear Discriminant Analysis
EEC	: Easy Ensemble Classifier
RSC	: Random Sample Consensus
BBC	: Balanced Bagging Classifier
BRFC	: Balanced Random Forest Classifier
MÖ	: Makine öğrenmesi
DÖ	: derin öğrenme
SNN	: Siamese Neural Network
AAC	: Amino acid composition
DPC	: Dipeptide Composition
GAAC	: Grouped Amino Acid Composition
DDE	: Dipeptide Deviation from Expected Mean
PseAAC	: Pseudo Amino Acid Composition
PsePSSM	: Pseudo Position-Specific Scoring Matrix
CKSAAGP	: Composition of k-Spaced Amino Acid Group Pairs
GDPC	: Grouped Dipeptide Composition
GTPC	: Grouped Tripeptide Composition
SMOTE-Tomek	: Synthetic Minority Over-sampling Technique – Tomek Links
Ada-CNN	: Adaptive Condensed Nearest Neighbor
ENN	: Edited Nearest Neighbour

RENN : Repeated Edited Nearest Neighbour
NCL : Neighbourhood Cleaning Rule
ONESS : One-Sided Selection
All-KNN : All k-nearest neighbors
K-NN : k-Nearest Neighbors
SMOTE : Synthetic Minority Over-sampling Technique
ADASYN : Adaptive Synthetic Sampling
SMOTEENN : SMOTE Edited Nearest Neighbours
Borderline-SMOTE : Borderline Synthetic Minority Over-Sampling Technique
SMOTE-NC : SMOTE for Nominal and Continuous
SMOTEN : Synthetic Minority Over-Sampling Technique for Nominal Data
QDA : Quadratic Discriminant Analysis
ROC : Receiver Operating Characteristic
AUC : Area Under the Curve
SCC-KNN : Siamese -ClusterCentroids-Nearest Neighbors
SCN-KNN : Siamese -ConNearNeigh-Nearest Neighbors
SCC-SVM : Siamese -ClusterCentroids-Linear SVM
SCN-SVM : Siamese -ConNearNeigh-Linear SVM
SCC-DT : Siamese -ClusterCentroids-Decision Tree
SCN-DT : Siamese -ConNearNeigh-Decision Tree
SCC-RF : Siamese -ClusterCentroids-Random Forest
SCN-RF : Siamese -ConNearNeigh-Random Forest
SCC-NN : Siamese -ClusterCentroids-Neural Net
SCN-NN : Siamese -ConNearNeigh-Neural Net
SCC-AB : Siamese -ClusterCentroids-AdaBoost
SCN-AB : Siamese -ConNearNeigh-AdaBoost
SCC-NB : Siamese -ClusterCentroids-Naive Bayes
SCN-NB : Siamese -ConNearNeigh-Naive Bayes
SCC-QDA : Siamese -ClusterCentroids-QDA
SCN-QDA : Siamese -ConNearNeigh-QDA
SCC-LDA : Siamese -ClusterCentroids-LDA
SCN-LDA : Siamese -ConNearNeigh-LDA
SCC-EEC : Siamese -ClusterCentroids-EEC
SCN-EEC : Siamese -ConNearNeigh-EEC

SCC-RSC : Siamese -ClusterCentroids-RSC
SCN-RSC : Siamese -ConNearNeigh-RSC
SCC-BBC : Siamese -ClusterCentroids-BBC
SCN-BBC : Siamese -ConNearNeigh-BBC
SCC-BRFC : Siamese -ClusterCentroids-BRFC
SCN-BRFC : Siamese -ConNearNeigh-BRFC
SEN-KNN : Siamese -EditNearNeighb-Nearest Neighbors
SRN-KNN : Siamese -ReEditNearNeighb-Nearest Neighbors
SEN-SVM : Siamese -EditNearNeighb-Linear SVM
SRN-SVM : Siamese -ReEditNearNeighb-Linear SVM
SEN-DT : Siamese -EditNearNeighb-Decision Tree
SRN-DT : Siamese -ReEditNearNeighb-Decision Tree
SEN-RF : Siamese -EditNearNeighb-Random Forest
SRN-RF : Siamese -ReEditNearNeighb-Random Forest
SEN-NN : Siamese -EditNearNeighb-Neural Net
SRN-NN : Siamese -ReEditNearNeighb-Neural Net
SEN-AB : Siamese -EditNearNeighb-AdaBoost
SRN-AB : Siamese -ReEditNearNeighb-AdaBoost
SEN-NB : Siamese -EditNearNeighb-Naive Bayes
SRN-NB : Siamese -ReEditNearNeighb-Naive Bayes
SEN-QDA : Siamese -EditNearNeighb-QDA
SRN-QDA : Siamese -ReEditNearNeighb-QDA
SEN-LDA : Siamese -EditNearNeighb-LDA
SRN-LDA : Siamese -ReEditNearNeighb-LDA
SEN-EEC : Siamese -EditNearNeighb-EEC
SRN-EEC : Siamese -ReEditNearNeighb-EEC
SEN-RSC : Siamese -EditNearNeighb-RSC
SRN-RSC : Siamese -ReEditNearNeighb-RSC
SEN-BBC : Siamese -EditNearNeighb-BBC
SRN-BBC : Siamese -ReEditNearNeighb-BBC
SEN-BRFC : Siamese -EditNearNeighb-BRFC
SRN-BRFC : Siamese -ReEditNearNeighb-BRFC
SAKNN-KNN : Siamese -AllKNN-Nearest Neighbors
SNC-KNN : Siamese -NeighbCleanRule-Nearest Neighbors

SAKNN-SVM : Siamese -AllKNN-Linear SVM
SNC-SVM : Siamese -NeighbCleanRule-Linear SVM
SAKNN-DT : Siamese -AllKNN-Decision Tree
SNC-DT : Siamese -NeighbCleanRule-Decision Tree
SAKNN-RF : Siamese -AllKNN-Random Forest
SNC-RF : Siamese -NeighbCleanRule-Random Forest
SAKNN-NN : Siamese -AllKNN-Neural Network
SNC-NN : Siamese -NeighbCleanRule-Neural Net
SAKNN-AB : Siamese -AllKNN-AdaBoost
SNC-AB : Siamese -NeighbCleanRule-AdaBoost
SAKNN-NB : Siamese -AllKNN-Naive Bayes
SNC-NB : Siamese -NeighbCleanRule-Naive Bayes
SAKNN-QDA : Siamese -AllKNN-QDA
SNC-QDA : Siamese -NeighbCleanRule-QDA
SAKNN-LDA : Siamese -AllKNN-LDA
SNC-LDA : Siamese -NeighbCleanRule-LDA
SAKNN-EEC : Siamese -AllKNN-EEC
SNC-EEC : Siamese -NeighbCleanRule-EEC
SAKNN-RSC : Siamese -AllKNN-RSC
SNC-RSC : Siamese -NeighbCleanRule-RSC
SAKNN-BBC : Siamese -AllKNN-BBC
SNC-BBC : Siamese -NeighbCleanRule-BBC
SAKNN-BRFC : Siamese -AllKNN-BRFC
SOS-KNN : Siamese-OneSidedSel-Nearest Neighbors
SRUS-KNN : Siamese-RandUnderSam-Nearest Neighbors
SOS-SVM : Siamese-OneSidedSel-Linear SVM
SRUS-SVM : Siamese-RandUnderSam-Linear SVM
SOS-DT : Siamese-OneSidedSel-Decision Tree
SRUS-DT : Siamese-RandUnderSam-Decision Tree
SOS-RF : Siamese-OneSidedSel-Random Forest
SRUS-RF : Siamese-RandUnderSam-Random Forest
SOS-NN : Siamese-OneSidedSel-Neural Net
SRUS-NN : Siamese-RandUnderSam-Neural Net
SOS-AB : Siamese-OneSidedSel-AdaBoost

SRUS-AB : Siamese-RandUnderSam-AdaBoost
SOS-NB : Siamese-OneSidedSel-Naive Bayes
SRUS-NB : Siamese-RandUnderSam-Naive Bayes
SOS-QDA : Siamese-OneSidedSel-QDA
SRUS-QDA : Siamese-RandUnderSam-QDA
SOS-LDA : Siamese-OneSidedSel-LDA
SRUS-LDA : Siamese-RandUnderSam-LDA
SOS-EEC : Siamese-OneSidedSel-EEC
SRUS-EEC : Siamese-RandUnderSam-EEC
SOS-RSC : Siamese-OneSidedSel-RSC
SRUS-RSC : Siamese-RandUnderSam-RSC
SOS-BBC : Siamese-OneSidedSel-BBC
SRUS-BBC : Siamese-RandUnderSam-BBC
SOS-BRFC : Siamese-OneSidedSel-BRFC
SRUS-BRFC : Siamese-RandUnderSam-BRFC
STL-KNN : Siamese-TomekLinks-Nearest Neighbors
SROS-KNN : Siamese-RandOverSampler-Nearest Neighbors
STL-SVM : Siamese-TomekLinks-Linear SVM
SROS-SVM : Siamese-RandOverSampler-Linear SVM
STL-DT : Siamese-TomekLinks-Decision Tree
SROS-DT : Siamese-RandOverSampler-Decision Tree
STL-RF : Siamese-TomekLinks-Random Forest
SROS-RF : Siamese-RandOverSampler-Random Forest
STL-NN : Siamese-TomekLinks-Neural Net
SROS-NN : Siamese-RandOverSampler-Neural Net
STL-AB : Siamese-TomekLinks-AdaBoost
SROS-AB : Siamese-RandOverSampler-AdaBoost
STL-NB : Siamese-TomekLinks-Naive Bayes
SROS-NB : Siamese-RandOverSampler-Naive Bayes
STL-QDA : Siamese-TomekLinks-QDA
SROS-QDA : Siamese-RandOverSampler-QDA
STL-LDA : Siamese-TomekLinks-LDA
SROS-LDA : Siamese-RandOverSampler-LDA
STL-EEC : Siamese-TomekLinks-EEC

SROS-EEC : Siamese-RandOverSampler-EEC
STL-RSC : Siamese-TomekLinks-RSC
SROS-RSC : Siamese-RandOverSampler-RSC
STL-BBC : Siamese-TomekLinks-BBC
SROS-BBC : Siamese-RandOverSampler-BBC
STL-BRFC : Siamese-TomekLinks-BRFC
SROS-BRFC : Siamese-RandOverSampler-BRFC
SSM-KNN : Siamese-SMOTE-Nearest Neighbors
SSNC-KNN : Siamese-SMOTENC-Nearest Neighbors
SSM-SVM : Siamese-SMOTE-Linear SVM
SSNC-SVM : Siamese-SMOTENC-Linear SVM
SSM-DT : Siamese-SMOTE-Decision Tree
SSNC-DT : Siamese-SMOTENC-Decision Tree
SSM-RF : Siamese-SMOTE-Random Forest
SSNC-RF : Siamese-SMOTENC-Random Forest
SSM-NN : Siamese-SMOTE-Neural Net
SSNC-NN : Siamese-SMOTENC-Neural Net
SSM-AB : Siamese-SMOTE-AdaBoost
SSNC-AB : Siamese-SMOTENC-AdaBoost
SSM-NB : Siamese-SMOTE-Naive Bayes
SSNC-NB : Siamese-SMOTENC-Naive Bayes
SSM-QDA : Siamese-SMOTE-QDA
SSNC-QDA : Siamese-SMOTENC-QDA
SSM-LDA : Siamese-SMOTE-LDA
SSNC-LDA : Siamese-SMOTENC-LDA
SSM-EEC : Siamese-SMOTE-EEC
SSNC-EEC : Siamese-SMOTENC-EEC
SSM-RSC : Siamese-SMOTE-RSC
SSNC-RSC : Siamese-SMOTENC-RSC
SSM-BBC : Siamese-SMOTE-BBC
SSNC-BBC : Siamese-SMOTENC-BBC
SSM-BRFC : Siamese-SMOTE-BRFC
SSNC-BRFC : Siamese-SMOTENC-BRFC
SAD-KN : Siamese -ADASYN-Nearest Neighbors

SBS-KNN : Siamese -BorderlineSMOTE-Nearest Neighbors
SAD-SVM : Siamese -ADASYN-Linear SVM
SBS-SVM : Siamese -BorderlineSMOTE-Linear SVM
SAD-DT : Siamese -ADASYN-Decision Tree
SBS-DT : Siamese -BorderlineSMOTE-Decision Tree
SAD-RF : Siamese -ADASYN-Random Forest
SBS-RF : Siamese -BorderlineSMOTE-Random Forest
SAD-NN : Siamese -ADASYN-Neural Net
SBS-NN : Siamese -BorderlineSMOTE-Neural Net
SAD-AB : Siamese -ADASYN-AdaBoost
SBS-AB : Siamese -BorderlineSMOTE-AdaBoost
SAD-NB : Siamese -ADASYN-Naive Bayes
SBS-NB : Siamese -BorderlineSMOTE-Naive Bayes
SAD-QDA : Siamese -ADASYN-QDA
SBS-QDA : Siamese -BorderlineSMOTE-QDA
SAD-LDA : Siamese -ADASYN-LDA
SBS-LDA : Siamese -BorderlineSMOTE-LDA
SAD-EEC : Siamese -ADASYN-EEC
SBS-EEC : Siamese -BorderlineSMOTE-EEC
SAD-RSC : Siamese -ADASYN-RSC
SBS-RSC : Siamese -BorderlineSMOTE-RSC
SAD-BBC : Siamese -ADASYN-BBC
SBS-BBC : Siamese -BorderlineSMOTE-BBC
SAD-BRFC : Siamese -ADASYN-BRFC
SBS-BRFC : Siamese -BorderlineSMOTE-BRFC
SSVM-KNN : Siamese -SVMSMOTE-Nearest Neighbors
SSTK-KNN : Siamese -SMOTETomek-Nearest Neighbors
SSVM-SVM : Siamese -SVMSMOTE-Linear SVM
SSTK-SVM : Siamese -SMOTETomek-Linear SVM
SSVM-DT : Siamese -SVMSMOTE-Decision Tree
SSTK-DT : Siamese -SMOTETomek-Decision Tree
SSVM-RF : Siamese -SVMSMOTE-Random Forest
SSTK-RF : Siamese -SMOTETomek-Random Forest
SSVM-NN : Siamese -SVMSMOTE-Neural Net

SSTK-NN : Siamese -SMOTETomek-Neural Net
SSVM-AB : Siamese -SVMSMOTE-AdaBoost
SSTK-AB : Siamese -SMOTETomek-AdaBoost
SSVM-NB : Siamese -SVMSMOTE-Naive Bayes
SSTK-NB : Siamese -SMOTETomek-Naive Bayes
SSVM-QDA : Siamese -SVMSMOTE-QDA
SSTK-QDA : Siamese -SMOTETomek-QDA
SSVM-LDA : Siamese -SVMSMOTE-LDA
SSTK-LDA : Siamese -SMOTETomek-LDA
SSVM-EEC : Siamese -SVMSMOTE-EEC
SSTK-EEC : Siamese -SMOTETomek-EEC
SSVM-RSC : Siamese -SVMSMOTE-RSC
SSTK-RSC : Siamese -SMOTETomek-RSC
SSVM-BBC : Siamese -SVMSMOTE-BBC
SSTK-BBC : Siamese -SMOTETomek-BBC
SSVM-BRFC : Siamese -SVMSMOTE-BRFC
SSTK-BRFC : Siamese -SMOTETomek-BRFC
SSN-KNN : Siamese -SMOTEN-Nearest Neighbors
SSN-SVM : Siamese -SMOTEN-Linear SVM
SSN-DT : Siamese -SMOTEN-Decision Tree
SSN-RF : Siamese -SMOTEN-Random Forest
SSN-NN : Siamese -SMOTEN-Neural Net
SSN-AB : Siamese -SMOTEN-AdaBoost
SSN-NB : Siamese -SMOTEN-Naive Bayes
SSN-QDA : Siamese -SMOTEN-QDA
SSN-LDA : Siamese -SMOTEN-LDA
SSN-EEC : Siamese -SMOTEN-EEC
SSN-RSC : Siamese -SMOTEN-RSC
SSN-BBC : Siamese -SMOTEN-BBC
SSN-BRFC : Siamese -SMOTEN-BRFC

ÖZET

DOKTORA TEZİ

İlaç – Hedef Etkileşiminde (DTI) proteinlerin Etkili Özelliklerinin Belirlenmesinde Yapay Zeka Algoritmaları Tabanlı Bir Çerçeve Geliştirme

Sara NAGHİB ZADEH

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Prof. Dr. Zümrüt ECEVİT SATI

II. Danışman : Doç. Dr. Ali GHANBARİ

Bu çalışma, ilaç-hedef etkileşimi (İHE) (Drug Target Interaction (DTI)) tahmininin iyileştirilmesine odaklanmaktadır. Bu alan, ilaç keşfi sürecinde temel bir öneme sahiptir. Geleneksel deneysel yöntemlerin karmaşık ve yüksek maliyetli doğası ile biyolojik verilerin hızla artan hacmi dikkate alındığında, makine öğrenmesi ve derin öğrenme yaklaşımlarının kullanımı kaçınılmaz hale gelmiştir. Bu bağlamda, bu çalışmada ilaç-hedef etkileşimlerinin tahmin doğruluğunu artırmayı amaçlayan çok aşamalı bir yaklaşımı içeren bütünleşik bir çerçeve önerilmektedir.

Bu çalışma kapsamında önerilen yöntemsel çerçeve, ham verilerin toplanması, ilaç ve protein verilerinden anlamlı özelliklerin çıkarılması, özellik seçimi ve boyut indirgeme, veri dengesinin sağlanması ve makine öğrenmesi algoritmalarının uygulanması aşamalarını içermektedir. Çalışmada dört temel protein sınıfı (Enzim (Enzyme (EN)) , iyon kanalları (Ion Channel (IC)) , G proteinine bağlı reseptörler (G Protein-Coupled Receptor (GPCR)) , Nükleer Reseptör (

Nuclear Receptor (NR))) incelenmiştir. Veriler güvenilir biyoinformatik veri tabanlarından elde edilmiş; ilaç özellikleri FP2 moleküler parmak izi [1]ile, protein özellikleri ise (Amino Acid Composition (AAC)), Dipeptit Kompozisyonu (Dipeptide Composition (DPC)) ve Pseudo-Konum-Özel Skorlama Matrisi (Pseudo Position-Specific Scoring Matrix (PsePSSM)) gibi istatistiksel tanımlayıcılar aracılığıyla temsil edilmiştir. Bu özelliklerin bütünleştirilmesi sonucunda, ilaç ve hedef protein arasındaki biyokimyasal ilişkileri yansıtan birleşik özellik vektörleri oluşturulmuştur.

Boyut indirgeme ve etkili özellik seçimi amacıyla Temel Bileşenler Analizi (TBA) (Principal Component Analysis (PCA)) yöntemi uygulanmıştır. Bu yöntem, verideki en yüksek varyansı koruyarak özellik uzayını 100 boyuta indirgemiş ve böylece modelin öğrenme hızı ile genel verimliliğini anlamlı ölçüde artırmıştır.

İlaç-hedef etkileşimlerinin tahmininde karşılaşılan temel zorluklardan biri veri dengesizliğidir; bilinen etkileşimlerin sayısı, bilinmeyen etkileşimlere kıyasla oldukça sınırlıdır. Bu sorunun üstesinden gelmek amacıyla, bu tezde Siyam Sinir Ağı (SSA) (Siamese Neural Network (SNN)) ve SMOTE-Tomek yöntemlerini birleştiren hibrit bir yaklaşım önerilmiştir. Önerilen bu kombinasyon, verilerin istatistiksel dengesini sağlarken aynı zamanda örnekler arasındaki karmaşık benzerlik örüntülerinin etkin bir biçimde öğrenilmesine olanak tanımaktadır. Böylece geliştirilen model, baskın sınıfa yönelik yanlılığı azaltarak gerçek ilaç-hedef etkileşimlerini yüksek doğrulukla ayırt edebilmiştir.

Ön işleme ve veri dengeleme aşamalarının ardından, nihai sınıflandırıcı olarak Destek Vektör Makineleri (Support Vector Machine (SVM)) algoritması seçilmiştir. Bu seçim, Random Forest, Decision Tree ve sinir ağı tabanlı alternatif modellerle yapılan kapsamlı karşılaştırmalar sonucunda gerçekleştirilmiştir.

Önerilen yaklaşımın performansı, doğruluk (Accuracy), kesinlik (Precision), duyarlılık (Recall), F1 skoru ve ROC-AUC gibi farklı değerlendirme ölçütleriyle analiz edilmiştir. Elde edilen bulgular, önerilen akıllı çerçevenin tüm veri kümelerinde yüksek doğrulukla çalıştığını ve bazı kombinasyonlarda %100'e yakın sonuçlar verdiğini ortaya koymuştur. En başarılı sonuçlar, NR ve IC veri kümelerinde, PSEAAC-PSEPSSM ve GAAC-DDE özellik kombinasyonlarının Destek Vektör Makineleri ve Rastgele Orman algoritmalarıyla birlikte

[1] FP2 moleküler parmak izi, moleküllerin yapısını ikili kodlarla temsil eden bir yöntemdir. Bu sayede moleküller hızlıca karşılaştırılıp benzerlikleri analiz edilir.

kullanılmasıyla elde edilmiştir. Ayrıca, GPCR ve EN veri kümeleri de yüksek F1 ve AUC değerleriyle güçlü genellenebilirlik ve kararlılık sergilemiştir.

Ağustos 2025, 242 sayfa.

Anahtar kelimeler: İlaç-hedef etkileşimi, Siyam sinir ağı, SMOTE-Tomek tekniği, Veri dengeleme teknikleri, Protein özellik çıkarımı

SUMMARY

Ph.D. THESIS

Development of a Framework Based on Artificial Intelligence Algorithms to Determine the Effective Properties of Proteins in the Drug-Target Interaction (DTI)

Sara NAGHIB ZADEH

İstanbul University

Institute of Graduate Studies in Sciences

Department of Informatics

Supervisor : Prof. Dr. Zümrüt ECEVİT SATI

Co-Supervisor : Assoc. Prof. Dr. Ali GHANBARİ

This thesis study focuses on improving the prediction of drug–target interactions (DTI), a field that holds fundamental importance in the drug discovery process. Given the complexity and high cost of traditional experimental methods, along with the rapidly growing volume of biological data, the application of machine learning and deep learning approaches has become indispensable. In this context, an integrated multi-stage framework is proposed to enhance the accuracy of drug–target interaction predictions.

The proposed methodological framework includes several key stages: data collection, extraction of meaningful features from drug and protein data, feature selection and dimensionality reduction, data balancing, and the application of machine learning algorithms. Four major classes of target proteins (Enzymes (EN), Ion Channels (IC), and G Protein-Coupled Receptors (GPCR), Nuclear Receptor (NR)) were analyzed. Data were obtained from reliable

bioinformatics databases. Drug features were represented using FP2 molecular fingerprints, while protein features were described through statistical descriptors such as Amino Acid Composition (AAC), Dipeptide Composition (DPC), and Pseudo Position-Specific Scoring Matrix (PsePSSM). The integration of these features produced unified feature vectors that effectively capture the biochemical relationships between drugs and their target proteins.

To achieve dimensionality reduction and efficient feature selection, Principal Component Analysis (PCA) was applied. This method reduced the feature space to 100 dimensions while preserving the highest variance in the data, thereby improving both the learning speed and overall efficiency of the model.

One of the main challenges in DTI prediction is data imbalance, as the number of known interactions is significantly smaller than that of unknown ones. To address this issue, a hybrid approach combining Siamese Neural Networks (SNN) with the SMOTE-Tomek technique was developed. This combination not only ensured statistical balance in the dataset but also enabled the model to effectively learn complex similarity patterns among samples. Consequently, the proposed model successfully reduced bias toward the dominant class and accurately distinguished true drug–target interactions.

After preprocessing and data balancing, the Support Vector Machine (SVM) algorithm was selected as the final classifier, following extensive comparisons with alternative models, including Random Forest, Decision Tree, and neural network–based methods.

The performance of the proposed approach was evaluated using multiple metrics, including Accuracy, Precision, Recall, F1-score, and ROC-AUC. The results demonstrated that the proposed intelligent framework achieved high accuracy across all datasets, reaching near-perfect performance in certain combinations. The best outcomes were obtained for the NR and IC datasets when the PseAAC–PsePSSM and GAAC–DDE feature combinations were used in conjunction with Support Vector Machines and Random Forest algorithms. Furthermore, the GPCR and EN datasets also exhibited high F1 and AUC values, indicating strong generalization ability and stability of the proposed method.

August 2025, .242 pages.

Keywords: Drug target interaction, Siamese neural network, SMOTE-Tomek technique, Data balancing techniques, Protein feature extraction

1. GİRİŞ

Tıp tarihi, daima hastalıklar için etkili tedavilerin bulunması ve insan yaşam kalitesinin yükseltilmesi çabalarıyla iç içe olmuştur. İlk deneysel bitkisel ilaçlardan ileri teknolojilere dayanan modern ilaç ürünlere kadar, ilaç keşfi yaşam bilimlerinin en zorlu ve en hayati alanlarından biri olarak öne çıkmaktadır. Biyoteknolojideki ilerlemeler, biyolojik verilerin olağanüstü bir hızla çoğalması ve moleküler yapıların daha derinlemesine anlaşılması sayesinde, geleneksel yaklaşımlar giderek hesaplamalı yöntemlere ve veri madenciliğine evrilmiştir(Patel ve diğ., 2023). Bu bağlamda, ilaç tasarımı ve geliştirme süreçlerinin temel aşamalarından biri olan ilaç-hedef etkileşimlerinin (DTI) belirlenmesi ve tahmini, günümüzde artan bir ilgi ve önem kazanmaktadır (Lambrinidis ve Tsantili-Kakoulidou, 2021).

Geçmişte ilaç keşfi, büyük ölçüde doğal ya da kimyasal bileşiklerin taranmasına dayanan deneysel yöntemlerle yürütülmekteydi. Bu süreçler, antibiyotikler, kanser ilaçları ve antiviral tedavilerin geliştirilmesinde önemli başarılar sağlamış olsa da, yüksek maliyet, uzun hazırlık süreleri ve yüksek oranda başarısızlık gibi ciddi sınırlamalarla karşı karşıya kalmıştır. Nitekim ilaç adaylarının yalnızca çok küçük bir yüzdesi klinik deneylerin son aşamalarına ulaşarak onay alabilmektedir. Ayrıca, insan biyolojik tepkilerinin karmaşıklığı, bireysel genetik farklılıklar ve beklenmedik yan etkilerin ortaya çıkması, ilaç geliştirme sürecini son derece güç, riskli ve öngörülemez bir hale getirmiştir (Rifaioglu ve diğ., 2019).

İlaç-hedef etkileşimi, bir ilaç bileşiğinin vücuttaki biyolojik bir hedefe bağlanarak, bu hedefin işlevinde bir değişikliğe yol açmasını ifade eder. Biyolojik hedefler genellikle biyolojik yolların düzenlenmesinde ve hücre sinyal iletiminde kilit rol oynayan proteinler, enzimler, membran reseptörleri veya nükleik asitler gibi yapılardan oluşur. İlaç, bu yapılara bağlanarak istenen tedavi edici etkileri ortaya çıkarabilir veya zararlı reaksiyonların meydana gelmesini engelleyebilir (Ezzat ve diğ., 2019).

Bu zorluklara yanıt olarak, hesaplamalı yaklaşımlar son yıllarda önemli ölçüde ilerleme kaydetmiştir. Makine öğrenimi algoritmaları, sinir ağları, benzerlik tabanlı yöntemler ve özellik analizi tekniklerinin kullanımı, araştırmacıların ilaç-hedef etkileşimlerini daha doğru ve hızlı bir şekilde tahmin etmelerine olanak tanımaktadır. Bu yöntemler yalnızca laboratuvar maliyetlerini azaltmakla kalmayıp, aynı zamanda potansiyel olarak etkili bileşikleri erken

aşamada filtreleyerek arama kapsamını büyük ölçüde daraltabilmektedir (Mohamed ve diğ., 2020).

Hesaplamalı yaklaşımlar üç temel kategoriye ayrılır: (1) ilaçların yapısal benzerliğinden yararlanarak etkileşimleri tahmin eden ligand tabanlı yöntemler, (2) ilaç-hedef etkileşimini üç boyutlu olarak simüle eden moleküler yerleştirme (docking) tabanlı yöntemler ve (3) öğrenme sürecinde hem ilaç hem de hedefe ait birleşik bilgileri kullanan kimyagenomik yaklaşımlar. Bu yöntemlerin her biri, doğruluklarını ve genellenebilirliklerini artırmak amacıyla destek vektör makineleri (SVM), rastgele ormanlar (RF), sinir ağları (NN) ya da derin öğrenme gibi çeşitli makine öğrenimi algoritmalarıyla entegre edilebilir (Sachdev ve Gupta, 2019).

Bu ilerlemelere rağmen, bazı sorunlar hâlâ varlığını sürdürmektedir. Bu zorluklar arasında, pozitif ve negatif veriler (etkileşimli ve etkileşimsiz sınıflar) arasında ciddi bir dengesizlik, biyolojik verilerdeki gürültü, güvenilir laboratuvar verilerine erişimdeki kısıtlılık ve büyük veri kümelerini analiz ederken ortaya çıkan yüksek hesaplama maliyeti yer almaktadır. Ayrıca, protein yapılarının karmaşıklığı ile ilaçların kimyasal bileşimlerindeki çeşitlilik, makine öğrenimi modellerini eğitmek için doğru ve etkili öznitelikler elde etmeyi önemli ölçüde güçleştirmektedir (Pirgazi ve diğ., 2018).

Bu nedenle, bu araştırmanın amacı; ileri düzey kimyasal ve biyolojik özellik çıkarımı, akıllı veri dengeleme teknikleri ve optimal sınıflandırma yöntemlerinin birleşimini kullanarak, ilaç-protein etkileşimlerini yüksek doğrulukla tahmin edebilen bir hesaplamalı yöntem geliştirmektir. Bu yaklaşımda, öncelikle protein ve ilaç verilerinden istatistiksel özellikler çıkarılarak çeşitli veri kümeleri oluşturulmaktadır. Dengesiz veri kümesi, ilk aşamada bir Siyam sinir ağı yardımıyla işlenmekte; ardından SMOTE-Tomek yöntemiyle dengelenmekte ve eğitime hazır hâle getirilmektedir. Bir sonraki adımda, veriler Doğrusal Destek Vektör Makinesi (Linear SVM) modeliyle eğitilmektedir. Önerilen bu çerçeve, verilerin biyolojik boyutlarını ve istatistiksel öğrenmenin zorluklarını birlikte ele almayı hedeflemektedir.

Bu araştırmanın temel amacı, mevcut verilerden ve gelişmiş hesaplama araçlarından yararlanarak zaman alıcı ve maliyetli laboratuvar süreçlerini daha hızlı ve hedef odaklı bir biçimde yürütmektir. Ayrıca, ilaçlar ile protein hedefleri arasındaki yeni etkileşimleri yüksek doğrulukla tahmin edebilen bir modelin geliştirilmesi, yeni ilaçların keşfini hızlandırabilir,

mevcut ilaçların yeniden değerlendirilmesine olanak tanıyabilir ve nihayetinde ilaç tedavilerinin etkinliğini önemli ölçüde artırabilir.

Bu tez çalışması bu bölüm dahil toplamda 5 bölümden oluşmaktadır. Bu bölümden sonraki genel kısımlar bölümünde, tezin kavramsal altyapısını oluşturmak amacıyla İlaç-Protein Etkileşimi, Etkileşim Tahmininde Kullanılan Geleneksel ve Modern Yaklaşımlar ile Karşılaşılan Temel Zorluklar konularına ve alt başlıklarına değinilmiştir. Malzeme ve yöntem bölümünde, tez kapsamında geliştirilen çerçeve ayrıntılı olarak açıklanmış; Özellik Çıkarımı Süreçleri, Veri Dengeleme Teknikleri ve Makine Öğrenimi Modellerinin Tasarımı ele alınmıştır. Yapılan deneyler ve gerçekleştirilen istatistiksel analizler sonucu elde edilen bulgular detaylı olarak bulgular bölümünde sunulmuştur. Son olarak, elde edilen bulguların değerlendirilmesi, mevcut literatürle karşılaştırılması ve gelecekte yapılabilecek çalışmalara dair öneriler tartışma ve sonuç bölümünde kapsamlı şekilde paylaşılmıştır.

1.1. ÇALIŞMANIN AMACI

Bu tezin temel amacı, ilaç ve protein arasındaki etkileşimin tahminine ilişkin zorlukların giderilmesidir. Bu çalışmada özellikle dikkat edilen ilk konu, mevcut önerilen yöntemlerde etkili özellik çıkarımına yeterince önem verilmemiş olmasıdır. Oysa bu durum, modelin doğruluğunu ve çıktıların kalitesini önemli ölçüde etkilemekte ve nihayetinde elde edilen sonuçların kalitesinin artmasına katkı sağlamaktadır.

Sınıf dengesizliği, bu alanda hâlihazırda var olan bir diğer önemli sorundur. Şöyle ki, doğru bir şekilde pozitif etkileşim olarak tanımlanmış örneklerin sayısından çok daha fazla sayıda bilinmeyen etkileşim bulunmaktadır. Sonuç olarak, bu iki sınıf arasındaki dengesizlik, ele alınması ve çözülmesi gereken temel bir zorluk teşkil etmektedir.

Genel olarak, bu tezin hedefleri şu şekilde dört kategoriye ayrılmaktadır:

- İlaç-protein etkileşimlerinin ilaç geliştirilmeden önce tahmin edilmesinin gerekliliğini göstermek
- Farklı ve çeşitli özelliklerin çıkarılmasıyla ayırt edilebilirlik ve tanıma gücünü artırmak
- Dengesiz sınıf dağılımının yol açtığı tahmin hatalarını azaltmak amacıyla, pozitif ve negatif örnekler arasındaki veri dengesizliğini gidermek ve sınıflar arası öğrenme sürecini optimize etmek

- İlaç-protein etkileşimlerinin tahmin hızını ve doğruluğunu artırarak ilaç geliştirme sürecini hızlandırmak ve iyileştirmek

1.2. ARAŞTIRMA HEDEFLERİNİN GERÇEKLEŞTİRİLMESİNE YÖNELİK ÖNERİLEN ADIMLAR

Bu tezde 1-5 bölümünde belirtilen hedeflere dayanarak, her bir hedefe ulaşmak için gerekli adımlar bu bölümde açıklanmıştır.

- "İlaç-protein etkileşimlerinin ilaç geliştirilmeden önce tahmin edilmesinin gerekliliği" hedefine ulaşmak için yaptığımız işlem, doğru bir şekilde pozitif olarak tanımlanan etkileşimlerin bulunmasıdır. Bu işlem sayesinde, özellikle COVID-19 pandemisi gibi durumlarda, hastalıkların tedavisinde kullanılacak ilaçların geliştirilme hızı artırılmaktadır.
- "Farklı ve çeşitli özelliklerin çıkarılmasıyla ayırt edilebilirlik ve tanıma gücünün artırılması" hedefine ulaşmak için yaptığımız işlem, ilaçların (parmak izi kullanılarak) ve proteinlerin (amino asit dizileri kullanılarak) çeşitli özelliklerinin çıkarılmasıdır. Bu işlemle elde edilen sonuç, ilaç ve protein etkileşimlerinde ayırt edici özelliklerin tanımlanması ve ilaçların ve proteinlerin özgün özelliklerinin ortaya konmasıdır.
- "İlaç ve hedef etkileşimleri için dengeli bir veri seti oluşturma" hedefine ulaşmak için yaptığımız işlem,derin öğrenme tabanlı Siamese ağı ve over-sampling ile under-sampling yöntemlerinin birleşik bir yöntemi geliştirilmiştir. Bu birleşik yöntem, Siamese algoritması ve over-sampling ile under-sampling tekniklerinin avantajlarını birleştirir. Siamese algoritması, ilaç ve protein etkileşimlerini doğru şekilde tespit ederken, veri dengeleme yöntemleri önemli bilgilerin korunmasına ve sınıf önyargılarının azaltılmasına yardımcı olur. Sonuç olarak, veri setinin verimliliği artar, bilgi kaybı azalır ve tahmin doğruluğu yükselir. Bu yöntem, ilaç-protein etkileşimlerinin simülasyonunu iyileştirerek ilaç keşif sürecini hızlandırır.

1.3. ÇALIŞMANIN VARSAYIMLARI

Önerilen çalışma kapsamında, araştırmanın sürdürülebilirliği için bazı varsayımlar gereklidir. Bunlar şunlardır:

- Uygun özelliklerin çıkarılması, önerilen modelin verimliliğini artırabilir.

- Çok modlu (multimodal) bir modelin sunulması, ilaç ve protein etkileşimlerinin tanınmasını artırabilir.
- Veri dengeleme, modelin performansını artırır: Bu varsayım, verilerdeki pozitif ve negatif örneklerin sayısı dengelendiğinde, modelin ilaç-protein etkileşimlerini doğru bir şekilde öğrenme ve tahmin etme yeteneğinin daha iyi olacağına inanır.
- Siamese sinir ağı, örnekler arasındaki ilişkileri simüle etmek için uygundur: Siamese sinir ağının, mimarisi kullanılarak, pozitif ve negatif örnekler arasındaki benzerlik ve farklılık ilişkilerini doğru bir şekilde modelleyebileceği ve böylece daha doğru tahminler sunabileceği varsayılmaktadır.
- Farklı özelliklerin birleştirilmesi, tahmin doğruluğunu artırmaya yardımcı olur: Bu varsayımda, çeşitli özelliklerin kullanılması ve bunların ilaçlar ve proteinlerden çıkarılması, modelin daha karmaşık etkileşimleri öğrenmesini sağlar ve bu da daha doğru tahminler sunabilir.

2. GENEL KISIMLAR

2.1 GİRİŞ

Tüm canlı organizmaların bedenleri, biyolojik işlevlerin yürütülmesinde kritik rol oynayan protein moleküllerini içeren hücrelerden oluşur. Vücutta meydana gelen her olumlu ya da olumsuz tepki, bu proteinlerin işlevleri sonucunda ortaya çıkar. İlaçlar, bu proteinlere bağlanarak onların faaliyetlerini ya etkinleştirir ya da baskılar; böylece biyolojik süreçler üzerinde doğrudan etkili olabilir. Her ilaç, özgül olarak belirli bir protein grubunu hedef alır (Wanat, 2020).

Günümüzde ilaçlar, diğer temel ihtiyaçlarımızla birlikte vazgeçilmez bir gereklilik hâline gelmiştir. Hastalıkların tedavisi ve iyileştirilmesinde ilaçlar önemli bir rol oynamaktadır. Bireyler, hastalıkları teşhis etmek, tedavi etmek veya önlemek amacıyla ilaç kullanmaktadır. Ayrıca, bir kaza ya da cerrahi müdahale sonrasında da iyileşme sürecini desteklemek için ilaçlardan faydalanılmaktadır.İlaçlar farklı şekillerde üretilmekte ve çeşitli yöntemlerle vücuda alınmaktadır. Antibiyotikler, ilaçlara verilebilecek en iyi örneklerden biridir ve özellikle geçtiğimiz yüzyılda insan ömrünün uzamasına kayda değer bir katkı sağlamıştır.İnsan sağlığını tehdit eden on binlerce hastalık tanımlanmış olup, her yıl bu listeye yenileri eklenmektedir. Bu hastalıklar arasında yeni ortaya çıkanlar (örneğin COVID-19) ile uzunyıllardır toplumda yaygın olarak görülen ve henüz tedavisi bulunamayan Parkinson ve Alzheimer gibi hastalıklar yer almaktadır (Xu ve diğ., 2019) .Bu hastalıkları etkili bir biçimde tedavi edebilecek ilaçların hızlı ve doğru bir şekilde keşfedilmesi, toplum sağlığının geliştirilmesi açısından büyük önem taşımaktadır. Ancak laboratuvar ortamında yürütülen ilaç keşfi ve klinik öncesi çalışmaların genel başarı oranı oldukça düşüktür; bu oran yaklaşık olarak %0,05 ila %0,1 arasında değişmektedir. Üstelik aday bileşiklerin yalnızca %1'inden daha azı beklenen etkiyi göstererek klinik deneme aşamasına geçebilmektedir.Bu bağlamda, ilaç-hedef (drug-target) etkileşimlerinin araştırılması, ilaç keşif sürecinde kritik bir aşamayı oluşturmaktadır ve yeni ilaçların başarıyla geliştirilme olasılığını artırma potansiyeline sahiptir (Chen ve diğ., 2019).

İlaç keşfi, yeni bir ilaç adayının yararlı tedavi edici etkilerinin veya tıbbi uygulamalarının belirlendiği süreçtir. İlaç keşfi ve geliştirmesi, hedef belirleme ve doğrulamanın ilk aşamasından, bileşik ipuçlarının tanımlanmasından, doğrulama ve optimizasyondan ve Gıda ve İlaç İdaresi (Food and Drug Administration (FDA)) tarafından nihai onaya kadar çeşitli klinik

öncesi ve klinik deney türlerinden geçer (Paul ve diğ., 2010). İlaçlar farklı moleküler hedeflerle etkileşerek etki gösterirler. Bu tür etkileşimlere ilaç-hedef etkileşimleri (DTI'ler) denir. Proteinler bu hedeflerin yararlı bir grubudur. Bağlanma yoluyla ilaçlar, proteinler tarafından gerçekleştirilen işlevleri artırabilir veya kontrol edebilir ve böylece hastalık koşullarını etkileyebilir (Santos ve diğ., 2017).

İlaç-protein etkileşimlerinin belirlenmesi, ilaç keşfi sürecinin temel ve kritik aşamalarından biridir. Bu etkileşimler, kimyasal bileşikler ile insan vücudundaki hedef proteinler arasındaki fiziksel temasın saptanmasını ifade etmektedir. İlaç keşfi, biyoaktif bileşiklerin biyolojik rollerinin tanımlanmasıyla başlayan ve bu bileşiklerin potansiyel tedavi edici etkilerine yönelik yeni ilaç adaylarının geliştirilmesiyle devam eden çok aşamalı bir süreçtir. Geleneksel ilaç araştırma ve geliştirme süreçleri, genellikle belirli bir hastalığa ya da tedavi hedeflerine yönelik biyomoleküler yapıların tanımlanmasıyla başlamakta; ardından, bu hedefler doğrultusunda yüksek kapasiteli sanal tarama yöntemleriyle uygun bileşiklerin seçilmesiyle sürdürülmektedir. Bu bağlamda, belirlenen hedef moleküller için uygun biyoaktif bileşikler ve bunların biyolojik aktivite düzeyleri sistematik olarak analiz edilmektedir. Yüksek kapasiteli tarama tekniklerindeki teknolojik gelişmeler sayesinde, günümüzde binlerce farklı bileşiğin kısa sürede taranması ve hedef proteinler üzerindeki etkilerinin değerlendirilmesi mümkün hâle gelmiştir. Bu gelişmeler, ilaç adaylarının daha hızlı ve etkin biçimde belirlenmesine olanak sağlamaktadır (Hertzberg ve Pope, 2000) .

2.2 BİYOLOJİK KAVRAMLAR

Belirtilen zorluklarla başa çıkma yöntemlerini sunmadan önce, konunun daha iyi anlaşılması için ilgili bileşenleri tanıtacağız ve bu alanda araştırmacıların benimsediği yaklaşımları kısaca gözden geçireceğiz. Daha sonra, konunun daha da netleşmesiyle birlikte, ilaç-protein etkileşimlerini tahmin etme zorluklarını çözmek için önerilen yöntemleri inceleyeceğiz.

2.2.1 Gen

Gen, genetik bilgiyi ebeveynlerden bir sonraki nesle aktaran ve çocuklarda ebeveynlerine benzer özelliklerin ortaya çıkmasına neden olan DNA molekülünün bir parçasıdır. Genler, DNA'dan oluşan kalıtımın temel fiziksel ve işlevsel birimleridir. Bazı genler, protein adı verilen moleküllerin üretiminde bir kılavuz görevi görürken, birçok gen ise herhangi bir proteini

kodlamaz. İnsan genomunda yaklaşık 20.000 ila 25.000 arasında gen bulunduğu tahmin edilmektedir. Her birey, her bir genin iki kopyasına sahiptir; bunlardan biri anneden, diğeri ise babadan kalıtım yoluyla alınır. Genlerin büyük çoğunluğu tüm insanlarda benzer yapıdadır; ancak, az sayıda gen bireyler arasında küçük farklılıklar gösterebilir. Bu küçük genetik farklılıklar, her insanın kendine özgü fiziksel özelliklere sahip olmasına katkı sağlar (Gen ve Cheng, 2000).

Genler, RNA üretimi ve sentezi için gerekli bilgileri hücrenin diğer mekanizmalarına aktaran genetik kodları taşır (Papantonis ve Cook, 2013) .Bu süreçte, genler öncelikle hücresel enzimler aracılığıyla transkribe edilir ve ardından işlevsel ürünlerin sentezine olanak sağlayan çeviriler gerçekleştirilir.Genlerin tanımlanabilmesi için her biri belirli bir sistem dâhilinde adlandırılır. Ancak gen adlarının genellikle uzun olması nedeniyle, bu genlere kısa harf kombinasyonlarından ve bazen sayılardan oluşan semboller atanmıştır. Bu semboller, gen adlarının kısaltılmış versiyonlarını temsil eder (Grimplet ve diğ., 2014).

Genler, temel işlevlerini kaybetmeden ve çoğalma sürecine girmeden ek bir işlev de üstlenebilirler. Bu olgu, gen paylaşımı olarak adlandırılır (Magadum ve diğ., 2013). Gen paylaşımı, bir proteinin birden fazla biyolojik işlev üstlenebileceği anlamına gelir ve bu durum, farklı görevlerde kullanılabilen proteinlerin oluşumunu mümkün kılar. Söz konusu proteinler arasında enzimler, hücresel reseptörler ve iyon kanalları yer alır.Gen ekspresyonu ise şu adımlarla gerçekleşir: Öncelikle, genin transkripsiyonu ile RNA zincirleri sentezlenir. Ardından bu RNA'lar, çeviri (translasyon) süreciyle amino asit zincirlerine dönüştürülür. Bu zincirler, katlanma (folding) aşamasından geçerek proteinlerin üç boyutlu yapısını ve işlevsel özelliklerini kazanır (Thangamani ve Kalaiselvi, 2017).

2.2.2 Amino Asit

Amino asitler, amino ve karboksilik asit gruplarını içeren organik moleküllerdir ve proteinlerin yapı taşları olarak kabul edilirler (Belitz Grosch, 1999). Bu bileşikler, mRNA'nın yönlendirdiği çeviri sürecinde birbirlerine bağlanarak protein yapılarını oluştururlar. Her bir amino asit, kendine özgü özelliklerini belirleyen özel bir yan zincire (R grubu) sahiptir. Amino asitlerin önemi, protein sentezindeki temel rolleri sayesinde, canlı organizmaların hayati işlevlerini ve anatomik ile fizyolojik özelliklerini düzenlemelerinden kaynaklanmaktadır. Doğada 100'den fazla amino asit bulunmasına rağmen, insan vücudu normal işlevlerini sürdürebilmek için

yalnızca 20 standart amino aside ihtiyaç duyar. Bu standart amino asitlerin yaklaşık yarısı vücut tarafından sentezlenemediğinden, dışarıdan beslenme yoluyla alınmaları gerekir ve bu nedenle "esansiyel amino asitler" olarak adlandırılırlar (Hoffer, 2016).

2.2.3 Protein

Proteinler, yaşamın temel ve vazgeçilmez yapı taşlarıdır. Bu moleküller, sahip oldukları çeşitli işlevler sayesinde organizmadaki pek çok biyolojik süreçte anahtar rol oynarlar. Kimyasal reaksiyonların katalizlenmesinden madde taşınmasına, sinyal iletiminden kas hareketine ve bağışıklık savunmasından yabancı maddelere karşı korunmaya kadar çok sayıda işlevi yerine getirirler (Chatterjee, 2023).

Proteinlerin incelenmesi, hem bilimsel hem de endüstriyel alanlarda büyük bir öneme sahiptir. Bu moleküller, fizyolojik ve patolojik süreçlerdeki hayati rolleri nedeniyle, işlevlerinin istenilen şekilde değiştirilmesi, ilaç tasarımında önemli bir hedef oluşturur . İlaç tasarımının temel kavramlarından biri, hastalıkla ilişkili proteinlerin, küçük moleküller kullanılarak hedeflenmesi ve engellenmesidir; bu yaklaşım, tedavi edici etkilerin elde edilmesini sağlar. Ayrıca, proteinler biyoteknolojik ilaçlar veya biyolojik ilaçlar olarak da kullanılır; bu tür ilaçların en bilinen örneği ise antikorlardır (Hay ve diğ., 2014).

Biyoteknoloji alanında ise proteinler, sağladıkları özel işlevler ve çeşitlilikleri ile önemli bir yer tutar. Özellikle, kimyasal reaksiyonları hızlandıran ve yönlendiren enzimler, biyoteknolojik süreçlerde biyokatalizör olarak yaygın biçimde kullanılmaktadır . Bunun yanı sıra, proteinlerin özellikleri endüstriyel uygulamalara özel olarak uyarlanabilir ve optimize edilebilir . Bu doğrultuda, protein mühendisliği teknikleri kullanılarak enzimsel aktivitenin artırılması, daha geniş bir alt tabaka yelpazesi elde edilmesi ve proteinlerin daha yüksek kararlılıkla çalışması sağlanmaktadır (Li ve diğ., 2020).

Proteinler, bir veya daha fazla amino asit zincirinden oluşan büyük biyomoleküllerdir ve vücutta hayati işlevler üstlenirler.İlaç hedefi (Ayrıca hedef, hedef protein, protein, hedef gen veya gen olarak da adlandırılır.), ilaçlarla etkileşime girebilen moleküler bir yapı olarak kabul edilir (Imming ve diğ., 2006) . Proteinler, bir veya daha fazla amino asit zincirinden oluşan ve vücutta temel işlevleri yerine getiren büyük biyomoleküllerdir.proteinler biyomakromoleküllerdir ve enzimatik, yapısal ve hormonal roller gibi temel işlevleri yerine getirirler. Proteinler, 20 farklı amino asidin farklı dizilimlerde bir araya gelmesiyle oluşan

doğrusal polimerlerdir. Tüm amino asitler, bir alfa karbon atomu, bir amin grubu (-NH_2), bir karboksil grubu (-COOH) ve her bir amino aside özgü değişken bir yan zincir (R grubu) içeren ortak bir yapıya sahiptir (Dewangan ve diğ., 2023)

Bir polipeptit zincirindeki amino asitler, peptit bağlarıyla birbirine bağlanır. Bir amino asit, protein zincirine katıldığında, artık "kalıntı" (rezidü) olarak adlandırılır. Birbirine bağlanan karbon, azot ve oksijen atomları ise proteinin ana zinciri veya omurgası olarak bilinir.Son yıllarda, kriyo-elektron mikroskobu (Cryo-EM) kullanımı ve küçük protein bölgeleri için protein yapısının tahmini, atomik doğrulukla proteinlerin dizilimini belirlemede iki yöntem olarak öne çıkmaktadır (Fennema ve diğ., 2017).

Amino asit dizileri, protein tanımlayıcılarını tahmin etmek için kullanılabilir. Amino asitler, fizikokimyasal özelliklerine göre küçük molekül tanımlayıcılarına benzer şekilde tarif edilebilirler. Farmakoloji alanında biyolojik bir hedef (protein), bir organizmada ligandların veya ilaçların bağlandığı herhangi bir şey olarak tanımlanabilir. İnsan genomu 20.000'den fazla gen içerir ve bunların yaklaşık %80'i bir veya daha fazla proteini kodlayabilir. Sadece birkaç protein farmakolojik olarak aktif tanımlanmıştır ve şu anda onaylanmış ilaçların hedefidir (Cheng ve diğ., 2021). Şu anda hedef moleküllerin çoğu, dört protein ailesinin [enzimler, G protein-bağlı reseptörler (GPCR'ler), iyon kanalları ve nükleer reseptörler] oluşturduğu proteinlerdir. Hedef moleküllerin %44'ünü proteinler oluşturmaktadır ve halihazırda geliştirilen ilaçların %70'i bu dört protein ailesi ile etkileşimlerden oluşmaktadır (Mahmud ve diğ., 2020).Mevcut ilaçlar ve hedefler arasında yeni bağlantılar bulmak veya belirli bir ilaç için yeni hedefler bulmak, ilaç geliştirmede önemli bir rol oynar (Murase ve diğ., 2003).

2.2.4 İlaç Ve Kimyasal Yapısı

Tıp biliminde ilaç, hastalıkların tedavisi, semptomların hafifletilmesi, teşhisi veya önlenmesi amacıyla kullanılan ve vücuda girdikten sonra organizmanın yapısını veya işlevini iyileştiren her türlü maddeyi ifade eder (Volkova ve Volkov, 2023) . Bu maddeler, hücre içi, hücre dışı ya da hücre zarındaki reseptörler üzerinde etkili olarak belirli işlevlerin başlamasına ya da engellenmesine neden olabilir. İlacın etkisi, bu etkileşimlerin sayısı ve şiddetiyle doğrudan ilişkilidir. İlaçlar genellikle biyokimyasal veya fizyolojik etkileri olan kimyasal bileşiklerdir ve canlıların fiziksel ve ruhsal sağlığını geliştirmek amacıyla kullanılırlar (Alswedani ve diğ., 2023).

İlaçlar bitkisel, hayvansal ya da sentetik kökenli olabilir ve genellikle laboratuvar ortamında eczacılar veya hekimler tarafından keşfedilip gerekli araştırma ve onay süreçlerinin ardından ilaç fabrikalarında üretilir. İlacın kimyasal yapısı, çözünürlük, yüzey gerilimi, protein bağlanması, kompleks oluşumu ve eş-yerlilik gibi fiziko-kimyasal özelliklerini belirler ve bunun sonucunda emilim, dağılım, metabolizma, atılım ve toksisite gibi ilaç hareketbilimi süreçleri etkiler. İlacın kullanım amaçları arasında vücudun ihtiyaç duyduğu maddelerin temini, enfeksiyonların önlenmesi, zehirsizleştirme , bazı doğal işlevlerin geçici olarak engellenmesi ve bozulmuş işlevlerin düzeltilmesi yer alır (Wishart ve diğ., 2008).

2.3 İLAÇ GELİŞTİRME AŞAMALARI

İlaç geliştirme süreci genellikle üç ana aşamada gerçekleştirilir ve her biri özel bir öneme sahip olup titizlikle yürütülmelidir. Birinci aşama, yeni bir ilacın tanımlanması ve onun etkileşimde bulunduğu hedef proteinin belirlenmesini amaçlayan keşif aşamasıdır. Bu aşamada, araştırmacılar yeni bir ilacı tanımlamaya ve ilgili protein hedefini belirlemeye çalışırlar. Bu süreç, ilacın biyolojik ve kimyasal mekanizmalarını anlamaya yönelik kapsamlı ve karmaşık araştırmaları içermektedir . Ardından, geliştirme aşaması olarak adlandırılan ikinci aşamaya geçilir. Bu aşamada, ilacın insanlar üzerinde kullanılmadan önce güvenli olup olmadığı değerlendirilir. Bu amaçla, ilaçlar hayvan modelleri üzerinde test edilerek istenen tedavi edici etkileri gösterip göstermediği ve zehirlilik düzeyleri araştırılır (Suruliandi ve diğ., 2024)

Üçüncü aşama ise klinik insan denemeleri ve pazarlama sürecidir. Bu aşamada, ilaçlar ilk olarak küçük insan grupları üzerinde test edilir. Eğer olumlu sonuçlar elde edilirse, denemeler daha geniş insan gruplarına yayılır ve ilacın toplumda kullanımı başlatılır. Bu aşamanın başarıyla tamamlanmasının ardından ilaç pazara sunulmak üzere ruhsatlandırılır (Suruliandi ve diğ., 2024). Bu üç aşama birlikte, ilacın bilimsel açıdan geçerli, yeterli güvenlik profiline sahip ve etkili bir pazarlama stratejisiyle piyasaya sürülmesini garanti eder. Tahminlere göre, birinci faz klinik çalışmaların maliyeti yaklaşık 3,4 milyon dolar, ikinci faz 8,6 milyon dolar ve üçüncü faz ise yaklaşık 21,4 milyon dolardır (Martin ve diğ., 2017).

2.4. PROTEİN ÖZELLİKLERİNİN TAHMİNİ VE UYGULAMALARI

Bu çalışma, ilaç keşfi ve protein mühendisliği alanlarında kullanılan hesaplamalı yöntemleri derinlemesine incelemektedir. Bu çerçevede, belirli bir proteine ait yapısal, işlevsel, kimyasal, biyolojik, farmakolojik ve dinamik gibi çeşitli nitelikler, genel olarak “protein özellikleri”

kapsamında ele alınmaktadır. Çalışmada, yaşam bilimlerine dair mevcut zorlukların çözümüne katkı sunabilecek tüm öngörülen özellikler kapsamlı biçimde değerlendirilmiştir. Bu özellikler arasında protein fonksiyonu, potansiyel bağlanma bölgeleri, bağlanma ortakları, biyomoleküllerin bağlanma afinitesi, esneklik, hidrofobisite ve kararlılık gibi unsurlar yer almaktadır. Söz konusu özelliklerin doğru ve bütüncül bir şekilde anlaşılması; biyoloji ve kimya alanlarındaki bilgi birikiminin derinleşmesine, aynı zamanda biyoteknoloji ve eczacılık gibi uygulamalı bilimlerde kayda değer ilerlemelerin sağlanmasına olanak tanıyabilir (Poluri ve diğ., 2021).

Örneğin, proteinlerin işlevlerinin öngörülmesi, yaşamın moleküler düzeyde daha derinlemesine kavranmasına önemli ölçüde katkı sunabilir. Bir proteinin moleküler özellikleri, biyolojik rolleri ve etkinlikleri, biyolojik sistemlerin ve organizmaların modellenmesinde sağlam ve güvenilir bir temel oluşturur. İşlevsel özelliklerdeki çeşitlilik ise, hücre içi yerleşim yerlerinin belirlenmesinden, enzimlerin kataliz ettiği özgül kimyasal tepkimelerin tahmin edilmesine kadar uzanan geniş bir yelpazede çıkarımlara olanak tanır (Lee ve diğ., 2007).

İlaç endüstrisi ve bilimsel araştırmalarda, proteinlerin diğer moleküllerle olan etkileşimlerinin öngörülmesi, son derece kritik bir rol oynamaktadır. Yapıya dayalı ilaç keşfi süreçlerinde, tedaviye yönelik protein hedefleri için yeni ilaç adayları geliştirirken ilk adım, hedef proteinin yapısında yer alan uygun bağlanma bölgelerinin belirlenmesidir. Bu aşamayı takiben, söz konusu bağlanma bölgelerine yüksek etkinlikle bağlanabilecek küçük molekülleri tanımlamak amacıyla sanal tarama gibi hesaplamalı yöntemlerden yararlanılır. Bu tür tahminsel yaklaşımlar, ilaç keşfinin hemen her aşamasında ve farklı evrelerinde yaygın olarak kullanılmakta, sürecin etkinliğini ve başarısını önemli ölçüde artırmaktadır (Soleymani ve diğ., 2022).

2.5 DTI TAHMİN MODELLERİ

İlaç keşfi süreçlerinde, aday bileşiklerin hedef proteinlerle olan etkileşimlerini doğru bir şekilde öngörebilmek büyük önem taşımaktadır. Bu doğrultuda geliştirilen DTI tahmin modelleri, hem zaman hem de maliyet açısından oldukça avantajlı olup, deneysel doğrulamalara yön veren güçlü araçlar olarak öne çıkmaktadır. Farklı hesaplamalı yaklaşımlar, hedef proteinlerin ve ilaç adaylarının yapısal, kimyasal ve biyolojik özelliklerini dikkate alarak bu etkileşimleri tahmin

etmeye çalışır. Bu bölümde, literatürde sıkça kullanılan başlıca DTI tahmin modelleri ayrıntılı biçimde ele alınacak .

2.5.1 Kantitatif yapı-aktivite ilişkisi modeli (Quantitative Structure–Activity Relationship (QSAR))

QSAR modeli, bileşiklerin kimyasal yapıları ile biyolojik aktiviteleri arasındaki ilişkiyi nicel verilerle inceleyen sistematik bir yaklaşımdır. Bu model, araştırmacıların moleküler özellikleri analiz ederek ilaçların biyolojik etkilerini önceden tahmin etmelerine olanak tanır. QSAR modellemesi, biyolojik açıdan etkili bileşikleri belirlemek ve bu bileşiklerin etkinliğini deneysel yöntemlerle doğrulamak amacıyla geliştirilen çeşitli hesaplamalı yöntemleri kapsamaktadır. Modelin temel amacı, ilaç bileşiklerinin yapısal özelliklerinde değişiklikler yaparak etkinliği artırmak, yan etkileri en aza indirmek ve böylece daha güvenli ve etkili tedavi seçenekleri ortaya koymaktır (Sheridan ve diğ., 2015).

QSAR yöntemleri genel olarak iki ana gruba ayrılır: 2D-QSAR ve 3D-QSAR modelleri. 2D-QSAR yaklaşımı, farmakolojik bileşiklerin iki boyutlu yapılarından yola çıkarak fizikokimyasal tanımlayıcılar üretir ve bu veriler doğrultusunda biyolojik aktiviteleri tahmin eder. Buna karşılık, 3D-QSAR modeli moleküllerin üç boyutlu geometrik özelliklerini de hesaba katarak daha hassas tahminler sunar ve bağlanıcı molekül-hedef etkileşimlerini daha ayrıntılı biçimde analiz etme imkânı sağlar. 2D-QSAR modelleri, 3D-QSAR'a göre daha basit ve hesaplama açısından daha hızlıdır. Ancak, aynı bileşik ailesiyle sınırlı kalabilme eğilimi ve rastlantısal korelasyonlara açık olması gibi bazı kısıtlamaları da bulunmaktadır. Son yıllarda, bir bileşiğin birden fazla biyolojik hedefle olan etkileşimini değerlendirebilen çoklu hedefli QSAR modelleri geliştirilmiş ve bu alanda yeni bir araştırma yönü oluşturmuştur (Jana ve diğ., 2018).

QSAR modelleri, ilaç endüstrisinde ilaç aktivitelerinin , farmakokinetik özellikleri , toksisitenin, sulu çözünürlüğün ve ilaç dağılımının tahmin edilmesinde yaygın olarak kullanılmaktadır (Jenkins, 2012).

2.5.2 Proteokimimetrik Model (Proteochemometric Modeling (PCM))

Proteokimimetrik modelleme , ilaç-hedef etkileşimlerinin öngörülmesinde kullanılan QSAR yönteminin daha gelişmiş ve kapsamlı bir versiyonudur. Bu yaklaşım, benzer yapıya sahip

ligandları ve hedef proteinleri eşzamanlı olarak değerlendirerek, etkileşimlerin tek bir model çerçevesinde tahmin edilmesini mümkün kılar. PCM, bilinen bileşiklerin biyolojik aktivite verilerini kullanarak yeni bileşikler için genelleştirilmiş ve güvenilir öngörüler sunabilir; hatta hedef yapısının bilinmediği durumlarda dahi etkili sonuçlar sağlayabilir. Bununla birlikte, geniş ölçekli veri eksikliği ve modelin uygulanabilirlik sınırlarının netleştirilmesi gibi bazı zorluklar, özel yöntem ve stratejiler gerektirmektedir. PCM yalnızca ilaç-hedef etkileşimlerini tahmin etmekle kalmaz; başarısız ilaç bileşiklerinin yeniden değerlendirilmesi ve ligandı henüz tanımlanmamış biyolojik hedefler için yeni ligandların keşfinde de etkili bir yaklaşım sunar. Ayrıca bu model, bileşiklerin hedef dışı etkilerini analiz etme ve seçicilik düzeylerini belirlemede de önemli bir rol üstlenmektedir. Genel olarak, PCM farmakolojik araştırmaların birçok aşamasında güçlü ve umut vaat eden bir araç olarak kabul edilmekte olup, son yıllarda bu alanda kayda değer ilerlemeler kaydedilmiştir (Cortés-Ciriano ve diğ., 2015).

2.5.3 Moleküler yerleştirme (**Molecular docking**)

Moleküler Yerleştirme, ilaç keşfinde sanal taramanın (virtual screening) temel yöntemlerinden biri olup, yüksek verimli tarama (High-Throughput Screening (HTS)) sürecinde gerekli olan bağlanıcı molekül sayısını azaltarak bu süreci daha verimli hale getirir. Bu teknik, ilaç-hedef etkileşimini simüle ederek, bilinen üç boyutlu yapıya sahip bir hedef proteinin aktif bölgesine bir ligandın bağlanmasını öngörür. Skorlama fonksiyonları aracılığıyla ligandların protein hedeflerine olan bağlanma eğilimleri değerlendirilir ve sonraki aşamalarda kullanılmak üzere en uygun bileşikler seçilir (Furlan ve diğ., 2018).

Önemli avantajlarına rağmen, bu yöntem yüksek hesaplama maliyeti ve skorlama fonksiyonlarının bağlanma eğilimini tahmin etmedeki sınırlı doğruluğu gibi bazı zorluklarla karşı karşıyadır. Ayrıca, birçok proteinin üç boyutlu yapısına erişilememesi, bu yöntemin kullanımını kısıtlamaktadır. Bununla birlikte, son yıllarda makine öğrenmesindeki ilerlemeler sayesinde her bir protein için özel olarak geliştirilen skorlama fonksiyonları aracılığıyla tahmin doğruluğu artırılmıştır (Vasudevan ve Churchill, 2009).

Buna ek olarak, ters yerleştirme (inverse docking) ve ters sanal tarama (Inverse Virtual Screening, IVS) gibi ters yaklaşımlar, bir ligandın olası hedeflerini veritabanlarında tanımlamak için kullanılmakta ve ilaç başarısını artırma potansiyeli taşımaktadır (Xing ve diğ., 2011). Ayrıca, RNA'ya dayalı ilaçların artan kullanımıyla birlikte, moleküler yerleştirme RNA

hedefleri için de uygulanmaya başlanmış ve bileşik kütüphanelerinin tasarımı, istenmeyen yan etkilerin öngörülmesi ve ilaç yeniden konumlandırması gibi alanlarda önemli roller üstlenmiştir. Genel olarak, moleküler yerleştirme, ilaç keşfinde güçlü ve yaygın olarak kullanılan bir araç olmaya devam etmektedir (Furlan ve diğ., 2018).

2.6 TEMEL HESAPLAMALI YAKLAŞIMLARLA DTI TAHMİNİ

İlaç-hedef etkileşimlerini hesaplamalı olarak tahmin etmek için iki ana yaklaşım vardır: yerleştirme (docking) simülasyonları ve makine öğrenimi yöntemleri (Peska ve diğ., 2017).Bu yöntemler, deneysel yöntemlerin zaman ve maliyet açısından sınırlılıklarını aşmak için etkili alternatifler sunmaktadır. Ayrıca, bu hesaplamalı yaklaşımlar, erken aşama ilaç keşfi süreçlerinde aday bileşiklerin belirlenmesinde önemli rol oynamaktadır.

2.6.1 Yerleştirme Simülasyonları

Bu yöntemde, ilaç moleküllerinin ve hedef proteinlerin üç boyutlu yapıları analiz edilerek potansiyel bağlanma bölgeleri belirlenir. Bu teknik biyolojik olarak güvenilir ve kabul görmüş olsa da, uygulanması zaman alıcıdır. Ayrıca, hedef proteinin üç boyutlu yapısı bilinmediğinde bu yöntem kullanılamaz. Örneğin, G proteini ile eşleşmiş reseptörler (GPCR) olarak adlandırılan birçok protein henüz tam olarak kristalize edilmemiştir. Bu nedenle, bu proteinler için yerleştirme yöntemleri kullanılamaz (Yamanishi ve diğ., 2008).

Bu zorlukların üstesinden gelmek için kimyagenomik alternatif bir yaklaşım olarak önerilmiştir. Bu yöntemde, her proteinin ayrı ayrı incelenmesi yerine, kimyasal ve genomik alanlar aynı anda analiz edilerek olası etkileşimler belirlenir. kimyagenomik çalışmaların temel amacı, görüntüleme görüntüleme araçları veya yeni ön ilaçlar olarak kullanılamaya keşfetmektir (Jacob ve Vert, 2008).

Kemogenomik yaklaşımlar; ligand bazlı, hedef bazlı ve ligand-hedef bazlı hibrit yöntemler olmak üzere üç kategoriye ayrılmaktadır. Bu yöntemler, proteinler ve ilaç bileşikleri arasındaki benzerlikleri temel alır. Bu benzerliklere dayalı yaklaşımlar, makine öğrenimi tekniklerinin ilaç-hedef etkileşimlerini tahmin etmek için kullanılmasını sağlamaktadır (Pharmacology ve diğ., 2007).

2.6.2 Makine Öğrenimi Yöntemleri

Bu yaklaşımda, onaylanmış ilaçlar, hedef proteinler ve bunlar arasındaki etkileşimlere ilişkin veriler kullanılarak makine öğrenmesi modelleri eğitilmektedir. Eğitilen modeller, daha sonra yeni ilaçlar ile potansiyel hedefler arasındaki etkileşimleri tahmin etmek amacıyla kullanılmaktadır. Bu yöntemin temel dayanağı birkaç ana varsayıma dayanır: Kimyasal olarak benzer ilaçların belirli bir hedefle etkileşime girme olasılığı daha yüksektir; yapısal olarak benzer proteinler belirli bir ilaçla reaksiyona girebilir; ve benzer ilaçlar benzer proteinlerle etkileşime girebilir. İlaç bileşikleri ile proteinler arasındaki benzerliği ölçmek için çekirdek fonksiyonları (kernel functions) gibi özel yöntemler kullanılmaktadır (Nath ve diğ., 2018).

Mevcut ilaç ve hedef bilgilerine dayanarak, ilaç-hedef etkileşimini tahmin etme problemi dört ana kategoriye ayrılabilir: (1) bilinen bir ilaç ile bilinen bir hedef arasındaki etkileşimin tahmini, (2) bilinen bir ilacın yeni bir hedefle etkileşiminin tahmini, (3) yeni bir ilacın bilinen bir hedefle etkileşiminin tahmini ve (4) yeni bir ilaç ile yeni bir hedef arasındaki etkileşimin tahmini. Bu yöntemler, yeni ilaçların geliştirilmesi süreçlerinde ve ilaç keşif sürecinin optimize edilmesinde önemli bir rol oynamaktadır (Chen ve diğ., 2016).

2.7 DTI TAHMİNİNDE KULLANILAN MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Bu bölümde, ilaç-hedef etkileşimlerini tahmin etmek için kullanılan en önemli makine öğrenmesi yöntemleri tanıtılmakta ve incelenmektedir. Bu yöntemler; benzerlik veya mesafe tabanlı yaklaşımlar, özellik tabanlı yöntemler, matris ayrıştırma teknikleri, ağ tabanlı modeller, derin öğrenme yöntemleri ve ayrıca birleşik (kombine) yaklaşımları içermektedir. Her bir yaklaşım, veri türüne ve analiz amacına bağlı olarak kendine özgü avantajlar ve sınırlılıklar taşımaktadır.

2.7.1.Benzerlik/Mesafe Tabanlı Yöntemler

İlaç-hedef etkileşimlerini tahmin etmek için en sık kullanılan yöntemlerden biri, ilaçlar ve hedefler arasındaki benzerliğin analiz edilmesine dayanır. Bu yaklaşımda, ilaçlar ve farklı hedefler arasındaki yakınlık derecesini değerlendirmek için benzerlik veya uzaklık ölçütleri kullanılır.Genellikle bu yöntemler, ilaç, protein veya bu iki çiftin kombinasyonu arasındaki benzerliğe dayalı bir puanlama sistemini kullanır. Ayrıca, yeni bir ilacın daha önce bilinen ilaçlara ve hedeflere ne kadar yakın olduğunu belirleyen bir mesafe fonksiyonu aracılığıyla

benzerlik derecesi ölçülebilir.Bu fonksiyonlar, ilaçların farmakolojik benzerliklerine, protein dizilerindeki genetik benzerliklere ve ilaçlarla hedef proteinler arasındaki ilişkiyi gösteren karmaşık ağlardaki yapısal örüntülere dayanarak tanımlanabilir (Shi ve diğ., 2015; Zhang ve diğ., 2017).

2.7.2.Özellik Tabanlı Yöntemler

İlaç-hedef etkileşimlerini tahmin etmek için kullanılan makine öğrenmesi algoritmalarının çoğu bu kategoriye girmektedir. Bu grup, destek vektör makineleri (SVM), karar ağacı tabanlı modeller ve diğer çekirdek tabanlı teknikler gibi çeşitli yöntemleri içerir.Bu yaklaşımlarda her ilaç ve hedef, belirli boyutlara sahip özellik vektörleri olarak temsil edilir (Today ve diğ., 2015). Tipik olarak bu vektörler, ilaç ile ilgi duyulan hedef arasında bir ilişki olup olmadığını gösteren ikili etiketlere sahiptir. Bu yöntemin temel zorluklarından biri, hücre zarı proteinlerinin üç boyutlu yapılarına erişimin olmamasıdır (Wang ve diğ., 2018). Bu sınırlama, temel özelliklerin çıkarılmasını zorlaştırır ve sonuç olarak tahmin modellerinin doğruluğunu azaltır. Bu bilgi mevcutsa, makine öğrenimi modellerinin performansının artmasını bekleyebiliriz (Sharma ve diğ., 2018).

2.7.3 Matris Ayrıştırma Yöntemleri

Diğer makine öğrenmesi teknikleriyle karşılaştırıldığında, matris ayrıştırma yöntemleri ilaç-hedef etkileşimlerini tahmin etmede daha iyi performans göstermektedir. İlaçların iki boyutlu (2D) yapısal benzerliklerine dayanan birçok geleneksel modelin aksine, bu yöntemlerden bazıları doğrudan kimyasal bilgiye ihtiyaç duymadan çalışabilmektedir.Bu alandaki yaygın yaklaşımlardan biri, bir örneği olasılık matrisi faktörizasyonu (Probabilistic Matrix Factorization (PMF)) olan işbirlikçi filtrelemedir (Cobanoglu ve diğ., 2013) . Diğer bazı yöntemler Düşük Dereceli Gömme (Low-Rank Embedding (LRE)) modellerinden esinlenmiştir . Bu tekniklerde asıl amaç, orijinal bilginin korunduğu ve yeniden yapılandırma hatasının en aza indirildiği şekilde veri kümesinin düşük boyutlu bir gösterimini bulmaktır (Li ve diğ., 2017).

Bu modellerde ilaçların ve hedeflerin ortak bir uzayda yer aldığı ve aralarındaki etkileşimin gücünün bu uzaydaki mesafelerine göre tahmin edilebileceği varsayılmaktadır. Bu yaklaşım, her iki grubun (ilaçlar ve hedefler) belirli koşullar ve kısıtlamalar altında da olsa düşük boyutlu bir alt uzaya yerleştirilmesine olanak tanır (Wang ve diğ., 2018).

Bu yöntemlerin daha yüksek doğruluk ve verimliliğe sahip olmasına rağmen, biyolojik ve farmakolojik verilerin hacmi ve çeşitliliğinin sürekli artması ciddi bir zorluk yaratmıştır. Bu nedenle matris tabanlı modeller, bu kadar büyük veri hacmiyle karşı karşıya kaldıklarında ölçeklenebilirlik ve verimlilik açısından sınırlamalarla karşılaşabilirler (Ezzat ve diğ., 2015).

2.7.4 Ağ Tabanlı Yöntemler

Ağ tabanlı yöntemler, ilaç-hedef etkileşimlerini analiz etmek ve tahmin etmek için kullanılan grafik tabanlı teknikleri ifade eder. Bu yöntemlerin en basit ama en etkili olanlarından biri, basit yapısına rağmen en güvenilir benzerlik tabanlı yöntemlerden biri olarak kabul edilen ağ tabanlı çıkarımdır (Luo ve diğ., 2017).

Bu yöntemlerden bazıları, protein-protein benzerlik ağı, ilaç-ilaç benzerlik ağı ve ilaçlar ile hedefler arasındaki doğrulanmış etkileşim ağı olmak üzere üç farklı ağ türünü heterojen bir yapıda birleştirir.Bu çerçevede, kimyasal olarak benzer ilaçların genellikle ortak protein hedeflerini hedef aldığı varsayılmaktadır.Ayrıca, tahmin algoritmalarını eğitmek için iki katmanlı, yönlendirilmemiş bir grafik modeli kullanılabilir. Bu yöntemle hem doğrudan ilaç-hedef etkileşimleri (esas olarak ligandın proteine bağlanmasıyla ilgili) hem de dolaylı etkileşimler analiz edilebilmektedir. Ayrıca protein bağlanması, aktivasyonu ve engellenmesi gibi ilaç etki mekanizmalarını simüle ederek ilaçların nasıl çalıştığını daha iyi anlamamıza yardımcı oluyor (Huang ve diğ., 2020).

2.7.5 Derin Öğrenme Yöntemleri

Derin öğrenme, doğal dil işleme, konuşma tanıma ve görüntü tanıma gibi alanlardaki olağanüstü performansı nedeniyle veri bilimi ve biyobilgisayar alanında en yaygın kullanılan yöntemlerden biri haline gelmiştir. Son yıllarda, derin öğrenmenin ilaç keşfindeki rolü önemli ölçüde artmıştır (You ve diğ., 2019).

Bu yöntemlerin ilaç-hedef etkileşimlerini tahmin etmedeki temel avantajlarından biri, bazı geleneksel makine öğrenmesi yöntemlerinde meydana gelen temel bilgi kaybının azaltılmasıdır. Ancak bu tekniklerin zorluklarından biri, modelleri eğitmek için yeterli ve kaliteli verinin olmamasıdır. Bu sorunu ele almak için son araştırmalar, DTI tahmini ve ilaç yeniden tedariki alanlarında karmaşık ve çalkantılı verileri analiz etmek için derin öğrenme algoritmaları geliştirdi (Gao ve diğ., 2018; Wang ve diğ., 2018).

DTI tahmini için derin öğrenme modelleri genellikle iki temel adımdan oluşur: verilerin özellik vektörlerine dönüştürülmesi ve mevcut ilaç-hedef etkileşimlerine derin öğrenme algoritmalarının uygulanması. Bu süreçte, ilaçlardan ve protein hedeflerinden elde edilen üç ana özellik türü — biyolojik özellikler (örneğin moleküler etkileşimler ve biyolojik yollar), topolojik özellikler (örneğin biyolojik ağlardaki bağlantılar) ve fizikokimyasal özellikler (örneğin ilaçların kimyasal ve yapısal özellikleri) — özellik vektörlerinin oluşturulmasında kullanılır. Bu yöntemler, hem tahminlerin doğruluğunu artırmakta hem de ilaç-hedef etkileşimlerinin daha iyi anlaşılmasına katkı sağlayarak yeni ilaçların geliştirilmesinde önemli bir rol oynamaktadır (Gao ve diğ., 2018).

2.7.6 Kombine Yöntemler

Kombine yöntemler, tahmin doğruluğunu artırmak için özellik tabanlı modeller, matris ayrıştırma, derin öğrenme ve ağ analizi gibi birkaç farklı tekniği birleştiren yaklaşımlardır. Bu yöntemler, farklı veri kaynaklarını entegre ederek modellerin ilaç-hedef etkileşimlerini belirleme kapasitesini artırır Genel olarak, bu yöntemler iki temel amaca hizmet eder: birincisi, DTI'ları tahmin ederken bilinmeyen etkileşimlerin zorluklarını ele almak; ikincisi ise birden fazla makine öğrenmesi yönteminin avantajlarından aynı anda yararlanarak daha etkili ve doğru sonuçlar elde etmektir (Yuan ve diğ, 2016).

Birden fazla tekniğin tek bir modelde birleştirilmesi, her yöntemin diğerinin zayıf yönlerini kapatabilmesi nedeniyle tahmin doğruluğunu artırabilir. Ancak birden fazla yöntemin entegre edilmesinin, hesaplama karmaşıklığının ve işlem maliyetinin artmasına yol açabileceği, bu nedenle veri ve hesaplama kaynaklarının optimum şekilde yönetilmesi gerektiği unutulmamalıdır (Shi ve diğ, 2019).

2.8 DTI ÜZERINE YAPILAN ÇALIŞMALARIN İNCELENMESI

Önceki bölümlerde de belirtildiği gibi, ilaç-protein etkileşimlerini tahmin etme çalışmalarının temeli, özellik temelli yöntemler, çekirdek temelli yöntemler ve benzerlik temelli yöntemler gibi makine öğrenmesi yaklaşımlarından biri olabilir.Şimdiye kadar ilaç ve protein etkileşimlerini tahmin etmek için çeşitli hesaplamalı yöntemler geliştirilmiştir. 2021 yılında Peng ve çalışma arkadaşları, grafik temsiline dayalı yeni bir çerçeve sunmuşlardır (Peng ve diğ., 2021).

Çekirdek (kernel) tabanlı yöntemler, makine öğrenmesinin önemli bir alt dalını oluşturmakta olup, bu alanda birçok araştırmacı çalışmalar yürütmüştür. Mohammad ve çalışma arkadaşları, Bayesyen matris ayrıştırması kullanarak, onkoloji alanında entegre ve özel QSAR yaklaşımlarını analiz etmişlerdir (Din ve diğ., 2014). 2018 yılında gerçekleştirilen bir diğer çalışmada, Anna Cichonska ve meslektaşları, eğitim süresi ve bellek kullanımını daha verimli hâle getiren çift çekirdekli çoklu bir yöntem önermişlerdir (Cichonska ve diğ., 2018).Önemli bir diğer yöntem grubu ise benzerlik tabanlı yöntemlerdir (Sridhar ve diğ., 2016). Bu doğrultuda, Thafar ve çalışma arkadaşları, yeni ilaçlar ile bilinen proteinler arasındaki etkileşimleri belirlemek amacıyla benzerlik takibine dayalı bir yöntem önermişlerdir (Thafar ve diğ., 2020). 2021 yılında, ilaç-protein etkileşimlerini tahmin etmek için benzerlik tabanlı bir model sunulmuştur. Bu model, ilaçlar ve proteinler için oluşturulan iki benzerlik matrisine karşılık gelen sütun vektörlerinin dış çarpımları üzerinde iki boyutlu evrişimsel sinir ağlarını (2D CNN) kullanmaktadır (Shim ve diğ., 2021).

Özellik tabanlı yöntemler, makine öğrenmesi tekniklerinden biridir ve bu alanda birçok araştırma yapılmıştır. İlaç ve protein etkileşimlerini tahmin etmeye yönelik özellik tabanlı çalışmalar, genellikle özellik çıkarımı ve seçimi, yeni sınıflandırma yöntemleri ve veri dengeleme teknikleri gibi dört temel yeniliğe odaklanmaktadır (Sorkhi ve diğ., 2021).Özellik çıkarımı alanında, Cheng Wang ve çalışma arkadaşları, 2020 yılında hedef proteinlerin parmak izi özelliklerini kullanarak, ilaçların elektro-topolojik durumunu ve yarı amfipatik amino asit bileşimlerini belirlemiştir (Wang ve diğ., 2020). Ayrıca, 2021 yılında, dengesiz veri kümeleriyle çalışan FastUS algoritması önerilmiştir (Mahmud ve diğ., 2021) . Zheng Yang ve çalışma arkadaşları, PSHOG gradyanı ve PSSM matrisi ile özellik çıkarmak için yeni bir hesaplamalı model kullanmıştır (Zhao ve diğ., 2021). Li ve arkadaşları , PPI (Protein-Protein Etkileşimleri) tahmini için LBP (Local Binary Patterns) özellik çıkarma yöntemi ve Rotation Forest sınıflandırıcısı kullanan verimli bir hesaplamalı yöntem önerilmiştir (Li, ve diğ., 2019).

Son yıllarda, derin öğrenmenin biyoinformatik alanındaki dikkat çekici başarıları göz önüne alındığında, bu yöntemlerin topolojik verilerden değerli bilgiler çıkarabildiği ve heterojen veriler arasındaki karmaşık etkileşimleri analiz edebildiği ortaya konmuştur (Xiao, ve diğ., 2021). Bu doğrultuda, derin öğrenme yöntemleri ilaç-protein etkileşimlerinin tahmininde yaygın olarak kullanılmaktadır; çünkü bu yöntemler, büyük veri kümelerinde ilaçlar ve proteinler arasındaki gizli desenleri tanımlayabilmekte ve ilaç ile protein ağlarından yeni

bilgiler elde edebilmektedir. Bu yaklaşımlarda, ilaç ve proteinlerden uygun özellikler çıkarılmakta ve bu özellik çiftleri bir sınıflandırıcıyı eğitmek amacıyla kullanılmaktadır; böylece ilaç-protein etkileşimleri öngörülebilmektedir (Rifaioglu ve diğ., 2019). CNN-DTI yöntemi, ilaçlar ve proteinlerin birleşiminden oluşan birleşik özellik vektörünü giriş olarak alır ve konvolüsyonel sinir ağlarını (CNN) bir sınıflandırma modeli olarak kullanır (Huang ve diğ., 2021).

Peng ve arkadaşları, özellik temsil öğrenmesi ile derin öğrenme tabanlı bir yöntem sunulmuştur (Peng ve diğ., 2020). Mahdaddi ve arkadaşları, derin öğrenmeye dayalı hiperparametreler incelenmiştir (Mahdaddi ve diğ., 2021). Wang ve arkadaşları, derin hibrit modeller için Evrimsel (Evolutionary Algorithm(EA)) yöntemi tanıtmaktadır(Wang ve diğ., 2022). Song ve arkadaşları , bilgi grafikleri tabanlı bir derin öğrenme yöntemi tartışmaktadır (Song ve diğ., 2022). Hasan ve arkadaşları , derin öğrenmeye dayalı çok ölçekli özellik kombinasyonu sunmuştur (Hasan Mahmud ve diğ., 2020). Zhang ve arkadaşları, çok katmanlı bir grafik yöntemi önermektedir (Zhang ve diğ., 2022). Wang ve arkadaşları , belirli bir tür derin ağ kullanan bir hesaplama yöntemi tanıtmaktadır (Wang ve diğ., 2018).

İlaç-hedef etkileşimlerinin tahmin sürecinde, makine öğrenmesi ve derin öğrenme algoritmalarının doğruluğu, doğrudan giriş verilerinin kalitesine bağlıdır. Bu alandaki temel zorluklardan biri, sınıf dengesizliğidir ve bu durum, tahmin modellerinin performansını olumsuz yönde etkileyebilir. Bu sorunu gidermek için, genellikle sınıf dağılımını dengelemek amacıyla aşırı örnekleme (oversampling) ve düşük örnekleme (undersampling) tekniklerinden yararlanılmaktadır. Rastgele Aşırı Örnekleme yöntemi, azınlık sınıfına rastgele yeni örnekler ekleyerek denge sağlar. Öte yandan, Rastgele Düşük Örnekleme yöntemi ise çoğunluk sınıfından rastgele örnekleri kaldırır (Hasanin ve diğ., 2019).

Sınıf dengesizliği ve yüksek boyutluluk gibi iki temel zorluğu aynı anda çözmek amacıyla Redkar ve arkadaşları (2020), azınlık sınıf örneklerini artırmak için SMOTE tekniğini ve önemli özelliklerin seçimi ile boyut indirgeme için Wrapper algoritmasını bir arada kullanmışlardır (Redkar ve diğ., 2020).

Mahmud ve arkadaşları (2021), PreDTIs adlı yeni bir yaklaşım geliştirerek DTI tahmini için yenilikçi bir yöntem sunmuşlardır. Bu yöntemde, proteinle ilgili özellikler çeşitli analiz teknikleriyle çıkarılmış, ilaçlar ise MACCS parmak izi algoritmasıyla kodlanmıştır. Sınıf

dengesizliği sorununu çözmek için geliştirilmiş bir düşük örnekleme tekniği olan FastUS algoritması kullanılmıştır. Bu algoritma, geleneksel yöntemlere kıyasla önemli özelliklerin korunmasına ve hayati verilerin silinmesinin önlenmesine odaklanarak daha etkili bir dengeleme sağlar. Verilerin hazırlanmasının ardından, tahmin işlemi için LightGBM modeli uygulanmıştır (Mahmud ve diğ., 2020).

Ayrıca, Prasetyo ve Anggraeni (2024), DTI tahmini için ilaç ve proteine ait açıklayıcı özelliklerin Random Forest (Rastgele Orman) algoritmasıyla birlikte kullanıldığı yenilikçi bir yaklaşım geliştirmişlerdir. Özellik boyutlarının azaltılması için Rastgele Yansıtma (Random Projection) yöntemi ve veri dengesi sağlamak amacıyla NearMiss tekniği uygulanmıştır. NearMiss, karar sınırına yakın olan çoğunluk sınıfı örneklerinin seçici şekilde kaldırılmasıyla sınıf dengesizliğini azaltır ve modelin tahmin doğruluğunu artırır.

3. MALZEME VE YÖNTEM

Bu bölümde çalışmanın hedefler doğrultusunda sırasıyla, protein ve ilaçlardan özellik çıkarma, veri seti oluşturma, veri dengeleme yöntemler, model oluşturmaya ilişkin bilgiler yer almaktadır.

3.1 ARAŞTIRMANIN MODELİ

Bu çalışmada, ilaç-hedef etkileşimini tahmin etmek amacıyla yenilikçi bir yöntem önerilmektedir. İlk adımda, ilaçların kimyasal yapıları (SMILES formatında) ve protein dizileri (FASTA formatında), güvenilir veri tabanlarından belirli erişim tanımlayıcıları aracılığıyla toplanmaktadır. Verilerin toplanmasının ardından, ilaçlar ve proteinlerden, çeşitli özellik çıkarma teknikleri (parmak izleri ve istatistiksel tanımlayıcılar dahil) kullanılarak farklı özellikler elde edilmektedir. Elde edilen bu özellikler ve bilinen etkileşimler kullanılarak, ilaç-hedef çift vektörleri oluşturulmaktadır**.**

Ön işleme aşamasında, veri setleri MinMaxScaler yöntemi kullanılarak normalize edilmiş ve böylece veri ölçekleri birbiriyle uyumlu hâle getirilmiştir. Daha sonra, yüksek boyutlu verilerin karmaşıklığını azaltmak amacıyla Temel Bileşenler Analizi (TBA) (Principal Component Analysis (PCA)) uygulanmıştır. Bu yöntem, verideki en yüksek varyansı koruyarak tekrarlayan ve yüksek korelasyonlu özellikleri ortadan kaldırmış, dolayısıyla veri gürültüsünü azaltmıştır. PCA'nın uygulanması sonucunda, özellik sayısı 100 temel bileşene düşürülmüş; bu sayı, temel bilgilerin korunmasını sağlarken öğrenme hızını artırmış, hesaplama verimliliğini geliştirmiş ve modelin genellenebilirlik gücünü yükseltmiştir.

Bu çalışmada, DTI tahmininde karşılaşılan en önemli zorluklardan biri, veri setlerindeki sınıf dengesizliğidir; çünkü etkileşimli (pozitif) örneklerin sayısı, etkileşimsiz (negatif) örneklere kıyasla oldukça azdır. Bu sorunu gidermek amacıyla, Siyam sinir ağı ve SMOTE-Tomek'ten oluşan iki aşamalı hibrit bir yöntem kullanılmıştır. İlk aşamada, veriyi dengelemek için Siyam sinir ağı, ikili veri karşılaştırmaları üzerinden uygulanmıştır. Bu yaklaşımda, her bir çapa (Anchor) örneği için, etkileşimli veri setinden en yüksek benzerliğe sahip bir pozitif örnek ve etkileşimsiz veri setinden en düşük benzerliğe sahip üç negatif örnek seçilir; böylece eğitim setinde pozitif ve negatif veri oranı 1:3 olacak şekilde dengelenir.

Bu yaklaşımda, Siyam ağı bir temel ağ (Base Network) içerir ve bu ağ, girişleri Conv1D ve Dense katmanlarıyla işleyerek her örnek için 256 boyutlu bir özellik vektörü (Embedding) üretir. Ağ, her veri çifti (çapa ve karşılaştırma örneği) için ortak olarak kullanılır ve böylece özellikler benzer şekilde çıkarılır. Daha sonra Lambda katmanı ile veri çiftlerinin özellik vektörleri arasındaki Öklid mesafesi hesaplanır. Contrastive Loss kayıp fonksiyonu ağı, pozitif çiftler arasındaki mesafeyi azaltacak, negatif çiftler arasındaki mesafeyi ise artıracak şekilde yönlendirir. Siyam ağı, bu dengelenmiş verilerle 200 epoch boyunca eğitilmiş ve model her aşamada pozitif ve negatif örneklerin seçiminde daha optimal kararlar alarak pozitif ve negatif verileri daha doğru bir şekilde ayırt etme yeteneği kazanmıştır. Bu süreç, verilerin sonraki modelleme aşamaları için dengeli ve verimli bir şekilde hazırlanmasını sağlamıştır.

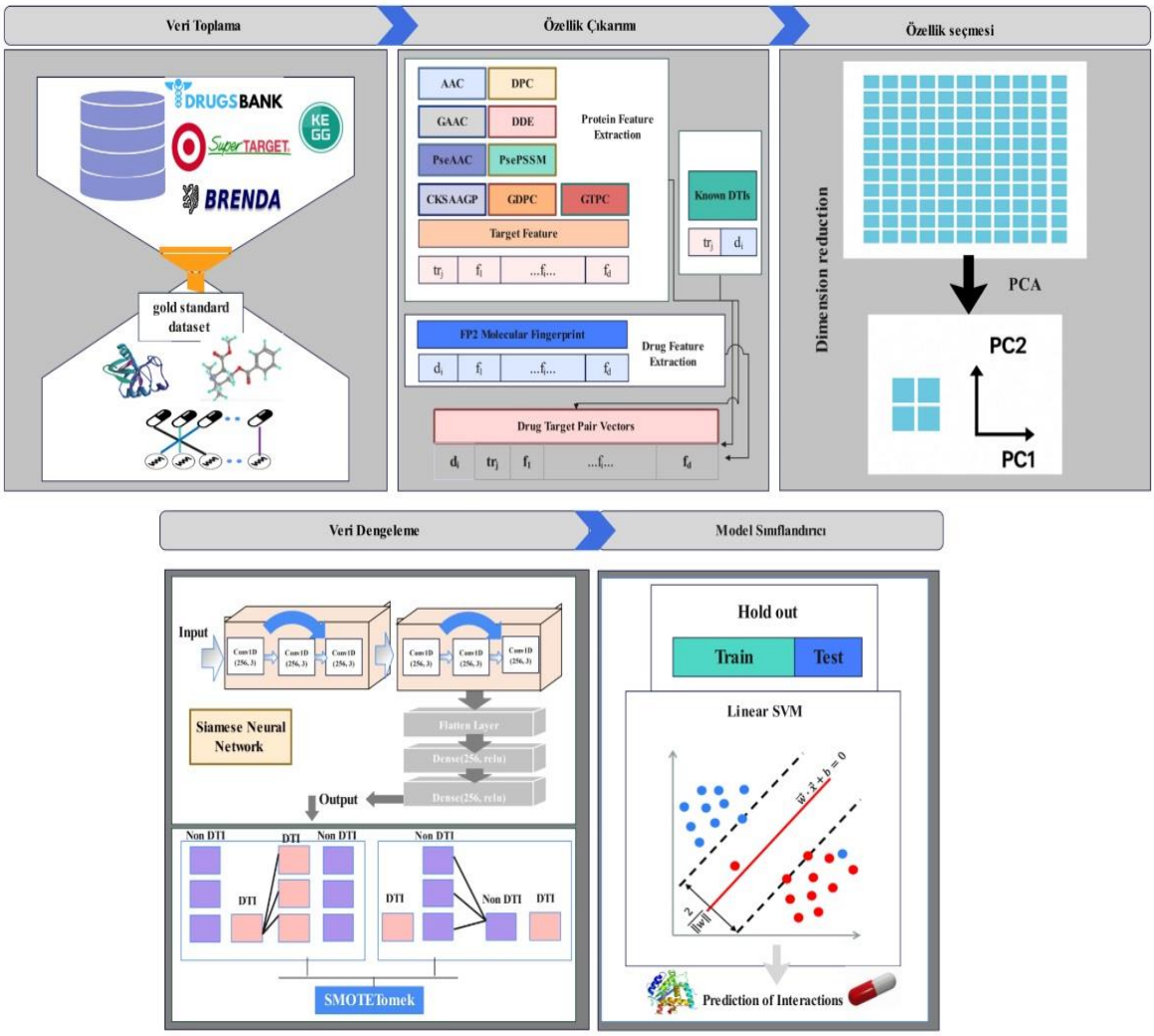
Veriyi dengeleyici bir ön işlem olarak Siamese sinir ağı kullanıldı. Bu bağlamda, veriler "anchor", "pozitif" ve "negatif" örnekler şeklinde üçlü yapılar hâlinde organize edildi. Her bir anchor görüntüsü için, en çok benzeyen etkileşimli örnek (pozitif) ve en az benzeyen etkileşimsiz örnek (negatif) seçildi. Böylece her bir anchor için biri pozitif ve üçü negatif olmak üzere dengeli ve yapay olarak zenginleştirilmiş bir veri seti oluşturuldu. Bu yapı sayesinde, etkileşimli ve etkileşimsiz çiftler arasında 1:3 oranında örneklem elde edildi. Siamese ağı, bu üçlü örnekler üzerinden benzerlikleri ve farklılıkları öğrenmeye yönelik olarak eğitildi. Model, anchor ile pozitif örneklerin benzerliğini yüksek, anchor ile negatif örneklerin benzerliğini ise düşük olarak öğrenmeyi hedefledi. Mesafe hesaplamalarında, anchor ile diğer örnekler arasındaki öznitelik farkları temel alındı. En büyük mesafeye sahip negatif örnek, anchor'dan en fazla farklı olan veri olarak seçildi ve modele bu şekilde sunuldu. Bu strateji, hem pozitif hem de negatif veri kümelerinin dengeli şekilde oluşturulmasını sağladı hem de modelin, etkileşimli ve etkileşimsiz çiftleri ayırt edebilmesini mümkün kıldı. Elde edilen pozitif örnekler, anchor'a en yakın etkileşimli görüntülerden oluştuğundan, yüksek düzeyde benzerlik içerdi ve modelin etkileşim tahmini yapmadaki başarısını artırdı.

Tam dengeli bir veri seti elde etmek amacıyla, çeşitli aşırı örnekleme, alt örnekleme ve hibrit teknikler değerlendirilmiş ve uygulanmıştır. Bu teknikler arasında ClusterCentroids, CondensedNearestNeighbour, EditedNearestNeighbour, RepeatedEditedNearestNeighbours, AllKNN, NeighbourhoodCleaningRule, OneSidedSelection, RandomUnderSampler, TomekLinks, RandomOverSampler, SMOTE, SMOTENC, SMOTEN, ADASYN, BorderlineSMOTE, SVMSMOTE ve SMOTETomek yer almaktadır. Bu yöntemler arasından,

SMOTE-Tomek (Tomek bağlantıları ile birlikte Sentetik Azınlık Aşırı Örnekleme Tekniği) en yüksek dengeleme performansını göstermiş ve bu nedenle nihai seçim olarak belirlenmiştir. SMOTE-Tomek yöntemi, azınlık sınıfına ait (bilinen etkileşimler) sentetik örnekler üretirken, çoğunluk sınıfına ait (etkileşimde olmayanlar) gürültülü ve sınırda kalan verileri veri setinden ayıklayarak hem dengeli hem de yüksek kaliteli bir örneklem oluşturmuştur. Bu ön işleme ve dengeleme adımları, nihai sınıflandırma modelini eğitmeden önce uygulanmış ve modelin genel performansını ile birlikte sağlamlığını artırmada kritik rol oynamıştır.

Veri seti dengeleyip iyileştirildikten sonra, son hâli üzerinde bir Linear SVM sınıflandırıcısı eğitilmiştir. En uygun öğrenme algoritmasını belirlemek amacıyla, Nearest Neighbors, Linear SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA, LDA, EEC, RSC **ve** BBC olmak üzere tüm yöntemler üzerinde kapsamlı bir değerlendirme gerçekleştirilmiştir. Yapılan analizler sonucunda, Linear SVM sınıflandırıcısı en yüksek başarıyı göstermiş ve bu nedenle model geliştirme sürecinde tercih edilmiştir. Bu sınıflandırıcı, ilaç ve proteinlerden çıkarılan özniteliklere dayanarak, ilaç-hedef etkileşimlerini tahmin etmek amacıyla kullanılmıştır. Böylece, dengelenmiş ve optimize edilmiş veri seti üzerinden yüksek doğrulukta tahminleme yapılması sağlanmıştır.

Özellik çıkarma, veri ön işleme, veri dengeleme ve sınıflandırma adımları, bütüncül bir çerçevede entegre şekilde uygulanmıştır. Bu birleşik yapı, önerilen modelin etkinliğini en üst düzeye çıkararak, ilaç-hedef etkileşimi tahminleri için sağlam ve sistematik bir yaklaşım ortaya koymuştur. Modelin genel mimarisini gösteren diyagram, veri toplama aşamasından tahmin sürecine kadar uzanan entegre yaklaşımı görsel olarak sunmakta olup, Şekil 3.1'de yer almaktadır.

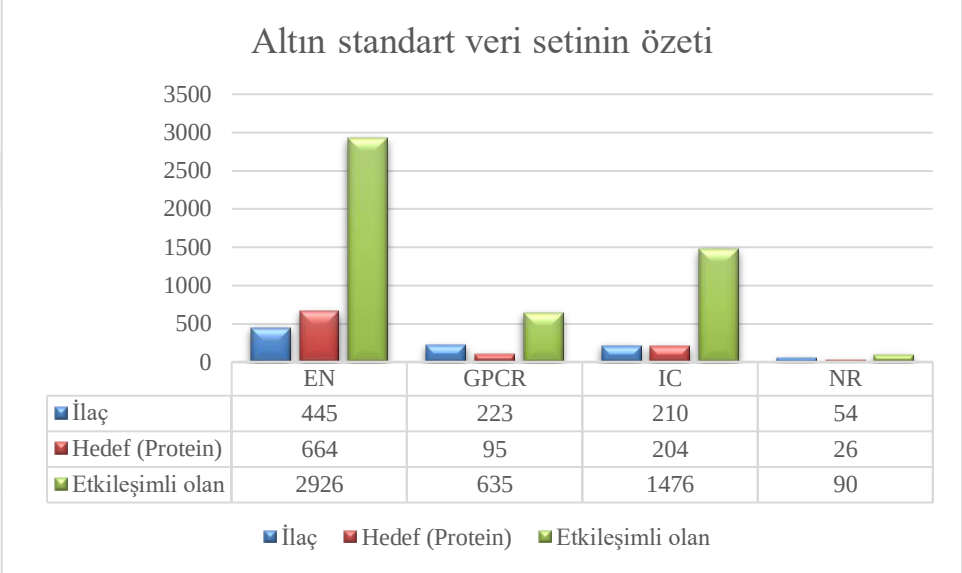


Şekil 3. 1: İlaç-hedef etkileşimlerini tahmin etmek için önerilen modelin süreç akışı (Yazar tarafından oluşturulmuştur).

3.2 İLAÇ-HEDEF VERI KÜMELERI

Bu çalışmada, ilaçlar, hedefler (proteinler) ve ilaç-hedef etkileşimleriyle ilgili veriler, Yamanishi ve arkadaşları (Yamanishi ve diğ., 2008) tarafından derlenen veri kümelerinden alınmıştır. Bu veri kümeleri, güncel ve herkese açık olan KEGG BRITE (Kanehisa ve diğ., 2012) , BRENDA (Schomburg ve diğ., 2004) , SuperTarget (Günther ve diğ., 2008) ve DrugBank (Wishart ve diğ., 2006) veritabanlarından türetilmiş olup, literatürde "altın standart veri kümesi" olarak kabul edilmektedir.

Altın standart veri kümesi, diğer veri kümelerinin kalitesini değerlendirmek için en doğru ve güvenilir referans noktası olarak kullanılmaktadır. Bu veri kümesi, insanlarda dört farklı ilaç-hedef etkileşim ağı kategorisini içermektedir: enzimler (EN), iyon kanalları (IC), G-protein bağlı reseptörler (GPCR) ve nükleer reseptörler (NR). Bu kategoriler içinde ilaçların yapısal benzerliği, hedef proteinlerin dizilim benzerliği ve ilaç-hedef etkileşim ağının topolojisi arasında önemli bir korelasyon gözlemlenmiştir. Şekil 3.2, bu çalışmada kullanılan altın standart veri kümesinin özetini sunmaktadır.



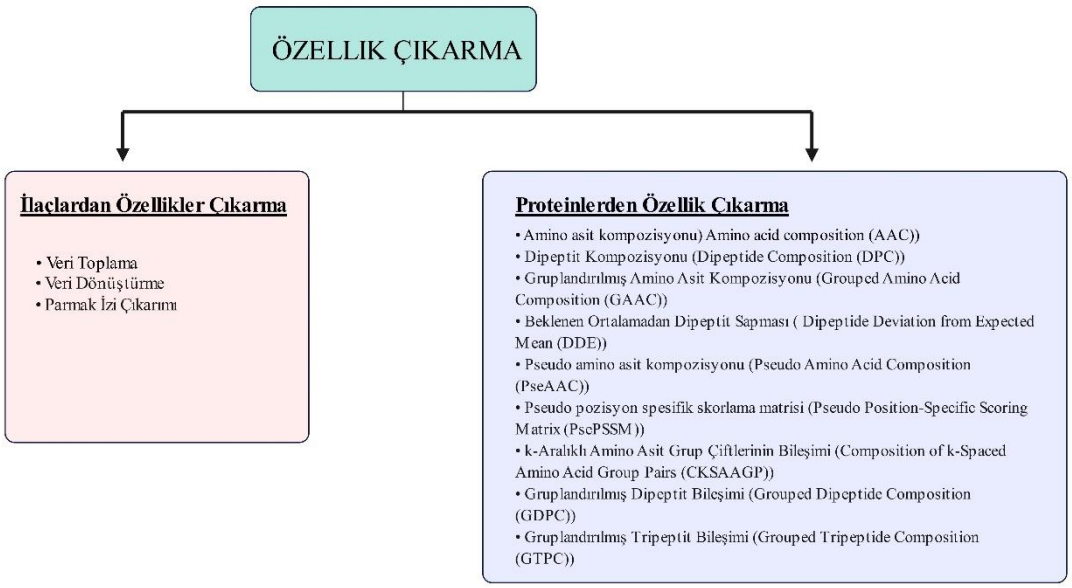
Şekil 3. 2: Altın standart veri kümesinin özeti

(http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/)

3.3 ÖZELLIK ÇIKARMA YÖNTEMLERI

İlaç-protein etkileşimlerinin (DTI) daha doğru ve güvenilir bir şekilde tahmin edilebilmesi için, hem ilaçlara hem de protein hedeflerine ait çeşitli özelliklerin ayrıntılı biçimde çıkarılması büyük önem taşımaktadır. Bu özellik çıkarımı süreci, bilinen etkileşimlere dair mevcut verileri derinleştirerek modelin öğrenme kapasitesini artırmakta ve etkileşim tespitindeki doğruluk oranını önemli ölçüde iyileştirmektedir. Şekil 3.3, ilaçlara ve hedef proteinlere ait on farklı özellik grubunu kapsayan genel bir özet sunmaktadır. Söz konusu özellikler, temel olarak iki ana başlık altında sınıflandırılmıştır: ilaç temelli özellikler ve hedef temelli özellikler. Buna ek olarak, her iki kategori altındaki özellikler, işlevsel benzerliklerine ve yapısal niteliklerine göre çeşitli alt gruplara ayrılmıştır. Bu grupların her biri, etkileşim tahmin modellerinin

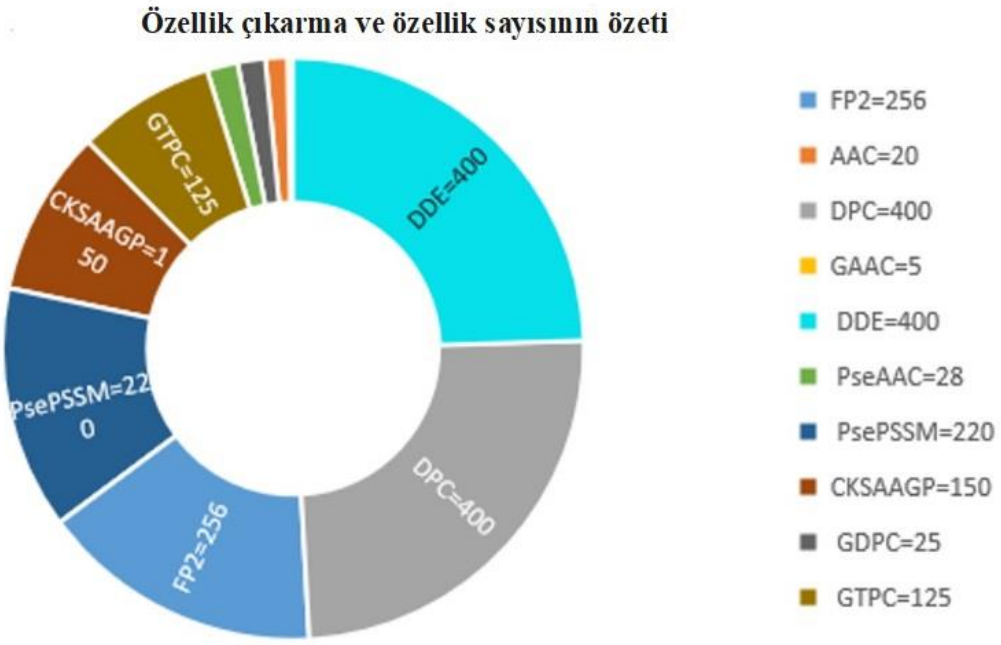
performansına katkı sağlayan özgün ve tamamlayıcı bilgiler içermektedir. Şekil 3.3'te, her bir özellik grubunun içeriği ve modelleme sürecine olan katkıları ayrıntılı biçimde açıklanmaktadır.



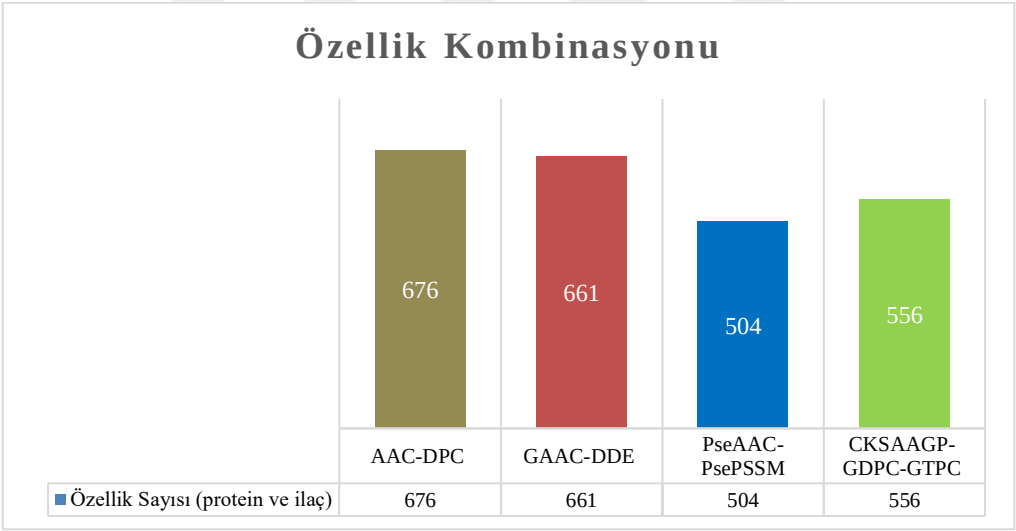
Şekil 3. 3: İlaç–protein etkileşim verilerinde özellik çıkarma yöntemlerinin sınıflandırılması

3.3.1 Veri Çeşitliliğinin Önemi Ve Özelliklerin Alt Kümeleri

İlaç-protein etkileşimlerinin doğru şekilde tahmin edilebilmesi için veri çeşitliliğinin sağlanması büyük önem taşımaktadır. Bu doğrultuda, özellikle proteinlere ait çok boyutlu ve tamamlayıcı bilgileri yansıtacak şekilde dört birleşik özellik alt kümesi oluşturulmuştur. Söz konusu alt kümeler şunlardır: AAC-DPC, GAAC-DDE, PseAAC-PsePSSM ve CKSAAGP-GDPC-GTPC. Bu dört grup, protein dizilimlerinden istatistiksel, kimyasal, evrimsel ve sıralı düzeyde çeşitli bilgilerin çıkarılmasıyla elde edilmiştir. Ayrıca, bu protein temelli özellik kümelerine ilaçlara ait yapısal temsilleri yansıtan moleküler parmak izi tanımlayıcıları da eklenmiştir. Böylece model hem hedef proteine ait derin temsil bilgilerini hem de ilaç moleküllerinin kimyasal yapılarını birlikte değerlendirme yeteneği kazanmıştır. Özellik çıkarımının genel yapısı ve elde edilen toplam özellik sayıları Şekil 3.3'te, dört özellik alt kümesinin yapısı ise Şekil 3.4'te özetlenmektedir.



Şekil 3. 4: Özellik Çıkarımı ve Özellik Sayısının Özeti.



Şekil 3. 5: İlaç ve hedef özelliklerinden türetilen dört özellik alt kümesi.

İlk özellik alt kümesi olan AAC-DPC, istatistiksel ve sıralı bilgileri bir arada sunar. AAC, protein dizilimindeki 20 farklı amino asidin frekansını hesaplayarak genel bileşim bilgisi sağlar. Buna karşılık, DPC, amino asit çiftlerinin sıralı örüntülerini analiz ederek yerel düzeydeki dizilim özelliklerini ortaya koyar. Bu iki tanımlayıcının birleşimi, proteinlerin hem içerik hem de yapı bakımından dengeli bir şekilde temsil edilmesini sağlar.

İkinci özellik birleşimi GAAC-DDE, kimyasal ve fiziksel temsilleri istatistiksel sapmalarla harmanlar. GAAC, amino asitleri belirli gruplara ayırarak dizilimin özetlenmesini sağlar. DDE ise amino asit çiftlerinin gözlenen frekanslarını beklenen değerlerle karşılaştırarak istatistiksel anomalileri ve potansiyel biyolojik işaretleri tespit eder. Bu birleşim, gürültülü veride dahi ayırt edici örüntülerin saptanmasını kolaylaştırır.

Üçüncü grup olan PseAAC-PsePSSM, hem dizilim temelli hem de evrimsel bilgiyi birleştiren güçlü bir yapıya sahiptir. PseAAC, amino asitlerin sıralama düzeni ve fizikokimyasal özelliklerini de hesaba katarak zengin bir temsil sunar. PsePSSM ise konumsal skor matrislerinden türetilen evrimsel bilgileri yansıtır. Bu kombinasyon sayesinde proteinlerin hem yapısal hem de evrimsel düzeydeki bilgileri eş zamanlı olarak modele sunulmuş olur.

Son grup olan CKSAAGP-GDPC-GTPC, sıralı-grupsal temsillerin en kapsamlı halidir. CKSAAGP, belirli mesafelerdeki amino asit çiftlerini analiz ederek kısa ve uzun menzilli dizilim düzenlerini yakalar. GDPC, çift amino asitleri gruplar üzerinden özetlerken, GTPC bu analiz kapsamını üçlü dizilimlere genişleterek daha karmaşık yapıları temsil eder. Bu kombinasyon, dizilimlerin farklı ölçeklerdeki örüntülerini ortaya çıkararak modelin genel öğrenme kapasitesini artırır.

Bu dört birleşik tanımlayıcının seçimi rastgele yapılmamış; aksine, her birinin farklı boyutlarda (istatistiksel, kimyasal, sıralı, evrimsel) tamamlayıcı bilgiler sunduğu dikkate alınarak planlı biçimde belirlenmiştir. Böylece modellerin öngörü gücü artırılırken, bilgi kaybı en aza indirilmiş ve hesaplama verimliliği de korunmuştur.

3.3.2 İlaçlardan Özellikler Çıkarma

Tıbbi kimya ve sentetik organik kimyadaki gelişmelerle birlikte ilaç moleküllerinin çeşitliliği artmıştır. Kemoinformatik, moleküler parmak izleri kullanarak hızlı karşılaştırmalar yapmayı, yapı-aktivite ilişkisi çalışmalarını ve sanal taramayı kolaylaştırmaktadır (Caron ve diğ., 2020; Naveja ve Franco, 2017). Moleküler parmak izleri, hidrojen atomlarının sayısı ve belirli moleküler alt yapılar gibi farklı karmaşıklık seviyelerindeki özellikleri kodlayan ikili dizilimlerdir (Alpay ve diğ., 2022).

İlaç keşfi ve geliştirme süreçlerinde, ilaçların moleküler yapısını dijital olarak temsil etmek amacıyla FP2, FP3, FP4 ve MACCS gibi parmak izi (fingerprint) tanımlayıcıları yaygın olarak

kullanılmaktadır (Dong ve diğ., 2018). Bu çalışmada, ilaç moleküllerini temsil etmek için FP2 formatı seçilmiştir. Bir ilacın moleküler parmak izini çıkarmak için aşağıdaki adımlar uygulanmıştır:

1. Veri Toplama: İlk olarak, ilacın KEGG veritabanındaki (http://www.kegg.jp/kegg/) kodu kullanılarak, kimyasal yapı bilgilerini içeren Mol dosyası elde edilir.
2. Veri Dönüştürme: Açık kaynaklı OpenBabel yazılımı (http://openbabel.org/) kullanılarak kimyasal veriler farklı formatlara dönüştürülüp işlenir.
3. Parmak İzi Çıkarımı: Son olarak, OpenBabel yazılımı kullanılarak ilacın Mol dosyası FP2 moleküler parmak izi formatına dönüştürülür.

FP2 parmak izi, 256 karakter uzunluğunda onaltılık bir dizi olarak üretilir ve 0 ile 15 arasında değişen ondalık değerlere çevrilebilir. Bu süreç, ilaç molekülünü temsil eden 256 boyutlu bir vektör oluşturur (Shi ve diğ., 2019).

3.3.3 Proteinlerden Özellikler Çıkarma

Proteinler ile ilaçlar arasındaki etkileşimlerin doğru bir şekilde tahmin edilebilmesi, protein dizilimlerinden anlamlı ve ayırt edici özelliklerin çıkarılmasına bağlıdır. Bu bağlamda, bu çalışmada proteinlerin biyolojik, kimyasal, yapısal ve evrimsel özelliklerini temsil edebilecek çok çeşitli tanımlayıcılar kullanılarak özellik çıkarımı gerçekleştirilmiştir. Elde edilen bu özellikler, AAC, DPC, GAAC, DDE, PsePSSM, PseAAC, CKSAAGP, GDPC ve GTPC olmak üzere dokuz farklı tanımlayıcıdan oluşmaktadır. Her bir tanımlayıcı, protein dizisinin farklı bir boyutunu modelleyerek etkileşim tahmini sürecine katkı sağlamaktadır. Bu tanımlayıcıların her biri ve çıkarım yöntemleri aşağıda ayrıntılı olarak açıklanmıştır.

A)Amino asit kompozisyonu (Amino acid composition (AAC))

Amino asit kompozisyonu, protein dizisi analizinde önemli bir özelliktir ve protein dizilerinde 20 doğal amino asidin her birinin sıklığını inceleyen bir yöntemdir. Bu özellik, her boyutun belirli bir amino asidin sıklığına karşılık geldiği 20 boyutlu bir vektör olarak temsil edilir. Bu kompozisyonda dikkate alınan 20 doğal amino asit, ACDEFGHIKLMNPQRSTVWY'yi içerir (Charoenkwan ve diğ., 2021; Guo ve diğ., 2020; Sun ve diğ., 2020; Wang ve diğ., 2011) . Her bir amino asidin sıklığı Formül 3.1 kullanılarak hesaplanır:

$$f_t = \frac{N(t)}{N} \quad , \; t \in \{A, C, D, \dots, Y\} \tag{3.1}$$

Formül 3.1'de , N(t)/N, protein dizisinde amino asit türü t'nin kaç kez tekrarlandığını temsil ederken, N, protein dizisinin toplam uzunluğunu belirtir. Bu özellik, proteinlerdeki belirli desenleri tanımlamakta değerli olduğu kanıtlanmış ve termofilik proteinleri mezofilik proteinlerden ayırt etmek gibi prediktif analizlerde önemli bir rol oynamaktadır. Bu tür kompozisyonlara dayalı özellik çıkarımı ve model oluşturma için iLearnPlus gibi platformların kullanımı, çeşitli çalışmalarda gösterilmiştir (Ai ve diğ., 2018; Guo ve diğ., 2020).

A)Dipeptit Kompozisyonu (Dipeptide Composition (DPC))

Dipeptit Kompozisyonu , peptit dizilerindeki dipeptit çiftlerinin göreli sıklığını hesaplama yöntemidir. Bu yöntemde, amino asitlerin sırası dikkate alınmaz. Bu özellik, özellikle homolog verilerle protein dizisi analizinde faydalıdır ve dipeptit sayımlarını dengelemek için 400'e kadar özellik sunabilir. DPC hesaplamak için kullanılan formül şu şekildedir:

$$DPC(s,t) = \frac{N(s,t)}{L-1} \quad , \quad s,t \; \in \{A, C, D, \dots, Y\} \tag{3.2}$$

Formül 3.2'de , N(s,t), s ve t amino asitlerinin kombinasyonuyla oluşturulan dipeptitlerin sayısını temsil ederken, L peptit dizisinin uzunluğudur (Chen ve diğ., 2018; Saravanan ve Gautham, 2015). Bu özelliğin dahil edilmesi, analizlerin doğruluğunu ve verimliliğini artırır ve dizileri anlamlı özelliklere dönüştürmeye yardımcı olur. Ayrıca, amino asitlerin fiziksel ve kimyasal özellikleri, alifatik gruplar (G1: A, G, L, I, M, V), aromatik gruplar (G2: F, W, Y), pozitif yüklü gruplar (G3: H, R, K), negatif yüklü gruplar (G4: D, E) ve yüklü olmayan gruplar (G5: C, N, P, Q, S, T) analizlerde tamamlayıcı kriterler olarak kullanılır (Lee ve diğ., 2011). Bu yaklaşım, ilaç keşfi, protein mühendisliği ve proteinlerin fonksiyonel bölgelerinin tanımlanmasında geniş bir uygulama alanına sahiptir, biyolojik aktiviteler ve proteinlerin fonksiyonel özelliklerinin anlaşılmasına katkı sağlar.

B) Gruplandırılmış Amino Asit Kompozisyonu (Grouped Amino Acid Composition (GAAC))

Genetik Amino Asit Kompozisyonu deskriptoru, amino asitlerin hidroforbilite, moleküler boyut ve yük gibi fiziko-kimyasal özelliklerine dayalı olarak geliştirilmiş bir araçtır. Bu özelliklere göre, 20 tür amino asit beş ana gruba ayrılır: alifatik grup (g1: GAVLMI), aromatik

grup (g2: FYW), pozitif yüklü grup (g3: KRH), negatif yüklü grup (g4: DE) ve nötr grup (g5: STCPNQ). Bu sınıflandırma, bu gruplar arasındaki karmaşık ve dinamik etkileşimlere dayalıdır ve peptitlerin yapı ve fonksiyonunda önemli rol oynar. Her amino asit, tek harfli kodu (örneğin, glisin için G, alanin için A vb.) ile tanımlanır ve her amino asit grubunun sıklığı, GAAC deskriptoru olarak hesaplanır. Bu deskriptor, her grup içindeki amino asitlerin sayısının protein dizisinin toplam uzunluğuna oranını temsil eder (Dutta ve diğ., 2024) . Bu yaklaşım, protein dizileriyle ilgili sınıflandırma ve tahmin görevlerinde derin öğrenmede özellikle faydalıdır ve polipeptitlerin yapı ve fonksiyonel özelliklerinin daha derin bir şekilde anlaşılmasına yardımcı olur (Khawaja ve diğ., 2024) .

$$f(g) = \frac{N(g)}{N} \quad , \quad t \in \{g1, g2, g3, g4, g5\} \tag{3.3}$$

$$N(g_t) = \sum_{t \in g} N(t) \ , \ t \in g \tag{3.4}$$

Formül 3.3 , 3.4'te :

- N(g) grup g'deki amino asitlerin sayısını temsil eder,
- N(t) amino asit türü t'nin sayısını temsil eder,
- N ise protein dizisinin uzunluğudur.

C) Beklenen Ortalamadan Dipeptit Sapması (Dipeptide Deviation from Expected Mean (DDE))

DDE özellik vektörü, üç parametreyi hesaplayarak oluşturulur: dipeptit kompozisyonu (Dc), teorik ortalama (Tm) ve teorik varyans (Tv). Bu kodlama yöntemi, protein dizilerindeki dipeptitlerin kompozisyonel özelliklerini, her bir dipeptitin gözlemlenen sıklığının beklenen ortalama değerinden sapmasını ölçerek değerlendirir. Bu işlem, 400 boyutlu bir özellik vektörü oluşturur. Bu vektör, dipeptitlerin kompozisyonel özelliklerini etkili bir şekilde yakalar, çünkü 20 standart amino asit tarafından oluşturulan 400 olası dipeptiti kapsar (Yan ve diğ., 2024). Hesaplama prosedürü aşağıdaki adımları içerir:

$$Dc(i) = \frac{n_i}{N} \tag{3.5}$$

$$TM(i) = \frac{C_{i1}}{CN} \times \frac{C_{i2}}{CN} \tag{3.6}$$

$$TV(i) = \frac{TM(i)\times(1-TM(i))}{N} \quad (3.7)$$

$$\mathrm{DDE(i)} = \frac{DC(i)-TM(i)}{\sqrt{TV(i)}} \quad (3.8)$$

$$DDE_P = \{DDE(1), DDE(2), \ldots, DDE(400)\} \quad (3.9)$$

Bu bağlamda, Ni, dizideki dipeptitinin sıklığını temsil ederken, N dizideki toplam dipeptit sayısını (dizi uzunluğundan 1 çıkarılarak hesaplanır) ifade eder. C_{i1} ve C_{i2}, sırasıyla dipeptit i'nin birinci ve ikinci amino asitlerini kodlayan kodon(codon) sayısını ifade eder. CN, üç Sonlandırma kodonu hariç olmak üzere toplam kodon sayısını (yani, 61 kodon) temsil eder (Arif ve diğ., 2020).

D) Pseudo amino asit kompozisyonu (Pseudo Amino Acid Composition (PseAAC))

Pseudo amino asit kompozisyonu, proteinin hücre içi lokalizasyonu ve membran protein türlerinin tahminini geliştirmek amacıyla tanıtılmıştır (Jiang ve diğ., 2021). Bu kompozisyon, bir protein dizisinin sıralama bilgilerini kaybetmeden, onu ayrık bir model olarak temsil edilmesini sağlar. PseAA kompozisyonu, 20'den fazla ayrık faktörden oluşur, ilk 20 faktör geleneksel amino asit kompozisyonunu temsil ederken, ek faktörler, sıralama bilgilerini korelasyonlar veya diğer ilişkili faktörler aracılığıyla entegre eder (Jan ve diğ., 2022). PseAA kompozisyonunun matematiksel modeli şu şekilde temsil edilir.

$$P = [p_1, \ldots p_{20}, p_{20+1}, \ldots p_{20+\lambda},]^{T} \quad (3.10)$$

Formül 3.10'da , P1, P2, ..., ve P20, 20 standart amino asidin normalize edilmiş değerlerine karşılık gelir. Ayrıca, korelasyon faktörleri, her amino asidin hidrofobiklik, hidrofiliklik, sertlik, esneklik ve yerine geçebilirlik özelliklerine dayalı olarak tanımlanır.

$$\tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} J_{i,i+1}$$

$$\tau_2 = \frac{1}{L-2}\sum_{i=1}^{L-2} J_{i,i+2}$$

$$\tau_3 = \frac{1}{L-3}\sum_{i=1}^{L-3} J_{i,i+3}$$

$$\tau_4 = \frac{1}{L-4}\sum_{i=1}^{L-4} J_{i,i+4}$$

$$\tau_5 = \frac{1}{L-5}\sum_{i=1}^{L-5} J_{i,i+5}$$

… … … ..

$$\tau_\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} J_{i,i+\lambda} \quad (3.11)$$

Formül 3.11'de , L protein dizisinin uzunluğunu ve birinci, ikinci ve son sıralama korelasyon faktörlerini temsil eder. Ampirik gözlemlerden yola çıkarak, λ=21 olarak seçilmiştir, yani PseAA'nın ilk 21 sıralama korelasyon derecesini dikkate alıyoruz (Jan ve diğ., 2022).

E) Pseudo pozisyon spesifik skorlama matrisi (Pseudo Position-Specific Scoring Matrix (PsePSSM))

Bu çalışmada, submitokondriyal protein dizileri FASTA formatında analiz edilmiş ve PSI-BLAST programı kullanılarak değerlendirilmiştir. Analiz sırasında, NCBI tarafından sağlanan ve 85 milyondan fazla protein dizisi içeren tekrar etmeyen (non-redundant) protein veritabanı (nr) referans olarak kullanılmıştır. Bu veritabanına ftp://ftp.ncbi.nih.gov/blast/db/nr adresinden erişilebilmektedir. PSI-BLAST analizi, 3 iterasyon ve 0.001 e-değeri eşik değeri ile gerçekleştirilmiş, diğer parametreler ise varsayılan ayarlarda bırakılmıştır. Bu ayarlarla her bir protein dizisi için pozisyon-spesifik skorlama matrisleri (Position-Specific Scoring Matrix (PSSM)) elde edilmiştir. Böylece proteinlerin dizi benzerlikleri ve işlevsel özellikleri hakkında daha derinlemesine bilgi edinilmiştir (Altschul ve diğ., 1997).

$$P_{PSSM} = \begin{bmatrix} M_{1,1} & M_{1,2} & \dots & M_{1,20} \\ M_{2,1} & M_{2,2} & \dots & M_{2,20} \\ . & . & . & . \\ . & . & . & . \\ M_{i,1} & M_{i,2} & \dots & M_{i,20} \\ . & . & . & . \\ . & . & . & . \\ M_{L,1} & M_{L,2} & \dots & M_{L,20} \end{bmatrix} \quad (3.12)$$

Formül 3.12'de ,PPSSM matrisinde, boyutu L × 20 (L, hedef protein dizisindeki toplam kalıntı sayısıdır) olan her satır, sorgu dizisindeki karşılık gelen pozisyonlardaki amino asit yer değiştirmelerinin log-likelihood skorunu temsil eder. Mi,j, protein dizisinin ith pozisyonundaki bir amino asit kalıntısının, evrimsel süreçte j. amino asit türüne değiştirilmesi için skoru gösterir. PSSM değerleri, protein dizileri için -9 ile 11 arasında değişir. İndeksin değişen boyutları nedeniyle istatistiksel analizde herhangi bir önyargıyı önlemek amacıyla, bu makalede PSSM matrisi normalize edilmiştir. Matrisin içindeki değerler, aşağıdaki denklem kullanılarak 0 ile 1 arasında ölçeklenir:

$$f(x) = \frac{1}{1+ e^{-x}} \tag{3.13}$$

Proteinlerin dizilim verilerini kullanarak ve Chou'nun PseAAC kavramını PSSM'ye uygulayarak, bir proteinin PPSSM'si bir özellik vektörüne dönüştürülür, bu vektör PPsePSSM olarak adlandırılır ve protein dizisini temsil eder. Her bir protein dizisinin uzunluğu (L) farklı olabileceğinden, her dizilim L × 20 boyutlarında bir matrisle sonuçlanır. Farklı protein dizilerinden elde edilen PSSM matrislerinde tutarlı boyutlar sağlamak amacıyla, bu matrislerin aynı uzunlukta vektörlere dönüştürülmesi gerekir. Bu işlem, Formül 3.14 ve 3.15 gösterildiği üzere PsePSSM'nin elde edilmesini sağlar.

$$\bar{M}_j = \frac{1}{L} \sum_{i=1}^{L} M_{i,j} (j = 1,2,\dots,20) \tag{3.14}$$

$$G_j^{\xi} = \frac{1}{L-\xi} \sum_{i=1}^{L-\xi} [M_{i,j} - M_{(i+\xi),j}]^2 \quad (j = 1,2,\dots,20 \, , \xi < \mathrm{L}) \tag{3.15}$$

Sonuç olarak, PsePSSM yöntemi protein dizilerinden özellikler çıkarmak için kullanılır Formüle 3.16 ulaşılır.

$$P_{PsePSSM} = \left[\bar{M}_1 \bar{M}_2 \dots \bar{M}_{20} G_1^{\xi} G_2^{\xi} \dots G_{20}^{\xi}\right]^T (\xi = 0,1,2, \dots, \mathrm{L}-1) \tag{3.16}$$

Formül 3.16'da , $\bar{M}$ j, bir protein dizisinin ortalama skorunu ifade eder; burada, i. pozisyondaki amino asit kalıntısı evrimsel süreçte j türdeki amino aside mutasyona uğrar. G^{ξ} j, amino asit kalıntılarının PSSM skorları arasındaki ilişkiyi temsil eden korelasyon faktörüdür ve bu kalıntılar i − 1 pozisyonları ile ayrılmıştır. ξ'nin değeri kullanılan veri setine bağlı olarak değişir. ξ = 0 olduğunda, PPsePSSM 20 boyutlu bir özellik vektörüne dönüştürülür, bu da Formül 3.9'da

gösterilmiştir. Bir protein dizisi daha sonra 20 + 20 × ξ boyutlarında bir vektör olarak temsil edilebilir ve bu, protein bileşimini yakalarken sıralama bilgilerini de korur (Qiu ve diğ., 2018).

F) k-Aralıklı Amino Asit Grup Çiftlerinin Bileşimi (Composition of k-Spaced Amino Acid Group Pairs (CKSAAGP))

CKSAAP yöntemi, k amino asit uzaklıkla ayrılan amino asit çiftlerinin oluşum sıklığını belirlemek için kullanıldı. Örneğin, k = 0 olduğunda, arasında boşluk bulunmayan kalıntı çiftleri şu şekilde gösterilir: AA, AC, AD, ..., YY. Eğer k = 1 ise, bir amino asitle ayrılan kalıntı çiftleri şu şekilde gösterilir: AxA, AxC, AxD, ..., YxY. CKSAAP'ler aşağıdaki şekilde tanımlanır:

$$k = 0 \left(\frac{N[AA]}{N_0}, \frac{N[AC]}{N_0}, \frac{N[AD]}{N_0}, \ldots\ldots\ldots, \frac{N[YY]}{N_0}\right) 400$$

$$k = 1 \left(\frac{N[AxA]}{N_1}, \frac{N[AxC]}{N_1}, \frac{N[AxD]}{N_1}, \ldots\ldots\ldots, \frac{N[YxY]}{N_1}\right) 400$$

$$k = 2 \left(\frac{N[AxxA]}{N_2}, \frac{N[AxxC]}{N_2}, \frac{N[AxxD]}{N_2}, \ldots\ldots\ldots, \frac{N[YxxY]}{N_2}\right) 400 \quad (3.17)$$

Formül 3.17'de , "x" herhangi bir 20 amino asitten birini temsil eder. N_k, N_k=L−(k+1)) formülüyle hesaplanır; burada L, protein dizisinin uzunluğunu ifade eder ve k, 1, 2, 3 vb. değerler alabilir. Son özellik vektörünü elde etmek için bireysel özellik vektörleri birleştirilir. Örneğin, k = 5 olduğunda, ortaya çıkan vektör 2.400 boyuta sahip olacaktır (400 × 6).

Amino asit kalıntıları, yan zincirlerinin kimyasal özelliklerine göre beş gruba ayrılır:

- Alifatik (g1: GAVLMI)
- Aromatik (g2: FYW)
- Pozitif yüklü (g3: KRH)
- Negatif yüklü (g4: DE)
- Nötr (g5: STCPNQ)

CKSAAP yöntemi , bu grupların k amino asit aralığında ayrılmış çiftlerinin frekansına dayanarak türetilir(Wang ve diğ., 2021) . k = 0 olduğunda, aralarında boşluk bulunmayan grup çiftleri Formül 3.18'de gösterilir:

$$k = 0 \left(\frac{N[g1g1]}{N_0}, \frac{N[g1g2]}{N_0}, \frac{N[g1g3]}{N_0}, \ldots\ldots\ldots, \frac{N[g5g5]}{N_0}\right) 25 \quad (3.18)$$

G) Gruplandırılmış Dipeptit Bileşimi (Grouped Dipeptide Composition (GDPC))

GDPC yöntemi kodlama tekniği, geleneksel Dipeptit Bileşimi yönteminin gelişmiş bir versiyonudur (Chen ve diğ., 2022) . Bu teknik, amino asitleri fizikokimyasal özelliklerine göre beş kategoriye ayırır:

- **Aromatik:** W, Y, F
- **Pozitif yüklü:** H, K, R
- **Alifatik:** A, I, M, G, L, V
- **Yüksüz:** C, T, P, S, Q, N
- **Negatif yüklü:** D, E, G

GDPC yöntemi, bu gruplar tarafından oluşturulan dipeptitlerin oluşum frekanslarını hesaplayarak 25 farklı tanımlayıcı içeren bir özellik kümesi oluşturur. Bu tanımlayıcılar Formül 3.19 kullanılarak belirlenir:

$$GDPC = \frac{N_{R_i R_j}}{L-1} \tag{3.19}$$

Formül 3.19'da , N_{RiRj} $(1 \leq i,j \leq 5)$, belirli bir kalıntı çifti RiRj'nin dizide kaç kez geçtiğini ifade eder. Burada Ri ve Rj , önceden tanımlanmış beş gruptan birine aittir. L, protein veya peptit dizisinin uzunluğunu temsil eder. Beş gruptan gelen kalıntı çiftlerinin 5 × 5 = 25 farklı kombinasyonu olduğu için, GDPC yöntemi 25 boyutlu bir özellik vektörü üretir (Yan ve diğ., 2024) .

GDPC yöntemi, amino asitleri kimyasal özelliklerine göre sınıflandırmakla kalmaz, aynı zamanda bu gruplar arasındaki etkileşimleri göz önünde bulundurarak özellik çıkarma işleminin karmaşıklığını azaltır. Bu yöntem, protein dizileri içindeki yapısal ve yerel dizilim desenlerini keşfetmeye olanak tanırken, hesaplama yükünü makul bir seviyede tutar (Sikander ve diğ., 2022).

H) Gruplandırılmış Tripeptit Bileşimi (Grouped Tripeptide Composition (GTPC))

GTPC yöntemi, tripeptitleri ortak fizikokimyasal özelliklerine göre kategorilere ayırarak analiz sürecini geliştirir. Bu yaklaşım, protein dizilerindeki daha yüksek düzeydeki yapısal ve fonksiyonel desenleri belirlemeye olanak tanıyarak, biyolojik rollerine dair daha derin bir anlayış sunar (Ferdous ve diğ., 2024) .

GTPC yöntemi, protein dizilerini 125 boyutlu bir özellik vektörü olarak temsil eder. Her boyut, üç ardışık kalıntıdan oluşan tripeptitlerin, önceden tanımlanmış sınıflara yapısal benzerlikleri veya fizikokimyasal özelliklerine göre gruplanmasına karşılık gelir (Ferdous ve diğ., 2024). Bu sınıflandırma, amino asitlerin dizilim düzenini kapsamlı bir şekilde analiz ederek, polipeptit zincirlerindeki üç ardışık kalıntıdan türetilen desenlere odaklanır (Khawaja ve diğ., 2024) .

Derin öğrenme tekniklerinden faydalanarak, GTPC yöntemi tripeptitlerin fonksiyonel özelliklerini hassas bir şekilde tahmin etmeyi sağlar. Bu sayede, proteinlerin biyolojik aktivitesi ve tedavi edici potansiyeli hakkında önemli içgörüler sunar. Yenilikçi yaklaşımı sayesinde, proteinlerde işlevsel olarak kritik bölgelerin belirlenmesinde etkili olduğu kanıtlanmış ve ilaç keşfi ile protein mühendisliği alanlarında önemli bir potansiyele sahip olduğu gösterilmiştir (Khawaja ve diğ., 2024) . GTPC yöntemi aşağıdaki şekilde ifade edilir:

$$f(r,s) = \frac{N_{rst}}{N-2} \quad ,r,s,t \in \{g_1, g_2, g_3, g_4, g_5\} \tag{3.20}$$

Formül 3.20' de , N_{rst}, r ,s ve t tipi amino asitlerden oluşan dipeptitlerin sayısını ifade ederken, N protein dizisinin toplam uzunluğunu temsil eder.

3.4 TEMEL BİLEŞENLER ANALİZİ (PRINCIPAL COMPONENTS ANALYSİS (PCA))

Temel Bileşen Analizi , İkinci Dünya Savaşı'ndan önce tanıtılmış en eski ve yaygın çok değişkenli istatistik yöntemlerinden biridir; ancak 1960'lı yıllardan itibaren, doğal ve sosyal bilimlerdeki "nicel devrim" ile birlikte geniş çapta kullanılmaya başlanmıştır (Jolliffe ve diğ., 2016). PCA, boyut indirgeme yöntemleri ailesine aittir ve özellikle veri setinde çok sayıda değişken, yüksek gözlem hacmi ve değişkenler arasında yüksek korelasyon olduğunda kullanışlıdır. Bu yöntemin temel amacı, orijinal verileri düşük boyutlu bir alt uzayda en az bilgi kaybıyla temsil edebilecek daha küçük bir özellik kümesini belirlemektir. Verileri özetlemenin yanı sıra, PCA bireysel farklılıklar hakkında değerli bilgiler çıkarma imkânı da sunar ve bu nedenle büyük veri çağında ve kişiselleştirilmiş tıpta özel bir öneme sahiptir (Gewers ve diğ., 2021).

PCA, verileri yüksek boyuttan daha düşük boyutlara dönüştürür ve mümkün olan en fazla bilgiyi korurken hesaplama karmaşıklığını azaltır. Bu yöntem, genetik ve biyoinformatik

alanında gen ekspresyon verilerinin analizinde, hastalık alt türlerinin tanımlanmasında, genler arasındaki ilişkilerin anlaşılmasında ve yüksek boyutlu biyolojik verilerde gürültünün azaltılmasında kullanılır (Greenacre ve diğ., 2022).

PCA'nın avantajları arasında verilerin basitleştirilmesi ve algoritma verimliliğinin artırılması amacıyla boyut indirgeme, verilerdeki değişimlerden anlamlı ve etkili özelliklerin belirlenmesini sağlayan özellik çıkarımı, alakasız değişikliklerin ortadan kaldırılması ve sinyal/gürültü oranının iyileştirilmesini sağlayan gürültü azaltma, çok boyutlu verilerin iki veya üç boyutlu alanlarda gösterimi ile desenlerin ve kümelerin görünür hale getirilmesini sağlayan görselleştirme ve değişkenler arasındaki güçlü korelasyon sorunlarının çözülmesini sağlayan çoklu doğrusal ilişki tespiti yer almaktadır. Öte yandan, PCA'nın sınırlamaları arasında non-lineer ilişkilerde zayıf performans gösteren doğrusal varsayım, bileşenlerin orijinal değişkenlerin birleşimi olması nedeniyle yorumlamasının zor olması, aykırı verilere karşı duyarlılık, boyut indirgeme sürecinde kaçınılmaz olan bilgi kaybı ve bileşen sayısının seçiminin öznel bir karar olması nedeniyle eksik veya gereksiz karmaşıklığa yol açabilmesi bulunmaktadır (Kherif ve Latypova, 2020).

PCA, verilerdeki ortak gizli yapıyı tanımlayarak, veri değişkenliğinin büyük bir kısmını özetleyen az sayıda ana bileşen çıkaran parametrik olmayan bir yöntemdir (Aït-Sahalia ve Xiu, 2019). Bu bileşenler, diğer öğrenme modellerinde, örneğin Ana Bileşen Regresyonu gibi, orijinal değişkenlerin yerine kullanılabilir. PCA, birbirine bağlı değişkenleri birbirine dik (ortogonal) bir değişken setine dönüştürür; ana bilgi çoğunlukla ilk bileşende yoğunlaşır ve sonraki bileşenler azalan sırayla kalan bilgiyi açıklar (Shafizadeh-Moghadam, 2021).

Bu yapı, PCA'yı boyut indirgeme, özellik çıkarımı, veri sıkıştırma ve karmaşık veri görselleştirme için güçlü bir araç haline getirir ve veri bilimi ile biyoinformatik alanlarında geniş bir uygulama alanı sunar.

3.5 VERI DENGELEME TEKNIĞI

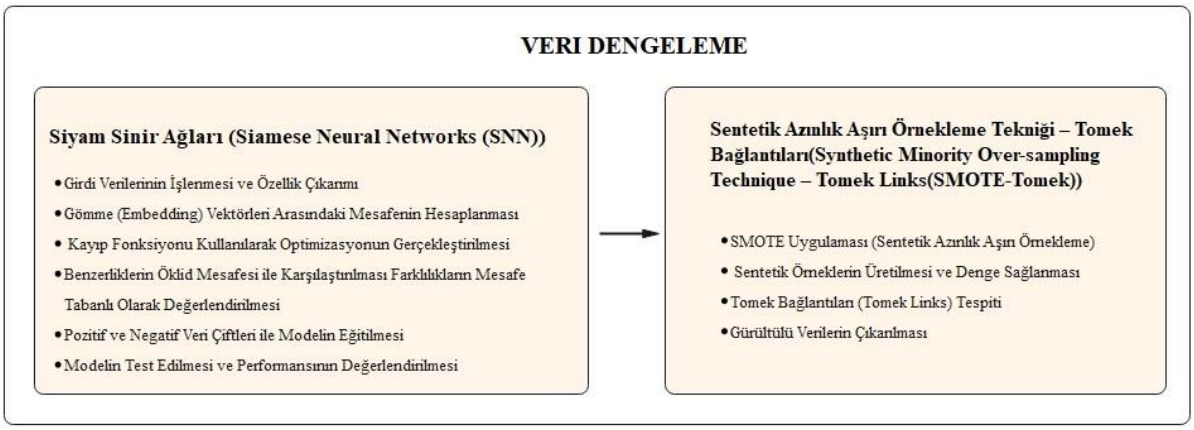
Bilişimsel biyolojide ilaç-hedef etkileşimlerini tahmin etmede karşılaşılan en büyük zorluklardan biri, veri setindeki etkileşimli (pozitif) ve etkileşimsiz (negatif) çiftler arasındaki dengesizliktir. Genellikle etkileşimli çiftlerin sayısı daha az olduğundan, bu durum modellerin etkileşimsiz örneklere yönelmesine ve etkileşimli çiftleri doğru tahmin edememesine neden

olur. Bu nedenle, modelin doğruluğunu artırmak için veri setinin dengelenmesi ve modelin uygun şekilde eğitilmesi büyük önem taşır (Redkar ve diğ., 2020).

Sınıfları dengelemek için yaygın olarak rastgele örnekleme gibi yöntemler kullanılır. Ancak bu yöntemler, genellikle çoğunluk sınıfından belirli bir kısmın çıkarılmasını gerektirir ve bu da modelin genel doğruluğunu olumsuz etkileyebilir. Bu nedenle, modelin performansını artırmak için yalnızca veri kümesini dengelemek değil, aynı zamanda mevcut bilgiyi koruyarak etkileşimli verileri etkin bir şekilde kullanmak da kritik önem taşımaktadır (Ezzat ve diğ., 2016).

Veri dengesizliği sorununu çözmek için üç temel yaklaşım kullanılmaktadır. Maliyet-duyarlı yaklaşımda, eğitim sürecinde çoğunluk sınıfının etkisi azaltılarak her iki sınıf için yanlış sınıflandırma maliyetleri hesaplanır ve azınlık sınıfının daha iyi temsil edilmesi sağlanır. Algoritma-düzeyinde yaklaşımda ise model, az rastlanan örnekleri daha fazla vurgulayacak şekilde ayarlanır ve böylece azınlık sınıfı örneklerine daha fazla ağırlık verilerek dengesizlik azaltılır. Veri-düzeyinde işleme yaklaşımında ise azaltmalı örnekleme ile çoğunluk sınıfındaki bazı örnekler çıkarılarak denge sağlanırken, bu yöntem yararlı verilerin kaybına neden olabilir; öte yandan, arttırmalı örnekleme ile azınlık sınıfındaki örnekler çoğaltılarak denge kurulsa da, bu durum modelin aşırı öğrenmesine yol açabilir (Javaid ve diğ., 2021).

Bu çalışmada önerilen yöntem, Şekil 3.6'da gösterildiği üzere, DTI tahmininde veri dengesizliğine karşı etkili ve bütüncül bir çözüm sunmaktadır. Özellikle karmaşık biyolojik verilerle başa çıkmak üzere tasarlanmış olan Siamese Ağı, örnekler arasındaki derin benzerlik ve farklılıkları mesafe metriklerine dayalı olarak analiz ederek, dengesiz veri kümesini daha dengeli bir yapıya (örneğin 1:3 oranına) dönüştürmede etkili olmuştur. Bu aşamanın ardından uygulanan SMOTE-Tomek algoritması ise, bir yandan azınlık sınıf için sentetik örnekler üreterek temsil gücünü artırmakta, diğer yandan Tomek Links yöntemiyle çoğunluk sınıfındaki örtüşen ve gürültülü verileri ayıklayarak veri setinin saflığını ve denge düzeyini daha da iyileştirmektedir. Böylece, yalnızca sınıflar arası dağılım dengelenmekle kalmamış, aynı zamanda örneklerin kalitesi ve modelin genellenebilirliği de kayda değer biçimde artmıştır. Bu iki yöntemin birleşimi, modelin doğruluk oranını ve performansını belirgin ölçüde yükseltmiş; ayrıca, bu alanda ender rastlanan Siamese Ağı'nın entegrasyonu, araştırmaya hem yöntemsel hem de sonuçsal açıdan özgünlük kazandırarak önemli bir katkı sunmuştur.



Şekil 3. 6: DTI tahmininde veri dengesizliğine çözüm tekniği

3.5.1 Siyam Sinir Ağları (Siamese Neural Networks (SNN))

Siyam Sinir Ağları , veriler arasındaki benzerliği öğrenmek amacıyla tasarlanmış özel bir yapay sinir ağı mimarisidir. “Siyam” terimi, özdeş ikizler kavramından türetilmiş olup, bu ağların simetrik yapısına ve ağırlık paylaşımına atıfta bulunur . Bu mimari genellikle aynı yapıya ve aynı ağırlıklara sahip iki veya daha fazla alt ağdan oluşur. Böyle bir yapı, giriş verilerinin bağımsız olarak işlenmesini ve elde edilen çıktılarının ortak bir gömme (embedding) uzayında karşılaştırılmasını mümkün kılar (Ilina ve diğ., 2022).

Bu ağların klasik mimarilere göre en önemli avantajlarından biri, az miktarda etiketli veriyle etkili bir şekilde öğrenme yapabilmesidir. Geleneksel sınıflandırma modelleri her sınıf için çok sayıda örneğe ihtiyaç duyarken, Siyam Sinir Ağları veri çiftlerinden doğrudan benzerlik öğrenerek bu ihtiyacı azaltır .Bu çift tabanlı yaklaşım sayesinde SNN, yalnızca daha önce görülen sınıflarla değil, aynı zamanda hiç karşılaşılmamış veya sınırlı verilere sahip kategorilerle de başarılı bir şekilde çalışabilir. Bu özellik, sınırlı veri bulunan ya da çok sayıda sınıf içeren problemler için SNN’yi vazgeçilmez bir araç haline getirir (Ondrasovic ve Tarabek, 2021).

Bu ağlardaki öğrenme süreci tamamen uçtan uca (end-to-end) gerçekleşir. Eğitim sırasında ağ, giriş verilerinin özelliklerini çıkarırken aynı zamanda bu veriler arasındaki benzerliği de öğrenir (Li ve diğ., 2023). Bu amaçla genellikle Öklid yen mesafesi veya kontrastif kayıp (Contrastive Loss) gibi mesafe ölçümleri kullanılır. Eğer iki giriş aynı sınıfa aitse, ağ bu örneklerin çıktılarının gömme uzayında birbirine yakın olmasını sağlar; aksi halde aralarındaki mesafeyi

artırır . Bu yöntem, yalnızca sınıf farklarını ayırt etmeyi değil, aynı zamanda bu farkları anlamsal olarak da temsil etmeyi mümkün kılar (Sahito ve diğ., 2019).

Siyam ağlarının kullanım alanı oldukça geniştir. Yüz tanıma, nesne takibi, görüntü eşleştirme, panoramik görüntü birleştirme, anormallik tespiti, öneri sistemleri ve anlamsal benzerlik analizi gibi pek çok alanda başarıyla kullanılmaktadır. Bu mimari ilk kez 1993 yılında Bromley ve arkadaşları tarafından dijital imza doğrulama amacıyla önerilmiştir. O tarihten bu yana, derin öğrenme tekniklerinin gelişmesi ve artan hesaplama gücü sayesinde biyoloji, kimya, eczacılık ve bilgisayarla görü gibi birçok alanda yaygın olarak uygulanmaktadır (Chicco, 2021) .

Dikkat edilmesi gereken bir diğer önemli nokta ise, Siyam mimarisinde kullanılan alt ağların sadece basit yapay sinir ağları ile sınırlı olmamasıdır. Özellikle görüntü işleme gibi karmaşık görevlerde, verilerden daha derin özellikler çıkarabilmek adına Evrişimli Sinir Ağları (Convolutional Neural Networks(CNN)) veya Çok Katmanlı Algılayıcılar (Multilayer Perceptrons(MLP)) tercih edilmektedir (Chicco, 2021). Bu yüksek düzeydeki esneklik, Siyam Sinir Ağları'nı yalnızca temiz ve düzenli veri kümeleri ile değil; aynı zamanda gürültülü, sınırlı ya da daha önce karşılaşılmamış verilerle başa çıkabilen güçlü bir çözüm haline getirir (Bansal ve diğ., 2025) .

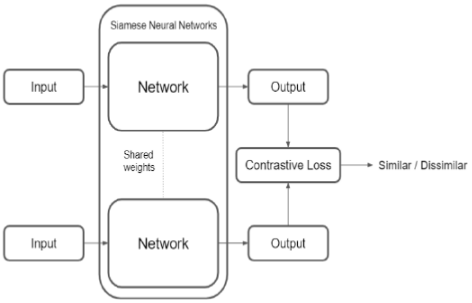
3.5.1.1 Siyam Sinir Ağı Mimarisi

İki giriş örneğini karşılaştırmak ve bunların aynı sınıfa ait olup olmadığını belirlemek amacıyla bir Siyam Sinir Ağı tasarlanmış ve eğitilmiştir (Chicco, 2021) .Bu amaçla klasik sinir ağlarından farklı olarak, özel bir mimari yapıya ihtiyaç duyulmuştur. Bu ağın yapısı, klasik sinir ağı mimarilerinden belirgin şekilde farklıdır. Şekil 3.5'te görüldüğü gibi, Siyam ağı genellikle iki özdeş sinir ağından oluşur; bu alt ağlar çoğunlukla evrişimli sinir ağları biçimindedir ve çıktıları sonunda bir kayıp fonksiyonuna yönlendirilir. Bu fonksiyon, ağın eğitim sürecinde kritik bir rol üstlenir. Çünkü nihai çıktının hedef değere ne kadar yakın olduğunu belirleyerek, daha doğru tahminler elde edebilmek için ağın katmanlarındaki ağırlık ve parametrelerin nasıl güncellenmesi gerektiğine yön verir (Chen, ve diğ., 2022).

Kayıp fonksiyonu, özünde modelin öğrenme hedefini temsil eder; yani modelin eğitim boyunca minimize etmeye çalıştığı değerdir. Örneğin, Çapraz Entropi (Cross-Entropy) kayıp fonksiyonu, modelin tahminlerindeki hata oranını ölçmede en yaygın kullanılan yöntemlerden biridir. Bu fonksiyonun değeri düştükçe, modelin doğruluk oranı artar ve teşhisler daha isabetli

hale gelir. Siyam ağlarında ise genellikle daha özel kayıp fonksiyonları, özellikle de Kontrastif Kayıp (Contrastive Loss) fonksiyonu tercih edilir. Bu fonksiyon, gömme uzayında (embedding space) benzer örneklerin birbirine yakın, farklı örneklerin ise belirli bir mesafede konumlanmasını sağlar (Chicco, 2021) .

Siyam ağlarının eğitilmesindeki nihai amaç, örnekler arasındaki anlamsal ilişkileri doğru biçimde yansıtan, doğrusal olmayan bir yerleştirme alanı oluşturmaktır. Bu alanda, birbirine benzeyen veriler kısa mesafeli vektörlerle temsil edilirken; benzemeyen veriler, aralarında belirli bir uzaklık olacak şekilde konumlandırılır. Bu yapı sayesinde, model yüz tanıma, görüntü eşleştirme ya da imza doğrulama gibi karşılaştırmaya dayalı problemlerde daha tutarlı ve isabetli kararlar alabilir (Habibi ve diğ., 2020).



Şekil 3. 7: Siyam sinir ağı mimarisi (Utkin ve diğ., 2021)

Siyam sinir ağlarında, kontrast hata fonksiyonu Formül 3.21'e göre hesaplanır:

$$L(W,(YX_1,X_2)^i) = (1-Y)L_P\,(E_W\,(X_1.X_2)^i + YL_N(E_W(X_1.X_2)^i) \quad (3.21)$$

Formül 3.21'de :

- $E_W(X_1,X_2)$ girişlerinin yerleştirme (embedding) uzayındaki uzaklığını ifade eder.
- Y etiketi, bu iki örneğin aynı sınıfa ait olup olmadığını belirtir. Eğer Y=0 ise örnekler benzer, Y=1 ise benzer değildir.
- L_P, benzer çiftler (positive pair) için kullanılan kısmi kayıp fonksiyonudur.
- L_N, benzer olmayan çiftler (negative pair) için tanımlanmış kayıp fonksiyonudur.

- Bu yapıda, X_1 genellikle "çapa" (anchor) veri olarak kabul edilirken, X_2'ye yakın bir komşu ya da uzak bir veri noktası olabilir.

Bu kayıp fonksiyonunun temel amacı şudur:

- Eğer iki veri örneği benzerse (Y=0) ve aralarındaki mesafe büyükse, L_P değeri artar ve bu durum modelin bu verileri birbirine daha fazla yakınlaştırması yönünde itici güç oluşturur.
- Tersine, örnekler benzer değilse (Y=1) ve aralarındaki mesafe küçükse, L_N değeri yükselir ve modelin bu verileri birbirinden uzaklaştırması teşvik edilir.

Sonuç olarak, model embedding uzayında benzer verileri birbirine yakın, farklı verileri ise daha uzak yerleştirmeyi öğrenir. Bu mekanizma, yüz tanıma, görüntü eşleştirme, anlamsal analiz ve ürün önerisi gibi benzerlik temelli uygulamalarda son derece etkili bir rol oynar (Chopra ve diğ., 2005).

Ağ ağırlıklarının güncellenmesinde Adam Optimizasyon Algoritması kullanılmaktadır. Bu algoritma, stokastik gradyan inişi yönteminin gelişmiş bir versiyonudur. Çalışma prensibine göre, kayıp fonksiyonunun artması mı yoksa azalması mı gerektiği belirlendikten sonra, ağın parametreleri çıkış katmanından giriş katmanına doğru güncellenir. Bu süreçte, zincir kuralı (chain rule) aracılığıyla türevler geriye doğru aktarılır ve her ileri geçişin (forward pass) ardından ağırlıklar ile önyargılar (bias) yeniden düzenlenir. Adam algoritmasının klasik optimizasyon yöntemlerine kıyasla en belirgin avantajı, her bir parametreye özel adaptif öğrenme hızı belirlemesidir. Bu özellik, modelin eğitim sürecini hem daha kararlı hem de daha verimli hâle getirir (Sahito ve diğ., 2019).

3.5.1.2 Mesafe Fonksiyonu ve Nihai Kayıp Fonksiyonu

Siyam Sinir Ağları'nda , iki giriş örneği arasındaki benzerliğin karşılaştırılması ve değerlendirilmesi temel bir rol oynamaktadır. Bu benzerliğin ölçülmesi için genellikle Öklid mesafesi kullanılır. Bu mesafe fonksiyonu, özdeş alt ağlar tarafından üretilen gömülü (embedding) vektörler arasındaki uzaklığı hesaplar (He ve diğ., 2018). Öklid mesafesi Formül 3.22'e göre tanımlanır:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.22}$$

Pratikte, hesaplama kolaylığı ve türev alma işlemlerinde sadelik sağlamak amacıyla çoğunlukla kök içermeyen hali, yani $d^2(x,y)$ biçimi tercih edilir. Bu formülde, x ve y iki gömme vektörünü, x_i ve y_i bu vektörlerin i'inci bileşenlerini ve n ise vektör boyutunu ifade eder.

Mesafe hesaplandıktan sonra, modelin başarımını değerlendirmek için uygun bir kayıp fonksiyonu kullanılır. Siyam ağlarında bu amaçla genellikle Kontrastif Kayıp Fonksiyonu (Contrastive Loss) tercih edilir. Bu fonksiyon, benzer örneklerin gömme uzayında birbirine yakın, benzer olmayan örneklerin ise birbirinden uzak konumlanmasını sağlayacak şekilde tasarlanmıştır (Ondrasovic ve Tarabek, 2021). Kontrastif kayıp fonksiyonu aşağıdaki şekilde tanımlanır:

$$Loss = Y \cdot d^2 + (1 - Y) \cdot max\,(0\,, m - d)^2 \tag{3.23}$$

Formül 3.23'te, Y iki örnek arasındaki benzerlik etiketini ifade eder; Y=0 ise örnekler benzer, Y=1 ise farklıdır. ddd, iki gömme vektör arasındaki Öklid mesafesini ve mmm, benzer olmayan örnekler için beklenen minimum mesafeyi (margin) gösterir.

Fonksiyonun çalışma mantığı, iki örnek benzer olduğunda (Y=0) ve aralarındaki mesafe büyükse ikinci terimin büyük bir değer almasıyla modelin bu örnekleri gömme uzayında birbirine daha da yaklaştıracak şekilde güncellenmesi, tersine örnekler farklı olduğunda (Y=1) ve aralarındaki mesafe küçükse birinci terimin artmasıyla modelin bu örnekleri birbirinden uzaklaştırması esasına dayanır (Chicco, 2021).

Bu kayıp fonksiyonunun nihai amacı, eğitim süreci boyunca gömme uzayında anlamlı bir yerleşim elde etmektir; benzer örnekler yakın kümelenirken, farklı örnekler arasında belirgin bir mesafe bulunur. Bu özellik, özellikle yüz tanıma, imza doğrulama, ilaç-protein benzerliği analizi ve öneri sistemleri gibi uygulamalarda oldukça etkili ve faydalıdır . Kontrastif kayıp fonksiyonunun, geleneksel sınıflandırma fonksiyonları (örneğin Cross-Entropy) yerine kullanılması, modelin yalnızca sınıf etiketi tahmini için değil; aynı zamanda örnekler arasındaki anlamsal ilişkileri öğrenmesi için de bir fırsat sunar. Bu durum, özellikle az veriyle ya da çok sayıda sınıfa sahip problemlerde modelin genel performansını büyük ölçüde artırır (Serrano ve Bellogín, 2023).

3.5.1.3 Siyam Ağı'nda Temel Mimari ve Eğitim Süreci

Bir Siyam Sinir Ağı , iki girdi örneğini karşılaştırmak ve aralarındaki benzerlik düzeyini değerlendirmek için kullanılır. Örneğin, bir ilacın ve bir proteinin birbiriyle etkileşime girip girmediğini analiz etmek gibi. Ağ, tamamen simetrik ve benzer olan iki alt koldan (yoldan) oluşur; bu da her iki kolun aynı katman yapısını ve aynı ağırlıkları kullandığı anlamına gelir. Ancak bu iki kol, farklı girişler üzerinde çalışır. Örneğin, bir kola ilaç görüntüsü, diğer kola ise protein görüntüsü verilir (Chicco, 2021).

Her bir kol, girdiden özellik çıkarımı (feature extraction) yaparak onu sayısal bir özellik vektörüne (Embedding Vector) dönüştüren tam bir yol içerir. Bu vektör, girişe ait en anlamlı bilgilerin yoğun bir özetidir ve diğer örneklerle karşılaştırmak için kullanılabilir (Bansal ve diğ., 2025).

Her yolun yapısı şu şekildedir:

A) Conv1D Katmanları

Her yol, sırasıyla 256 ve 128 filtreye sahip iki adet bir boyutlu evrişim (Conv1D) katmanı içerir. Bu katmanlar, sıralı verilerdeki yerel desenleri (örüntüleri) tanımak için uygundur. Sıralı verilere örnek olarak şunlar verilebilir:

- Proteinlerin amino asit dizileri
- İlaçların SMILES gösterimleri (kimyasal yapılarını tanımlar)

Bu katmanlar, filtreleri veri üzerinde kaydırarak, modelin önemli ve tekrar eden yerel desenleri öğrenmesini sağlar (Chicco, 2021).

B) Flatten Katmanı

Flatten katmanı, Conv1D katmanlarından gelen çok boyutlu çıktıları, bir boyutlu bir vektöre dönüştürür. Bu vektör, Dense (tam bağlantılı) katmanlarda işlenebilir hale gelir. Bu işlem, karmaşık bir görüntünün sıkıştırılarak düz bir sayı dizisine çevrilmesine benzetilebilir (Zhou ve diğ., 2025)

C) Dense Katmanları

Daha sonra, her biri 256 nörona sahip üç adet Dense (tam bağlantılı) katman bulunur. Bu katmanlar şu amaçlarla tasarlanmıştır:

- Özellikler arasındaki doğrusal olmayan ilişkileri öğrenmek
- Yüksek anlam taşıyan, girişe ait temsili bir özellik vektörü oluşturmak (Javaid ve diğ., 2021)

D) Nihai Çıktı: Özellik Vektörü

Sonuç olarak, her yolun çıktısı 256 boyutlu sayısal bir vektördür. Bu vektör, modelin girişten çıkardığı bilgilerin yoğun bir özetidir. İki vektör (biri ilaç, diğeri protein için) Öklid uzaklığı gibi bir mesafe fonksiyonu kullanılarak karşılaştırılır ve aralarındaki benzerlik seviyesi hesaplanır (Bedi ve diğ., 2020).

3.5.1.4 Siyam Sinir Ağlarında Optimizasyon ve Aşırı Öğrenmenin Önlenmesi

Siamese ağları, girdiler arasındaki benzerliği öğrenmek için tasarlanmış, paralel çalışan ve ortak ağırlıkları paylaşan iki özdeş daldan oluşan bir yapıyı kullanır (He ve diğ., 2018). Bu ağın temel amacı, iki girdinin gömülü temsillerinin (embeddings) Öklid mesafesini hesaplamak olup, bu mesafe girdilerin benzerliğini veya farklılığını temsil eder. Öklid mesafesini hesaplamak için Formül 3.24 kullanılır.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.24}$$

Burada:

- x ve y, iki farklı girdi örneğidir.
- n, her bir girdinin boyutudur (örneğin, her bir gömme vektörünün boyutu).
- X_i ve Y_i, sırasıyla x ve y vektörlerinin her bir bileşenidir.

Formül 3.24'te, x ve y, girdilerin özellik vektörlerini temsil eder . d(x,y), bu girdiler arasındaki Öklid mesafesini ifade eder (Serrano ve Bellogín, 2023). Bu çalışmada, bu formülden elde edilen mesafe değerleri, ilaçlar ile proteinler arasındaki etkileşim veya etkileşim olmayan durumları tahmin etmek için kullanılmaktadır.

Bu çalışmada tasarlanan temel ağ, girdilerden özellik çıkarımı için çeşitli katmanlar içermektedir. Girdi türüne bağlı olarak ağ, sıralı veriler için konvolüsyonel katmanlar veya tamamen bağlı katmanlar içerebilir. Konvolüsyonel veriler için, 256 ve 128 filtreye sahip iki Conv1D katmanı kullanılır; bu katmanlar, karmaşık desenlerin çıkarılmasına yardımcı olur. Bu katmanlardan sonra, verilerin düzleştirilmesi için bir Flatten katmanı uygulanır ve ardından her biri 256 nöron içeren üç ardışık Dense katmanı yer alır; bu katmanlar nihai gömülü temsili (embedding) oluşturmayı amaçlar. Temel ağın nihai çıkışı, girdinin özelliklerini kompakt ve hassas bir şekilde temsil eden 256 boyutlu bir gömme (embedding) olur (Li ve diğ., 2022).

İki girdi arasındaki karşılaştırma için, temel ağ tarafından oluşturulan gömme (embedding) vektörleri arasındaki öklid mesafesi hesaplanır. Bu mesafe, iki girdi arasındaki benzerliği temsil eder: daha küçük değerler daha yüksek benzerliği, daha büyük değerler ise daha büyük farklılıkları ifade eder. Modeli eğitmek için kullanılan kontrastif kayıp fonksiyonu (contrastive loss) Formül 3.25'e göre tanımlanır:

$$Loss = Y \cdot d^2 + (1 - Y) \cdot max\,(0\,, m - d)^2 \tag{3.25}$$

Formül 3.25'te , Y gerçek etiketi (1 etkileşim için ve 0 etkileşim olmayanlar için), d hesaplanan Öklid mesafesini ve m ise marjin değerini temsil eder. Bu kayıp fonksiyonu, benzer örneklerin birbirine daha yakınlaştırılmasını, farklı örneklerin ise daha uzaklaştırılmasını sağlar .

Tasarlanan model, pozitif (ilaç-protein etkileşimi) ve negatif (etkileşim olmayan) çiftlerinden oluşan eğitim verisi kullanılarak eğitildi. Doğruluğu artırmak için, RMSprop optimizer kullanıldı; bu, eğitim süreci boyunca optimal öğrenme oranı ve kararlılık sağlar. Her iterasyonda veriler mini veri kümeleri halinde işlendi ve modelin performansı çapraz doğrulama (cross-validation) kullanılarak değerlendirildi (Serrano ve Bellogín, 2023).

3.4.1.5 Kullanılan Siyam Modeli Mimarisi Ve Veri Dengeleme Stratejisi

Bu çalışmada, ilaç-hedef etkileşimini öngörmek amacıyla Siyam ağına dayalı bir mimari tasarlanmış ve uygulanmıştır. Bu model, giriş olarak verilen örnek çiftleri arasındaki benzerliği öğrenebilmek ve etkileşim ile etkileşimsizlik durumlarını hassas bir şekilde ayırt edebilmek için geliştirilmiştir. Modelin temel yapısı, ağırlıkları ortak olan bir temel ağdan yararlanan iki paralel koldan oluşmaktadır. Bu ortak ağ, her bir girişin özellik vektörünü ortak bir temsilde sunar ve böylece anlamlı bir karşılaştırma yapılmasını mümkün kılar. Veri dengesizliğinin

önüne geçmek amacıyla, model eğitilmeden önce giriş verileri sistematik bir şekilde düzenlenmiş ve etkileşimli (pozitif) örnekler ile etkileşimsiz (negatif) örnekler arasında bire üç oran korunmuştur. Bu oranlama, yapay veri üretimi yoluyla değil, mevcut örneklerin uygun biçimde seçilip birleştirilmesiyle gerçekleştirilmiştir. Siyam ağı, 200 epoch boyunca bu dengelenmiş veri kümesi üzerinde eğitilmiş ve ortak özellik temsili öğrenme ile benzerlik ölçme sürecini gerçekleştirmiştir .

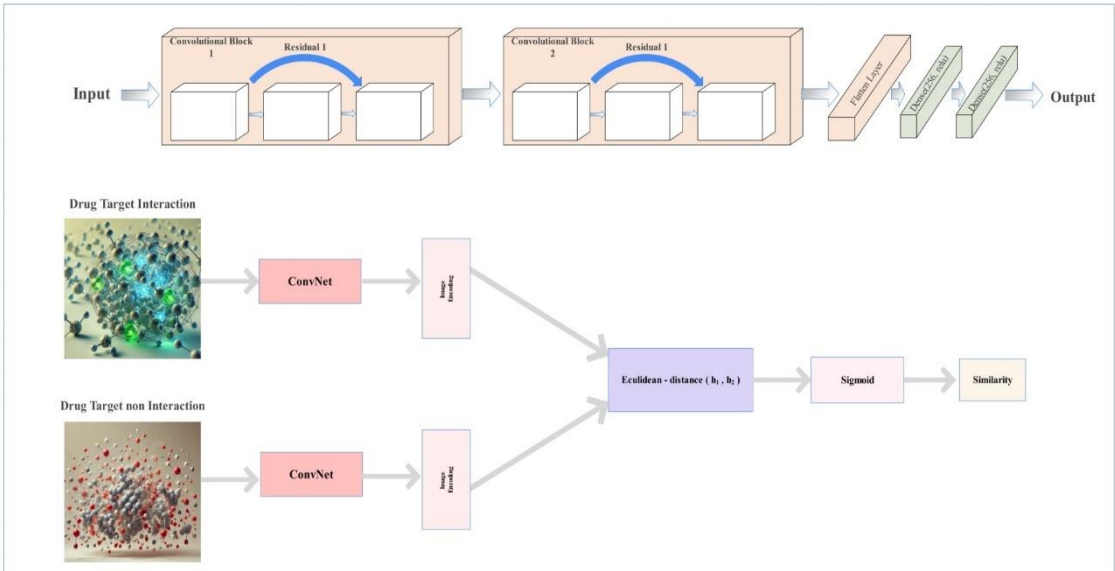
Tasarlanan temel ağ, konvolüsyonel modda ilk aşamada 256 ve 128 filtreye sahip, çekirdek boyutu 3 olan iki adet bir boyutlu konvolüsyon katmanı kullanmakta ve giriş dizilerindeki önemli yerel desenleri çıkarmayı hedeflemektedir. Bu katmanların çıktısı, bir Flatten katmanı ile tek boyutlu bir vektöre dönüştürülmekte ve ardından her biri 256 nörondan oluşan ve ReLU aktivasyon fonksiyonuna sahip üç ardışık yoğun katman, verilerden daha üst düzey temsillerin öğrenilmesinden sorumlu olmaktadır. Son katman olan gömme katmanı, her bir giriş için nihai vektörü üretmekte ve bu vektörler örnek çiftleri arasındaki mesafe karşılaştırmasının temelini oluşturmaktadır.

İki örnek arasındaki benzerliği ölçmek amacıyla Öklid mesafesi kullanılmıştır. Bu mesafe, iki embedding vektörü arasındaki sayısal uzaklığı ifade eder ve örneklerin öğrenilmiş özellikler uzayında birbirine ne kadar yakın veya uzak olduklarını gösterir. Mesafe azaldıkça, ilaç ile hedef arasında etkileşim olasılığı daha yüksek kabul edilmektedir. Modeli eğitmek için Contrastive Loss kayıp fonksiyonundan yararlanılmıştır; bu fonksiyon, etkileşimli örneklerin mesafesini en aza indirmeyi, etkileşimsiz örneklerin mesafesini ise belirli bir eşik (margin) değerinin üstünde tutmayı hedefler. Böylece model, pozitif (etkileşimli) çiftleri birbirine yaklaştırmayı ve negatif (etkileşimsiz) çiftleri birbirinden uzaklaştırmayı öğrenir.

Bu çalışmada dikkat çekici bir husus, Siyam ağı yapısının yardımıyla sistematik şekilde örnek çiftleri üretilerek verinin dengelenmesidir. Pek çok DTI probleminde etiketli veriler dengesiz biçimde dağılmıştır; yani negatif örneklerin sayısı pozitif örneklerden çok daha fazladır. Bu çalışmada eğitim yapısı tasarlanarak ve örnek çiftleri seçilerek veri oranı bire üç biçiminde kesin olarak ayarlanmış, böylece model hem negatif verilerin çeşitliliğinden faydalanmış hem de negatif sınıfın baskın hale gelmesi engellenmiştir. Bu yöntem, klasik yeniden örnekleme teknikleri (undersampling veya oversampling gibi) yerine ya da onlara ek olarak kullanılabilecek, Siyam ağına dayalı dengelenmiş çift veri üretimi yaklaşımı olarak değerlendirilebilir.

Modelin uygulama aşamasında veriler iki boyutlu olarak düzenlenmiştir; her eğitim örneğinde iki giriş vektörü (biri ilaç, diğeri hedef) ve etkileşim ya da etkileşimsizlik etiket bilgisi bulunmaktadır. Ağ eğitimi RMSprop optimizasyon algoritması kullanılarak, epoch sayısı 200, batch boyutu 512 ve doğrulama oranı %25 olacak şekilde titizlikle gerçekleştirilmiştir. Ayrıca hata eğilim grafikleri ve doğrulama kaybının ortalama değerleri çizilerek modelin zaman içindeki davranışı gözlemlenmiş ve doğru şekilde yakınsadığı doğrulanmıştır.

Sonuç olarak, konvolüsyonel mimariye ve Contrastive Loss fonksiyonuna sahip bu Siyam ağ çerçevesi, dizisel verilerden anlamlı vektörel temsiller üretme kabiliyetine sahip olmuş ve aynı zamanda sınıf dengesizliği sorununu yöneterek verileri bire üç oranında sistematik biçimde dengelemiştir. Bu özellik, modelin ilaç-hedef etkileşimlerini yüksek doğrulukla ayırt etmesine ve gerçek veriler üzerinde tatmin edici bir performans göstermesine olanak tanımıştır.Bu çalışmada kullanılan siyam ağının iç yapısı Şekil 3.6'da gösterilmiştir.



Şekil 3. 8: Veriye Benzerlik Hesaplama için Siamese Ağının İç Yapısı (yazar tarafından oluşturulmuştur).

Siyam Sinir Ağı , bu problemde verilerin ilk dengelemesinde önemli bir rol oynar. Bu ağ, yeni bir özellik uzayı (embedding) öğrenerek pozitif örnekleri (ilaç-hedef etkileşim örnekleri) ve negatif örnekleri (etkileşim olmayan örnekler) öyle bir şekilde temsil etmeye çalışır ki, aralarındaki mesafe anlamlı ve kolayca ayrılabilir hale gelsin. Başka bir deyişle, çift dallı bu ağ, aynı sınıfa ait örnek çiftlerini vektör uzayında birbirine yakınlaştırmayı, farklı sınıflara ait

olanları ise birbirinden uzaklaştırmayı öğrenir. Bu süreç, verilerin doğal dengesizliğinin bir kısmının azalmasına ve başlangıçta oldukça dengesiz olan oranların daha dengeli bir seviyeye, yaklaşık 1'e 3 oranına gelmesine katkıda bulunur. Ayrıca, bu şekilde oluşturulan yeni vektör temsili sayesinde benzer veriler daha düzenli ve kümelenmiş bir yapı kazanır. Ancak tüm bu avantajlarına rağmen, Siyam ağı tek başına gerçek yapay örnekler üretip azınlık sınıfındaki veri sayısını arttıramaz; yalnızca sınıfların oranını veya ağırlığını nispeten dengeler. Bunun yanında, yeni özellik uzayı öğrenildikten sonra bile veride gürültü ve sınıflar arasındaki örtüşme olasılığı devam eder ve azınlık sınıfındaki örneklerin kesin sayısı yine de tam bir denge için yeterli değildir.

Bu nedenle, Siyam ağıyla yapılan ilk temsilleme ve dengelemeden sonra SMOTE-Tomek tekniğinin kullanılması gerekli hale gelir ki kalan problemler de çözülebilsin. Her ne kadar Siyam ağı sayesinde sınıfların oranı yaklaşık 1'e 3 düzeyine getirilmiş olsa da, azınlık sınıfındaki gerçek örneklerin sayısı hâlâ ideal seviyenin altındadır ve modelin doğru karar sınırları öğrenebilmesi için daha fazla veriye ihtiyaç vardır. SMOTE tekniği bu sorunu, azınlık sınıfındaki en yakın komşular arasında yapılan enterpolasyon aracılığıyla gerçek yapay örnekler üreterek çözer ve oranı yaklaşık 1'e 1'e yaklaştırır. Sonraki aşamada ise Tomek Links, bir temizleme aracı olarak devreye girer ve karar sınırına yakın konumlanmış, çoğunluk sınıfına ait yanıltıcı ve gürültülü örnekleri tespit edip veriden çıkarır. Bu iki tamamlayıcı aşama sayesinde hem azınlık sınıfının veri hacmi artar hem de sınıflar arasındaki karar sınırı daha belirgin hale gelir; böylece model dengeli ve temiz bir veri kümesi üzerinde daha yüksek doğruluk ve genellenebilirlikle eğitilebilir.

3.5.2 Sentetik Azınlık Aşırı Örnekleme Tekniği – Tomek Bağlantıları(Synthetic Minority Over-sampling Technique – Tomek Links(SMOTE-Tomek))

SMOTE-Tomek tekniği, makine öğrenmesi modellerinde veri dengesizliği sorunlarını ele almak için SMOTE ile Tomek Bağlantılarını birleştiren hibrit bir örnekleme yöntemidir. SMOTE, en yakın komşular arasındaki doğrusal enterpolasyon kullanarak azınlık sınıfı için sentetik örnekler üretir. Ancak, SMOTE'un bir sınırlaması, sentetik örneklerin çoğunluk sınıfıyla örtüşme potansiyelidir; bu durum aşırı uyum ve net karar sınırlarının tanımlanmasında belirsizliğe yol açabilir. Bu sorunu hafifletmek için, sonraki işleme(post-processing) adımı olarak Tomek Bağlantıları uygulanır. Tomek Bağlantıları, zıt sınıflardan olan ve birbirine yakın

olan örnek çiftlerini tespit eder ve bu tür çiftleri çoğunluk sınıfından çıkarır, böylece veriyi daha temiz hale getirir, aşırı uyum riskini azaltır ve model doğruluğunu artırır (Hairani ve diğ., 2017).

SMOTE-Tomek süreci iki ana aşamadan oluşur. İlk aşamada, SMOTE uygulanır; burada, azınlık sınıfından örnekler rastgele seçilir ve en yakın komşularına olan mesafeler hesaplanır. Daha sonra, enterpolasyon yoluyla sentetik örnekler üretilir ve bu işlem, sınıflar arasındaki istenilen denge sağlanana kadar tekrarlanır. İkinci aşamada, Tomek Bağlantıları kullanılarak, azınlık sınıfına yakın olan ve çoğunluk sınıfından gürültülü örnekler tespit edilip çıkarılır. Bu gürültülü örnekler, sınıf sınırlarını bozdukları ve model performansını düşürdükleri için sorunlu olarak kabul edilir. SMOTE ve Tomek Bağlantıları'nı birleştirerek, bu yöntem hem azınlık sınıfının temsilini hem de veri kalitesini geliştirir, daha doğru ve optimize edilmiş bir yaklaşım sunar ve nihayetinde modelin genelleme yeteneğini artırır (Wang ve diğ., 2024).

Araştırmalar, SMOTE-Tomek'in, bu tekniklerden her birinin bağımsız olarak kullanılmasıyla karşılaştırıldığında, veri dengesizliğini ele alma ve model performansını iyileştirme konusunda daha etkili olduğunu göstermektedir (Jonathan ve diğ., 2020). SMOTE, azınlık sınıfı için sentetik örnekler üretmede önemli bir rol oynarken, Tomek Bağlantıları, öğrenme sürecini zorlaştırabilecek çoğunluk sınıfındaki gürültülü verilerin olumsuz etkilerini ortadan kaldırmaya yardımcı olur (Koloszyc ve diğ., 2023). Veriyi bu şekilde iyileştirerek, SMOTE-Tomek, modelin dengeli ve temiz bir veri seti üzerinde eğitilmesini sağlar ve daha iyi sınıflandırma sonuçları elde edilmesine yol açar (Prasetyaningrum ve diğ., 2024). Bu hibrit yaklaşım, görüntü sınıflandırma, dolandırıcılık tespiti ve biyoinformatik gibi veri dengesizliğinin önemli zorluklar oluşturduğu alanlarda yaygın olarak uygulanmaktadır (Tyagi ve diğ., 2024).

3.6 DEĞERLENDİRİLEN VERİ DENGELEME YÖNTEMLERİ VE ÖRNEKLEME TEKNİKLERİ

Bu bölümün geri kalanında , önerilen yöntemin performansının karşılaştırıldığı çeşitli veri dengeleme yöntemleri kısaca tanıtılmakta ve açıklanmaktadır. Bu yöntemler; SMOTE, ADASYN ve Borderline-SMOTE gibi farklı oversampling tekniklerini, RandomUnderSampler ve TomekLinks gibi undersampling yaklaşımlarını ve ayrıca SMOTE-Tomek ile SMOTEENN gibi hibrit yöntemleri kapsamaktadır. Her bir yöntem, veri dengesini artırmak ve veri kümesindeki gürültüyü azaltmak için farklı bir strateji sunmaktadır. Bunun yanı sıra, Linear SVM, Random Forest, Neural Net, Decision Tree, AdaBoost, Naive Bayes, QDA, LDA, EEC,

RSC ve BBC gibi çeşitli makine öğrenimi algoritmaları da kısaca açıklanmış ve önerilen modelin diğer sınıflandırıcılarla karşılaştırılmasına temel oluşturacak bilgiler verilmiştir. Bu karşılaştırmalar, ilaç-hedef etkileşimlerini tahmin etme alanında önerilen yaklaşımın avantajlarının ve güçlü yönlerinin daha iyi anlaşılmasına katkı sağlamaktadır.

3.6.1 Kümeleme Merkezi Algoritması(Cluster Centroid Algorithm)

Kümeleme Merkezi algoritması, K-ortalamalar yönteminin geliştirilmiş bir versiyonudur. Bu algoritma, veri noktaları ile küme merkezleri arasındaki mesafeyi en aza indirmenin yanı sıra, amaç fonksiyonuna bir ceza fonksiyonu ekleyerek ağırlık dengesini de gözetir. Böylece her veri noktası, kümenin ağırlık dağılımına olan etkisine göre atanır (Abdalameer ve diğ, 2022).

Kümeleme Merkezi, çoğunluk sınıfına ait örneklerin sayısını azaltmak amacıyla kümeleme yaklaşımını kullanan, fakat önemli bilgileri tamamen kaybetmeden bunu gerçekleştiren akıllı bir alt örnekleme yöntemidir. Bu yöntemde çoğunluk sınıfına ait veriler kümeleme algoritmasıyla gruplara ayrılır ve her grubun ağırlık merkezi, o kümenin temsilcisi olarak seçilir. Sonuç olarak, tüm örnekler yerine yalnızca bu merkez noktalar veri setinde kalır, geri kalan örnekler ise elenir (Ongko ve Abdullah, 2015).

Veri dengesizliği problemleriyle başa çıkmada, özellikle coğrafi verilerde veya her örneğin farklı ağırlıklara sahip olduğu durumlarda, bu merkez noktaları dengeli kümeleri tanımlamak için oldukça etkili bir araçtır. Küme Merkezli algoritmalar yalnızca verileri ağırlıklı merkezler aracılığıyla sınıflandırmakla kalmaz, aynı zamanda kümeler arasında daha adil ve dengeli bir dağılım oluşturur (Uykan, 2023).

3.6.2 Uyarlanabilir Yoğunlaştırılmış En Yakın Komşu (Adaptive Condensed Nearest Neighbor (Ada-CNN))

Ada-CNN ،klasik CNN algoritmasının geliştirilmiş bir versiyonudur ve temel amacı eğitim verilerinin boyutunu küçültmektir. Bu yöntem, çoğunluk sınıfındaki gereksiz verileri ortadan kaldırır ve yalnızca karar sınırının belirlenmesinde kritik öneme sahip örnekleri seçerek akıllı bir alt örnekleme uygular. Azınlık sınıfı için yeni örnekler üretmez; mevcut verileri sıkıştırır ve seçici olarak kullanır (Siddappa ve Kampalappa, 2019).

Ada-CNN, veri indirgeme ve sınıfa özgü ağırlıklandırmayı birleştirerek özellikle gürültülü ortamlarda dengesizlik problemlerine karşı etkili bir çözüm sunar. Yöntem, verileri üç kategoriye ayırır: aykırı değerler (sınıflandırılamayan örnekler), prototipler (temsilci örnekler) ve emilen noktalar (prototipler sayesinde doğru sınıflandırılan örnekler). Algoritma, veri alt kümelerini iteratif olarak oluşturur; her adımda örnekler seçilip sınıflandırılır, doğru sınıflandırılamayanlar elenir. Böylece, yalnızca karar sınırını tanımlayan kritik veriler korunur ve model daha hızlı ve verimli çalışır (Siddappa ve Kampalappa, 2019; Bhavsar ve diğ., 2024).

3.6.3 Düzenlenmiş En Yakın Komşu (Edited Nearest Neighbour (ENN))

ENN yöntemi, ilk olarak Wilson (1972) tarafından geliştirilmiş ve eğitim veri setindeki gürültülü ve yanıltıcı örnekleri temizlemeye yönelik bir alt örnekleme tekniğidir (Gu ve diğ., 2008). ENN, özellikle sınıflar arası örtüşmenin yoğun olduğu veri kümelerinde, çoğunluk sınıfındaki örneklerin model üzerindeki baskısını azaltmak amacıyla kullanılır. Yöntem, her örneğin en yakın k komşusunu (genellikle k=3) belirleyerek, komşuların çoğunluğu farklı sınıfa aitse o örneği veri setinden çıkarır. Böylece, karar sınırlarında bulunan yanıltıcı çoğunluk sınıfı örnekleri temizlenmiş olur (Orman, 2024).

Bu yaklaşım sayesinde, azınlık sınıfı daha belirgin hale gelir ve model, daha doğru ve dengeli karar sınırları oluşturabilir. ENN özellikle yüksek sınıf örtüşmesine sahip veri setlerinde etkilidir; çünkü karar sınırındaki gürültülü ve baskın sınıf örneklerini ortadan kaldırarak modelin hata yapma olasılığını azaltır ve öğrenme sürecini iyileştirir (Mamilla ve diğ., 2025).

3.6.4 Tekrarlamalı Düzenlenmiş En Yakın Komşu (Repeated Edited Nearest Neighbour (RENN))

RENN yöntemi, ENN algoritmasının yinelemeli geliştirilmiş bir versiyonu olup, gürültülü veya karar sınırında bulunan hatalı örnekleri ardışık olarak temizleyerek daha küçük, dengeli ve temiz bir veri seti oluşturmayı amaçlar (Orman, 2024) . ENN'den farklı olarak RENN bu işlemi tek sefer değil, veri kümesinde yeni bir örnek kalmayana kadar tekrar tekrar uygulayarak çoğunluk sınıfındaki tutarsız örneklerin kademeli şekilde elenmesini sağlar (Panjainam ve Kanjanawattana, 2024). Bu yaklaşım, özellikle sınıf örtüşmesi ve veri dengesizliğinin problem oluşturduğu sınıflandırma senaryolarında etkili bir alt örnekleme stratejisi olarak öne çıkar.

RENN, yapay veri üretimi gerçekleştiren aşırı örnekleme tekniklerinin aksine sentetik örnek oluşturmaz; yalnızca mevcut veri kümesindeki zararlı ve hatalı verileri ortadan kaldırır (Barberi ve diğ., 2012). Algoritmanın temel adımları; her örneğin k en yakın komşusunun belirlenmesi, sınıf çoğunluğu farklı ise örneğin çıkarılması ve bu işlemin veri kararlı hâle gelene kadar tekrarlanmasıdır. Bu sayede modelin doğruluk ve genellenebilirlik kapasitesi artırılırken, veri kümesinin doğal yapısı da korunur (Kanj ve diğ., 2016).

3.6.5 Komşuluk Temizleme Kuralı (Neighbourhood Cleaning Rule (NCL))

Mahalle Temizleme Kuralı algoritması, eğitim verilerindeki sınıf dengesizliği sorununu ele almak üzere geliştirilmiş, sofistike ve akıllı bir alt örnekleme yöntemidir. Çoğunluk sınıfına ait örnekleri rastgele kaldıran basit yöntemlerin aksine, NCL algoritması; veri kümesindeki potansiyel hatalı ya da gürültülü örnekleri mahalle analizi yoluyla tespit eder ve bunları veri kümesinden ayıklamayı hedefler. Bu yaklaşım, Düzenlenmiş En Yakın Komşular tekniğine dayanarak çalışır ve komşularının sınıflarıyla tutarsızlık gösteren örnekleri sistemli bir şekilde ortadan kaldırır. Böylece, geride kalan verilerin hem güvenilirliği hem de temsil gücü artar. Bu süreç sonucunda, modelin daha az hatalı ve daha temiz veriyle eğitilmesi mümkün hâle gelir; bu da dengesiz veri kümeleri karşısında makine öğrenimi modellerinin performansını gözle görülür biçimde artırır (Panahi ve diğ., 2020).

NCL algoritmasının uygulanması üç temel adımdan oluşur. İlk olarak, veri kümesindeki her bir örnek için en yakın üç komşu, genellikle öklid uzaklığı gibi bir ölçütle belirlenir. Ardından, örneklemin sınıfına göre bir değerlendirme yapılır: Eğer örnek, çoğunluk sınıfına aitse ve komşularının yarısından fazlası azınlık sınıfındansa, bu örnek muhtemel bir hata olarak değerlendirilir ve veri kümesinden çıkarılır. Öte yandan, eğer örnek azınlık sınıfındansa fakat komşularının çoğu çoğunluk sınıfındansa, bu durumda örnek korunur; ancak yanıltıcı olabilecek komşular ayıklanır. Çünkü bu komşular, karar sınırlarını hatalı şekillendirebilir. Sonuç olarak, bu dikkatli temizlik süreci; daha dengeli, daha küçük fakat niteliği yüksek bir veri kümesi elde edilmesini sağlar ve bu da daha etkili ve isabetli makine öğrenimi modellerinin geliştirilmesine zemin hazırlar (Ependi ve diğ., 2023).

3.6.6 Tek Taraflı Seçim(One-Sided Selection (ONESS))

ONESS algoritması, Kubat ve Matwin tarafından sınıf dengesizliği sorununu azaltmak amacıyla geliştirilmiş bir alt örnekleme yöntemidir (Batista ve diğ., 2000).Bu yöntem, çoğunluk sınıfındaki örnekleri rastgele kaldırmak yerine, yalnızca karar sınırında bulunan, gürültülü veya önemsiz verileri ayıklayarak daha dengeli bir veri seti oluşturur. ONESS, 1-En Yakın Komşu (1-NN) filtreleme ve Tomek Bağlantıları kavramlarını birleştirerek, yanlış sınıflandırılan veya karar sınırında yer alan çoğunluk örneklerini sistemden çıkarır (Kuncheva ve diğ., 2019).

Algoritmada önce çoğunluk sınıfından rastgele bir örnek seçilip azınlık sınıfıyla birleştirilir ve bu küme 1-NN ile değerlendirilir. Yanlış sınıflandırılan örnekler C kümesine eklenir, ardından Tomek Bağlantılarındaki çoğunluk sınıfı örnekleri temizlenir. Bu işlemler, istenen denge sağlanana kadar tekrarlanır ve sonuçta daha temiz, dengeli ve kaliteli bir veri kümesi elde edilir (Gangrade ve diğ., 2021).

3.6.7 Tüm k-en yakın komşular (All k-nearest neighbors (All-KNN))

All-KNN algoritması, dengesiz veri kümelerini dengelemek amacıyla geliştirilmiş, K-En Yakın Komşular algoritmasına dayalı, akıllı bir alt örnekleme yöntemidir. Basit yöntemlerin çoğunluk sınıfına ait verileri rastgele kaldırmasının aksine, All-KNN analitik bir yaklaşımla çalışır; modelin öğrenme sürecine anlamlı katkı sağlamayacak, yani bilgi açısından düşük değerli örnekleri ayıklamayı hedefler. Algoritma, örneklerin mahalle yapısını inceleyerek hangi çoğunluk sınıfı verilerinin korunacağına ve hangilerinin eleneceğine karar verir. Bu sayede makine öğrenimi modellerinin, özellikle azınlık sınıfını tespit etmedeki başarısı artar; çünkü elde kalan veriler daha temsili olup, gürültüden arındırılmış bir yapıya sahiptir (Triguero ve diğ., 2019).

All-KNN'in uygulama süreci, öncelikle k değerinin (komşu sayısı) belirlenmesiyle başlar. Bu değer, veri kümesinin boyutu ve çeşitliliğine bağlı olarak genellikle 3 ya da 5 olarak seçilir. Ardından, çoğunluk sınıfına ait her örnek için, Öklid uzaklığı gibi bir ölçüte göre en yakın komşular hesaplanır. Her bir örnek, komşularının sınıf dağılımına göre değerlendirilir: eğer çoğu komşu azınlık sınıfındansa, örnek yararlı sayılır ve korunur; aksi hâlde, düşük bilgi değeri taşıdığı varsayılır ve veri kümesinden çıkarılır. Bu işlem, sınıflar arasındaki dengesizlik giderilene veya artık elenebilecek örnek kalmayana kadar yinelemeli şekilde sürdürülür. Bu

adım adım yaklaşım, daha dengeli, sade ve model eğitimi açısından daha elverişli bir veri kümesi elde edilmesini sağlar (Chatzimilioudis ve diğ., 2012).

3.6.8.RandomOverSampler

RandomOverSampler algoritması, sınıf dengesizliğini gidermek için en basit ve yaygın aşırı örnekleme yöntemlerinden biridir ve Imbalanced-Learn kütüphanesinde yer alır. Bu yöntemde azınlık sınıfına ait örnekler rastgele seçilip kopyalanır ve bu işlem, örnek sayısı çoğunluk sınıfına eşitlenene kadar devam eder. Yöntemin basitliği, hızlı çalışmasını ve farklı makine öğrenimi modelleriyle kolayca entegre edilmesini sağlar. Ayrıca örnekleme oranının esnek biçimde ayarlanabilmesi, birden fazla azınlık sınıfında da kullanılmasına imkân tanır (Nhita ve diğ., 2023). Ancak, yalnızca mevcut örneklerin kopyalanması yeni bilgi sağlamaz ve yüksek oranlı örnekleme durumunda aşırı öğrenme riskini artırır. Bu nedenle RandomOverSampler genelde ilk dengeleme adımı olarak kullanılır ve sonrasında SMOTE veya ADASYN gibi daha gelişmiş tekniklerle desteklenir (Mamun ve diğ., 2024).

3.6.9 RandomUnderSampler

RandomUnderSampler algoritması, sınıf dengesizliğini gidermek için kullanılan en basit ve yaygın yöntemlerden biridir ve Imbalanced-Learn kütüphanesinde bulunur. Bu teknikte çoğunluk sınıfından rastgele örnekler çıkarılarak sınıflar dengelenir. Yöntemin basitliği ve hızı, onu birçok projede pratik bir çözüm hâline getirmiştir (Konstantinov ve diğ., 2025). Ancak örneklerin rastgele silinmesi, çoğunluk sınıfına ait önemli bilgilerin kaybına ve karar sınırlarının öğrenilmesinde zorluklara yol açabilir. Bu nedenle RandomUnderSampler genellikle daha gelişmiş tekniklerle birlikte veya stratejik yaklaşımların bir parçası olarak kullanılır (Borchiellini ve diğ., 2024).

3.6.10 Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique(SMOTE))

İlk olarak 2002 yılında Chawla ve arkadaşları tarafından önerilen SMOTE algoritması, makine öğrenmesi problemlerinde sınıf dengesizliği sorununa karşı en yaygın ve etkili çözümlerden biridir (Chawla, 2009). Bu yöntem, azınlık sınıfındaki örneklerin basitçe kopyalanmasının neden olduğu aşırı uyumu (overfitting) önlemeyi amaçlar. Bunun yerine, azınlık sınıfına ait

gerçek bir örnek ile en yakın k komşularından biri arasında özellik uzayında doğrusal enterpolasyon yoluyla sentetik örnekler üretir. İki örnek arasındaki fark vektörü sıfır ile bir arasında rastgele seçilen bir sayıyla çarpılır ve başlangıç örneğine eklenerek yeni veri noktası elde edilir. Böylece oluşturulan yapay örnekler mevcut verilerin tekrarları değil, azınlık sınıfının potansiyel dağılımını temsil eden noktalar olur (Leevy ve diğ., 2018).

SMOTE'un en önemli avantajı, verileri çoğaltmak yerine karar bölgelerini genişletmesi ve modelin azınlık sınıfını daha iyi öğrenmesine katkı sağlamasıdır. Ancak yüksek boyutlu veri kümelerinde aşırı genelleme veya varyans artışı gibi istenmeyen etkiler ortaya çıkabilir. Ayrıca, çoğunluk sınıfının dağılımını dikkate almadığı için karar sınırlarına veya gürültülü bölgelere yakın hatalı örnekler oluşturabilir (Viana ve diğ., 2024).

3.6.11 Uyarlanabilir Sentetik Örnekleme (Adaptive Synthetic Sampling(ADASYN))

İlk olarak 2008 yılında He ve arkadaşları tarafından geliştirilen ADASYN (Uyarlanabilir Sentetik Örnekleme) algoritması, SMOTE'un daha gelişmiş ve akıllı bir versiyonudur . Bu yöntemin temel amacı, öğrenilmesi daha zor olan azınlık sınıfı örneklerine odaklanarak sınıf dengesizliğinden kaynaklanan önyargıyı azaltmaktır. SMOTE tüm azınlık örnekleri için eşit sayıda sentetik veri üretirken, ADASYN her örneğin yoğunluk dağılımını ve öğrenme zorluğunu dikkate alır. Başka bir deyişle, çoğunluk sınıfına yakın veya karar sınırlarında bulunan azınlık örnekleri için daha fazla sentetik veri oluşturur (Brandt ve Lanzén, 2021).

ADASYN'in en büyük avantajlarından biri, modelin karmaşık bölgelerdeki örüntülere daha fazla odaklanmasını sağlamasıdır. Bu sayede dolandırıcılık tespiti, biyomedikal veri analizi gibi alanlarda yüksek başarı elde edilmektedir. Uyarlanabilir yapısıyla yalnızca Geri Çağırma (Recall) oranını yükseltmekle kalmaz, karar sınırlarının daha optimal şekillenmesini sağlar (Muley ve di., 2025). Bununla birlikte, bazı sınırlamaları da vardır. Veri setinde gürültü veya aykırı değerlerin bulunması durumunda, uygun olmayan sentetik örnekler üretilebilir ve bu durum modelin doğruluğunu düşürebilir. Ayrıca SMOTE'a göre daha yüksek hesaplama maliyeti vardır ve parametreler yanlış ayarlandığında modelin belirli örneklere aşırı odaklanarak genelleştirilebilirliğinin azalmasına neden olabilir. Tüm bu sınırlamalara rağmen, ADASYN günümüzde sınıf dengesizliğini ele almak için en etkili ve ileri düzey aşırı örnekleme tekniklerinden biri olarak kabul edilmektedir (Niaz ve diğ., 2022).

3.6.13 SMOTE ve Düzenlenmiş En Yakın Komşular (SMOTE Edited Nearest Neighbours (SMOTEENN))

SMOTEENN algoritması, veri örnekleme alanında birbirini tamamlayan iki yaklaşımın akıllıca harmanlanmasıyla oluşturulmuş hibrit bir yöntemdir: aşırı örnekleme için SMOTE ve az örnekleme için ENN . Bu yöntem, sınıflar arasındaki dengesizliği azaltmak ve makine öğrenimi modellerinin doğruluğunu artırmak amacıyla geliştirilmiştir. SMOTEENN, ilk olarak SMOTE algoritmasıyla azınlık sınıfına ait sentetik örnekler üreterek veri setindeki dengesizliği azaltır. Ardından ENN yöntemi devreye girerek, ister çoğunluk ister azınlık sınıfına ait olsun, yakın komşularına göre yanlış sınıflandırılan örnekleri veri setinden ayıklar. Bu süreç, Tomek Links gibi diğer temizleme tekniklerine kıyasla daha derinlemesine bir veri arındırma sağlar ve daha fazla gürültülü veya uyumsuz örneğin elenmesine olanak tanır (Batista ve diğ., 2004).

SMOTEENN'in en belirgin avantajlarından biri, hem azınlık sınıfının kapsama alanını artırma hem de belirsiz ve tutarsız verileri ortadan kaldırma yeteneğidir. Bu yöntem, modelin karar sınırlarını daha net ve kararlı hale getirirken, sentetik verilere aşırı uyum sağlama riskini de azaltır. SMOTE veya ENN yöntemlerinin tek başına uygulanmasına kıyasla, SMOTEENN özellikle sınıf dağılımında ciddi dengesizliklerin bulunduğu problemler için daha başarılı sınıflandırma sonuçları üretmektedir. Bu yönüyle SMOTEENN, dengesiz veri kümeleriyle çalışırken etkili bir ön işleme yöntemi olarak öne çıkmaktadır (Gupta ve Sahare, 2012).

3.6.14 Sınırda Sentetik Azınlık Aşırı Örnekleme Tekniği (Borderline Synthetic Minority Over-sampling Technique (Borderline-SMOTE))

Borderline-SMOTE algoritması, azınlık sınıfına ait sınır örneklerine odaklanarak, örtüşme ve girişim içeren veri kümelerinde daha etkili sonuçlar elde etmeyi amaçlayan, SMOTE yönteminin geliştirilmiş bir versiyonudur (Gosain ve Sardana, 2017). Klasik SMOTE, tüm azınlık sınıfı örneklerini eşit biçimde örneklerken, Borderline-SMOTE yalnızca karar sınırına yakın, yani sınıflandırma açısından daha riskli bölgelerdeki örnekleri hedef alır. Bu yaklaşımın temel mantığı, bu kritik bölgelerdeki verilerin yanlış sınıflandırılma olasılığının daha yüksek olmasıdır. Algoritma, her azınlık örneğinin k en yakın komşusunu analiz ederek, bu komşuların sınıf dağılımına göre örnekleri "güvenli" , "tehlikeli" ve "gürültülü" olarak üçe ayırır. Yalnızca tehlikeli gruptaki örnekler üzerinden sentetik veriler üretilir; çünkü bu örnekler, modelin daha

sağlam ve doğru karar sınırları oluşturabilmesi için desteklenmeye ihtiyaç duyar (Carvalho ve diğ., 2025) .

Borderline-SMOTE, gürültülü verileri (yani tamamen çoğunluk sınıfıyla çevrili ve öğrenmeye katkı sağlamayan örnekleri) dışlayarak, gereksiz veya zararlı sentetik veri üretimini önler. Bu sayede sınıflandırma hataları azaltılır. Özellikle, azınlık sınıfının sınırlı sayıda ve dağınık şekilde yer aldığı, karar sınırlarının belirsiz olduğu durumlarda klasik SMOTE yetersiz kalabilirken, Borderline-SMOTE bu gibi senaryolarda daha dengeli ve anlamlı örnekler üreterek model performansını belirgin şekilde artırır. Sonuç olarak bu algoritma, sınıf dengesizliği sorununa karşı geliştirilen en etkili veri artırma stratejilerinden biri olarak, makine öğrenimi modellerinin hem doğruluğunu hem de genelleme kabiliyetini güçlendirir (Bourdonnaye ve Daniel, 2022).

3.6.15 SVM-SMOTE

SVM-SMOTE algoritması, Destek Vektör Makinesi algoritmasının gücünden faydalanarak, azınlık sınıfı için daha hedefli ve etkili yapay örnekler üretmeyi amaçlayan gelişmiş bir aşırı örnekleme yöntemidir. Geleneksel SMOTE yönteminin rastgele biçimde azınlık sınıfı örnekleri arasında veri üretmesine karşılık, SVM-SMOTE sınıflar arasındaki karar sınırlarına odaklanır. Çünkü bu bölgeler, modelin sınıflandırma doğruluğu açısından kritik öneme sahiptir; zira sınıflandırma hatalarının büyük kısmı tam da bu hassas alanlarda gerçekleşmektedir (Jiang ve diğ., 2025).

Bu yöntemde, öncelikle sınıflar arasındaki karar sınırlarını belirlemek üzere bir SVM modeli eğitilir. Ardından, bu sınıra yakın konumda yer alan azınlık sınıfı örnekleri tespit edilerek yapay örnek üretimi için aday olarak seçilir. Yeni sentetik veriler, bu örnekleri en yakın komşularına bağlayan vektörler boyunca oluşturulur. Bu sayede model, özellikle sınır bölgelerinde daha isabetli kararlar verebilecek bir yapıya kavuşur ve sınıflandırma hataları belirgin şekilde azalır (Khan ve Malim, 2023)

Genel olarak değerlendirildiğinde, SVM-SMOTE algoritması sınıflar arasında ayrım yapmanın zor olduğu, dengesiz ve seyrek veri içeren problemlerde oldukça etkili bir çözümdür. Karar sınırlarındaki veri yoğunluğunu artırarak, modelin genelleme yeteneğini güçlendirir ve makine öğrenimi sistemlerinin performansını kayda değer biçimde iyileştirir (Paoletti ve diğ., 2023).

3.6.16 Tomek Bağlantıları (Tomek Links)

Tomek Links algoritması, makine öğrenmesinde kullanılan etkili bir veri temizleme yöntemidir ve temel amacı sınıflar arası örtüşmeyi azaltarak veri setinin kalitesini artırmaktır. Algoritma, farklı sınıflara ait iki örneğin birbirinin en yakın komşusu olduğu durumları "Tomek bağlantısı" olarak tanımlar. Bu tür örneklerin, sınıf sınırına çok yakın ya da gürültülü veri olduğu varsayılır ve genellikle çoğunluk sınıfına ait olan örnek veri setinden çıkarılır. Bu sayede sınıflar arasındaki karar sınırları daha belirgin hâle gelir (Yousefimehr ve diğ., 2025).

Tomek Links'in en büyük avantajı, yeni veri üretmeden mevcut veri yapısını sadeleştirerek modelin doğruluğunu artırmasıdır. Özellikle gürültülü veya yoğun örtüşen veri kümelerinde, şüpheli örnekleri temizleyerek sınıflandırma hatalarını azaltır. Bu yöntem, genellikle SMOTE gibi aşırı örnekleme algoritmalarıyla birlikte kullanılarak, sentetik veriler sonrası kalan karışık veya anlamlı olmayan çoğunluk örneklerini ayıklayıp daha dengeli ve net bir veri kümesi oluşturulmasına yardımcı olur (Altalhan ve diğ., 2025).

3.6.17. SMOTE Nominal ve Sürekli Değişkenler için (SMOTE for Nominal and Continuous (SMOTE-NC))

SMOTE-NC algoritması, hem sayısal (sürekli) hem de kategorik (nominal) değişkenler içeren karmaşık veri kümeleri için geliştirilmiş klasik SMOTE yönteminin ileri düzey bir versiyonudur. Geleneksel SMOTE yalnızca sayısal verilerle çalışabilirken, SMOTE-NC algoritması her iki değişken türünü dikkate alarak daha gerçekçi ve anlamlı sentetik örnekler üretmektedir. Bu yöntemde, kategorik özellikler arasındaki farklılıklar da mesafe ölçümüne dâhil edilerek anlamsız veya geçersiz örneklerin oluşturulması önlenir. Sayısal veriler için Öklid mesafesi kullanılırken, kategorik değişkenler için ceza katsayısı eklenmiş birleşik bir mesafe ölçütü uygulanır. Yeni örnekler, sayısal alanlarda interpolasyon yoluyla; kategorik alanlarda ise en sık karşılaşılan değer (mod) temel alınarak üretilir (Mukherjee ve Khushi, 2021).

SMOTE-NC algoritması, özellikle tıbbi veri analizleri, sosyal bilimler araştırmaları ve müşteri davranışlarının modellenmesi gibi hem sayısal hem de kategorik değişkenler içeren uygulama alanlarında etkili sonuçlar sunmaktadır. Bu yöntem, veri yapısını bozmadan sınıf dengesizliği problemini hassas biçimde ele alır ve anlamlı yapay veriler oluşturarak makine öğrenmesi

modellerinin azınlık sınıflarına ait örüntüleri daha doğru biçimde tanımasını sağlar. Özellikle verilerin hem dengesiz hem de çeşitlilik gösterdiği durumlarda, SMOTE-NC daha güvenilir ve dengeli modelleme imkânı sunar (Rahmayanti ve diğ., 2021).

3.6.18 Nominal Veriler için Sentetik Azınlık Aşırı Örnekleme TekniğiSMOTEN (Synthetic Minority Over-sampling Technique for Nominal Data(SMOTEN))

SMOTEN algoritması, tamamen kategorik (nominal) özelliklerden oluşan veri kümeleri için tasarlanmış klasik SMOTE yönteminin özel bir uyarlamasıdır. Bu algoritma, Öklid mesafesi yerine kategorik veriler arasındaki benzerliği ölçmek için "değer farkı metriği"ni kullanır. Böylece, yeni sentetik örnekler oluşturulurken özellik değerleri, komşular arasında en sık rastlanan (mod) değerlere göre belirlenir ve verilerin anlamsal tutarlılığı korunur. SMOTEN, sosyal bilimler, müşteri analizleri gibi büyük ölçüde kategorik değişkenler içeren veri setlerinde sınıf dengesizliğini gidermek amacıyla etkili bir şekilde kullanılmaktadır (Shri ve diğ., 2024).

Bununla birlikte, SMOTEN yalnızca nominal verilere uygun olduğundan, sayısal ya da karma veri kümelerinde etkili değildir. Ayrıca rastgele komşu seçiminden dolayı bazı durumlarda tekrar eden veya anlamsız veriler üretme riski taşır ve bu durum modelin doğruluğunu olumsuz etkileyebilir. Ancak veri tamamen kategorik yapıdaysa, SMOTEN bu tür verilerle daha iyi çalışan Naive Bayes veya karar ağaçları gibi algoritmaların performansını artırmak için güçlü ve uygun bir araç olarak öne çıkar (Altalhan ve diğ., 2025) .

3.7 DEĞERLENDİRİLEN MAKİNE ÖĞRENMESİ ALGORİTMALARI

Bu bölümde, önerilen modelin performansını kapsamlı bir şekilde ortaya koymak amacıyla, klasik ve ileri düzey makine öğrenmesi algoritmalarından oluşan bir dizi yöntem kısaca tanıtılmış ve incelenmiştir. Bu algoritmalar; doğrusal yöntemler, ağaç tabanlı yaklaşımlar, topluluk (Ensemble) modelleri, sinir ağları ve çeşitli istatistiksel modelleri kapsamaktadır. Her birinin,veri kümelerinin işlenmesi ve ilaç-hedef etkileşimlerinin tahmin edilmesi konusunda kendine özgü avantajları ve sınırlamaları bulunmaktadır. Bu algoritmaların gözden geçirilmesi ve sonuçlarının geliştirilen modelle karşılaştırılması, önerilen çalışmanın güçlü yönlerinin, performans iyileştirmelerinin ve yenilikçi unsurlarının daha ayrıntılı değerlendirilmesi için uygun bir zemin hazırlamaktadır.

3.7.1 K-En Yakın Komşu (k-Nearest Neighbors (K-NN)) Algoritması

K-en Yakın Komşu (k-NN) algoritması, örnekler arasındaki uzaklık veya benzerlik ölçümüne dayanan, en basit ve en yaygın kullanılan gözetimli makine öğrenmesi yöntemlerinden biridir. Bu algoritma, tüm eğitim örneklerini ve etiketlerini belleğe kaydeder ve yeni bir veri örneği geldiğinde, bu örneğin tüm eğitim verilerine olan uzaklıklarını hesaplayarak en yakın k komşuyu seçer. Ardından, bu k komşu arasındaki çoğunluk oylamasına göre yeni verinin sınıfı belirlenir. Özellikle k=1 olduğunda, en yakın komşunun sınıfı doğrudan tahmin edilir. Doğru k değerinin seçimi oldukça önemlidir; zira çok küçük bir k değeri, modelin aşırı uyum göstermesine, çok büyük bir k değeri ise modelin gürültülü verilere karşı hassas hale gelmesine ve dolayısıyla tahmin doğruluğunun azalmasına neden olabilir (Bogale, 2025).

Bu algoritmada, genellikle tüm sınıfların varyanslarının eşit olduğu varsayıldığında, örnekler arasındaki mesafe Öklid mesafesi ile ölçülür. Bu durumda, iki örnek arasındaki mesafe(x_{ai} ,x_{bi}), özellikler arasındaki farkların karelerinin toplamının karekökü alınarak hesaplanır; Formül 3.26'da p, özellik uzayının boyut sayısını ifade eder (Dai, 2019).

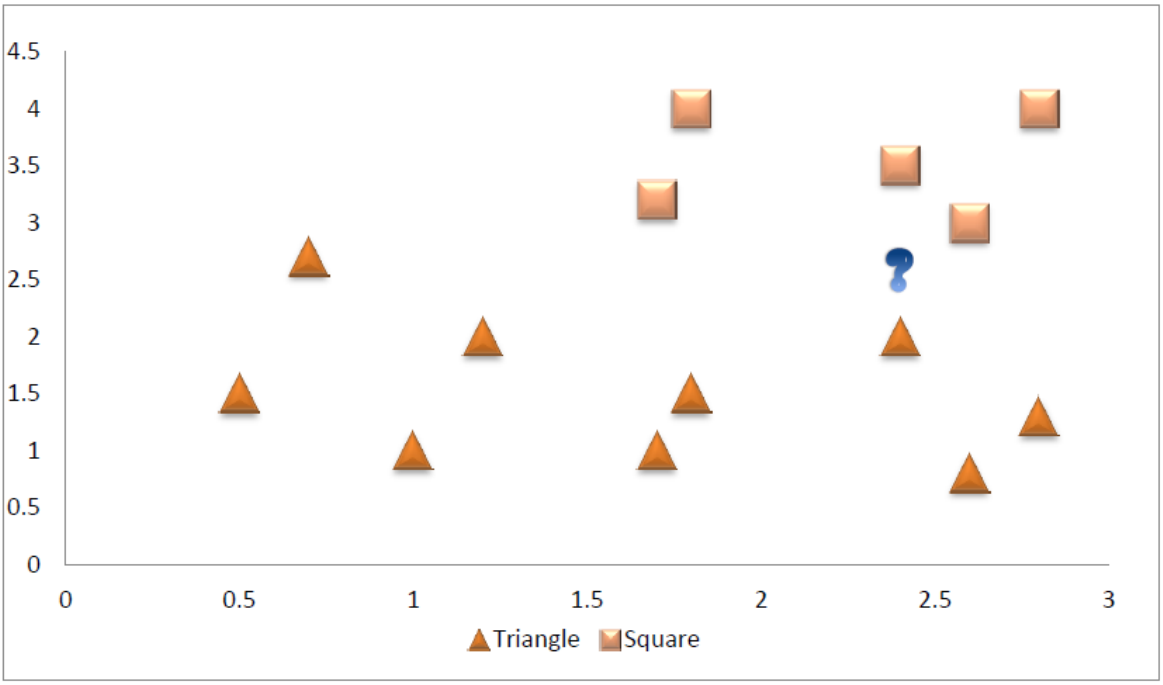
$$d = \sqrt{\sum_{i=1}^{p}(x_{ai} - x_{bi})^2} \tag{3.26}$$

Ancak, farklı sınıflar için varyanslar farklılık gösteriyorsa, Manhattan mesafesinin kullanılması daha uygun olur. Manhattan mesafesinde Formül 3.27'e göre , örneklerin ortalama değerleri (m_a ve m_b) arasındaki fark ve örnekler (x_a ,x_b)arasındaki kovaryans (C) hesaba katılır. Mesafe kriterinin doğru seçimi, k-NN algoritmasının farklı veri yapılarında daha yüksek doğrulukla çalışmasına ve veri dağılımından kaynaklanan sapmaların daha iyi kontrol edilmesine yardımcı olur (Dai, 2019).

$$d = \sqrt{(x_a - m_a)C^{-1}(x_b - m_b)} \tag{3.27}$$

Parametrik olmayan yapısı ve karmaşık modellemeye ihtiyaç duymaması nedeniyle, k-NN algoritması finansal dolandırıcılık tespiti, içeriden gelen tehditlerin belirlenmesi ve tıbbi veri sınıflandırması gibi pek çok farklı uygulamada popüler bir yöntemdir. Uygulaması oldukça basit olan bu algoritma, yeni verilere kolaylıkla uyum sağlayabilmektedir. Bununla birlikte, k-NN algoritmasında test aşamasının, tüm eğitim örneklerine uzaklık hesaplaması gerektiğinden,

eğitim aşamasına göre daha yavaş olduğu unutulmamalıdır. Ayrıca, verilerde alakasız veya gürültülü özellikler bulunması, algoritmanın performansını ciddi şekilde olumsuz etkileyebilir. Bu nedenle, k-NN performansını artırmak için özellik seçimi, veri normalizasyonu ve optimum kkk değerinin belirlenmesi gibi teknikler, genellikle çapraz doğrulama yöntemleri kullanılarak uygulanır (Zathry ve diğ., 2025). k-en yakın komşu yönteminin gösterimi Şekil 3.7'de gösterilmiştir.

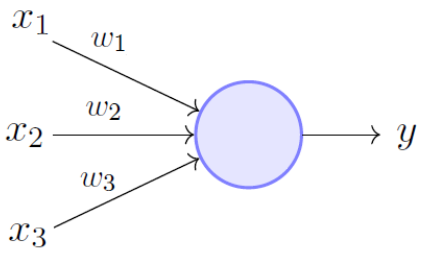


Şekil 3. 9: k-en yakın komşu yönteminin gösterimi

3.7.2 Yapay Sinir Ağları

Yapay sinir ağlarının geliştirilmesi, bilgisayarların hız, doğruluk ve tarafsızlık gibi temel özelliklerini koruyarak, onların insan düşüncesine ve anlayışına daha da yakınlaşmasını sağlayan önemli bir adım olmuştur. Bu alandaki temel dallardan biri olan Derin Öğrenme , yapay nöronlar adı verilen basit algoritmalar aracılığıyla karmaşık problemleri çözme yeteneğini kazanmıştır. Yapay nöronların ilk modeli olan perceptron, Frank Rosenblatt tarafından geliştirilmiş olup, ikili girdileri ağırlıklandırarak, bu ağırlıklı toplamı bir eşik değeri (bias) ile karşılaştırarak ikili bir çıktı üretmektedir (Park ve Lek, 2016). Şekil 3.8'de Perceptron

şeması ve Formül 3.28, perceptron modelinin ağırlıklı girdileri eşik değeriyle karşılaştırarak ikili bir çıktı oluşturma sürecini göstermektedir.



Şekil 3. 10: Perceptron şeması

$$output = \begin{cases} 0\,, if\ \sum_j w_i x_i \leq threshold \\ 1\,,\ otherwise \end{cases} \tag{3.28}$$

Zaman içerisinde, yapay nöronların daha gelişmiş modelleri tasarlanmıştır. Bunlardan biri, günümüz sinir ağlarında sıklıkla kullanılan sigmoid nörondur. Farklı ağırlık ve eşik değerlerine sahip çok sayıda nöronun çeşitli katmanlarda birleştirilmesi, güçlü ve çeşitli ağların oluşmasını mümkün kılmıştır. Derin öğrenme, çok katmanlı ağların kullanılması anlamına gelir ve bu yapı, verinin daha soyut ve üst düzey temsillerinin öğrenilmesini sağlar. Bu bağlamda, giriş, gizli ve çıkış katmanlarından oluşan, her katmanın ağırlıklar ve aktivasyon fonksiyonları yardımıyla veri işleyip özelliklerini çıkardığı çok katmanlı algılayıcılar en sık kullanılan yapılardan biridir (Park ve Lek, 2016).

Bu ağlarda öğrenme süreci, ağırlıklar için en uygun değerlerin bulunmasıyla gerçekleştirilir. Başlangıçta rastgele belirlenen ağırlıklar, hata geri yayılımı (Backpropagation) ve Adam gibi optimize edici algoritmalar kullanılarak güncellenir. Ağın performansı, tahmin edilen değerler ile gerçek değerler arasındaki farkı ölçen kayıp fonksiyonları yardımıyla değerlendirilir. Eğitim sürecinde, aşırı uyumu önlemek için Dropout, eğitim sürecini aşırı uzatmamak için Early Stopping gibi teknikler ve solan gradyan problemini aşmak için ReLU gibi doğrusal olmayan aktivasyon fonksiyonlarının kullanımı, ağların başarısında kilit rol oynamaktadır (Reyad ve diğ., 2023).

3.7.3 Naive Bayes Algoritması

Naive Bayes algoritması, Bayes teoremine dayanan ve özellikler arasında bağımsızlık varsayan, veri sınıflandırması için klasik ancak etkili yöntemlerden biridir. Bu algoritmada, koşullu olasılıklar hesaplanarak her bir örneğin farklı sınıflara ait olma olasılığı tahmin edilir ve örnek, en yüksek olasılığa sahip sınıfa atanır. Algoritmanın hafif yapısı ve yüksek işlem hızı, Naive Bayes'i büyük veri boyutlarına sahip veya hızlı tahminler gerektiren projeler için ideal bir seçim haline getirmektedir (Xu, 2018).

Naive Bayes algoritmasının temelini oluşturan Formül 3.29'da gösterilmiştir:

$$\frac{P(Y)\times P(X_1,X_2,\ldots,X_n\,|Y)}{P(X_1,X_2,\ldots..X_n)} = P(Y|X_1,X_2,\ldots,X_n) \tag{3.29}$$

Burada:

- Y sınıf etiketini,
- X1,X2,...,Xn ise örneklemdeki özellikleri temsil etmektedir.

Özelliklerin birbirinden bağımsız olduğu varsayıldığından, P(X1,X2,...,Xn|Y) ifadesi bireysel olasılıkların çarpımı olarak yeniden yazılabilir, bu da Formül 3.30'de aşağıdaki gibi gösterilebilir

$$P(X_K\,|Y)\prod_{k=1}^{n}\times P(Y) \propto P(Y|X_1,X_2,\ldots..,X_n) \tag{3.30}$$

Bu basitleştirme, hesaplamaları çok daha hızlı ve kolay hale getirmekte ve bu durum, Naive Bayes'in veri odaklı projelerde bu kadar popüler olmasının temel sebeplerinden biri olarak öne çıkmaktadır (Ismail ve diğ., 2020).

Naive Bayes algoritması, veri türüne bağlı olarak üç ana versiyonda sunulmaktadır:

- **Bernoulli Naïve Bayes**: Özelliklerin ikili değerlere sahip olduğu (örneğin 0 veya 1) veri türleri için uygundur ve özelliklerin Bernoulli dağılımını izlediği varsayılır.
- **Multinomial Naïve Bayes**: Özelliklerin ortaya çıkma sıklığının (örneğin bir metindeki kelime frekansı gibi) önemli olduğu verilerde kullanılır. Bu modelde, olasılık tahminlerini iyileştirmek için genellikle Laplace düzeltmesi gibi teknikler uygulanır.

- **Gaussian Naïve Bayes**: Özelliklerin sürekli olduğu ve normal (Gauss) dağılım varsayımına dayandığı durumlar için kullanılır. Bu durumda, her özellik için ortalama (μ\muμ) ve standart sapma (σ\sigmaσ) hesaplanır ve bu değerler üzerinden olasılıklar tahmin edilir (Ismail ve diğ., 2020).

Naive Bayes algoritmasının uygulama alanı oldukça geniştir; ilaç etkileşim tahmini, içeriden gelen tehditlerin tespiti, metin sınıflandırması, spam algılama ve duygu analizi gibi birçok farklı alanda kullanılmaktadır. Özellikle metin tabanlı projelerde, "kelime torbası" (Bag of Words) gibi tekniklerle birlikte modelin basitliği ve yüksek verimliliği sayesinde yaygın şekilde tercih edilmektedir (Uddin ve diğ., 2022).

Genel olarak, Naive Bayes algoritması; uygulamadaki basitliği, eğitim verilerine olan düşük ihtiyaç, parametre tahminlerindeki verimlilik ve farklı veri türlerine olan uyum yeteneği sayesinde birçok makine öğrenmesi probleminde en popüler ve etkili algoritmalardan biri olarak kabul edilmektedir (Xu, 2018).

3.7.4 Destek Vektör Makinesi (Support Vector Machine (SVM))

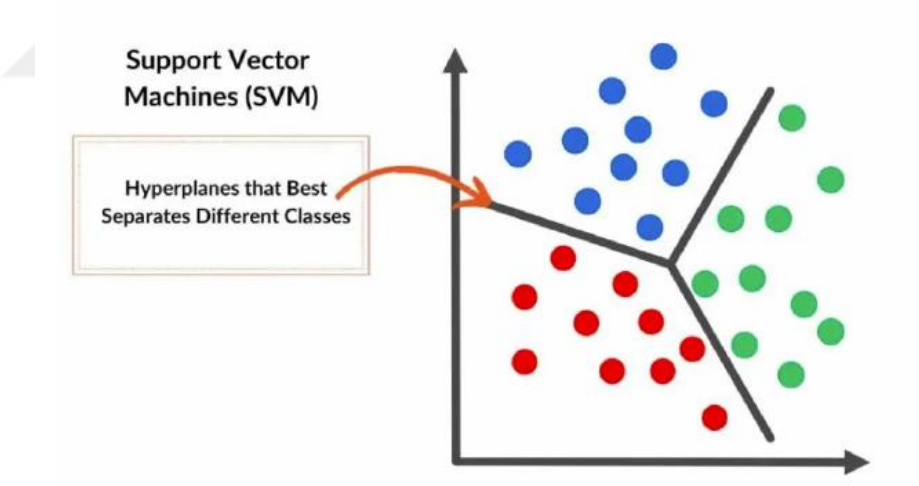
Destek Vektör Makinesi, gözetimli makine öğrenimi alanında sıkça kullanılan güçlü bir algoritmadır. Genellikle sınıflandırma görevlerinde tercih edilmekle birlikte, zaman zaman regresyon problemlerinde de etkili bir şekilde kullanılabilir. Bu algoritmanın temel amacı, iki farklı sınıfa ait verileri en geniş marjla ayırabilen optimal bir hiper düzlem belirlemektir. Bu özellik, modelin genelleştirilebilirliğini artırır ve daha önce karşılaşılmamış veriler üzerinde yüksek performans göstermesini sağlar (Shetty ve diğ., 2022). SVM'deki hiper düzlemler Şekil 3,9'da gösterilmiştir.

SVM'in temel yapısında, eğitim verileri ikili etiketlere sahip örneklerden oluşur. Algoritma, yapılandırılmış bir optimizasyon problemini çözerek, yalnızca sınıflandırma hatalarını en aza indirmeyi değil, aynı zamanda sınıflar arasındaki mesafeyi de maksimize etmeyi hedefler. Modelin esnekliğini artırmak adına, belirli ölçüde sınıflandırma hatasına izin verilir. Bununla birlikte, bu hataları sınırlamak ve modelin genel sağlamlığını korumak için çeşitli düzenleyici ve ceza mekanizmaları da sürece dâhil edilir (Amarappa ve Sathyanarayana, 2014).

SVM'nin en önemli avantajlarından biri, çekirdek (kernel) fonksiyonlarının kullanımına olanak tanımasıdır. Bu fonksiyonlar, verileri daha yüksek boyutlu bir özellik uzayına dönüştürerek,

doğrusal olmayan sınıf ayrımlarının bile mümkün hale gelmesini sağlar. Yaygın olarak kullanılan çekirdekler arasında doğrusal (linear) çekirdek, radyal tabanlı fonksiyon (RBF), polinom, sigmoid ve önceden hesaplanmış çekirdekler yer alır. Modelin başarısı üzerinde doğrudan etkili olan doğru çekirdeğin seçimi, genellikle deneme-yanılma yöntemleri ya da hiperparametre ayarlamalarıyla gerçekleştirilir (Patle ve Chouhan, 2013).

Yüksek boyutlu uzaylarda güçlü performans sergilemesi, aşırı uyuma karşı dirençli yapısı ve özellik sayısının örnek sayısından fazla olduğu durumlarda sağladığı verimlilik sayesinde SVM, konuşma tanıma, görüntü analizi, saldırı tespiti, finansal dolandırıcılık algılama ve zararlı URL sınıflandırması gibi birçok alanda yaygın olarak kullanılmaktadır. Ancak bu algoritma tamamen kusursuz değildir. SVM'nin başlıca dezavantajları arasında, 10.000'den fazla örnek içeren büyük veri kümelerinde verimliliği olumsuz etkileyen ikinci dereceden zaman karmaşıklığı yer alır. Ayrıca, SVM modelleri varsayılan olarak sınıflandırma olasılıklarını hesaplayamaz; bu nedenle, olasılıksal çıktıların gerekli olduğu durumlarda tamamlayıcı yöntemlerin kullanılmasına ihtiyaç duyulur (Gaye ve diğ., 2021).



Şekil 3. 11: SVM'deki hiper düzlemler (Otten, 2024)

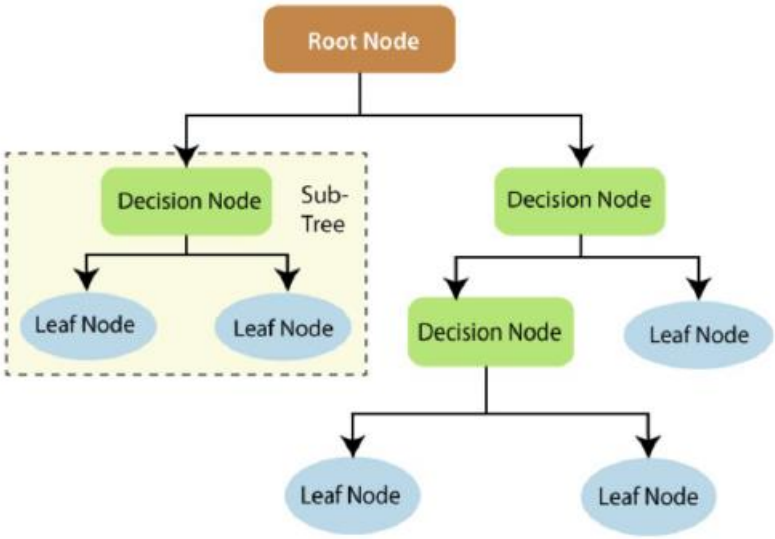
3.7.5 Karar Ağacı Algoritması

Karar ağacı algoritması, gözetimli makine öğrenmesinde en yaygın kullanılan yöntemlerden biri olup, sınıflandırma ve regresyon problemlerinde etkin bir şekilde kullanılmaktadır. Bu algoritma, hiyerarşik yapısıyla karar alma sürecini bir dizi "eğer–öyleyse" kuralı biçiminde modellemektedir. Bu modelde düğümler özellikleri, yapraklar ise sınıflandırma sonuçlarını

temsil eder. En çok bilinen karar ağacı türleri arasında; özellikleri ayırmak için entropi, Gini indeksi veya ki-kare testi gibi farklı ölçütlerden yararlanan ID3, C4.5, CART, CHAID ve QUEST algoritmaları yer almaktadır (Gaye ve diğ., 2021).

Ağaç oluşturma sürecinde, veri kümesindeki düzensizliği ölçmek için entropi kullanılırken, bilgi kazancı ise en uygun özelliği belirlemek adına hesaplanır. Özellik seçimi, bilgi karmaşasını en aza indirecek biçimde yapılır. Karar ağaçları; yorumlaması kolay olmaları, sayısal ve kategorik verilerle çalışabilmeleri ve istatistiksel varsayımlara gerek duymamaları nedeniyle önemli avantajlar sunar. Ancak, bu algoritmanın bazı sınırlamaları da mevcuttur. Aykırı değerlere karşı duyarlılık, aşırı öğrenme (overfitting) riski ve büyük veri kümelerinde hesaplama maliyetinin artması bu dezavantajlar arasında yer almaktadır (Mathan ve diğ., 2018).

Karar ağaçlarının sade yapısı ve yorumlanabilirlik özellikleri, onları genellikle Rastgele Orman (Random Forest) ve XGBoost gibi daha gelişmiş algoritmalar için temel bir yapı taşı hâline getirir. Bu gelişmiş yöntemler, birden fazla karar ağacını bir araya getirerek hem tahmin doğruluğunu artırmayı hem de aşırı öğrenme gibi sorunları azaltmayı hedefler. Bu yönüyle, karar ağaçları yalnızca tek başına güçlü bir modelleme aracı olmakla kalmaz, aynı zamanda hibrit yaklaşımlar ve karmaşık yapay zeka sistemlerinin geliştirilmesinde de kilit bir rol oynar (Sagi ve Rokach, 2021).Örnek bir karar ağacı Şekil 3.10'da gösterilmiştir.



Şekil 3. 12: Örnek bir karar ağacı (Wang ve Saddik, 2023)

3.7.6 Kareli Ayrımcı Analiz (Quadratic Discriminant Analysis(QDA))

QDA algoritması, istatistik ve olasılık temelli çalışan, makine öğrenmesinde sınıflandırma görevleri için kullanılan klasik ve güçlü yöntemlerden biridir. Doğrusal ayırıcı analiz (Linear Discriminant Analysis (LDA)) algoritmasından farklı olarak (ki bu algoritma tüm sınıfların kovaryans matrislerinin eşit olduğunu varsayar) QDA'da her sınıf için farklı kovaryans matrislerinin kullanılması mümkündür. Bu farklılık, QDA'ya sınıflar arasındaki karmaşık karar sınırlarını modelleme açısından önemli bir esneklik kazandırır; çünkü bu sınırlar doğrusal değil, ikinci dereceden eğriler şeklinde olabilir (Ghojogh ve diğ., 2019).

QDA süreci birkaç temel adımdan oluşur. İlk olarak, her sınıf için ortalama vektör ve kovaryans matrisini içeren bir Gauss olasılık yoğunluk fonksiyonu modellenir. Daha sonra, yeni bir veri örneği için bu fonksiyonlar yardımıyla her sınıfa ait olma olasılığı hesaplanır ve sınıflandırma kararı verilir. Hesaplamaları sadeleştirmek amacıyla genellikle logaritmik olasılık fonksiyonu kullanılır; bu da sonuçta özellik uzayında ikinci dereceden bir karar denklemi ortaya çıkarır. Her sınıfa ait bu fonksiyonlardan hangisi daha yüksek değer veriyorsa, veri o sınıfa atanır (Ghojogh ve diğ., 2019).

QDA'nın en güçlü yönlerinden biri, sınıflar arasındaki yapısal farklılıkları dikkate alabilmesidir. Bu sayede, özellikle veri dağılımının karmaşık olduğu durumlarda LDA'ya kıyasla daha yüksek sınıflandırma başarısı elde edebilir. Örneğin, prostat kanserinin ultrason verileri üzerinden teşhis edilmesinde, QDA'nın dağılım yapılarını daha esnek bir şekilde modelleyebilmesi sayesinde daha doğru sonuçlar verdiği gözlemlenmiştir. Ancak bu esneklik, özellikle özellik sayısının fazla ve örnek sayısının az olduğu durumlarda, parametre sayısının artması ve aşırı uyum riskini beraberinde getirebilir (Siqueira ve diğ., 2017).

Genel olarak QDA, Gauss dağılımlı ve sınıflar arası farklı kovaryans yapılarına sahip veri kümeleri için etkili bir sınıflandırma aracıdır. Veri dağılımları normalden sapma gösterdiğinde veya uç değerlerin yoğun olduğu durumlarda performansı düşebilir; ancak sınıflar arasında doğrusal olmayan ayrımların bulunduğu problemlerde hâlâ güvenilir bir alternatiftir. QDA ile LDA arasında yapılacak tercih, verinin istatistiksel yapısına bağlı olarak belirlenmeli ve bu seçim, modelin genel doğruluğu üzerinde belirleyici bir rol oynayabilir (Eya, 2024).

3.7.7 Doğrusal Ayırıcı Analiz (Linear Discriminant Analysis (LDA))

Doğrusal Ayırıcı Analiz , veri sınıfları arasındaki istatistiksel farklılıkları modellemek ve bu sınıflar arasındaki ayrımı maksimize etmek amacıyla geliştirilmiş, gözetimli öğrenme alanının en temel ve yaygın algoritmalarından biridir. Hem sınıflandırma hem de boyut indirgeme amacıyla kullanılan LDA, yalnızca verilerin genel varyansını dikkate alan PCA gibi yöntemlerin aksine, verilerin etiketli yapısını da göz önünde bulundurur (Zhao ve diğ., 2024).

LDA, orijinal özelliklerin doğrusal kombinasyonlarını bularak verileri daha düşük boyutlu bir uzaya projekte eder (Archana, 2015). Bu projeksiyon, sınıflar arası mesafeyi maksimuma çıkarırken, sınıf içi varyansı ise minimuma indirger. Bu sayede, yeni elde edilen özellik uzayında sınıflar birbirlerinden daha net bir şekilde ayrılabilir hâle gelir. Başka bir deyişle, LDA veri grupları arasında en güçlü ayrımı sağlayan doğrusal bir dönüşüm gerçekleştirir (Tharwat ve diğ., 2017).

Scikit-Learn gibi popüler kütüphaneler aracılığıyla uygulanan LDA, yalnızca etkili bir sınıflandırma algoritması olmakla kalmaz; aynı zamanda özellikle URL sınıflandırması ya da yüz tanıma gibi yüksek boyutlu ve karmaşık problemler için güçlü bir boyut indirgeme tekniği olarak da öne çıkar. Bu tür uygulamalarda, LDA'nın sınıfları en iyi şekilde ayıran özellik kombinasyonlarını bulma kabiliyeti büyük önem taşır (Corchado ve diğ., 2023).

Sınıflandırma sürecinde LDA'nın çıktısı genellikle modelin doğruluk, hassasiyet ve diğer performans ölçütlerinin değerlendirildiği bir karışıklık matrisi (confusion matrix) aracılığıyla sunulur. Algoritma, özellikle sınıflar arasında benzer varyansa ve normal dağılıma sahip veri kümelerinde oldukça başarılı sonuçlar verir. Ancak, verilerin bu istatistiksel varsayımlara uymadığı durumlarda LDA'nın performansı düşebilir (Choubey ve diğ., 2020).

3.7.8 Kolay Topluluk Sınıflandırıcısı (Easy Ensemble Classifier (EEC))

Kolay Topluluk Sınıflandırıcısı algoritması, sınıflandırma problemlerinde karşılaşılan sınıf dengesizliği sorununu ele almak için geliştirilen etkili ve güçlü yöntemlerden biridir (Yijing ve diğ., 2016). Bu algoritma, Rastgele Alt Örnekleme (Random Under-Sampling) ile AdaBoost tabanlı öğrenme tekniklerinin birleşimine dayanmaktadır. EEC yönteminde, çoğunluk sınıfına ait veriler rastgele bir şekilde birkaç alt kümeye ayrılır; her bir alt küme, azınlık sınıfının tüm örnekleriyle birlikte, çoğunluk sınıfından seçilen rastgele örnekleri içerir. Bu sayede, her model için dengeli bir eğitim veri kümesi elde edilir ve modelin çoğunluk sınıfına karşı önyargı geliştirmesi engellenmiş olur (Wang ve diğ., 2019).

Bu dengeli alt kümelerin her biri, AdaBoost sınıflandırıcısı tarafından ayrı ayrı eğitilmektedir. AdaBoost algoritması, önceki yinelemelerde doğru şekilde sınıflandırılamayan örneklere odaklanarak, modelin özellikle kritik karar sınırlarında kademeli olarak daha iyi performans göstermesini sağlar. Ardından, her bir güçlendirilmiş modelin çıktısı, genellikle çoğunluk oyu (majority voting) gibi bir toplama tekniği aracılığıyla birleştirilir ve nihai tahmin elde edilir. Bu çok katmanlı yapı sayesinde, EEC algoritması hem boosting hem de alt örnekleme yöntemlerinin güçlü yönlerini bir araya getirerek, sınıflandırma başarımını önemli ölçüde artırır (Monowar ve diğ., 2024).

Easy Ensemble algoritmasının, Bagging ya da klasik AdaBoost gibi diğer topluluk (ensemble) öğrenme yöntemlerine kıyasla dengesiz veri setleri üzerindeki en belirgin avantajı, azınlık sınıfını sınıflandırmadaki yüksek doğruluk oranı ve performans kararlılığıdır . Özellikle azınlık sınıfının örnek sayısının çoğunluk sınıfına göre oldukça düşük olduğu durumlarda, geleneksel algoritmalar bu sınıfı göz ardı etme eğilimindedir. Oysa EEC, her bir model eğitiminde dengeli alt veri kümeleri kullanarak ve zorlayıcı örnekleri ön plana çıkararak, azınlık sınıfına ait örneklerin doğru sınıflandırılma oranını önemli ölçüde artırmaktadır. Bu özelliği sayesinde,

hastalık teşhisi, dolandırıcılık tespiti gibi yüksek hassasiyet gerektiren medikal ve finansal uygulamalarda oldukça değerli bir araç haline gelmektedir (Mienye ve Sun, 2022).

Kolay Topluluk Sınıflandırıcı algoritması, genelleştirilebilirlik açısından üstünlüğü, aşırı uyum riskinin azaltılması ve dengesiz veri kümelerine karşı gösterdiği direnç sayesinde, özellikle tıbbi ve biyolojik veri analizlerinde sıkça tercih edilen yöntemlerden biri olarak öne çıkmaktadır (Yu ve Ni, 2014). Bagging veya Dengeli Rastgele Orman (Balanced Random Forest) gibi bazı yöntemler, çoğunluk sınıfına ait verilerin aşırı derecede azaltılması ya da karar sınırlarına yeterince odaklanılamaması nedeniyle zaman zaman düşük performans sergileyebilmektedir. Buna karşın, EEC algoritması önemli örnekleri koruyarak ve önyargıyı minimize ederek bu dengenin sağlanmasında başarılı bir yaklaşım sunar. AdaBoost'un adaptif öğrenme kabiliyeti ile kontrollü alt örneklemenin etkili bir biçimde bütünleştirilmesi, bu yöntemi özellikle hassasiyet gerektiren analizler için güçlü ve güvenilir bir seçenek haline getirmektedir (Rane ve diğ., 2024).

3.7.9 Dengeli Torba Sınıflandırıcı (Balanced Bagging Classifier (BBC))

Dengeli Torba Sınıflandırıcı algoritması, makine öğrenmesi problemlerinde sınıf dengesizliğiyle başa çıkmak için etkili yaklaşımlardan biridir(Lango ve Stefanowski, 2018). Bu yöntemin temelinde, sınıflandırma modelleri eğitilmeden önce dengeli alt veri kümelerinin oluşturulması yer alır. Başka bir ifadeyle, çoğunluk sınıfına ait örneklerin sayısı azaltılırken, azınlık sınıfına ait tüm örnekler korunarak dengeli bir veri kümesi elde edilir. Elde edilen bu alt kümeler, genellikle karar ağaçlarına dayalı çeşitli sınıflandırma modellerinin eğitilmesinde kullanılır. Bu yapı sayesinde, nihai model çoğunluk oylaması (majority voting) ile birden fazla modelin çıktısını birleştirerek karar verir ve bu sayede sınıf dengesizliği bulunan veri kümelerinde sınıflandırma doğruluğu önemli ölçüde artırılır (Bozcuk ve Yıldız, 2024).

Geleneksel Bagging yöntemleri genellikle yer değiştirmeli örnekleme esasına dayanır ve sınıf dağılımındaki dengesizliklere duyarsızdır. Buna karşın, Dengeli Bagging algoritması, sınıflar arasında denge oluşturmayı hedefleyerek özellikle zorlayıcı sınıflandırma problemleriyle başa çıkmak için daha etkili bir yaklaşım sunar. Bu yöntem, özellikle azınlık sınıfının veri setinde oldukça küçük bir orana sahip olduğu durumlarda, modelin kritik örüntüleri göz ardı etmesini engeller. Ayrıca, alt veri kümelerinin dikkatli biçimde seçilmesi ve eğitim verisinin miktarının

kontrollü bir şekilde belirlenmesi sayesinde, modelin belirli örneklere aşırı güvenmesi ya da veriye aşırı uyum sağlaması riski önemli ölçüde azaltılır (Roshan ve Asadi, 2020).

BBC algoritmasının öne çıkan yönlerinden biri, bileşen sınıflandırıcılarının sayısının ve temel öğrenme algoritmasının esnek biçimde ayarlanabilmesidir. Yapılan araştırmalar, nihai topluluk modelinin başarısının büyük ölçüde kullanılan temel algoritmanın türüne ve bileşen modellerin tahminlerindeki çeşitliliğe bağlı olduğunu ortaya koymuştur (Roshan ve Asadi, 2020). Bununla birlikte, BBC algoritması; sınıflar arasındaki örtüşme, karar sınırlarındaki dengesizlik ve seyrek azınlık sınıfı örnekleri gibi veri zorluğu unsurlarına karşı da tatmin edici bir performans sergilemiştir. Napierala ve Stefanowski gibi araştırmacıların ortaya koyduğu metodolojiler doğrultusunda geliştirilen Kaba Dengeli Bagging (Roughly Balanced Bagging) ve BBC'nin türevleri, karmaşık veri yapılarına uyum sağlama potansiyeline sahiptir (Milli ve diğ., 2024).

Dengesiz çoklu sınıflandırma problemleri için BBC algoritmasının geliştirilmesi, araştırmacılar için yeni bir perspektif sunmuştur. Azınlık sınıfının çoğunluk sınıfına karşı modellendiği ikili sınıflandırma problemlerinin aksine, birçok gerçek dünya uygulamasında birden fazla azınlık sınıfının ayrı ayrı tanımlanması gerekmektedir. Bu durum, dengeli örnekleme tekniklerinin uygulanmasını ve model sonuçlarının birleştirilmesini daha karmaşık bir hale getirmektedir. Ancak, verileri çoklu ikili alt problemlere dönüştürmek veya BBC algoritmasını daha karmaşık yapılarla uyumlu hale getirmek gibi yenilikçi yaklaşımlar, bu senaryolarda da topluluk tabanlı güçleri kullanmaya olanak sağlamıştır. Bu tür çabalar, nadir hastalıkların teşhisi, dolandırıcılığın tespiti ve genomik veri analizi gibi kritik alanlarda büyük önem taşımaktadır (Gabriel ve diğ., 2022; Milli ve diğ., 2024).

3.7.10 Uyarlanabilir Güçlendirme (Adaptive Boosting (AdaBoost))

AdaBoost algoritması , zayıf sınıflandırıcı modellerin performansını artırmak amacıyla geliştirilmiş ve makine öğrenmesi alanında yaygın olarak kullanılan etkili bir yöntemdir. Algoritmanın temelinde, tek bir karmaşık model yerine, birden fazla basit ya da zayıf sınıflandırıcının (örneğin basit karar ağaçları) ardışık biçimde bir araya getirilmesi fikri yatar. Her bir aşamada, önceki adımlarda doğru sınıflandırılamayan örnekler üzerine odaklanılarak, modelin başarısı kademeli olarak artırılır. Bu özelliği sayesinde, AdaBoost algoritması, sistemin zayıf yönlerini sistematik şekilde tespit edip telafi eder. Özellikle yüz tanıma gibi uygulamalarda yüksek performans göstermiştir. Bu tür uygulamalarda, özellikler özdeğerler

(eigenvalue) yardımıyla çıkarılır ve her bir aşamada modele yeni bir sınıflandırıcı entegre edilir (Gabriel ve diğ., 2022).

AdaBoost algoritması yapısal olarak yinelemeli ve ağırlık tabanlı bir yapıdadır. Eğitim sürecinin başında tüm örneklere eşit ağırlıklar atanır. Ardından, yineleme ilerledikçe doğru sınıflandırılan örneklerin ağırlıkları azaltılırken, yanlış sınıflandırılanlara daha fazla ağırlık verilir. Bu dinamik ağırlıklandırma, modelin zorlayıcı örnekler üzerine daha fazla yoğunlaşmasını sağlar. Sonuç olarak, her bir zayıf sınıflandırıcı, doğruluğuna göre belirlenen ağırlıklarla doğrusal biçimde birleştirilir ve böylece güçlü bir nihai sınıflandırıcı elde edilir. Bu yapı, sınıflar arasında ağırlıklı oy kullanımı prensibiyle nihai kararın oluşturulmasını sağlar (Liu ve diğ., 2022).

AdaBoost, sinir ağları ya da destek vektör makineleri gibi diğer öğrenme algoritmalarıyla kıyaslandığında, genellikle daha az hiperparametre gereksinimiyle tatmin edici doğruluk sonuçları sunar. Belirgin avantajlarından biri, veri kümesi gürültülü olduğunda performansı düşebilse de, çoğu durumda aşırı öğrenmeyi engelleyebilmesidir. Diğer olumlu yönleri arasında, uygulanmasının kolay olması ve özellikle basit sınıflandırıcılarla kullanıldığında eğitimin görece daha hızlı gerçekleşmesi yer alır. Yüz tanıma uygulamalarında, AdaBoost'un ConvMixer gibi derin öğrenme mimarileriyle entegre edilmesi, tanıma başarımını artırmış ve sınıflandırma hatalarını azaltmıştır (Rahman ve diğ., 2015).

Genel olarak AdaBoost, sınıflandırma problemleri için güçlü, ölçeklenebilir ve esnek bir algoritma olarak öne çıkar. Gücünü, karmaşık bir yapıya sahip olmaktan ziyade, zayıf sınıflandırıcıların akıllıca birleşimi ve zor örnekler üzerine uyarlanabilir biçimde yoğunlaşabilmesinden alır. Dinamik ağırlık mekanizması, yinelemeli yapısı ve çağdaş özellik çıkarma teknikleriyle uyumlu çalışabilme kapasitesi sayesinde, AdaBoost yüz tanıma gibi ileri düzey uygulamalarda tercih edilen bir yöntem haline gelmiştir (Shahraki ve diğ., 2020).

3.7.11 Rastgele Azaltma ve Güçlendirme (RandomUndersampling ve Boosting (RUSBoost))

RUSBoost algoritması, veri dengesizliği sorununu ele almak üzere baskın sınıftan rastgele örnekleme (undersampling) ile öğrenicilerin aşamalı olarak güçlendirilmesini (boosting) bir

araya getiren hibrit bir yaklaşımdır . Dengesiz veri setlerinde, geleneksel sınıflandırıcılar genellikle baskın sınıfa öncelik verirken, azınlık sınıfına ait örnekleri göz ardı etme eğilimindedir. RUSBoost algoritması, her yinelemede baskın sınıfa ait örneklerin sayısını rastgele azaltarak sınıflar arasında yapay bir denge oluşturur; ardından boosting yöntemiyle hatalı sınıflandırılan örneklere odaklanarak son sınıflandırıcıyı kademeli olarak eğitir. Bu yaklaşım, elde edilen modelin hem genel doğruluk hem de azınlık sınıfları tanıma yeteneği açısından daha yüksek performans sergilemesini sağlar (Seiffert ve diğ., 2010).

RUSBoost algoritması, aşamalı ve yinelemeli bir öğrenme sürecine sahiptir. Başlangıçta tüm örneklere eşit ağırlık atanır. Her yinelemede, rastgele örnekleme ve ağırlıklı dağılım yöntemleri kullanılarak daha dengeli bir alt veri kümesi oluşturulur. Bu geçici alt küme, genellikle zayıf bir karar ağacı gibi basit bir öğreniciyle eğitilir. Eğitimin ardından, oluşturulan sınıflandırıcının hatası sözde-kayıp (pseudo-loss) metriğiyle hesaplanır. Elde edilen hata değeri, bir sonraki adımda sınıflandırıcının katkı katsayısı (α) olarak kullanılır. Son aşamada, sınıflandırıcının başarı durumuna göre örnek ağırlıkları güncellenir; böylece algoritma, sonraki yinelemelerde hatalı sınıflandırılmış örneklere daha fazla odaklanır(Forslund, 2022).

RUSBoost'un dikkat çeken yönlerinden biri, undersampling ve boosting tekniklerini birlikte kullanarak her iki yöntemin dezavantajlarını dengelemesidir (Forslund, 2022). Geleneksel undersampling yöntemlerinde önemli veriler kaybolabilir; ancak RUSBoost'ta örnekleme her yinelemede farklı yapıldığından ve öğreniciler farklı alt kümeler üzerinde eğitildiğinden, model geniş bir veri yelpazesinden faydalanır. Öte yandan boosting, modeli adım adım daha güçlü hale getirerek özellikle sınıflandırılması zor örnekler üzerinde yoğunlaşmasını sağlar. Sonuç olarak, RUSBoost; uyarlanabilir öğrenme stratejisi, zayıf öğrenicilerin birleştirilmesi ve akıllı örnekleme yaklaşımıyla dengesiz veri kümelerinde üstün performans sergileyen, karmaşık sınıflandırma problemlerinde sıkça başvurulan etkili bir algoritmadır (Seiffert ve diğ., 2008) .

3.8 DEĞERLENDİRME ÖLÇÜTLERİ

Bir makine öğrenmesi algoritması eğitildikten sonra, bu algoritmanın hedeflenen fonksiyonu doğru bir biçimde öğrenip öğrenmediğinin denetlenmesi büyük önem taşır. Bu amacı gerçekleştirmek için genellikle daha önce eğitim sürecinde kullanılmamış, yeni ve bağımsız bir test veri setinden yararlanılır. Söz konusu veri seti, modelin genelleme yeteneğini ölçmek üzere

kullanılır. Bu veriler üzerinden elde edilen sonuçların incelenmesiyle çeşitli performans değerlendirme ölçütleri hesaplanabilir.

Gözetimli öğrenme kapsamında kullanılan sınıflandırma algoritmalarında, modelin başarımını ölçmek adına dört temel metrik dikkate alınır: gerçek pozitif (True Positive(TP)), gerçek negatif (True Negative(TN)), yanlış pozitif (False Positive(FP)) ve yanlış negatif (False Negative(FN)). TP, modelin bir örneği doğru sınıfa başarıyla atadığı durumları ifade ederken; TN, bir örneğin ait olmadığı sınıfa doğru şekilde dahil edilmediğini gösterir. Öte yandan FP, aslında bir sınıfa ait olmayan örneklerin model tarafından hatalı bir biçimde o sınıfa aitmiş gibi sınıflandırıldığı durumları yansıtır. FN ise, gerçekte belirli bir sınıfa ait olan ancak model tarafından yanlışlıkla o sınıfa ait değilmiş gibi değerlendirilmiş örnekleri temsil eder.

Makine öğrenmesinde, bir algoritmanın eğitildikten sonraki performansının doğru şekilde değerlendirilmesi son derece önemlidir. Bu değerlendirmenin temel amacı, geliştirilen modelin hedeflenen kalıpları doğru biçimde tanıyıp tanımadığını ve buna dayanarak güvenilir tahminler sunup sunamadığını belirlemektir. Bu amaçla genellikle, modelin eğitim sürecine dahil edilmemiş bağımsız bir test veri seti kullanılır. Bu veri seti, araştırmacılara modelin daha önce görmediği verilere genelleme yapabilme yeteneğini ölçme ve gerçek performansını değerlendirme imkanı tanır.

Model performansları hassaslık (Precision), geri çağırma (Recall), F1 skoru, genel doğruluk (Accuracy) ve ROC eğrisi altındaki alan (ROC-AUC) gibi ölçütlerle değerlendirilmiştir. Bu ölçütlerin her biri modelin farklı bir yönünü ortaya koymakta olup, daha kapsamlı bir analiz yapmaya olanak sağlamaktadır.Bu ölçütlerin Förmul 3.31-3.34 açıklanmıştır.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.31}$$

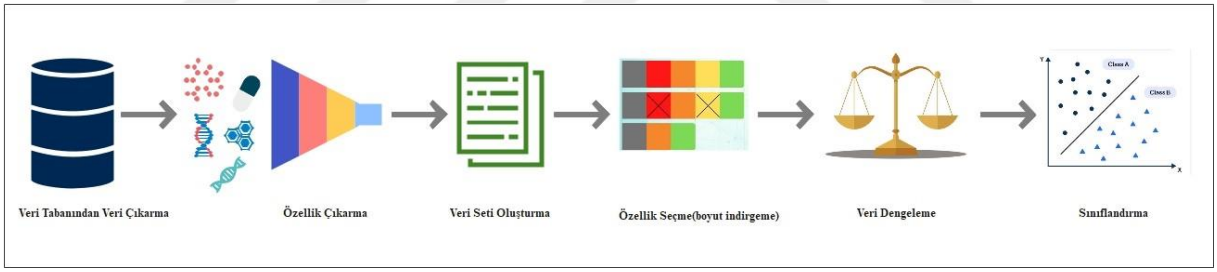
$$Precision = \frac{TP}{TP + FP} \tag{3.32}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.33}$$

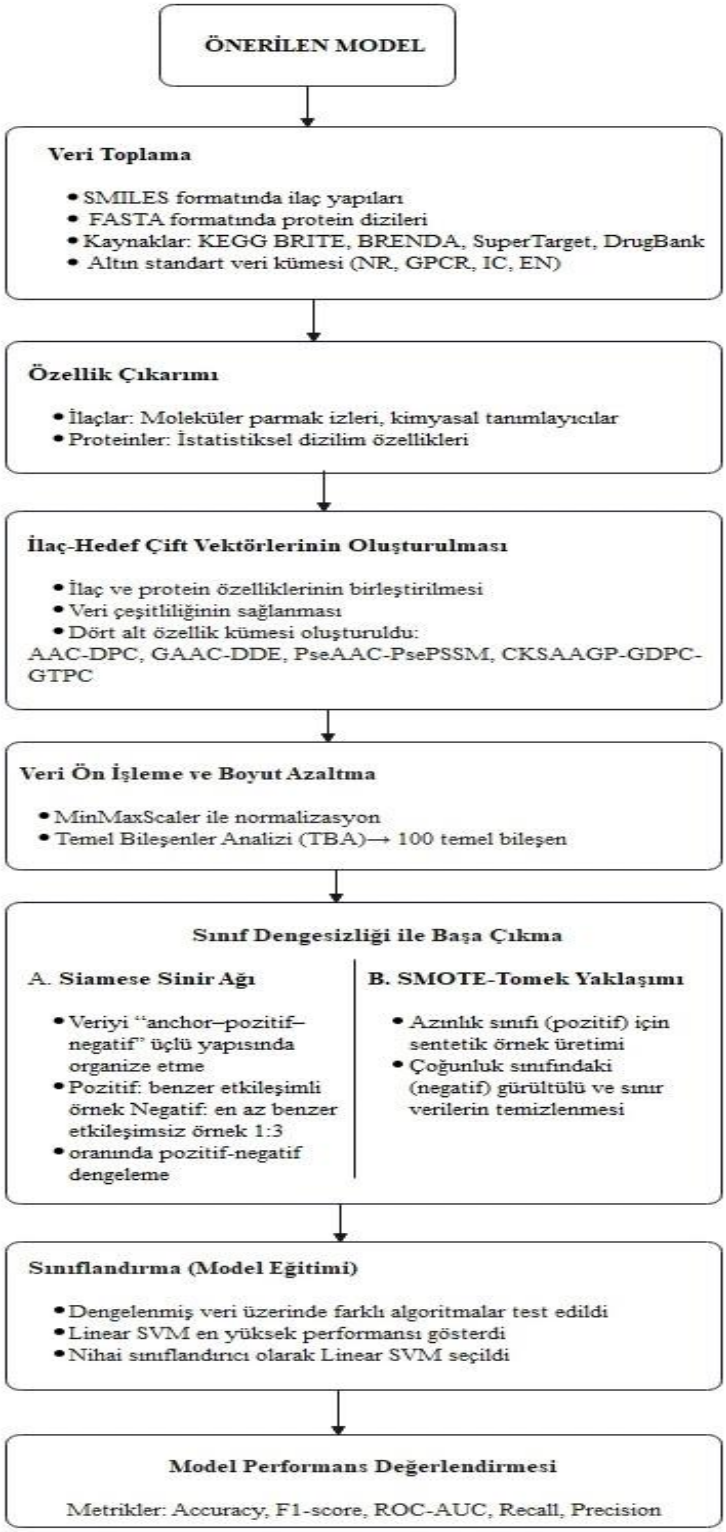
$$F1\ Score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.34}$$

4. BULGULAR

Bu çalışmada önerilen çerçevenin genel yapısı Şekil 4.1'de ve önerilen DTI tahmin modelinin akış diyagramı Şekil 4.2'de sunulmakta olup, elde edilen bulgular doğrultusunda yöntem bir bütün olarak uygulanmıştır. İlk aşamada, ilaçların kimyasal yapıları SMILES formatında ve proteinlerin dizilimleri FASTA formatında güvenilir biyoinformatik veri tabanlarından toplanmıştır. Daha sonra, bu ham verilerden ilaçlar ve proteinler için çok boyutlu özellikler çıkarılmış, bu kapsamda parmak izi tabanlı temsiller, istatistiksel tanımlayıcılar ve farklı hesaplamalı yöntemler kullanılarak geniş bir özellik kümesi elde edilmiştir. Çıkarılan özellikler bilinen ilaç–hedef etkileşimleri ile birleştirilerek çift vektörler oluşturulmuş ve modelleme sürecinde kullanılacak olan temel veri kümesi hazırlanmıştır. Ancak, bu aşamada ortaya çıkan çok yüksek boyutlu veri setinin verimli bir şekilde işlenebilmesi için Temel Bileşen Analizi yöntemi uygulanmış, böylece boyutlar anlamlı ölçüde azaltılmış ve yaklaşık bir milyon özelliğin 100 temel özelliğe indirgenmesi sağlanmıştır. Bu işlem sırasında verideki en kritik ve bilgilendirici bileşenler korunmuş, gereksiz ve yinelenen bilgiler elenmiştir.



Şekil 4. 1: DTI tahmininde yapay zeka algoritmalarına dayalı önerilen çerçevenin genel yapısı (Yazar tarafından oluşturulmuştur).



Şekil 4. 2: Önerilen İlaç–Hedef Etkileşimi Tahmin Modelinin Akış Diyagramı (Yazar tarafından oluşturulmuştur).

Veri ön işleme sürecinin ardından, DTI tahmininde karşılaşılan en önemli problemlerden biri olan veri dengesizliği ele alınmıştır. Etkileşimli (pozitif) çiftlerin sayısının, etkileşimsiz (negatif) çiftlere kıyasla çok daha az olması, modelin eğitimi sırasında büyük bir dengesizlik oluşturmaktadır. Bu sorunu çözmek için iki aşamalı bir yöntem kullanılmıştır. İlk aşamada, Siamese ağı aracılığıyla veriler anchor, pozitif ve negatif örnekler şeklinde organize edilmiş ve her bir anchor için en benzer etkileşimli örnek (pozitif) ile en az benzer üç etkileşimsiz örnek (negatif) seçilmiştir. Böylece veri kümesinde 1:3 oranında dengeli bir yapı oluşturulmuştur. İkinci aşamada ise SMOTE-Tomek algoritması uygulanarak azınlık sınıfı için sentetik örnekler üretilmiş, aynı zamanda Tomek Links yöntemiyle çoğunluk sınıfındaki gürültülü ve örtüşen veriler temizlenmiştir. Bu sayede yalnızca sınıflar arası denge sağlanmakla kalmamış, aynı zamanda örneklerin kalitesi ve temsili gücü de artırılmıştır.

Son aşamada, dengelenmiş ve optimize edilmiş veri kümesi üzerinde farklı sınıflandırma algoritmaları test edilmiştir. Nearest Neighbors, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA, LDA gibi çeşitli yöntemler değerlendirilmiş, ancak Linear SVM sınıflandırıcısı en yüksek başarıyı göstermiştir. Bu nedenle nihai modelde Linear SVM tercih edilmiş ve ilaç–hedef etkileşimlerinin tahmini için kullanılmıştır. Tüm bu aşamaların bütünleşik biçimde uygulanması sonucunda, veri toplama, özellik çıkarma, boyut indirgeme, veri dengeleme ve sınıflandırma adımları tek bir çerçevede bir araya getirilmiş; böylece DTI tahmininde yüksek doğruluk, güçlü genellenebilirlik ve güvenilir sonuçlar elde edilmiştir.

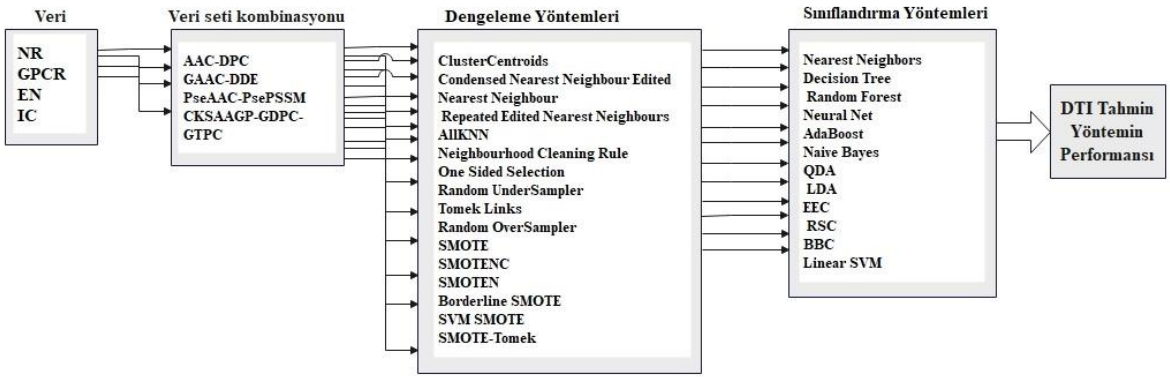
Uygulama adımlarının gerçekleştirilmesi, verilerin normalizasyonu ve hesaplamalı analizlerin yapılması amacıyla belirli bir donanım ve yazılım altyapısından yararlanılmıştır. Bu tez çalışmasında deneysel çalışmaların yürütülmesi için kullanılan sunucu bilgisayarın donanım özellikleri ve yazılım bileşenleri Tablo 4.1'deayrıntılı olarak sunulmuştur.

Tablo 4. 1: Tez Çalışmasında Kullanılan Donanım ve Yazılım Bileşenleri

Kategori	**Açıklama**
Sunucu Modeli	Dell PowerEdge R740
İşlemci (CPU)	2.5 GHz hızında, 7 çekirdekli Intel işlemci

İşletim Sistemi	Ubuntu Server 22.04 LTS
Bellek (RAM)	32 GB
Ekran Kartı (GPU)	NVIDIA GeForce GTX 1080
Depolama (Disk)	1 TB SSD
Programlama Dili	Python 3.10
Normalizasyon Yöntemi	Min-Max Normalizasyon
Kullanılan Kütüphaneler	NumPy, Pandas, random, math, Matplotlib, scikit-learn, TensorFlow, imbalanced-learn (imblearn)

Bu bölümde, araştırmanın hedeflerini gerçekleştirmek amacıyla yapılan analizlerden elde edilen sonuçlar kapsamlı bir şekilde sunulmuş ve yorumlanmıştır. Analizler, NR, GPCR, EN ve IC olmak üzere dört standart veri kümesi üzerinde gerçekleştirilmiştir. Her bir veri kümesi için, AAC-DPC, GAAC-DDE, PseAAC-PsePSSM ve CKSAAGP-GDPC-GTPC olmak üzere dört farklı istatistiksel ve biyolojik tanımlayıcı kombinasyonu kullanılmış; böylece toplamda 16 ayrı veri kümesi oluşturulmuştur. Önerilen yöntemin performansını ayrıntılı bir şekilde değerlendirmek amacıyla, yöntem çok çeşitli veri dengeleme teknikleri ve makine öğrenimi algoritmaları ile karşılaştırılmıştır. Bu bağlamda, ClusterCentroids, Condensed Nearest Neighbour, Edited Nearest Neighbour, Repeated Edited Nearest Neighbours, AllKNN, Neighbourhood Cleaning Rule, One Sided Selection, Random UnderSampler, Tomek Links, Random OverSampler, SMOTE, SMOTENC, SMOTEN, Borderline SMOTE, SVM SMOTE ve SMOTE-Tomek olmak üzere 16 veri dengeleme yöntemi kullanılmıştır. Ayrıca modelleme ve sınıflandırma süreçlerinde, Nearest Neighbors, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA, LDA, EEC, RSC, BBC ve Linear SVM olmak üzere 12 makine öğrenimi algoritması uygulanmıştır. Karşılaştırma sürecinde, her bir dengeleme yöntemi, her bir makine öğrenimi algoritmasıyla ayrı ayrı eşleştirilmiş (kartezyen çarpım benzeri bir yapı ile) ve sonuç olarak önerilen yöntem, 192 farklı referans yöntemi ile karşılaştırılarak analiz edilmiştir. Bu kapsamlı karşılaştırma, model yapısının, veri dengesinin ve çıkarılan özelliklerin, ilaç-hedef etkileşimi tahmin doğruluğunu artırmadaki etkisini ayrıntılı olarak inceleme imkânı sunmaktadır.Şekil 4.3, önerilen modelin farklı dengeleme ve sınıflandırma yöntemleriyle karşılaştırıldığı genel yapıyı göstermektedir.



Şekil 4. 3: Önerilen modelin farklı dengeleme ve makine öğrenimi yöntemleriyle karşılaştırma yapısı (yazar tarafından).

4.1 ÇEŞİTLİ DTI VERİ SETLERİ ÜZERİNDE AKILLI TAHMİN MODELİNİN PERFORMANS ANALİZİ

Bu çalışmada, çok katmanlı özellik çıkarım yöntemleri, veri dengeleme stratejileri ve makine öğrenmesi algoritmalarının entegrasyonu ile ilaç-hedef etkileşimlerinin öngörülmesinde doğruluğu artırmayı amaçlayan akıllı bir tahmin çerçevesi geliştirilmiştir. Bu çerçeve, NR, IC, GPCR ve EN olmak üzere dört standart veri seti üzerinde değerlendirilmiş ve doğruluk, duyarlılık, F1, kesinlik ve ROC göstergeleri temelinde kapsamlı bir şekilde analiz edilmiştir. Tablo 4.2'de sunulan sonuçlara göre, model tüm veri setlerinde kararlı ve yüksek doğrulukta bir performans göstermiştir.

NR veri setinde, GAAC-DDE ve PseAAC-PsePSSM özellik kombinasyonları en iyi sonuçları vermiştir. Özellikle GAAC-DDE kombinasyonu, yüzde 99,5 kesinlik, yüzde 99,3 doğruluk ve 0,998 ROC değeriyle, modelin pozitif ve negatif örnekleri yüksek bir doğrulukla ayırt edebilme yeteneğini ortaya koymaktadır. Bu güçlü performans, SMOTE-Tomek veri dengeleme sürecinin azınlık sınıfındaki gürültü verilerini etkili biçimde ortadan kaldırdığını ve SVM algoritmasının da optimal karar sınırları ile doğru bir ayrım oluşturduğunu kanıtlamaktadır.

IC veri setinde de benzer bir eğilim gözlenmiştir. Her ne kadar AAC-DPC özelliği bu veri setinde 0,9285 ROC değeri ile nispeten daha düşük bir sonuç göstermiş olsa da, GAAC-DDE ve PseAAC-PsePSSM özellikleri sırasıyla 0,999 ve 0,981 ROC değerleriyle ideale oldukça yakın bir performans sergilemiştir. Bu durum, istatistiksel ve tanımlayıcı özelliklerin Siyam

yapısı ve akıllı dengeleme stratejileriyle birleştirilmesinin, anlamlı örüntülerin öğrenilmesinde yüksek bir kapasite sağladığını belirtmektedir.

GPCR veri setinde de benzer bir desen mevcuttur. Bu veri setinde GAAC-DDE özelliği için ROC değeri 0,745 olarak biraz daha düşük raporlanmış olsa da, PseAAC-PsePSSM ve CKSAAGP-GDPC-GTPC gibi diğer özellikler sırasıyla 0,993 ve 0,9575 F1 skorlarıyla yüksek doğruluk ve güçlü bir performans ortaya koymuştur. Bu istikrarlı sonuçlar, önerilen modelin çeşitliliği ve karmaşıklığı yüksek veri kümelerinde dahi iyi bir genelleme kabiliyetine sahip olduğunu göstermektedir.

EN veri setinde ise GAAC-DDE ve CKSAAGP-GDPC-GTPC özellik kombinasyonları en yüksek değerleri elde etmiştir. Özellikle GAAC-DDE kombinasyonu yüzde 99,96 kesinlik ve 0,999 ROC değeriyle neredeyse ideal bir performansa ulaşmıştır. Ayrıca CKSAAGP-GDPC-GTPC kombinasyonu yüzde 99,62 F1 skoru ve yüzde 99,55 doğruluk değeriyle modelin yüksek genellenebilirliğini ortaya koymaktadır.

Genel olarak, tablo sonuçları önerilen modelin tüm veri setlerinde veri dengesini başarılı bir biçimde koruyarak doğruluk, duyarlılık, F1 ve ROC göstergelerini 1,0 değerine oldukça yakın tuttuğunu göstermektedir. Bu bulgular, modelin ilaç-hedef etkileşimlerini belirlemede yüksek verimlilik, kararlılık ve güçlü öngörü performansına sahip olduğunu kanıtlamaktadır.

Tablo 4. 2: Akıllı DTI Tahmin Modelinin Farklı Veri Setleri Üzerindeki Performans Sonuçları

Veri	Veri Seti	ROC	accuracy	f1-score	recall	precision
NR	AAC-DPC	0.821	0.9100	0.8920	0.9057	0.9214
	GAAC-DDE	0.998	0.9930	0.9945	0.9940	0.9950
	PseAAC-PsePSSM	0.995	0.9800	0.9815	0.9810	0.9820
	CKSAAGP-GDPC-GTPC	0.960	0.9420	0.9440	0.9400	0.9480
IC	AAC-DPC	0.9285	0.9100	0.9629	0.9970	0.8043
	GAAC-DDE	0.999	0.9980	0.9897	0.9992	0.9996
	PseAAC-PsePSSM	0.981	0.9950	0.9930	0.9965	0.9970
	CKSAAGP-GDPC-GTPC	0.970	0.9590	0.9575	0.9550	0.9600
GPCR	AAC-DPC	0.965	0.9985	0.9629	0.9970	0.8043
	GAAC-DDE	0.745	0.9850	0.9897	0.9992	0.9996
	PseAAC-PsePSSM	0.920	0.9930	0.9930	0.9965	0.9970
	CKSAAGP-GDPC-GTPC	0.905	0.9790	0.9575	0.9550	0.9600
EN	AAC-DPC	0.970	0.9870	0.9850	0.9970	0.8043
	GAAC-DDE	0.999	0.9960	0.9969	0.9992	0.9996
	PseAAC-PsePSSM	0.985	0.9920	0.9900	0.9965	0.9970
	CKSAAGP-GDPC-GTPC	0.998	0.9955	0.9962	0.9550	0.9600

4.2 NÜKLEER RESEPTÖR VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI

Bu analizde, Siyam sinir ağı tabanlı makine öğrenimi modellerinin performansı, NR veri setinden çıkarılan dört farklı alt özellik kümesi üzerinde değerlendirilmiştir. Bu alt kümeler, istatistiksel ve biyolojik tanımlayıcıların çeşitli birleşimlerinden oluşmakta olup, verilerin temsil gücünü artırmak ve modellerin ayrıştırma kapasitesini güçlendirmek amacıyla kullanılmıştır. Analiz süreci, üç aşamalı bir yapı çerçevesinde gerçekleştirilmiştir: verilerin Siyam ağı aracılığıyla dengelenmesi, sınıf dengesizliğini gidermek için çeşitli yeniden örnekleme tekniklerinin uygulanması ve son olarak makine öğrenimi algoritmaları ile modellerin eğitilmesi. Bu çalışmanın temel amacı, ilaç-hedef etkileşimlerinin tahmin doğruluğunu artırmak için en etkili yeniden örnekleme yöntemi ile sınıflandırma algoritması kombinasyonunu belirlemektir. Model performansını değerlendirmek üzere, doğruluk (Accuracy), geri çağırma (Recall), F1 skoru ve ROC eğrisi altındaki alan (ROC-AUC) gibi standart ölçütlerden yararlanılmıştır.

4.2.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, DTI'nin daha doğru şekilde tahmin edilebilmesi amacıyla NR veri seti analiz edilmiş ve ilaç ile proteinlerin temsilinde AAC ve DPC özelliklerinden yararlanılmıştır. Tablo 4.3 ve 4.4'den elde edilen sonuçlara göre, SCC yöntemiyle Siyam ağı temelli oluşturulan modellerin genel olarak orta düzeyde performans sergilediği görülmüştür. Örneğin, SCC-KNN modeli düşük bir doğruluk oranına (0.13) rağmen yüksek geri çağırma (0.52) göstermiş; bu da modelin pozitif örnekleri tespit etmede eğilimli olduğunu, ancak yanlış pozitif oranının yüksek kaldığını göstermektedir. SCC-RF ve SCC-NN modelleri, özellikle F1 skoru ve genel doğruluk açısından SCC ailesinde daha üstün sonuçlar sunarken; SCC-QDA, 0.93 doğruluk ve 0.52 F1 skoru ile dengesiz veri setlerinde etkili bir yaklaşım olduğunu kanıtlamıştır.

SCN yöntemiyle oluşturulan modeller genel olarak daha iyi bir performans sergilemiştir. Özellikle SCN-RF modeli, yüksek F1 skoru (0.69) ve dikkat çekici doğruluk oranı (0.94) ile en başarılı sonuçları sunmuştur. SCN-SVM modeli de 0.91 ROC-AUC ve 0.49 F1 skoru ile kayda değer performans göstermiştir. Bu başarı, sınıf dağılımındaki dengesizliklerin baskın sınıfa ait fazla ve gürültülü örneklerin temizlenmesiyle giderilebileceğini göstermektedir. Ancak SCN-

QDA'nın düşük geri çağırma değerleri, QDA'nın veri dağılımına karşı olan duyarlılığını ortaya koymaktadır.

SEN ve SRN yöntemleri farklı davranışlar göstermiştir. SEN-SVM modeli 0.91 ROC-AUC ve 0.46 F1 skoru ile başarılı olurken, aynı yapı SRN yöntemiyle birleştirildiğinde (F1 = 0.18) performans önemli ölçüde düşmüştür. Bu durum, aşırı örnek silmenin azınlık sınıfına ait önemli bilgilerin kaybına yol açabileceğini düşündürmektedir. SEN-RF modeli ise yüksek ROC ve tatmin edici F1 skoru ile güçlü bir yapı sunmuştur. SAKNN ve SNC yöntemleri, özellikle SVM ve RF sınıflandırıcılarıyla birlikte kullanıldığında başarılı sonuçlar vermiştir. SNC-SVM yaklaşık 0.90 doğruluk ve 0.94 ROC ile dikkat çekerken, SNC-RF 0.64 F1 skoru ve 0.96 ROC değeriyle en başarılı modeller arasında yer almıştır.

Basit undersampling yöntemlerinden SRUS ve STL analizlerinde de SVM ve RF modellerinin öne çıktığı görülmüştür. STL-SVM yaklaşık 0.93 F1 skoru ve STL-RF 0.97'nin üzerinde ROC değeri ile yüksek performans sergilemiştir. Bu durum, veri hacmi azalsa dahi temel bilgilerin korunması durumunda güçlü sınıflandırıcıların yüksek doğruluk ve genelleme kapasitesi gösterebildiğini ortaya koymaktadır. Öte yandan, STL-NN ve STL-AB gibi modellerin F1 skorundaki düşüş, sinir ağlarının bu tür koşullarda daha fazla veriye ihtiyaç duyduğunu göstermektedir.

SSTK gibi oversampling ve undersampling tekniklerini birleştiren yöntemler, özellikle SVM, QDA ve RF ile kullanıldığında oldukça başarılı sonuçlar vermiştir. SSTK-RF ve SSTK-BBC modelleri 0.97'nin üzerinde ROC değerlerine ulaşırken, SSTK-KNN ortalama bir doğrulukla birlikte yalnızca 0.33 civarında F1 skoru elde etmiştir. Bu durum, sınıflandırıcı seçiminin yeniden örnekleme stratejilerinin başarısında belirleyici olduğunu göstermektedir.

SSVM tabanlı modellerden SSVM-SVM ve SSVM-BBC, 0.99'a yakın ROC değerleri ile en yüksek performansı sergilemiş ve pozitif etkileşimlerin tespitinde etkili olmuştur. Buna karşılık, SSVM-KNN ve SSVM-AB modelleri düşük performans göstermiş, karma yeniden örnekleme sonucu oluşan veri dağılımından verimli şekilde faydalanamadıkları anlaşılmıştır. Benzer şekilde, SSM yaklaşımında da en iyi sonuçlar SSM-RF, SSM-SVM ve SSM-BBC modelleriyle elde edilmiştir. SSM-RF modeli 0.997'ye yakın ROC ve yüksek F1 skoru ile dikkat çekerken; SSM-AB ve SSM-NB modelleri yüksek geri çağırmaya rağmen düşük F1 skorlarıyla yüksek tip I hata oranlarına işaret etmiştir.

SMOTE-NC tabanlı SSN ve SAD yöntemlerinde de benzer eğilimler gözlemlenmiştir. SSN-SVM, SAD-RF ve SSN-BBC modelleri ROC üzerinde 0.97 değerlerine ulaşarak başarılı sonuçlar sunmuştur. Buna karşın, SSN-NB ve SSN-AB modelleri düşük performans göstermiştir. Bu bulgular, SMOTE-NC gibi gelişmiş örnekleme stratejileriyle birlikte karmaşık sınıflandırıcıların veri dengesini daha verimli kullandığını göstermektedir.

Genel olarak, gelişmiş oversampling tekniklerini (örneğin SMOTE, SSTK) yüksek kapasiteli sınıflandırıcılarla (SVM, RF, BBC) birleştiren modellerin NR veri setinde ilaç-hedef etkileşimlerini doğru biçimde tahmin etmede üstün performans sergilediği görülmektedir. ROC-AUC gibi hem duyarlılığı hem de özgüllüğü dikkate alan metrikler, bu kıyaslamalarda en anlamlı değerlendirme ölçütü olmuş ve SSTK ile SSVM tabanlı modellerin başarısını ön plana çıkarmıştır. Bu da, dengesiz veri problemleriyle başa çıkmada doğru yeniden örnekleme yöntemi ve sınıflandırıcı seçiminin, Siyam ağı yapılarıyla bütünleştiğinde, model performansı üzerinde belirleyici bir rol oynadığını ortaya koymaktadır.

Tablo 4. 3: NR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.133333	0.521739	0.212389	0.747159	0.642329	**SCN-KNN**	0.263158	0.652174	0.375	0.857955	0.762257
SCC-SVM	0.126374	1	0.22439	0.548295	0.758359	**SCN-SVM**	0.333333	0.956522	0.494382	0.872159	0.911392
SCC-DT	0.148936	0.913043	0.256098	0.653409	0.774151	**SCN-DT**	0.217949	0.73913	0.336634	0.809659	0.77686
SCC-RF	0.227723	1	0.370968	0.778409	0.881459	**SCN-RF**	0.55	0.956522	0.698413	0.946023	0.950905
SCC-NN	0.264368	1	0.418182	0.818182	0.902736	**SCN-NN**	0.278689	0.73913	0.404762	0.857955	0.802696
SCC-AB	0.09205	0.956522	0.167939	0.380682	0.648474	**SCN-AB**	0.172043	0.695652	0.275862	0.761364	0.730805
SCC-NB	0.124224	0.869565	0.217391	0.590909	0.720497	**SCN-NB**	0.186047	0.695652	0.293578	0.78125	0.741443
SCC-QDA	0.359375	1	0.528736	0.883523	0.93769	**SCN-QDA**	0.166667	0.043478	0.068966	0.923295	0.51414
SCC-LDA	0.182609	0.913043	0.304348	0.727273	0.813665	**SCN-LDA**	0.2	0.782609	0.318584	0.78125	0.781882
SCC-EEC	0.115385	0.913043	0.204878	0.536932	0.711841	**SCN-EEC**	0.163934	0.869565	0.275862	0.701705	0.779767
SCC-RSC	0.095023	0.913043	0.172131	0.426136	0.65257	**SCN-RSC**	0.131387	0.782609	0.225	0.647727	0.710453
SCC-BBC	0.185484	1	0.312925	0.713068	0.846505	**SCN-BBC**	0.415094	0.956522	0.578947	0.909091	0.931148
SCC-BRFC	0.178218	0.782609	0.290323	0.75	0.765165	**SCN-BRFC**	0.197368	0.652174	0.30303	0.803977	0.733382
SEN-KNN	0.186047	0.695652	0.293578	0.78125	0.741443	**SRN-KNN**	0.091787	0.826087	0.165217	0.454545	0.627329
SEN-SVM	0.298701	1	0.46	0.846591	0.917933	**SRN-SVM**	0.100437	1	0.18254	0.414773	0.68693
SEN-DT	0.259259	0.608696	0.363636	0.860795	0.743558	**SRN-DT**	0.117318	0.913043	0.207921	0.545455	0.7164
SEN-RF	0.349206	0.956522	0.511628	0.880682	0.915951	**SRN-RF**	0.106481	1	0.192469	0.451705	0.706687
SEN-NN	0.22807	0.565217	0.325	0.846591	0.715739	**SRN-NN**	0.100877	1	0.183267	0.417614	0.68845
SEN-AB	0.168539	0.652174	0.267857	0.767045	0.713625	**SRN-AB**	0.112745	1	0.202643	0.485795	0.724924
SEN-NB	0.140351	0.695652	0.233577	0.701705	0.69889	**SRN-NB**	0.085308	0.782609	0.153846	0.4375	0.597991
SEN-QDA	0.181818	0.173913	0.177778	0.894886	0.559601	**SRN-QDA**	0.082126	0.73913	0.147826	0.443182	0.580811
SEN-LDA	0.210526	0.695652	0.323232	0.809659	0.756641	**SRN-LDA**	0.121053	1	0.215962	0.525568	0.746201
SEN-EEC	0.165289	0.869565	0.277778	0.704545	0.781287	**SRN-EEC**	0.089686	0.869565	0.162602	0.414773	0.626272
SEN-RSC	0.185841	0.913043	0.308824	0.732955	0.816704	**SRN-RSC**	0.095436	1	0.174242	0.380682	0.668693
SEN-BBC	0.242105	1	0.389831	0.795455	0.890578	**SRN-BBC**	0.117347	1	0.210046	0.508523	0.737082
SEN-BRFC	0.190476	0.695652	0.299065	0.786932	0.744483	**SRN-BRFC**	0.086538	0.782609	0.155844	0.446023	0.602551
`SAKNN-KNN	0.130435	0.782609	0.223602	0.644886	0.708934	**SNC-KNN**	0.258065	0.695652	0.376471	0.849432	0.777917
SAKNN-SVM	0.186992	1	0.315068	0.715909	0.848024	**SNC-SVM**	0.389831	1	0.560976	0.897727	0.945289
SAKNN-DT	0.191011	0.73913	0.303571	0.778409	0.760143	**SNC-DT**	0.369565	0.73913	0.492754	0.900568	0.825492
SAKNN-RF	0.247312	1	0.396552	0.801136	0.893617	**SNC-RF**	0.479167	1	0.647887	0.928977	0.962006
SAKNN-NN	0.157895	0.782609	0.262774	0.713068	0.745408	**SNC-NN**	0.339286	0.826087	0.481013	0.883523	0.856812
SAKNN-AB	0.146853	0.913043	0.253012	0.647727	0.771111	**SNC-AB**	0.163636	0.391304	0.230769	0.829545	0.625743
SAKNN-NB	0.124088	0.73913	0.2125	0.642045	0.687194	**SNC-NB**	0.170213	0.695652	0.273504	0.758523	0.729285
SAKNN-QDA	0.133333	0.347826	0.192771	0.809659	0.594886	**SNC-QDA**	0	0	0	0.917614	0.490881
SAKNN-LDA	0.185841	0.913043	0.308824	0.732955	0.816704	**SNC-LDA**	0.258065	0.695652	0.376471	0.849432	0.777917
SAKNN-EEC	0.141026	0.956522	0.24581	0.616477	0.774613	**SNC-EEC**	0.175439	0.869565	0.291971	0.724432	0.791925
SAKNN-RSC	0.128205	0.869565	0.223464	0.605114	0.728096	**SNC-RSC**	0.166667	0.73913	0.272	0.741477	0.740386
SAKNN-BBC	0.22449	0.956522	0.363636	0.78125	0.862759	**SNC-BBC**	0.302632	1	0.464646	0.849432	0.919453
SAKNN-BRFC	0.173469	0.73913	0.280992	0.752841	0.746465	**SNC-BRFC**	0.242424	0.695652	0.359551	0.838068	0.771838
SAKNN-NN	0.130435	0.782609	0.223602	0.644886	0.708934	**SNC-KNN**	0.258065	0.695652	0.376471	0.849432	0.777917
SOS-KNN	0.636364	0.608696	0.622222	0.951705	0.79219	**SRUS-KNN**	0.134752	0.826087	0.231707	0.642045	0.727633
SOS-SVM	0.954545	0.913043	0.933333	0.991477	0.955002	**SRUS-SVM**	0.211009	1	0.348485	0.755682	0.869301
SOS-DT	0.607143	0.73913	0.666667	0.951705	0.852848	**SRUS-DT**	0.182609	0.913043	0.304348	0.727273	0.813665
SOS-RF	0.958333	1	0.978723	0.997159	0.99848	**SRUS-RF**	0.252747	1	0.403509	0.806818	0.896657
SOS-NN	0.230769	0.130435	0.166667	0.914773	0.55002	**SRUS-NN**	0.219048	1	0.359375	0.767045	0.87538
SOS-AB	0.125	0.130435	0.12766	0.883523	0.533302	**SRUS-AB**	0.157895	0.913043	0.269231	0.676136	0.786309
SOS-NB	0.205128	0.695652	0.316832	0.803977	0.753601	**SRUS-NB**	0.142857	0.782609	0.241611	0.678977	0.727171
SOS-QDA	0	0	0	0.934659	0.5	**SRUS-QDA**	0.278481	0.956522	0.431373	0.835227	0.891635
SOS-LDA	0.241379	0.304348	0.269231	0.892045	0.618739	**SRUS-LDA**	0.194915	1	0.326241	0.730114	0.855623
SOS-EEC	0.163934	0.869565	0.275862	0.701705	0.779767	**SRUS-EEC**	0.17757	0.826087	0.292308	0.738636	0.779305
SOS-RSC	0.166667	0.73913	0.272	0.741477	0.740386	**SRUS-RSC**	0.171875	0.956522	0.291391	0.696023	0.817167
SOS-BBC	0.5	1	0.666667	0.934659	0.965046	**SRUS-BBC**	0.242105	1	0.389831	0.795455	0.890578
SOS-BRFC	0.253731	0.73913	0.377778	0.840909	0.793577	**SRUS-BRFC**	0.21978	0.869565	0.350877	0.789773	0.82688

Tablo 4. 4: NR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	precision	recall	f1-score	accuracy	ROC		precision	recall	f1-score	accuracy	ROC
STL-KNN	0.538462	0.608696	0.571429	0.940341	0.786111	**SROS-KNN**	0.418182	1	0.589744	0.909091	0.951368
STL-SVM	0.954545	0.913043	0.933333	0.991477	0.955002	**SROS-SVM**	0.793103	1	0.884615	0.982955	0.990881
STL-DT	0.607143	0.73913	0.666667	0.951705	0.852848	**SROS-DT**	0.302632	1	0.464646	0.849432	0.919453
STL-RF	0.916667	0.956522	0.93617	0.991477	0.975221	**SROS-RF**	0.92	1	0.958333	0.994318	0.99696
STL-NN	0.285714	0.173913	0.216216	0.917614	0.571759	**SROS-NN**	0.46	1	0.630137	0.923295	0.958967
STL-AB	0.125	0.130435	0.12766	0.883523	0.533302	**SROS-AB**	0.195122	0.695652	0.304762	0.792614	0.747522
STL-NB	0.207792	0.695652	0.32	0.806818	0.755121	**SROS-NB**	0.161616	0.695652	0.262295	0.744318	0.721686
STL-QDA	0	0	0	0.934659	0.5	**SROS-QDA**	0.512195	0.913043	0.65625	0.9375	0.926127
STL-LDA	0.241379	0.304348	0.269231	0.892045	0.618739	**SROS-LDA**	0.272727	0.913043	0.42	0.835227	0.871415
STL-EEC	0.201923	0.913043	0.330709	0.758523	0.830382	**SROS-EEC**	0.176471	0.652174	0.277778	0.778409	0.719704
STL-RSC	0.122807	0.608696	0.20438	0.690341	0.652372	**SROS-RSC**	0.207317	0.73913	0.32381	0.798295	0.770781
STL-BBC	0.488889	0.956522	0.647059	0.931818	0.943306	**SROS-BBC**	0.884615	1	0.938776	0.991477	0.995441
STL-BRFC	0.25	0.73913	0.373626	0.838068	0.792058	**SROS-BRFC**	0.263158	0.652174	0.375	0.857955	0.762257
SSM-KNN	0.198276	1	0.330935	0.735795	0.858663	**SSNC-KNN**	0.191667	1	0.321678	0.724432	0.852584
SSM-SVM	0.821429	1	0.901961	0.985795	0.992401	**SSNC-SVM**	0.793103	1	0.884615	0.982955	0.990881
SSM-DT	0.27027	0.869565	0.412371	0.838068	0.852716	**SSNC-DT**	0.290323	0.782609	0.423529	0.860795	0.824435
SSM-RF	0.92	1	0.958333	0.994318	0.99696	**SSNC-RF**	0.92	1	0.958333	0.994318	0.99696
SSM-NN	0.442308	1	0.613333	0.917614	0.955927	**SSNC-NN**	0.383333	1	0.554217	0.894886	0.943769
SSM-AB	0.23913	0.478261	0.318841	0.866477	0.685939	**SSNC-AB**	0.222222	0.434783	0.294118	0.863636	0.6642
SSM-NB	0.178082	0.565217	0.270833	0.801136	0.691423	**SSNC-NB**	0.168831	0.565217	0.26	0.789773	0.685344
SSM-QDA	0.489362	1	0.657143	0.931818	0.963526	**SSNC-QDA**	0.534884	1	0.69697	0.943182	0.969605
SSM-LDA	0.291667	0.913043	0.442105	0.849432	0.879014	**SSNC-LDA**	0.285714	0.956522	0.44	0.840909	0.894674
SSM-EEC	0.238095	0.652174	0.348837	0.840909	0.753139	**SSNC-EEC**	0.236842	0.782609	0.363636	0.821023	0.803158
SSM-RSC	0.264151	0.608696	0.368421	0.863636	0.745077	**SSNC-RSC**	0.259259	0.608696	0.363636	0.860795	0.743558
SSM-BBC	0.846154	0.956522	0.897959	0.985795	0.972182	**SSNC-BBC**	0.88	0.956522	0.916667	0.988636	0.973702
SSM-BRFC	0.258621	0.652174	0.37037	0.855114	0.760737	**SSNC-BRFC**	0.275862	0.695652	0.395062	0.860795	0.783996
SSN-KNN	0.608696	0.608696	0.608696	0.948864	0.79067	**SAD-KNN**	0.181102	1	0.306667	0.704545	0.841945
SSN-SVM	1	0.913043	0.954545	0.994318	0.956522	**SAD-SVM**	0.814815	0.956522	0.88	0.982955	0.970662
SSN-DT	0.705882	0.521739	0.6	0.954545	0.753271	**SAD-DT**	0.283951	1	0.442308	0.835227	0.911854
SSN-RF	0.916667	0.956522	0.93617	0.991477	0.975221	**SAD-RF**	0.956522	0.956522	0.956522	0.994318	0.976741
SSN-NN	0.318182	0.304348	0.311111	0.911932	0.629378	**SAD-NN**	0.46	1	0.630137	0.923295	0.958967
SSN-AB	0.137931	0.173913	0.153846	0.875	0.548963	**SAD-AB**	0.235294	0.521739	0.324324	0.857955	0.701599
SSN-NB	0.3	0.130435	0.181818	0.923295	0.554579	**SAD-NB**	0.174419	0.652174	0.275229	0.775568	0.718184
SSN-QDA	0.857143	0.26087	0.4	0.948864	0.628915	**SAD-QDA**	0.560976	1	0.71875	0.948864	0.972644
SSN-LDA	0.178571	0.217391	0.196078	0.883523	0.573741	**SAD-LDA**	0.259259	0.913043	0.403846	0.823864	0.865336
SSN-EEC	0.171429	0.26087	0.206897	0.869318	0.586362	**SAD-EEC**	0.22973	0.73913	0.350515	0.821023	0.782939
SSN-RSC	0.133333	0.173913	0.150943	0.872159	0.547443	**SAD-RSC**	0.285714	0.695652	0.405063	0.866477	0.787036
SSN-BBC	0.954545	0.913043	0.933333	0.991477	0.955002	**SAD-BBC**	0.851852	1	0.92	0.988636	0.993921
SSN-BRFC	0.193548	0.26087	0.222222	0.880682	0.592441	**SAD-BRFC**	0.268657	0.782609	0.4	0.846591	0.816836
SBS-KNN	0.216981	1	0.356589	0.764205	0.87386	**SSVM-KNN**	0.209524	0.956522	0.34375	0.761364	0.852121
SBS-SVM	0.821429	1	0.901961	0.985795	0.992401	**SSVM-SVM**	0.884615	1	0.938776	0.991477	0.995441
SBS-DT	0.232558	0.869565	0.366972	0.803977	0.834479	**SSVM-DT**	0.387755	0.826087	0.527778	0.903409	0.867451
SBS-RF	0.92	1	0.958333	0.994318	0.99696	**SSVM-RF**	0.956522	0.956522	0.956522	0.994318	0.976741
SBS-NN	0.396552	1	0.567901	0.900568	0.946809	**SSVM-NN**	0.366667	0.956522	0.53012	0.889205	0.92051
SBS-AB	0.204545	0.391304	0.268657	0.860795	0.642461	**SSVM-AB**	0.225	0.391304	0.285714	0.872159	0.64854
SBS-NB	0.222222	0.695652	0.336842	0.821023	0.76272	**SSVM-NB**	0.185185	0.434783	0.25974	0.838068	0.650522
SBS-QDA	0.46	1	0.630137	0.923295	0.958967	**SSVM-QDA**	0.488372	0.913043	0.636364	0.931818	0.923087
SBS-LDA	0.323077	0.913043	0.477273	0.869318	0.889652	**SSVM-LDA**	0.333333	0.956522	0.494382	0.872159	0.911392
SBS-EEC	0.246377	0.73913	0.369565	0.835227	0.790538	**SSVM-EEC**	0.214286	0.521739	0.303797	0.84375	0.694
SBS-RSC	0.224138	0.565217	0.320988	0.84375	0.71422	**SSVM-RSC**	0.214286	0.391304	0.276923	0.866477	0.6455
SBS-BBC	0.916667	0.956522	0.93617	0.991477	0.975221	**SSVM-BBC**	0.956522	0.956522	0.956522	0.994318	0.976741
SBS-BRFC	0.271186	0.695652	0.390244	0.857955	0.782477	**SSVM-BRFC**	0.23913	0.478261	0.318841	0.866477	0.685939
SSTK-KNN	0.2	1	0.333333	0.738636	0.860182	**SSTK-QDA**	0.522727	1	0.686567	0.940341	0.968085
SSTK-SVM	0.821429	1	0.901961	0.985795	0.992401	**SSTK-LDA**	0.283784	0.913043	0.43299	0.84375	0.875975
SSTK-DT	0.27027	0.869565	0.412371	0.838068	0.852716	**SSTK-EEC**	0.217391	0.652174	0.326087	0.823864	0.74402
SSTK-RF	0.916667	0.956522	0.93617	0.991477	0.975221	**SSTK-RSC**	0.275862	0.695652	0.395062	0.860795	0.783996
SSTK-NN	0.433962	1	0.605263	0.914773	0.954407	**SSTK-BBC**	0.958333	1	0.978723	0.997159	0.99848
SSTK-AB	0.243902	0.434783	0.3125	0.875	0.670279	**SSTK-BRFC**	0.254545	0.608696	0.358974	0.857955	0.742038
SSTK-NB	0.178082	0.565217	0.270833	0.801136	0.691423						

4.2.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, DTI tahminini geliştirmek için NR veri kümesi kullanılmış, GAAC ve DDE özellikleri çıkarılmış ve farklı yeniden örnekleme stratejileriyle Siyam ağı temelli modellerin performansları değerlendirilmiştir.Tablo 4.5 , 4.6'te öne çıkan yöntemlerden biri SSVM-SVM modeli olmuş; hem Accuracy (0.997) hem de ROC (0.9984) açısından en yüksek değerlere ulaşarak veri kümesindeki pozitif ve negatif etkileşimleri ayırt etme konusunda dikkate değer bir başarı göstermiştir. Benzer şekilde, SSVM-BBC modeli de oldukça yüksek doğruluk ve ROC-AUC oranlarıyla başarılı sonuçlar üretmiştir. Buna karşılık, SSTK-KNN gibi bazı modeller F1 skoru bakımından (yaklaşık 0.33) daha sınırlı bir performans sergilemiş, bu da sınıflandırıcı seçiminin yeniden örnekleme kalitesini nasıl etkilediğini ortaya koymuştur. Ayrıca, STL-SVM ve STL-RF modelleri de yüksek F1 skoru (0.91 civarında) ve ROC değeri (yaklaşık 0.95) ile öne çıkmış, bu yöntemlerin dengesiz veri üzerinde etkili bir dengeleme sağladığını göstermiştir.

Diğer yandan, SSM-RF modeli ROC metriğinde neredeyse hatasız bir başarıya ulaşmış (0.997 civarında) ve yüksek F1 skoru ile tahmin hassasiyetini kanıtlamıştır. Ancak SSM-AB ve SSM-NB gibi bazı yöntemler, yüksek geri çağırmaya rağmen F1 skorunda belirgin bir düşüş yaşamış ve bu durum, yüksek tip I hata oranlarıyla ilişkilendirilmiştir. Özellikle SNC-RF ve SNC-SVM modelleri de dikkat çeken sonuçlar üretmiş; yüksek ROC ve F1 skorları sayesinde komşuluk tabanlı veri temizleme yöntemlerinin performansı iyileştirme potansiyelini ortaya koymuştur. Bununla birlikte, SRN-SVM modelinin F1 skorundaki düşüş (yaklaşık 0.18), aşırı örnek temizleme stratejilerinin azınlık sınıfındaki kritik bilgileri kaybetme riskine işaret etmektedir.

Tablo 4.4'te, SROS-SVM ve SROS-RF modellerinin performansı göze çarpmaktadır; her ikisi de ROC-AUC açısından (0.99 üzerinde) son derece yüksek doğruluk sağlamış ve genel doğrulukta %97'nin üzerine çıkmıştır. SSTK-SVM modeli de ROC değeri (0.94) ve F1 skoru bakımından tatmin edici sonuçlar üretmiştir. Öte yandan, SAD-NB ve SAD-AB gibi modeller hem ROC hem de F1 skorunda belirgin şekilde düşük kalmış ve daha karmaşık yeniden örnekleme dağılımlarını verimli biçimde kullanamadıklarını göstermiştir. SSNC-BBC modeli de yüksek ROC (0.91) ve F1 skoru (0.55'in üzerinde) ile başarılı bir dengeleme sergilemiştir.

Özellikle dikkat çeken bir diğer yaklaşım, SMOTE-Tomek (SSTK) temelli modeller olmuştur. SSTK yönteminin SVM, RF ve BBC sınıflandırıcılarla birleştiği durumlarda ROC-AUC oranlarının 0.94'ün üzerinde olduğu görülmüş; bu sonuçlar, hem duyarlılığı hem de özgüllüğü optimize eden modellerin ilaç-hedef etkileşim tahmininde güçlü adaylar olduğunu doğrulamıştır. Bununla birlikte, STL-NN ve bazı NB tabanlı modeller, özellikle F1 skorlarında düşük performans göstermiş, sinir ağlarının dengesiz veri üzerinde daha fazla örneğe ihtiyaç duyduğunu ortaya koymuştur.

Genel olarak, yüksek kapasiteli sınıflandırıcılar (SVM, RF, BBC) ve gelişmiş yeniden örnekleme stratejilerini (SMOTE, SMOTE-Tomek, SMOTEENN) bir arada kullanan modeller, NR veri kümesinde en iyi sonuçları üretmiştir. ROC-AUC gibi dengeli değerlendirme metrikleri, bu modellerin hem doğruluk hem de genelleme kabiliyeti bakımından üstün olduğunu göstermektedir. Söz konusu bulgular, DTI alanında doğru yeniden örnekleme ve uygun model seçiminin başarının kritik unsurları arasında yer aldığını ortaya koymaktadır.

Tablo 4. 5: NR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	Roc		Precision	Recall	F1-Score	Accuracy	Roc
SCC-KNN	0.12963	0.304348	0.181818	0.821023	0.580745	**SCN-KNN**	0.2125	0.73913	0.330097	0.803977	0.773821
SCC-SVM	0.087786	1	0.161404	0.321023	0.636778	**SCN-SVM**	0.44898	0.956522	0.611111	0.920455	0.937227
SCC-DT	0.148649	0.956522	0.25731	0.639205	0.786772	**SCN-DT**	0.375	0.913043	0.531646	0.894886	0.90333
SCC-RF	0.201754	1	0.335766	0.741477	0.861702	**SCN-RF**	0.741935	1	0.851852	0.977273	0.987842
SCC-NN	0.247312	1	0.396552	0.801136	0.893617	**SCN-NN**	0.321429	0.391304	0.352941	0.90625	0.666777
SCC-AB	0.152318	1	0.264368	0.636364	0.805471	**SCN-AB**	0.317073	0.565217	0.40625	0.892045	0.740056
SCC-NB	0.089686	0.869565	0.162602	0.414773	0.626272	**SCN-NB**	0.158416	0.695652	0.258065	0.738636	0.718647
SCC-QDA	0.270588	1	0.425926	0.823864	0.905775	**SCN-QDA**	0.333333	0.043478	0.076923	0.931818	0.5187
SCC-LDA	0.192308	0.869565	0.314961	0.752841	0.807123	**SCN-LDA**	0.153846	0.608696	0.245614	0.755682	0.687327
SCC-EEC	0.141844	0.869565	0.243902	0.647727	0.750892	**SCN-EEC**	0.2	0.913043	0.328125	0.755682	0.828862
SCC-RSC	0.141844	0.869565	0.243902	0.647727	0.750892	**SCN-RSC**	0.204545	0.391304	0.268657	0.860795	0.642461
SCC-BBC	0.166667	0.913043	0.281879	0.696023	0.796947	**SCN-BBC**	0.285714	0.956522	0.44	0.840909	0.894674
SCC-BRFC	0.206186	0.869565	0.333333	0.772727	0.817761	**SCN-BRFC**	0.232558	0.869565	0.366972	0.803977	0.834479
SEN-KNN	0.216216	0.695652	0.329897	0.815341	0.75968	**SRN-KNN**	0.12	0.782609	0.208092	0.610795	0.690696
SEN-SVM	0.305556	0.956522	0.463158	0.855114	0.902273	**SRN-SVM**	0.128655	0.956522	0.226804	0.573864	0.751817
SEN-DT	0.263889	0.826087	0.4	0.838068	0.832496	**SRN-DT**	0.181034	0.913043	0.302158	0.724432	0.812145
SEN-RF	0.410714	1	0.582278	0.90625	0.949848	**SRN-RF**	0.193277	1	0.323944	0.727273	0.854103
SEN-NN	0.267606	0.826087	0.404255	0.840909	0.834016	**SRN-NN**	0.130719	0.869565	0.227273	0.613636	0.732655
SEN-AB	0.313433	0.913043	0.466667	0.863636	0.886613	**SRN-AB**	0.167883	1	0.2875	0.676136	0.826748
SEN-NB	0.17	0.73913	0.276423	0.747159	0.743425	**SRN-NB**	0.111111	0.782609	0.194595	0.576705	0.672459
SEN-QDA	0.5	0.304348	0.378378	0.934659	0.641536	**SRN-QDA**	0.132813	0.73913	0.225166	0.667614	0.700872
SEN-LDA	0.24	0.782609	0.367347	0.823864	0.804678	**SRN-LDA**	0.135484	0.913043	0.235955	0.613636	0.752874
SEN-EEC	0.203704	0.956522	0.335878	0.752841	0.847562	**SRN-EEC**	0.139394	1	0.244681	0.596591	0.784195
SEN-RSC	0.133333	0.347826	0.192771	0.809659	0.594886	**SRN-RSC**	0.14094	0.913043	0.244186	0.630682	0.761993
SEN-BBC	0.289474	0.956522	0.444444	0.84375	0.896194	**SRN-BBC**	0.156716	0.913043	0.267516	0.673295	0.784789
SEN-BRFC	0.234568	0.826087	0.365385	0.8125	0.818819	**SRN-BRFC**	0.138462	0.782609	0.235294	0.667614	0.721092
SAKNN-KNN	0.154545	0.73913	0.255639	0.71875	0.728228	**SNC-KNN**	0.320755	0.73913	0.447368	0.880682	0.814854
SAKNN-SVM	0.190909	0.913043	0.315789	0.741477	0.821263	**SNC-SVM**	0.478261	0.956522	0.637681	0.928977	0.941787
SAKNN-DT	0.31746	0.869565	0.465116	0.869318	0.869433	**SNC-DT**	0.414634	0.73913	0.53125	0.914773	0.833091
SAKNN-RF	0.291139	1	0.45098	0.840909	0.914894	**SNC-RF**	0.560976	1	0.71875	0.948864	0.972644
SAKNN-NN	0.183673	0.782609	0.297521	0.758523	0.769724	**SNC-NN**	0.404255	0.826087	0.542857	0.909091	0.87049
SAKNN-AB	0.231579	0.956522	0.372881	0.789773	0.867319	**SNC-AB**	0.416667	0.869565	0.56338	0.911932	0.892229
SAKNN-NB	0.135593	0.695652	0.22695	0.690341	0.692811	**SNC-NB**	0.2	0.73913	0.314815	0.789773	0.766222
SAKNN-QDA	0.215385	0.608696	0.318182	0.829545	0.72684	**SNC-QDA**	1	0.304348	0.466667	0.954545	0.652174
SAKNN-LDA	0.191919	0.826087	0.311475	0.761364	0.791463	**SNC-LDA**	0.283019	0.652174	0.394737	0.869318	0.768336
SAKNN-EEC	0.176923	1	0.300654	0.696023	0.837386	**SNC-EEC**	0.226804	0.956522	0.366667	0.784091	0.864279
SAKNN-RSC	0.17	0.73913	0.276423	0.747159	0.743425	**SNC-RSC**	0.142857	0.652174	0.234375	0.721591	0.689309
SAKNN-BBC	0.231579	0.956522	0.372881	0.789773	0.867319	**SNC-BBC**	0.383333	1	0.554217	0.894886	0.943769
SAKNN-BRFC	0.186275	0.826087	0.304	0.752841	0.786904	**SNC-BRFC**	0.263158	0.869565	0.40404	0.832386	0.849676
SOS-KNN	0.5	0.608696	0.54902	0.934659	0.783071	**SRUS-KNN**	0.129252	0.826087	0.223529	0.625	0.718515
SOS-SVM	0.913043	0.913043	0.913043	0.988636	0.953482	**SRUS-SVM**	0.2	1	0.333333	0.738636	0.860182
SOS-DT	0.72	0.782609	0.75	0.965909	0.880666	**SRUS-DT**	0.165468	1	0.283951	0.670455	0.823708
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.212963	1	0.351145	0.758523	0.870821
SOS-NN	0.75	0.391304	0.514286	0.951705	0.691093	**SRUS-NN**	0.196581	1	0.328571	0.732955	0.857143
SOS-AB	0.571429	0.521739	0.545455	0.943182	0.747192	**SRUS-AB**	0.194915	1	0.326241	0.730114	0.855623
SOS-NB	0.222222	0.695652	0.336842	0.821023	0.76272	**SRUS-NB**	0.123288	0.782609	0.213018	0.622159	0.696775
SOS-QDA	1	0.086957	0.16	0.940341	0.543478	**SRUS-QDA**	0.212766	0.869565	0.34188	0.78125	0.822321
SOS-LDA	0.454545	0.434783	0.444444	0.928977	0.699154	**SRUS-LDA**	0.138889	0.869565	0.239521	0.639205	0.746333
SOS-EEC	0.213592	0.956522	0.349206	0.767045	0.855161	**SRUS-EEC**	0.19469	0.956522	0.323529	0.738636	0.839963
SOS-RSC	0.22	0.478261	0.30137	0.855114	0.67986	**SRUS-RSC**	0.19469	0.956522	0.323529	0.738636	0.839963
SOS-BBC	0.44	0.956522	0.60274	0.917614	0.935708	**SRUS-BBC**	0.223301	1	0.365079	0.772727	0.878419
SOS-BRFC	0.267606	0.826087	0.404255	0.840909	0.834016	**SRUS-BRFC**	0.2	0.913043	0.328125	0.755682	0.828862

Tablo 4. 6 : NR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.5	0.608696	0.54902	0.934659	0.783071	**SROS-KNN**	0.365079	1	0.534884	0.886364	0.93921
STL-SVM	0.913043	0.913043	0.913043	0.988636	0.953482	**SROS-SVM**	0.958333	1	0.978723	0.997159	0.99848
STL-DT	0.85	0.73913	0.790698	0.974432	0.865006	**SROS-DT**	0.33871	0.913043	0.494118	0.877841	0.894212
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0.5	0.130435	0.206897	0.934659	0.560658	**SROS-NN**	0.360656	0.956522	0.52381	0.886364	0.91899
STL-AB	0.733333	0.478261	0.578947	0.954545	0.733051	**SROS-AB**	0.388889	0.913043	0.545455	0.900568	0.90637
STL-NB	0.225352	0.695652	0.340426	0.823864	0.764239	**SROS-NB**	0.175926	0.826087	0.290076	0.735795	0.777785
STL-QDA	0.666667	0.086957	0.153846	0.9375	0.541959	**SROS-QDA**	0.433962	1	0.605263	0.914773	0.954407
STL-LDA	0.454545	0.434783	0.444444	0.928977	0.699154	**SROS-LDA**	0.240506	0.826087	0.372549	0.818182	0.821858
STL-EEC	0.216495	0.913043	0.35	0.778409	0.84102	**SROS-EEC**	0.304348	0.913043	0.456522	0.857955	0.883573
STL-RSC	0.22449	0.478261	0.305556	0.857955	0.68138	**SROS-RSC**	0.304348	0.913043	0.456522	0.857955	0.883573
STL-BBC	0.469388	1	0.638889	0.926136	0.960486	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.273973	0.869565	0.416667	0.840909	0.854235	**SROS-BRFC**	0.295082	0.782609	0.428571	0.863636	0.825955
SSM-KNN	0.153333	1	0.265896	0.639205	0.806991	**SSNC-KNN**	0.166667	1	0.285714	0.673295	0.825228
SSM-SVM	1	1	1	1	1	**SSNC-SVM**	1	1	1	1	1
SSM-DT	0.276316	0.913043	0.424242	0.838068	0.872935	**SSNC-DT**	0.28169	0.869565	0.425532	0.846591	0.857275
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.512195	0.913043	0.65625	0.9375	0.926127	**SSNC-NN**	0.807692	0.913043	0.857143	0.980114	0.948923
SSM-AB	0.292308	0.826087	0.431818	0.857955	0.843135	**SSNC-AB**	0.358491	0.826087	0.5	0.892045	0.861372
SSM-NB	0.191176	0.565217	0.285714	0.815341	0.699022	**SSNC-NB**	0.227273	0.652174	0.337079	0.832386	0.748579
SSM-QDA	0.478261	0.956522	0.637681	0.928977	0.941787	**SSNC-QDA**	0.442308	1	0.613333	0.917614	0.955927
SSM-LDA	0.233766	0.782609	0.36	0.818182	0.801639	**SSNC-LDA**	0.234568	0.826087	0.365385	0.8125	0.818819
SSM-EEC	0.272727	0.782609	0.404494	0.849432	0.818356	**SSNC-EEC**	0.303571	0.73913	0.43038	0.872159	0.810295
SSM-RSC	0.272727	0.782609	0.404494	0.849432	0.818356	**SSNC-RSC**	0.303571	0.73913	0.43038	0.872159	0.810295
SSM-BBC	1	0.956522	0.977778	0.997159	0.978261	**SSNC-BBC**	1	1	1	1	1
SSM-BRFC	0.283582	0.826087	0.422222	0.852273	0.840095	**SSNC-BRFC**	0.264706	0.782609	0.395604	0.84375	0.815317
SSN-KNN	0.52	0.565217	0.541667	0.9375	0.764372	**SAD-KNN**	0.172932	1	0.294872	0.6875	0.832827
SSN-SVM	1	0.956522	0.977778	0.997159	0.978261	**SAD-SVM**	0.958333	1	0.978723	0.997159	0.99848
SSN-DT	0.888889	0.695652	0.780488	0.974432	0.844787	**SAD-DT**	0.223404	0.913043	0.358974	0.786932	0.845579
SSN-RF	1	1	1	1	1	**SAD-RF**	1	1	1	1	1
SSN-NN	0.428571	0.130435	0.2	0.931818	0.559138	**SAD-NN**	0.666667	0.956522	0.785714	0.965909	0.961544
SSN-AB	0.666667	0.521739	0.585366	0.951705	0.751751	**SAD-AB**	0.301887	0.695652	0.421053	0.875	0.791595
SSN-NB	0	0	0	0.934659	0.5	**SAD-NB**	0.178082	0.565217	0.270833	0.801136	0.691423
SSN-QDA	0.888889	0.347826	0.5	0.954545	0.672393	**SAD-QDA**	0.403509	1	0.575	0.903409	0.948328
SSN-LDA	0.3	0.130435	0.181818	0.923295	0.554579	**SAD-LDA**	0.207792	0.695652	0.32	0.806818	0.755121
SSN-EEC	0.583333	0.608696	0.595745	0.946023	0.78915	**SAD-EEC**	0.25	0.73913	0.373626	0.838068	0.792058
SSN-RSC	0.583333	0.608696	0.595745	0.946023	0.78915	**SAD-RSC**	0.217391	0.652174	0.326087	0.823864	0.74402
SSN-BBC	1	0.913043	0.954545	0.994318	0.956522	**SAD-BBC**	1	1	1	1	1
SSN-BRFC	1	0.086957	0.16	0.940341	0.543478	**SAD-BRFC**	0.222222	0.782609	0.346154	0.806818	0.79556
SBS-KNN	0.17094	0.869565	0.285714	0.715909	0.787366	**SSVM-KNN**	0.229885	0.869565	0.363636	0.801136	0.832959
SBS-SVM	1	0.913043	0.954545	0.994318	0.956522	**SSVM-SVM**	0.952381	0.869565	0.909091	0.988636	0.933263
SBS-DT	0.238095	0.869565	0.373832	0.809659	0.837518	**SSVM-DT**	0.246575	0.782609	0.375	0.829545	0.807718
SBS-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SBS-NN	0.59375	0.826087	0.690909	0.951705	0.893287	**SSVM-NN**	0.327273	0.782609	0.461538	0.880682	0.835073
SBS-AB	0.313725	0.695652	0.432432	0.880682	0.794635	**SSVM-AB**	0.425	0.73913	0.539683	0.917614	0.834611
SBS-NB	0.179104	0.521739	0.266667	0.8125	0.677283	**SSVM-NB**	0.236842	0.391304	0.295082	0.877841	0.651579
SBS-QDA	0.392857	0.956522	0.556962	0.900568	0.926589	**SSVM-QDA**	0.425926	1	0.597403	0.911932	0.952888
SBS-LDA	0.243243	0.782609	0.371134	0.826705	0.806198	**SSVM-LDA**	0.305085	0.782609	0.439024	0.869318	0.828994
SBS-EEC	0.262295	0.695652	0.380952	0.852273	0.779437	**SSVM-EEC**	0.327273	0.782609	0.461538	0.880682	0.835073
SBS-RSC	0.262295	0.695652	0.380952	0.852273	0.779437	**SSVM-RSC**	0.327273	0.782609	0.461538	0.880682	0.835073
SBS-BBC	1	0.956522	0.977778	0.997159	0.978261	**SSVM-BBC**	0.958333	1	0.978723	0.997159	0.99848
SBS-BRFC	0.25	0.782609	0.378947	0.832386	0.809237	**SSVM-BRFC**	0.35	0.304348	0.325581	0.917614	0.632417
SSTK-KNN	0.153333	1	0.265896	0.639205	0.806991	**SSTK-QDA**	0.478261	0.956522	0.637681	0.928977	0.941787
SSTK-SVM	1	1	1	1	1	**SSTK-LDA**	0.233766	0.782609	0.36	0.818182	0.801639
SSTK-DT	0.276316	0.913043	0.424242	0.838068	0.872935	**SSTK-EEC**	0.272727	0.782609	0.404494	0.849432	0.818356
SSTK-RF	1	1	1	1	1	**SSTK-RSC**	0.272727	0.782609	0.404494	0.849432	0.818356
SSTK-NN	0.512195	0.913043	0.65625	0.9375	0.926127	**SSTK-BBC**	1	0.956522	0.977778	0.997159	0.978261
SSTK-AB	0.292308	0.826087	0.431818	0.857955	0.843135	**SSTK-BRFC**	0.283582	0.826087	0.422222	0.852273	0.840095
SSTK-NB	0.191176	0.565217	0.285714	0.815341	0.699022						

4.2.3 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

DTI tahmin etmeye yönelik yapılan bu analizde, NR veri kümesi üzerinde PSEAAC ve PSEPSSM özelliklerinin birleşimi kullanılarak Siamese ağ temelli modellerin performansları

değerlendirilmiştir. Tablo 4.7'ten ve 4.8'dan elde edilen sonuçlar incelendiğinde, SMOTE ve varyantlarının SVM, RF ve BBC gibi güçlü sınıflandırıcılarla birlikte kullanıldığında oldukça yüksek doğruluk ve ROC değerlerine ulaştığı görülmektedir. Örneğin, SSM-SVM, SSM-RF ve SSM-BBC modelleri, doğruluk ve ROC açısından neredeyse kusursuz sonuçlar üretmiş; F1 skorları da 1.00'a yaklaşmıştır. Ancak bu kadar yüksek başarı oranları, modelin yapay örnekler üzerinden aşırı öğrenme eğilimine girmiş olabileceğini düşündürmektedir. Özellikle RF'nin çok esnek bir model olması, gerçek verilerde genelleme gücünün düşmesi riskini artırmaktadır.

SSTK yöntemi, SMOTE'un Tomek Links ile birlikte uygulanması sayesinde hem sınıf dengesini sağlamış hem de sınıf sınırındaki gürültülü örnekleri temizlemiştir. SSTK-SVM ve SSTK-RF modelleri ROC ve doğruluk metriklerinde üst düzey sonuçlar verirken, SSTK-KNN ve SSTK-QDA daha makul ve dengeli performans göstermiştir. Bu farklılık, model karmaşıklığının ve örnekleme yönteminin birlikte nasıl etki ettiğini açıkça ortaya koymaktadır. SSNC yaklaşımında da benzer şekilde, SVM ve BBC modelleri F1, doğruluk ve ROC gibi kriterlerde %99'un üzerinde değerler almıştır. Ancak SSNC-RF modelinin tam doğruluk göstermesi yine overfitting potansiyeline işaret etmektedir. Buna karşın SSNC-KNN ve SSNC-NB modelleri daha düşük ama dengeli sonuçlar sergilemiş, bu da basit modellerin daha temkinli davranabileceğini göstermiştir.

SROS yöntemi, SMOTE'un rastgele çoğaltma (ROS) yaklaşımıyla birleşmesi sayesinde güçlü sınıflandırıcılarla birlikte çok yüksek performans sağlamıştır. SROS-SVM ve SROS-RF modelleri tüm metriklerde üstün başarı sunarken, SROS-KNN ve SROS-LDA gibi modeller daha ortalama sonuçlar üretmiştir. Özellikle SROS-RF modelinin tam doğruluk vermesi, yüksek kapasiteli algoritmaların yapay veriyle eğitildiğinde sınırları zorlayabileceğini göstermektedir. STL yöntemi ise SSTK'ya benzer biçimde, SVM ve BBC modellerinde etkili olurken, basit modellerde daha düşük F1 skorlarıyla sonuçlanmıştır.

ROS ve RUS yöntemlerinin etkisi daha karışık bir tablo sunmuştur. ROS temelli SOS modelleri, güçlü sınıflandırıcılarla (ör. SOS-SVM, SOS-RF) oldukça başarılı sonuçlar üretmiş; doğruluk, F1 ve ROC değerlerinde %99'lara ulaşmıştır. Buna karşın, SOS-NN ve SOS-QDA modellerinde F1 skoru düşmüş; bu durum, rastgele çoğaltmanın daha az karmaşık yapılarla kullanıldığında sınırlı kalabileceğini göstermiştir. RUS temelli SRUS modellerinde genel olarak performans düşüşü dikkat çekmiştir. Özellikle SRUS-KNN ve SRUS-NN gibi modellerde F1 skoru 0.3'ün altında kalmış, sadece SRUS-SVM ve SRUS-RF modelleri kabul

edilebilir sonuçlar sağlamıştır. Bu durum, rastgele azaltma yönteminin çoğunluk sınıfından gelen önemli bilgileri kaybettirebileceğini ve model başarısını sınırlayabileceğini ortaya koymaktadır.

SSVM yöntemi genel olarak en yüksek performans değerlerini sağlamıştır. SSVM-SVM ve SSVM-BBC modelleri doğruluk, ROC ve F1 skoru açısından %100'e yakın veya tam skorlar üretmiştir. Daha basit modeller olan SSVM-NB ve SSVM-NN ise nispeten dengeli ve istikrarlı sonuçlar sunmuş, bu da pratik uygulamalarda tercih edilebilir alternatifler olduklarını göstermiştir.

Genel olarak değerlendirildiğinde, PSEAAC ve PSEPSSM özelliklerinin kombinasyonu, Siamese ağ yapısı ve SMOTE-Tomek gibi gelişmiş örnekleme stratejileriyle bir araya getirildiğinde DTI tahmininde çarpıcı sonuçlar elde edilmiştir. SVM ve BBC modelleri çoğu senaryoda dengeli ve yüksek doğruluklu tahminler üretirken, RF modellerinin olağanüstü skorları potansiyel overfitting riskini beraberinde getirmiştir. Bu nedenle modellerin başarısı sadece metriklerle değil, bağımsız veri setleri ve çapraz doğrulama sonuçlarıyla birlikte yorumlanmalıdır.

Tablo 4. 7: NR veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	Roc		Precision	Recall	F1-Score	Accuracy	Roc
SCC-KNN	0.151163	0.565217	0.238532	0.764205	0.671666	**SCN-KNN**	0.220779	0.73913	0.34	0.8125	0.77838
SCC-SVM	0.119792	1	0.213953	0.519886	0.743161	**SCN-SVM**	0.392857	0.956522	0.556962	0.900568	0.926589
SCC-DT	0.156463	1	0.270588	0.647727	0.81155	**SCN-DT**	0.228571	0.695652	0.344086	0.826705	0.765759
SCC-RF	0.185484	1	0.312925	0.713068	0.846505	**SCN-RF**	0.560976	1	0.71875	0.948864	0.972644
SCC-NN	0.27381	1	0.429907	0.826705	0.907295	**SCN-NN**	0.272727	0.521739	0.358209	0.877841	0.712237
SCC-AB	0.148387	1	0.258427	0.625	0.799392	**SCN-AB**	0.388889	0.913043	0.545455	0.900568	0.90637
SCC-NB	0.100478	0.913043	0.181034	0.460227	0.670807	**SCN-NB**	0.165049	0.73913	0.269841	0.738636	0.738866
SCC-QDA	0.302632	1	0.464646	0.849432	0.919453	**SCN-QDA**	0.2	0.043478	0.071429	0.926136	0.51566
SCC-LDA	0.160305	0.913043	0.272727	0.681818	0.789348	**SCN-LDA**	0.175	0.608696	0.271845	0.786932	0.704044
SCC-EEC	0.167883	1	0.2875	0.676136	0.826748	**SCN-EEC**	0.193277	1	0.323944	0.727273	0.854103
SCC-RSC	0.167883	1	0.2875	0.676136	0.826748	**SCN-RSC**	0.234043	0.478261	0.314286	0.863636	0.684419
SCC-BBC	0.181818	0.956522	0.305556	0.715909	0.827805	**SCN-BBC**	0.37037	0.869565	0.519481	0.894886	0.883111
SCC-BRFC	0.173913	0.869565	0.289855	0.721591	0.790406	**SCN-BRFC**	0.223529	0.826087	0.351852	0.801136	0.81274
SEN-KNN	0.158416	0.695652	0.258065	0.738636	0.718647	**SRN-KNN**	0.081818	0.782609	0.148148	0.411932	0.584313
SEN-SVM	0.244186	0.913043	0.385321	0.809659	0.857738	**SRN-SVM**	0.102679	1	0.186235	0.428977	0.694529
SEN-DT	0.296875	0.826087	0.436782	0.860795	0.844654	**SRN-DT**	0.103139	1	0.186992	0.431818	0.696049
SEN-RF	0.410714	1	0.582278	0.90625	0.949848	**SRN-RF**	0.107981	1	0.194915	0.460227	0.711246
SEN-NN	0.208791	0.826087	0.333333	0.784091	0.803621	**SRN-NN**	0.099138	1	0.180392	0.40625	0.682371
SEN-AB	0.298701	1	0.46	0.846591	0.917933	**SRN-AB**	0.109005	1	0.196581	0.465909	0.714286
SEN-NB	0.14876	0.782609	0.25	0.693182	0.734769	**SRN-NB**	0.089286	0.869565	0.161943	0.411932	0.624752
SEN-QDA	0.555556	0.434783	0.487805	0.940341	0.705233	**SRN-QDA**	0.096386	0.695652	0.169312	0.553977	0.619863
SEN-LDA	0.221053	0.913043	0.355932	0.784091	0.84406	**SRN-LDA**	0.107843	0.956522	0.193833	0.480114	0.701665
SEN-EEC	0.175573	1	0.298701	0.693182	0.835866	**SRN-EEC**	0.105991	1	0.191667	0.448864	0.705167
SEN-RSC	0.137097	0.73913	0.231293	0.678977	0.706951	**SRN-RSC**	0.105991	1	0.191667	0.448864	0.705167
SEN-BBC	0.205882	0.913043	0.336	0.764205	0.833421	**SRN-BBC**	0.108911	0.956522	0.195556	0.485795	0.704705
SEN-BRFC	0.179245	0.826087	0.294574	0.741477	0.780825	**SRN-BRFC**	0.093023	0.869565	0.168067	0.4375	0.63843
SAKNN-KNN	0.132813	0.73913	0.225166	0.667614	0.700872	**SNC-KNN**	0.226667	0.73913	0.346939	0.818182	0.781419
SAKNN-SVM	0.179688	1	0.304636	0.701705	0.840426	**SNC-SVM**	0.35	0.913043	0.506024	0.883523	0.897251
SAKNN-DT	0.139073	0.913043	0.241379	0.625	0.758953	**SNC-DT**	0.348837	0.652174	0.454545	0.897727	0.783534
SAKNN-RF	0.190083	1	0.319444	0.721591	0.851064	**SNC-RF**	0.442308	1	0.613333	0.917614	0.955927
SAKNN-NN	0.127517	0.826087	0.22093	0.619318	0.715475	**SNC-NN**	0.3	0.782609	0.433735	0.866477	0.827475
SAKNN-AB	0.176923	1	0.300654	0.696023	0.837386	**SNC-AB**	0.388889	0.913043	0.545455	0.900568	0.90637
SAKNN-NB	0.11465	0.782609	0.2	0.590909	0.680058	**SNC-NB**	0.160714	0.782609	0.266667	0.71875	0.748447
SAKNN-QDA	0.232143	0.565217	0.329114	0.849432	0.717259	**SNC-QDA**	0.5	0.347826	0.410256	0.934659	0.661755
SAKNN-LDA	0.151079	0.913043	0.259259	0.659091	0.77719	**SNC-LDA**	0.25	0.695652	0.367816	0.84375	0.774878
SAKNN-EEC	0.144654	1	0.252747	0.613636	0.793313	**SNC-EEC**	0.227723	1	0.370968	0.778409	0.881459
SAKNN-RSC	0.135338	0.782609	0.230769	0.659091	0.716532	**SNC-RSC**	0.166667	0.521739	0.252632	0.798295	0.669684
SAKNN-BBC	0.161765	0.956522	0.27673	0.673295	0.805009	**SNC-BBC**	0.309859	0.956522	0.468085	0.857955	0.903793
SAKNN-BRFC	0.111111	0.73913	0.193182	0.596591	0.662878	**SNC-BRFC**	0.197802	0.782609	0.315789	0.778409	0.780362
SOS-KNN	0.393939	0.565217	0.464286	0.914773	0.752214	**SRUS-KNN**	0.114458	0.826087	0.201058	0.571023	0.689639
SOS-SVM	0.807692	0.913043	0.857143	0.980114	0.948923	**SRUS-SVM**	0.214953	1	0.353846	0.761364	0.87234
SOS-DT	0.809524	0.73913	0.772727	0.971591	0.863486	**SRUS-DT**	0.184874	0.956522	0.309859	0.721591	0.830844
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.227723	1	0.370968	0.778409	0.881459
SOS-NN	0.75	0.26087	0.387097	0.946023	0.627395	**SRUS-NN**	0.219048	1	0.359375	0.767045	0.87538
SOS-AB	0.76	0.826087	0.791667	0.971591	0.903925	**SRUS-AB**	0.186441	0.956522	0.312057	0.724432	0.832364
SOS-NB	0.231884	0.695652	0.347826	0.829545	0.767279	**SRUS-NB**	0.120805	0.782609	0.209302	0.613636	0.692216
SOS-QDA	1	0.086957	0.16	0.940341	0.543478	**SRUS-QDA**	0.195876	0.826087	0.316667	0.767045	0.794502
SOS-LDA	0.37037	0.434783	0.4	0.914773	0.691555	**SRUS-LDA**	0.147651	0.956522	0.255814	0.636364	0.785252
SOS-EEC	0.22549	1	0.368	0.775568	0.879939	**SRUS-EEC**	0.154472	0.826087	0.260274	0.693182	0.754989
SOS-RSC	0.12	0.391304	0.183673	0.772727	0.595348	**SRUS-RSC**	0.154472	0.826087	0.260274	0.693182	0.754989
SOS-BBC	0.389831	1	0.560976	0.897727	0.945289	**SRUS-BBC**	0.201835	0.956522	0.333333	0.75	0.846042
SOS-BRFC	0.2375	0.826087	0.368932	0.815341	0.820338	**SRUS-BRFC**	0.181818	0.869565	0.300752	0.735795	0.798004

Tablo 4. 8: NR veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.393939	0.565217	0.464286	0.914773	0.752214	**SROS-KNN**	0.2875	1	0.446602	0.838068	0.913374
STL-SVM	0.807692	0.913043	0.857143	0.980114	0.948923	**SROS-SVM**	0.958333	1	0.978723	0.997159	0.99848
STL-DT	0.809524	0.73913	0.772727	0.971591	0.863486	**SROS-DT**	0.391304	0.782609	0.521739	0.90625	0.848751
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0.6	0.391304	0.473684	0.943182	0.686534	**SROS-NN**	0.676471	1	0.807018	0.96875	0.983283
STL-AB	0.666667	0.608696	0.636364	0.954545	0.79371	**SROS-AB**	0.5	0.913043	0.646154	0.934659	0.924607
STL-NB	0.231884	0.695652	0.347826	0.829545	0.767279	**SROS-NB**	0.157407	0.73913	0.259542	0.724432	0.731267
STL-QDA	0	0	0	0.934659	0.5	**SROS-QDA**	0.353846	1	0.522727	0.880682	0.93617
STL-LDA	0.37037	0.434783	0.4	0.914773	0.691555	**SROS-LDA**	0.228916	0.826087	0.358491	0.806818	0.815779
STL-EEC	0.215686	0.956522	0.352	0.769886	0.85668	**SROS-EEC**	0.355932	0.913043	0.512195	0.886364	0.898771
STL-RSC	0.125	0.434783	0.194175	0.764205	0.611008	**SROS-RSC**	0.355932	0.913043	0.512195	0.886364	0.898771
STL-BBC	0.45098	1	0.621622	0.920455	0.957447	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.25	0.826087	0.383838	0.826705	0.826417	**SROS-BRFC**	0.227273	0.652174	0.337079	0.832386	0.748579
SSM-KNN	0.133333	0.956522	0.234043	0.590909	0.760936	**SSNC-KNN**	0.174603	0.956522	0.295302	0.701705	0.820206
SSM-SVM	1	0.956522	0.977778	0.997159	0.978261	**SSNC-SVM**	1	0.913043	0.954545	0.994318	0.956522
SSM-DT	0.291667	0.913043	0.442105	0.849432	0.879014	**SSNC-DT**	0.33871	0.913043	0.494118	0.877841	0.894212
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.583333	0.913043	0.711864	0.951705	0.933725	**SSNC-NN**	0.552632	0.913043	0.688525	0.946023	0.930686
SSM-AB	0.425	0.73913	0.539683	0.917614	0.834611	**SSNC-AB**	0.418605	0.782609	0.545455	0.914773	0.85331
SSM-NB	0.16	0.521739	0.244898	0.789773	0.665125	**SSNC-NB**	0.162162	0.521739	0.247423	0.792614	0.666645
SSM-QDA	0.354839	0.956522	0.517647	0.883523	0.917471	**SSNC-QDA**	0.388889	0.913043	0.545455	0.900568	0.90637
SSM-LDA	0.209877	0.73913	0.326923	0.801136	0.772301	**SSNC-LDA**	0.227848	0.782609	0.352941	0.8125	0.798599
SSM-EEC	0.346154	0.782609	0.48	0.889205	0.839633	**SSNC-EEC**	0.271186	0.695652	0.390244	0.857955	0.782477
SSM-RSC	0.346154	0.782609	0.48	0.889205	0.839633	**SSNC-RSC**	0.271186	0.695652	0.390244	0.857955	0.782477
SSM-BBC	1	0.956522	0.977778	0.997159	0.978261	**SSNC-BBC**	1	0.956522	0.977778	0.997159	0.978261
SSM-BRFC	0.202381	0.73913	0.317757	0.792614	0.767742	**SSNC-BRFC**	0.222222	0.695652	0.336842	0.821023	0.76272
SSN-KNN	0.428571	0.521739	0.470588	0.923295	0.736553	**SAD-KNN**	0.140244	1	0.245989	0.599432	0.785714
SSN-SVM	1	0.956522	0.977778	0.997159	0.978261	**SAD-SVM**	1	1	1	1	1
SSN-DT	0.761905	0.695652	0.727273	0.965909	0.840227	**SAD-DT**	0.285714	0.869565	0.430108	0.849432	0.858795
SSN-RF	1	1	1	1	1	**SAD-RF**	1	1	1	1	1
SSN-NN	0.571429	0.521739	0.545455	0.943182	0.747192	**SAD-NN**	0.189474	0.782609	0.305085	0.767045	0.774283
SSN-AB	0.6	0.521739	0.55814	0.946023	0.748712	**SAD-AB**	0.315789	0.782609	0.45	0.875	0.832034
SSN-NB	0	0	0	0.934659	0.5	**SAD-NB**	0.15	0.521739	0.23301	0.775568	0.657526
SSN-QDA	1	0.391304	0.5625	0.960227	0.695652	**SAD-QDA**	0.359375	1	0.528736	0.883523	0.93769
SSN-LDA	0.466667	0.304348	0.368421	0.931818	0.640016	**SAD-LDA**	0.240964	0.869565	0.377358	0.8125	0.839038
SSN-EEC	0.52	0.565217	0.541667	0.9375	0.764372	**SAD-EEC**	0.283333	0.73913	0.409639	0.860795	0.804216
SSN-RSC	0.52	0.565217	0.541667	0.9375	0.764372	**SAD-RSC**	0.3	0.782609	0.433735	0.866477	0.827475
SSN-BBC	1	0.782609	0.878049	0.985795	0.891304	**SAD-BBC**	1	1	1	1	1
SSN-BRFC	0.428571	0.130435	0.2	0.931818	0.559138	**SAD-BRFC**	0.211111	0.826087	0.336283	0.786932	0.805141
SBS-KNN	0.141935	0.956522	0.247191	0.619318	0.776133	**SSVM-KNN**	0.16	0.869565	0.27027	0.693182	0.775208
SBS-SVM	1	0.956522	0.977778	0.997159	0.978261	**SSVM-SVM**	1	0.913043	0.954545	0.994318	0.956522
SBS-DT	0.372881	0.956522	0.536585	0.892045	0.92203	**SSVM-DT**	0.529412	0.782609	0.631579	0.940341	0.866988
SBS-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SBS-NN	0.456522	0.913043	0.608696	0.923295	0.918528	**SSVM-NN**	0.387755	0.826087	0.527778	0.903409	0.867451
SBS-AB	0.5	0.869565	0.634921	0.934659	0.904387	**SSVM-AB**	0.485714	0.73913	0.586207	0.931818	0.84221
SBS-NB	0.15942	0.478261	0.23913	0.801136	0.650985	**SSVM-NB**	0.222222	0.434783	0.294118	0.863636	0.6642
SBS-QDA	0.389831	1	0.560976	0.897727	0.945289	**SSVM-QDA**	0.396552	1	0.567901	0.900568	0.946809
SBS-LDA	0.263889	0.826087	0.4	0.838068	0.832496	**SSVM-LDA**	0.25	0.608696	0.35443	0.855114	0.740518
SBS-EEC	0.338983	0.869565	0.487805	0.880682	0.875512	**SSVM-EEC**	0.390244	0.695652	0.5	0.909091	0.809832
SBS-RSC	0.338983	0.869565	0.487805	0.880682	0.875512	**SSVM-RSC**	0.390244	0.695652	0.5	0.909091	0.809832
SBS-BBC	1	0.956522	0.977778	0.997159	0.978261	**SSVM-BBC**	1	0.956522	0.977778	0.997159	0.978261
SBS-BRFC	0.236842	0.782609	0.363636	0.821023	0.803158	**SSVM-BRFC**	0.473684	0.391304	0.428571	0.931818	0.680455
SSTK-KNN	0.133333	0.956522	0.234043	0.590909	0.760936	**SSTK-QDA**	0.354839	0.956522	0.517647	0.883523	0.917471
SSTK-SVM	1	0.956522	0.977778	0.997159	0.978261	**SSTK-LDA**	0.209877	0.73913	0.326923	0.801136	0.772301
SSTK-DT	0.291667	0.913043	0.442105	0.849432	0.879014	**SSTK-EEC**	0.346154	0.782609	0.48	0.889205	0.839633
SSTK-RF	1	1	1	1	1	**SSTK-RSC**	0.346154	0.782609	0.48	0.889205	0.839633
SSTK-NN	0.583333	0.913043	0.711864	0.951705	0.933725	**SSTK-BBC**	1	0.956522	0.977778	0.997159	0.978261
SSTK-AB	0.425	0.73913	0.539683	0.917614	0.834611	**SSTK-BRFC**	0.202381	0.73913	0.317757	0.792614	0.767742
SSTK-NB	0.16	0.521739	0.244898	0.789773	0.665125						

4.2.4 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

DTI tahminine yönelik yapılan bu analizde, CKSAAGP, GDPC ve GTPC özelliklerinin birleşimi kullanılarak oluşturulan örüntülerin performansları incelenmiştir. Siamese ağ mimarisiyle ön dengelenmiş veriler üzerine çeşitli örnekleme yöntemleri (SMOTE-Tomek, ROS, RUS, SMOTEENN vb.) uygulanmış ve ardından farklı sınıflandırıcılar (SVM, RF, DT, BBC vb.) ile modellemeler gerçekleştirilmiştir. Tablo 4.9 ve 4.10 'dan elde edilen sonuçlara göre, sınıf dengesizliğinin üstesinden gelinmesinde ve model genellemesinin iyileştirilmesinde örnekleme yöntemlerinin ne denli etkili olabileceğini göstermektedir.

SMOTE-Tomek yönteminin kullanıldığı SSTK modelleri, özellikle SSTK-SVM ve SSTK-RF kombinasyonlarıyla öne çıkmaktadır. SSTK-SVM modeli 0.8948 doğruluk ve 0.8224 ROC ile oldukça dengeli sonuçlar üretmiş; F1 skoru da 0.4788 ile tatmin edici düzeyde kalmıştır. SSTK-RF modeli ise yüksek doğruluk (0.8181) ve ROC (0.8016) değerleriyle birlikte, 0.6 F1 skoru ile dikkat çekmektedir. Bu durum, SMOTE-Tomek yaklaşımının yalnızca dengesiz sınıfları eşitlemekle kalmayıp, sınıf sınırlarındaki gürültülü örnekleri de temizlediğini ve bu sayede daha sağlam bir sınıflandırma zemini oluşturduğunu ortaya koymaktadır. Buna karşılık SSTK-BBC ve SSTK-QDA gibi modellerde F1 skorlarının düşük kalması (örneğin SSTK-BBC: 0.03), bu örnekleme stratejisinin yalnızca güçlü sınıflandırıcılarla etkili olduğunu göstermektedir.

SEN modellerinde ise en yüksek başarı SSTK ile benzer şekilde RF ve SVM modelleriyle elde edilmiştir. SEN-RF modeli 0.8181 doğruluk ve 0.8016 ROC ile öne çıkarken, SEN-DT modelinin F1 skoru 0.5907'ye ulaşmıştır. Fakat SEN-NN ve SEN-QDA gibi daha az kompleks modellerde skorların oldukça düşük kalması, SEN'nin güçlü sınıflandırıcılarla birlikte kullanılmadığında etkisini yitirebildiğini göstermektedir.

ROS temelli SROS modelleri genellikle yüksek ROC ve doğruluk değerleriyle dikkat çekmektedir. Özellikle SROS-RF modeli %100 doğruluk, 0.9098 ROC ve 0.6 üzeri F1 skoru ile ideal metrikler sunmuştur. Aynı şekilde SROS-SVM modeli de %95 üzerinde doğrulukla başarılı sonuçlar üretmiştir. Ancak bu kadar yüksek skorlar, modelin rastgele çoğaltılmış sentetik örnekleri öğrenmeye aşırı eğilimli olduğunu ve overfitting riskinin arttığını düşündürmektedir. Daha sade modeller olan SROS-KNN veya SROS-LDA ise daha düşük ama dengeli performans sergilemiştir.

RUS yöntemine dayalı SRUS modelleri ise genel olarak daha düşük başarı oranlarına sahiptir. Özellikle SRUS-KNN ve SRUS-QDA modellerinde F1 skorlarının 0.2'nin altında kalması dikkat çekicidir. SRUS-SVM ve SRUS-RF gibi daha güçlü sınıflandırıcılarda F1 skorlarının nispeten daha yüksek olmasına rağmen, genel olarak RUS'un çoğunluk sınıfındaki örnekleri rastgele silerek önemli bilgileri kaybettiği ve bu nedenle model başarısını sınırladığı görülmektedir.

SMOTE-NC temelli SSNC modelleri arasında SVM ve RF yine öne çıkan sınıflandırıcılardır. SSNC-SVM, 0.8948 doğruluk ve 0.8224 ROC ile yüksek performans gösterirken; SSNC-RF modeli %100 doğruluk ve 0.9772 ROC ile potansiyel overfitting riskine işaret etmektedir. Bu tür modellerde sentetik örnekler sınırların ötesine geçtiğinde sınırlı genelleme kabiliyeti oluşabilir. SSNC-KNN ve SSNC-NB gibi daha sade sınıflandırıcılar ise daha ılımlı sonuçlar üretmiştir.

STL yaklaşımı, SMOTE kullanılmadan sadece Tomek Links ile örneklemeyi sadeleştirmiştir. Bu yapı özellikle STL-SVM, STL-BBC ve STL-RF modellerinde etkili olmuştur. STL-SVM modeli 0.9523 doğruluk, 0.8695 ROC ve 0.6 üzeri F1 skoru ile oldukça dengeli bir tablo sunmuştur. Ancak STL-KNN gibi modeller, F1 skorunda ciddi düşüşlerle (ör. 0.44) daha az güvenilir sonuçlar üretmiştir. Bu durum, daha düşük karmaşıklığa sahip algoritmaların Tomek filtrelemesinden yeterince faydalanamadığını düşündürmektedir.

SSVM yöntemi genelde en yüksek başarıyı yakalamış yöntem olarak öne çıkmaktadır. SSVM-SVM ve SSVM-BBC modelleri %91 doğruluk ve 0.95 üzeri ROC ile birlikte 0.6 seviyesindeki F1 skorlarıyla istikrarlı ve etkili sonuçlar vermiştir. Bunun yanında, SSVM-RF modeli 0.9449 doğruluk ve 0.8805 ROC gibi yüksek değerlerle performans göstermiştir; fakat bu kadar yüksek skorların overfitting riskini barındırdığını unutmamak gerekir.

ADASYN'e dayalı SAD modelleri, özellikle SAD-DT ve SAD-RF ile kabul edilebilir başarı sağlamış; SAD-DT 0.91 doğruluk ve 0.7 üzerinde ROC ile öne çıkmıştır. Ancak SAD-QDA, SAD-NB gibi modellerde F1 skoru oldukça düşüktür. Bu durum, ADASYN'in sınıf sınırlarını fazla zorlayan yapay örnekler üretmesi nedeniyle bazı modellerin bu karmaşıklığı yönetememesiyle açıklanabilir.

Sonuç olarak, CKSAAGP, GDPC ve GTPC özelliklerinin birleşimiyle oluşturulan bu analizde, SMOTE-Tomek (özellikle SSTK-SVM ve SSTK-RF) ve SMOTE-NC (SSNC-SVM, SSNC-

RF) modelleri genellikle en başarılı sonuçları sunmuştur. SVM ve RF gibi güçlü sınıflandırıcılar, çoğu örnekleme stratejisiyle birlikte çalışıldığında yüksek ROC ve doğruluk değerlerine ulaşabilmiştir. Ancak RF modellerinin olağanüstü skorları, gerçek dünya uygulamalarında overfitting riski doğurabileceğinden dikkatle değerlendirilmelidir. Dengeli sonuçlarıyla SVM ve BBC modelleri, uygulama senaryolarında daha güvenilir alternatifler olarak öne çıkmaktadır. Genel olarak, bu analiz örnekleme yöntemi ile sınıflandırıcı arasındaki etkileşimin model başarısını doğrudan etkilediğini göstermektedir.

Tablo 4. 9: NR veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.125	0.304348	0.177215	0.815341	0.577706	**SCN-KNN**	0.25	0.652174	0.361446	0.849432	0.757698
SCC-SVM	0.096639	1	0.176245	0.389205	0.673252	**SCN-SVM**	0.571429	0.869565	0.689655	0.948864	0.911986
SCC-DT	0.165468	1	0.283951	0.670455	0.823708	**SCN-DT**	0.257576	0.73913	0.382022	0.84375	0.795097
SCC-RF	0.172932	1	0.294872	0.6875	0.832827	**SCN-RF**	0.821429	1	0.901961	0.985795	0.992401
SCC-NN	0.277108	1	0.433962	0.829545	0.908815	**SCN-NN**	0.533333	0.347826	0.421053	0.9375	0.663275
SCC-AB	0.178295	1	0.302632	0.698864	0.838906	**SCN-AB**	0.605263	1	0.754098	0.957386	0.977204
SCC-NB	0.080402	0.695652	0.144144	0.460227	0.569711	**SCN-NB**	0.137931	0.521739	0.218182	0.755682	0.646888
SCC-QDA	0.306667	1	0.469388	0.852273	0.920973	**SCN-QDA**	0.833333	0.217391	0.344828	0.946023	0.607176
SCC-LDA	0.17094	0.869565	0.285714	0.715909	0.787366	**SCN-LDA**	0.161765	0.478261	0.241758	0.803977	0.652504
SCC-EEC	0.123596	0.956522	0.218905	0.553977	0.741179	**SCN-EEC**	0.161765	0.956522	0.27673	0.673295	0.805009
SCC-RSC	0.166667	1	0.285714	0.673295	0.825228	**SCN-RSC**	0.396552	1	0.567901	0.900568	0.946809
SCC-BBC	0.171642	1	0.292994	0.684659	0.831307	**SCN-BBC**	0.327869	0.869565	0.47619	0.875	0.872473
SCC-BRFC	0.155172	0.782609	0.258993	0.707386	0.742368	**SCN-BRFC**	0.192771	0.695652	0.301887	0.789773	0.746002
SEN-KNN	0.202532	0.695652	0.313725	0.801136	0.752081	**SRN-KNN**	0.095506	0.73913	0.169154	0.525568	0.624884
SEN-SVM	0.328125	0.913043	0.482759	0.872159	0.891172	**SRN-SVM**	0.128492	1	0.227723	0.556818	0.762918
SEN-DT	0.454545	0.869565	0.597015	0.923295	0.898308	**SRN-DT**	0.165354	0.913043	0.28	0.693182	0.795428
SEN-RF	0.389831	1	0.560976	0.897727	0.945289	**SRN-RF**	0.193277	1	0.323944	0.727273	0.854103
SEN-NN	0.316667	0.826087	0.457831	0.872159	0.850733	**SRN-NN**	0.139241	0.956522	0.243094	0.610795	0.771574
SEN-AB	0.425926	1	0.597403	0.911932	0.952888	**SRN-AB**	0.226804	0.956522	0.366667	0.784091	0.864279
SEN-NB	0.148936	0.608696	0.239316	0.747159	0.682767	**SRN-NB**	0.109677	0.73913	0.191011	0.590909	0.659839
SEN-QDA	0.666667	0.173913	0.275862	0.940341	0.583917	**SRN-QDA**	0.151515	0.434783	0.224719	0.803977	0.632285
SEN-LDA	0.229508	0.608696	0.333333	0.840909	0.732919	**SRN-LDA**	0.141104	1	0.247312	0.602273	0.787234
SEN-EEC	0.174242	1	0.296774	0.690341	0.834347	**SRN-EEC**	0.121693	1	0.216981	0.528409	0.74772
SEN-RSC	0.306667	1	0.469388	0.852273	0.920973	**SRN-RSC**	0.131429	1	0.232323	0.568182	0.768997
SEN-BBC	0.27381	1	0.429907	0.826705	0.907295	**SRN-BBC**	0.155405	1	0.269006	0.644886	0.81003
SEN-BRFC	0.188889	0.73913	0.300885	0.775568	0.758623	**SRN-BRFC**	0.11875	0.826087	0.20765	0.588068	0.698758
SSTK-KNN	1	0.956522	0.977778	0.997159	0.978261	**SNC-KNN**	0.254545	0.608696	0.358974	0.857955	0.742038
SSTK-SVM	0.354167	0.73913	0.478873	0.894886	0.822453	**SNC-SVM**	0.44	0.956522	0.60274	0.917614	0.935708
SSTK-DT	1	1	1	1	1	**SNC-DT**	0.454545	0.652174	0.535714	0.926136	0.798731
SSTK-RF	0.233766	0.782609	0.36	0.818182	0.801639	**SNC-RF**	0.605263	1	0.754098	0.957386	0.977204
SSTK-NN	0.164557	0.565217	0.254902	0.784091	0.682305	**SNC-NN**	0.258065	0.347826	0.296296	0.892045	0.638959
SSTK-AB	0.196721	0.521739	0.285714	0.829545	0.686401	**SNC-AB**	0.294118	0.217391	0.25	0.914773	0.590459
SSTK-NB	0.45098	1	0.621622	0.920455	0.957447	**SNC-NB**	0.202899	0.608696	0.304348	0.818182	0.720761
SSTK-QDA	0.226667	0.73913	0.346939	0.818182	0.781419	**SNC-QDA**	1	0.217391	0.357143	0.948864	0.608696
SSTK-LDA	0.297872	0.608696	0.4	0.880682	0.754196	**SNC-LDA**	0.28	0.608696	0.383562	0.872159	0.749637
SSTK-EEC	0.164557	0.565217	0.254902	0.784091	0.682305	**SNC-EEC**	0.214953	1	0.353846	0.761364	0.87234
SSTK-RSC	1	1	1	1	1	**SNC-RSC**	0.129412	0.478261	0.203704	0.755682	0.626668
SSTK-BBC	0.197368	0.652174	0.30303	0.803977	0.733382	**SNC-BBC**	0.305556	0.956522	0.463158	0.855114	0.902273
SSTK-BRFC	1	0.956522	0.977778	0.997159	0.978261	**SNC-BRFC**	0.2	0.695652	0.31068	0.798295	0.750562
SOS-KNN	0.351351	0.565217	0.433333	0.903409	0.746135	**SRUS-KNN**	0.096774	0.782609	0.172249	0.508523	0.635985
SOS-SVM	0.944444	0.73913	0.829268	0.980114	0.868045	**SRUS-SVM**	0.172932	1	0.294872	0.6875	0.832827
SOS-DT	1	0.304348	0.466667	0.954545	0.652174	**SRUS-DT**	0.15534	0.695652	0.253968	0.732955	0.715607
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.18254	1	0.308725	0.707386	0.843465
SOS-NN	0.9	0.391304	0.545455	0.957386	0.694132	**SRUS-NN**	0.166667	1	0.285714	0.673295	0.825228
SOS-AB	0.666667	0.173913	0.275862	0.940341	0.583917	**SRUS-AB**	0.132867	0.826087	0.228916	0.636364	0.724594
SOS-NB	0.207547	0.478261	0.289474	0.846591	0.675301	**SRUS-NB**	0.113636	0.652174	0.193548	0.644886	0.648275
SOS-QDA	1	0.086957	0.16	0.940341	0.543478	**SRUS-QDA**	0.140625	0.782609	0.238411	0.673295	0.724131
SOS-LDA	0.5625	0.391304	0.461538	0.940341	0.685014	**SRUS-LDA**	0.144737	0.956522	0.251429	0.627841	0.780692
SOS-EEC	0.181034	0.913043	0.302158	0.724432	0.812145	**SRUS-EEC**	0.191304	0.956522	0.318841	0.732955	0.836923
SOS-RSC	0.092784	0.391304	0.15	0.710227	0.561914	**SRUS-RSC**	0.132867	0.826087	0.228916	0.636364	0.724594
SOS-BBC	0.433962	1	0.605263	0.914773	0.954407	**SRUS-BBC**	0.156028	0.956522	0.268293	0.659091	0.79741
SOS-BRFC	0.210526	0.695652	0.323232	0.809659	0.756641	**SRUS-BRFC**	0.129032	0.695652	0.217687	0.673295	0.683692

Tablo 4. 10: NR veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.361111	0.565217	0.440678	0.90625	0.747654	**SROS-KNN**	0.359375	1	0.528736	0.883523	0.93769
STL-SVM	0.944444	0.73913	0.829268	0.980114	0.868045	**SROS-SVM**	1	1	1	1	1
STL-DT	1	0.304348	0.466667	0.954545	0.652174	**SROS-DT**	0.328358	0.956522	0.488889	0.869318	0.909872
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0.923077	0.521739	0.666667	0.965909	0.75935	**SROS-NN**	0.204819	0.73913	0.320755	0.795455	0.769261
STL-AB	0.666667	0.173913	0.275862	0.940341	0.583917	**SROS-AB**	0.168317	0.73913	0.274194	0.744318	0.741906
STL-NB	0.207547	0.478261	0.289474	0.846591	0.675301	**SROS-NB**	0.159091	0.608696	0.252252	0.764205	0.691886
STL-QDA	1	0.086957	0.16	0.940341	0.543478	**SROS-QDA**	0.396552	1	0.567901	0.900568	0.946809
STL-LDA	0.5625	0.391304	0.461538	0.940341	0.685014	**SROS-LDA**	0.215385	0.608696	0.318182	0.829545	0.72684
STL-EEC	0.183486	0.869565	0.30303	0.738636	0.799524	**SROS-EEC**	0.345455	0.826087	0.487179	0.886364	0.858332
STL-RSC	0.285714	0.173913	0.216216	0.917614	0.571759	**SROS-RSC**	0.168317	0.73913	0.274194	0.744318	0.741906
STL-BBC	0.469388	1	0.638889	0.926136	0.960486	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.239437	0.73913	0.361702	0.829545	0.787498	**SROS-BRFC**	0.197368	0.652174	0.30303	0.803977	0.733382
SSM-KNN	0.164286	1	0.282209	0.667614	0.822188	**SSNC-KNN**	0.165468	1	0.283951	0.670455	0.823708
SSM-SVM	1	0.956522	0.977778	0.997159	0.978261	**SSNC-SVM**	1	1	1	1	1
SSM-DT	0.354167	0.73913	0.478873	0.894886	0.822453	**SSNC-DT**	0.382979	0.782609	0.514286	0.903409	0.847231
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.233766	0.782609	0.36	0.818182	0.801639	**SSNC-NN**	0.219178	0.695652	0.333333	0.818182	0.7612
SSM-AB	0.164557	0.565217	0.254902	0.784091	0.682305	**SSNC-AB**	0.186047	0.695652	0.293578	0.78125	0.741443
SSM-NB	0.196721	0.521739	0.285714	0.829545	0.686401	**SSNC-NB**	0.193548	0.521739	0.282353	0.826705	0.684882
SSM-QDA	0.45098	1	0.621622	0.920455	0.957447	**SSNC-QDA**	0.468085	0.956522	0.628571	0.926136	0.940267
SSM-LDA	0.226667	0.73913	0.346939	0.818182	0.781419	**SSNC-LDA**	0.223881	0.652174	0.333333	0.829545	0.74706
SSM-EEC	0.297872	0.608696	0.4	0.880682	0.754196	**SSNC-EEC**	0.3	0.652174	0.410959	0.877841	0.772895
SSM-RSC	0.164557	0.565217	0.254902	0.784091	0.682305	**SSNC-RSC**	0.186047	0.695652	0.293578	0.78125	0.741443
SSM-BBC	1	1	1	1	1	**SSNC-BBC**	1	0.956522	0.977778	0.997159	0.978261
SSM-BRFC	0.197368	0.652174	0.30303	0.803977	0.733382	**SSNC-BRFC**	0.230769	0.652174	0.340909	0.835227	0.750099
SSN-KNN	0.382353	0.565217	0.45614	0.911932	0.750694	**SAD-KNN**	0.161972	1	0.278788	0.661932	0.819149
SSN-SVM	1	0.956522	0.977778	0.997159	0.978261	**SAD-SVM**	1	1	1	1	1
SSN-DT	0.909091	0.434783	0.588235	0.960227	0.715872	**SAD-DT**	0.418605	0.782609	0.545455	0.914773	0.85331
SSN-RF	1	1	1	1	1	**SAD-RF**	1	1	1	1	1
SSN-NN	0.333333	0.043478	0.076923	0.931818	0.5187	**SAD-NN**	0.225	0.782609	0.349515	0.809659	0.797079
SSN-AB	0.526316	0.434783	0.47619	0.9375	0.703713	**SAD-AB**	0.117647	0.608696	0.197183	0.676136	0.644773
SSN-NB	0	0	0	0.934659	0.5	**SAD-NB**	0.176471	0.521739	0.263736	0.809659	0.675763
SSN-QDA	1	0.130435	0.230769	0.943182	0.565217	**SAD-QDA**	0.469388	1	0.638889	0.926136	0.960486
SSN-LDA	0.5	0.086957	0.148148	0.934659	0.540439	**SAD-LDA**	0.220779	0.73913	0.34	0.8125	0.77838
SSN-EEC	0.636364	0.304348	0.411765	0.943182	0.646095	**SAD-EEC**	0.275862	0.695652	0.395062	0.860795	0.783996
SSN-RSC	0.526316	0.434783	0.47619	0.9375	0.703713	**SAD-RSC**	0.117647	0.608696	0.197183	0.676136	0.644773
SSN-BBC	1	1	1	1	1	**SAD-BBC**	1	1	1	1	1
SSN-BRFC	1	0.043478	0.083333	0.9375	0.521739	**SAD-BRFC**	0.175258	0.73913	0.283333	0.755682	0.747985
SBS-KNN	0.158333	0.826087	0.265734	0.701705	0.759548	**SSVM-KNN**	0.163636	0.782609	0.270677	0.724432	0.751487
SBS-SVM	1	0.869565	0.930233	0.991477	0.934783	**SSVM-SVM**	1	0.869565	0.930233	0.991477	0.934783
SBS-DT	0.413043	0.826087	0.550725	0.911932	0.87201	**SSVM-DT**	0.327586	0.826087	0.469136	0.877841	0.853773
SBS-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SBS-NN	0.266667	0.695652	0.385542	0.855114	0.780957	**SSVM-NN**	0.233333	0.608696	0.337349	0.84375	0.734439
SBS-AB	0.203125	0.565217	0.298851	0.826705	0.705101	**SSVM-AB**	0.179487	0.608696	0.277228	0.792614	0.707083
SBS-NB	0.209677	0.565217	0.305882	0.832386	0.708141	**SSVM-NB**	0.243902	0.434783	0.3125	0.875	0.670279
SBS-QDA	0.428571	0.913043	0.583333	0.914773	0.913969	**SSVM-QDA**	0.488372	0.913043	0.636364	0.931818	0.923087
SBS-LDA	0.245902	0.652174	0.357143	0.846591	0.756178	**SSVM-LDA**	0.326087	0.652174	0.434783	0.889205	0.778974
SBS-EEC	0.291667	0.608696	0.394366	0.877841	0.752676	**SSVM-EEC**	0.265306	0.565217	0.361111	0.869318	0.727897
SBS-RSC	0.203125	0.565217	0.298851	0.826705	0.705101	**SSVM-RSC**	0.179487	0.608696	0.277228	0.792614	0.707083
SBS-BBC	1	1	1	1	1	**SSVM-BBC**	1	1	1	1	1
SBS-BRFC	0.237288	0.608696	0.341463	0.846591	0.735959	**SSVM-BRFC**	0.392857	0.478261	0.431373	0.917614	0.713295

4.3 İYON KANALLARI VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI

Bu bölümde, önerilen modelin IC veri seti üzerindeki performansı, Siyam sinir ağı temelli bir yapı çerçevesinde incelenmiştir. Analizler, biyolojik verilerden çıkarılan ve her biri istatistiksel ve biyolojik tanımlayıcıların farklı bir kombinasyonunu içeren dört farklı alt özellik kümesi temelinde gerçekleştirilmiştir. Veriler ilk aşamada Siyam ağı kullanılarak dengeli bir biçimde yeniden temsil edilmiş, ardından çeşitli yeniden örnekleme yöntemleriyle işlenmiştir. Son olarak, ön işlemden geçirilen bu veriler farklı makine öğrenimi algoritmalarına sunularak ilaç-hedef etkileşimlerinin tahmini yapılmıştır. Bu çalışmanın amacı, Siyam mimarisi çerçevesinde en iyi performansı sağlayacak örnekleme yöntemi ile sınıflandırıcı kombinasyonunu belirlemektir. Model değerlendirmeleri; doğruluk (Accuracy), geri çağırma (Recall), F1 skoru ve ROC eğrisi altındaki alan (ROC-AUC) gibi standart metrikler kullanılarak yapılmıştır.

4.3.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, DTI tahmini için AAC ve DPC özellikleri kullanılarak oluşturulan modellerin performansı incelenmiştir. Tablo 4.11 ve 4.12'den elde edilen sonuçlara göre, SEN tabanlı modeller de RF ve SVM ile iyi performans sergilemiştir; SEN-RF neredeyse %100 doğruluk ve ROC puanı alırken, daha basit modeller olan SEN-NN ve SEN-QDA daha düşük skorlar elde etmiştir. Bu durum, SEN'in sınıflandırıcıların gücüne bağlı olduğunu göstermektedir. SEN tabanlı modeller de RF ve SVM ile iyi performans sergilemiştir; SEN-RF neredeyse %100 doğruluk ve ROC puanı alırken, daha basit modeller olan SEN-NN ve SEN-QDA daha düşük skorlar elde etmiştir. Bu durum, SEN'in sınıflandırıcıların gücüne bağlı olduğunu göstermektedir.

SROS yöntemde SROS-RF ve SROS-SVM modellerinde yüksek doğruluk ve ROC değerleriyle dikkat çekmiştir. Ancak bu yüksek performans, yapay örneklerin aşırı öğrenilmesine bağlı olabilir. Daha basit modeller bu yöntemde dengeli ancak daha düşük sonuçlar göstermiştir. SRUS modeli genellikle diğer yöntemlere göre zayıf performans sergilemiştir. Birçok sınıflandırıcıda F1 skorları %20'nin altında kalmış, bu da eğitim verisinin ciddi oranda azaldığını ve önemli bilgilerin kaybedildiğini göstermektedir. Yalnızca SRUS-RF ve SRUS-SVM gibi güçlü sınıflandırıcılarla nispeten iyileşme görülmüştür ancak genel olarak RUS yöntemi daha az başarılı olmuştur.

SSTK modeli SVM ve RF ile birleştiğinde oldukça başarılı olmuştur. SSTK-SVM, %97 doğruluk ve yaklaşık %99 ROC değerleriyle bu kombinasyonun etkileşimleri doğru tahmin etmedeki yüksek yeteneğini ortaya koymuştur. SSTK-RF de yüksek doğruluk ve ROC değerleriyle, SMOTE-Tomek yöntemi sayesinde gürültülü örneklerin elenip faydalı örneklerin korunduğunu göstermiştir. Ancak, SSTK-BBC ve SSTK-QDA gibi F1 skoru düşük modellerin yüksek skorları, yalnızca güçlü sınıflandırıcıların bu yöntemi tam anlamıyla kullanabildiğini göstermektedir.

SSNC modeli arasında SVM ve RF en iyi sınıflandırıcılar olmaya devam etmiştir. SSNC-SVM yaklaşık %90 doğruluk ve %80 üzeri ROC ile çok iyi performans göstermiş, SSNC-RF ise neredeyse %100 doğruluk ve ROC ile yüksek sonuçlar sunmuş ancak bu durum aşırı öğrenme riskini de işaret etmektedir. Daha basit sınıflandırıcılar bu grupta orta düzey performans göstermiştir.

Sadece Tomek Links kullanan STL yaklaşımı, STL-SVM ve STL-RF modellerinde başarılı olmuştur. STL-SVM'in yüksek doğruluk ve ROC skorları, sınır örneklerin çıkarılmasının veri dengesine tek başına katkı sağladığını göstermektedir. Ancak STL-KNN gibi daha basit modellerde düşük F1 skorları, bu yöntemin basit algoritmalar için sınırlı olduğunu ortaya koymaktadır.

SSVM modeli en iyi sonuçları sunmuştur. SSVM-SVM ve SSVM-BBC %90'ın üzerinde doğruluk ve çok iyi ROC skorlarıyla güçlü sınıflandırıcılar ile optimizasyonun bu alandaki başarının anahtarı olduğunu göstermektedir. Ancak aşırı öğrenme riski göz önünde bulundurulmalıdır.Güçlü yeniden örnekleme ve sınıflandırıcılarla oluşturulan SOS modelleri, SOS-RF ve SOS-AB gibi modeller neredeyse %100 doğruluk ve ROC skorları elde etmiş, ancak rastgele örnek silme temelli SRUS modelleri zayıf sonuçlar vermiştir ki bu veri kalitesindeki düşüşe bağlanmaktadır.

SCC ve SCN gibi diğer modeller farklı sonuçlar göstermiştir; genellikle yüksek Recall fakat düşük Precision değerlerine sahip olup pozitif örneklerin aşırı tahmin edilmesi ve doğrulukta düşüş olduğunu işaret etmektedir. Bu durum, bu grup için daha hassas ayarlamalar veya yeniden örnekleme yöntemlerinin geliştirilmesi gerekliliğine işaret etmektedir.Son olarak, artırım stratejileri bazlı SBS modelleri olumlu sonuçlar vermiştir. Özellikle SBS-RF ve SBS-AB yüksek doğruluk ve ROC skorlarıyla başarılı performans göstermiştir. Bu da uygun yeniden

örnekleme ve güçlendirme algoritmalarının tahmin kalitesine olumlu katkı sağladığını vurgulamaktadır.

Genel olarak, güçlü sınıflandırıcılar olan SVM, RF ve AdaBoost ile akıllı yeniden örnekleme yöntemleri olan SMOTE-Tomek, SMOTE-NC, ROS ve Boosting en iyi performansı sağlamıştır. Daha basit modeller ya da rastgele örnek çıkarma yöntemi olan RUS kullananlar doğruluk ve denge açısından zorluklarla karşılaşmıştır. Parametrelerin dikkatli ayarlanması ve algoritma seçiminin doğru yapılması, genelleştirilebilir ve doğru modeller için kritik öneme sahiptir.

Tablo 4. 11: IC veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.186275	0.730769	0.296875	0.868517	0.802363	**SCN-KNN**	0.243056	0.673077	0.357143	0.907962	0.795157
SCC-SVM	0.082803	1	0.152941	0.579255	0.781321	**SCN-SVM**	0.537634	0.961538	0.689655	0.967129	0.964444
SCC-DT	0.08557	0.980769	0.157407	0.601169	0.783475	**SCN-DT**	0.300885	0.653846	0.412121	0.929145	0.796931
SCC-RF	0.117117	1	0.209677	0.71366	0.851177	**SCN-RF**	0.684211	1	0.8125	0.982469	0.990888
SCC-NN	0.157088	0.788462	0.261981	0.831264	0.810708	**SCN-NN**	0.32	0.461538	0.377953	0.942294	0.711407
SCC-AB	0.075362	1	0.140162	0.533966	0.757783	**SCN-AB**	0.490566	1	0.658228	0.960555	0.979499
SCC-NB	0.068729	0.769231	0.126183	0.595325	0.678845	**SCN-NB**	0.13253	0.634615	0.219269	0.828342	0.735303
SCC-QDA	0.224299	0.923077	0.360902	0.875822	0.898516	**SCN-QDA**	0.355372	0.826923	0.49711	0.93645	0.883849
SCC-LDA	0.12894	0.865385	0.224439	0.772827	0.817278	**SCN-LDA**	0.163522	0.5	0.246445	0.883857	0.699506
SCC-EEC	0.069388	0.980769	0.129606	0.499635	0.730704	**SCN-EEC**	0.125668	0.903846	0.220657	0.757487	0.827777
SCC-RSC	0.081633	0.692308	0.146045	0.692476	0.692395	**SCN-RSC**	0.137566	0.5	0.215768	0.861943	0.688117
SCC-BBC	0.100402	0.961538	0.181818	0.671293	0.810686	**SCN-BBC**	0.195402	0.980769	0.325879	0.845873	0.910658
SCC-BRFC	0.090909	0.730769	0.161702	0.712199	0.721117	**SCN-BRFC**	0.155556	0.538462	0.241379	0.871439	0.711524
SEN-KNN	0.622642	0.634615	0.628571	0.971512	0.809715	**SRN-KNN**	0.435897	0.653846	0.523077	0.954711	0.810218
SEN-SVM	0.75	0.923077	0.827586	0.985391	0.955464	**SRN-SVM**	0.666667	0.923077	0.774194	0.979547	0.952427
SEN-DT	1	0.461538	0.631579	0.979547	0.730769	**SRN-DT**	1	0.461538	0.631579	0.979547	0.730769
SEN-RF	1	1	1	1	1	**SRN-RF**	1	1	1	1	1
SEN-NN	0	0	0	0.962016	0.5	**SRN-NN**	0	0	0	0.962016	0.5
SEN-AB	1	1	1	1	1	**SRN-AB**	0.896552	1	0.945455	0.995617	0.997722
SEN-NB	0.210084	0.480769	0.292398	0.911614	0.704697	**SRN-NB**	0.221311	0.519231	0.310345	0.912345	0.723549
SEN-QDA	0.522727	0.884615	0.657143	0.964938	0.926362	**SRN-QDA**	0.48913	0.865385	0.625	0.960555	0.914849
SEN-LDA	0.461538	0.346154	0.395604	0.959825	0.665104	**SRN-LDA**	0.45	0.346154	0.391304	0.959094	0.664725
SEN-EEC	0.155629	0.903846	0.265537	0.81008	0.855112	**SRN-EEC**	0.15016	0.903846	0.257534	0.802045	0.850936
SEN-RSC	0.154839	0.461538	0.231884	0.883857	0.681035	**SRN-RSC**	0.156863	0.461538	0.234146	0.885318	0.681794
SEN-BBC	0.361702	0.980769	0.528497	0.933528	0.956216	**SRN-BBC**	0.365672	0.942308	0.526882	0.93572	0.938884
SEN-BRFC	0.192053	0.557692	0.285714	0.894083	0.732529	**SRN-BRFC**	0.211679	0.557692	0.306878	0.90431	0.737844
SSTK-KNN	0.553191	1	0.712329	0.969321	0.984055	**SNC-KNN**	0.813953	0.673077	0.736842	0.981738	0.833501
SSTK-SVM	0.928571	1	0.962963	0.997078	0.998481	**SNC-SVM**	0.875	0.942308	0.907407	0.992695	0.968496
SSTK-DT	0.117493	0.865385	0.206897	0.747991	0.80437	**SNC-DT**	0.84375	0.519231	0.642857	0.978086	0.757717
SSTK-RF	1	1	1	1	1	**SNC-RF**	1	1	1	1	1
SSTK-NN	0.326087	0.865385	0.473684	0.926954	0.897385	**SNC-NN**	0	0	0	0.962016	0.5
SSTK-AB	1	1	1	1	1	**SNC-AB**	1	1	1	1	1
SSTK-NB	0.068354	0.519231	0.120805	0.712929	0.619904	**SNC-NB**	0.213675	0.480769	0.295858	0.913075	0.705457
SSTK-QDA	0.486111	0.673077	0.564516	0.960555	0.822491	**SNC-QDA**	0.555556	0.865385	0.676692	0.96859	0.919025
SSTK-LDA	0.141892	0.807692	0.241379	0.807159	0.807415	**SNC-LDA**	0.461538	0.346154	0.395604	0.959825	0.665104
SSTK-EEC	0.14094	0.807692	0.24	0.805698	0.806656	**SNC-EEC**	0.163763	0.903846	0.277286	0.821037	0.860807
SSTK-RSC	0.098131	0.807692	0.175	0.710738	0.757301	**SNC-RSC**	0.156863	0.461538	0.234146	0.885318	0.681794
SSTK-BBC	0.962264	0.980769	0.971429	0.997809	0.989625	**SNC-BBC**	0.376923	0.942308	0.538462	0.938641	0.940402
SSTK-BRFC	0.229008	0.576923	0.327869	0.910153	0.750117	**SNC-BRFC**	0.191083	0.576923	0.287081	0.891161	0.740246
SOS-KNN	0.891892	0.634615	0.741573	0.983199	0.815789	**SRUS-KNN**	0.077799	0.788462	0.141623	0.636961	0.709721
SOS-SVM	0.94	0.903846	0.921569	0.994156	0.950784	**SRUS-SVM**	0.171053	1	0.292135	0.815924	0.904328
SOS-DT	0.96	0.461538	0.623377	0.978817	0.73039	**SRUS-DT**	0.140625	0.865385	0.241935	0.79401	0.828288
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.273684	1	0.429752	0.899196	0.947608
SOS-NN	0.2	0.019231	0.035088	0.959825	0.508097	**SRUS-NN**	0.099338	0.865385	0.178218	0.696859	0.777795
SOS-AB	1	1	1	1	1	**SRUS-AB**	0.147929	0.961538	0.25641	0.788167	0.87143
SOS-NB	0.211864	0.480769	0.294118	0.912345	0.705077	**SRUS-NB**	0.109145	0.711538	0.189258	0.768444	0.741115
SOS-QDA	0.56962	0.865385	0.687023	0.970051	0.919784	**SRUS-QDA**	0.160714	0.865385	0.271084	0.823229	0.843474
SOS-LDA	0.461538	0.346154	0.395604	0.959825	0.665104	**SRUS-LDA**	0.10687	0.807692	0.188764	0.736304	0.770589
SOS-EEC	0.163265	0.923077	0.277457	0.817385	0.868144	**SRUS-EEC**	0.123596	0.846154	0.215686	0.766253	0.804626
SOS-RSC	0.2	0.403846	0.267516	0.915997	0.670032	**SRUS-RSC**	0.103015	0.788462	0.182222	0.731191	0.758695
SOS-BBC	0.404762	0.980769	0.573034	0.944485	0.961911	**SRUS-BBC**	0.207469	0.961538	0.341297	0.859021	0.908256
SOS-BRFC	0.225806	0.538462	0.318182	0.912345	0.732784	**SRUS-BRFC**	0.163934	0.576923	0.255319	0.872169	0.730375

Tablo 4. 12: IC veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.891892	0.634615	0.741573	0.983199	0.815789	**SROS-KNN**	0.712329	1	0.832	0.98466	0.992027
STL-SVM	0.94	0.903846	0.921569	0.994156	0.950784	**SROS-SVM**	0.928571	1	0.962963	0.997078	0.998481
STL-DT	0.96	0.461538	0.623377	0.978817	0.73039	**SROS-DT**	0.159574	0.865385	0.269461	0.821768	0.842715
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0	0	0	0.962016	0.5	**SROS-NN**	0.272222	0.942308	0.422414	0.902118	0.92142
STL-AB	0.981132	1	0.990476	0.99927	0.99962	**SROS-AB**	1	1	1	1	1
STL-NB	0.210084	0.480769	0.292398	0.911614	0.704697	**SROS-NB**	0.105572	0.692308	0.183206	0.765522	0.73036
STL-QDA	0.56962	0.865385	0.687023	0.970051	0.919784	**SROS-QDA**	0.55814	0.923077	0.695652	0.969321	0.947112
STL-LDA	0.461538	0.346154	0.395604	0.959825	0.665104	**SROS-LDA**	0.13522	0.826923	0.232432	0.792549	0.809058
STL-EEC	0.15894	0.923077	0.271186	0.811541	0.865107	**SROS-EEC**	0.194805	0.865385	0.318021	0.859021	0.862077
STL-RSC	0.230769	0.403846	0.293706	0.926224	0.675348	**SROS-RSC**	0.168103	0.75	0.274648	0.849525	0.801727
STL-BBC	0.413223	0.961538	0.578035	0.946676	0.953814	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.21374	0.538462	0.306011	0.907232	0.730127	**SROS-BRFC**	0.211679	0.557692	0.306878	0.90431	0.737844
SSM-KNN	0.553191	1	0.712329	0.969321	0.984055	**SSNC-KNN**	0.553191	1	0.712329	0.969321	0.984055
SSM-SVM	0.928571	1	0.962963	0.997078	0.998481	**SSNC-SVM**	0.928571	1	0.962963	0.997078	0.998481
SSM-DT	0.117493	0.865385	0.206897	0.747991	0.80437	**SSNC-DT**	0.117493	0.865385	0.206897	0.747991	0.80437
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.326087	0.865385	0.473684	0.926954	0.897385	**SSNC-NN**	0.326087	0.865385	0.473684	0.926954	0.897385
SSM-AB	1	1	1	1	1	**SSNC-AB**	1	1	1	1	1
SSM-NB	0.068354	0.519231	0.120805	0.712929	0.619904	**SSNC-NB**	0.068354	0.519231	0.120805	0.712929	0.619904
SSM-QDA	0.486111	0.673077	0.564516	0.960555	0.822491	**SSNC-QDA**	0.486111	0.673077	0.564516	0.960555	0.822491
SSM-LDA	0.141892	0.807692	0.241379	0.807159	0.807415	**SSNC-LDA**	0.141892	0.807692	0.241379	0.807159	0.807415
SSM-EEC	0.14094	0.807692	0.24	0.805698	0.806656	**SSNC-EEC**	0.14094	0.807692	0.24	0.805698	0.806656
SSM-RSC	0.098131	0.807692	0.175	0.710738	0.757301	**SSNC-RSC**	0.098131	0.807692	0.175	0.710738	0.757301
SSM-BBC	0.962264	0.980769	0.971429	0.997809	0.989625	**SSNC-BBC**	0.962264	0.980769	0.971429	0.997809	0.989625
SSM-BRFC	0.229008	0.576923	0.327869	0.910153	0.750117	**SSNC-BRFC**	0.229008	0.576923	0.327869	0.910153	0.750117
SSN-KNN	0.942857	0.634615	0.758621	0.98466	0.816548	**SSVM-KNN**	0.58209	0.75	0.655462	0.970051	0.86437
SSN-SVM	1	0.903846	0.949495	0.996348	0.951923	**SSVM-SVM**	0.962264	0.980769	0.971429	0.997809	0.989625
SSN-DT	0.941176	0.307692	0.463768	0.972973	0.653467	**SSVM-DT**	0.12069	0.807692	0.21	0.769175	0.787673
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0.333333	0.019231	0.036364	0.961286	0.508856	**SSVM-NN**	0.313433	0.807692	0.451613	0.925493	0.868918
SSN-AB	1	1	1	1	1	**SSVM-AB**	1	1	1	1	1
SSN-NB	0	0	0	0.962016	0.5	**SSVM-NB**	0.071429	0.346154	0.118421	0.804237	0.584239
SSN-QDA	0	0	0	0.962016	0.5	**SSVM-QDA**	0.714286	0.576923	0.638298	0.975164	0.783906
SSN-LDA	0.375	0.057692	0.1	0.960555	0.526948	**SSVM-LDA**	0.164502	0.730769	0.268551	0.848795	0.792112
SSN-EEC	0.5	0.057692	0.103448	0.962016	0.527707	**SSVM-EEC**	0.205714	0.692308	0.317181	0.886779	0.793382
SSN-RSC	0	0	0	0.957633	0.497722	**SSVM-RSC**	0.119863	0.673077	0.203488	0.799854	0.738968
SSN-BBC	1	0.903846	0.949495	0.996348	0.951923	**SSVM-BBC**	1	0.923077	0.96	0.997078	0.961538
SSN-BRFC	0	0	0	0.962016	0.5	**SSVM-BRFC**	0.193103	0.538462	0.284264	0.897005	0.724812
SBS-KNN	0.58209	0.75	0.655462	0.970051	0.86437	**SBS-QDA**	0.537313	0.692308	0.605042	0.965668	0.834385
SBS-SVM	0.962264	0.980769	0.971429	0.997809	0.989625	**SBS-LDA**	0.152174	0.807692	0.256098	0.821768	0.815008
SBS-DT	0.124661	0.884615	0.218527	0.759679	0.819681	**SBS-EEC**	0.163265	0.769231	0.26936	0.84149	0.806787
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.116608	0.634615	0.197015	0.803506	0.722395
SBS-NN	0.298611	0.826923	0.438776	0.919649	0.875117	**SBS-BBC**	0.98	0.942308	0.960784	0.997078	0.970774
SBS-AB	1	1	1	1	1	**SBS-BRFC**	0.271845	0.538462	0.36129	0.927684	0.740757
SBS-NB	0.072423	0.5	0.126521	0.737765	0.623576						

4.3.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Bu değerlendirmede, GAAC ve DDE özelliklerine dayalı olarak ilaç-hedef etkileşimlerinin tahmini amacıyla çeşitli yeniden örnekleme yöntemleri ile sınıflandırma algoritmalarının kombinasyonları incelenmiştir. Sadece başarılı modellerin çıktısına odaklanan yaklaşımların aksine, bu analiz güçlü, orta düzeyde ve zayıf performans gösteren modelleri kapsayan daha

geniş bir perspektif sunarak, modellerin farklı koşullardaki davranışlarına dair daha derin bir içgörü sağlamayı hedeflemektedir.

Tablo 4.13 ve 4.14'den elde edilen sonuçlara göre, yeniden örnekleme yöntemleri arasında, hem azınlık sınıf örneklerini artıran hem de sınırda kalan örnekleri temizleyen hibrit yapısıyla SMOTE-Tomek genellikle daha iyi sonuçlar vermiştir. Özellikle güçlü sınıflandırıcılar olan SVM ve RF ile birleştirildiğinde yüksek doğruluk oranları ve 1'e yakın ROC değerleri elde edilmiştir. Ancak QDA veya Naive Bayes gibi sınıflandırıcılarla birlikte kullanıldığında performans belirgin şekilde düşmüştür, bu da yöntemlerin tek başına yeterli olmadığını ve uygun algoritmalarla desteklenmesi gerektiğini göstermektedir.

Sayısal ve kategorik veriler için geliştirilmiş bir versiyon olan SSNC , sadece yüksek doğrusal olmayan öğrenme kapasitesine sahip modellerle birlikte kullanıldığında tatmin edici sonuçlar vermiştir. Aksi takdirde, F1 veya Precision gibi metriklerde dalgalanmalar gözlemlenmiştir. Bu istikrarsızlık, özellikle SSNC-KNN veya SSNC-NN gibi modellerde daha belirgin olup, yöntemin seçilen algoritma yapısına karşı hassasiyetini ortaya koymaktadır.

Buna karşılık, çoğunluk sınıf örneklerini rastgele silmeye dayanan daha basit bir yöntem olan SRUS, birçok kombinasyonda anlamlı performans düşüşlerine yol açmıştır. Eğitim verisinin ciddi şekilde azalması ve potansiyel olarak faydalı örneklerin silinmesi, özellikle SRUS-KNN veya SRUS-DT gibi modellerde F1 ve doğruluk oranlarının belirgin şekilde düşmesine neden olmuştur. Daha iyi sonuçlar beklenen SRUS-RF veya SRUS-SVM gibi modellerde bile performans çoğunlukla orta seviyede kalmıştır, bu da yöntemin karmaşık dengesiz veri senaryolarındaki sınırlamalarını ortaya koymaktadır.

Öte yandan, azınlık sınıf örneklerini rastgele çoğaltan SROS yöntemi, yüzeyde kabul edilebilir bir performans göstermiştir. Ancak, bazı modellerin detaylı incelemesinde (örneğin SROS-RF veya SROS-SVM) aşırı öğrenme belirtileri gözlenmiştir. SROC değerinin yüksek, F1 değerinin düşük olması, modelin sentetik olarak tekrar edilen örneklere aşırı odaklandığını gösterebilir. Ayrıca, SROS-QDA veya SROS-KNN gibi basit modellerdeki performans, model yapısı dikkate alınmadan yapılan yeniden örnekleme işlemlerinin her zaman fayda sağlamadığını ortaya koymaktadır.

Sıkıştırma veya temsili örneklerin akıllı seçimine dayanan SCC ve SCN gibi yöntemler ise farklı performans profilleri sergilemiştir. Bu yöntemlerde genellikle yüksek Recall ancak düşük Precision elde edilmiştir ki bu, azınlık sınıfının aşırı tahmin edilmesinin bir göstergesidir. Azınlık ve çoğunluk sınıfı örneklerinin örtüşmesinin fazla olduğu durumlarda, bu durum genel doğruluğun düşmesine neden olmaktadır. Bununla birlikte, SCN-RF veya SCN-AB gibi bazı modellerde nispi denge işaretleri gözlenmiş ve bu yöntemlerin daha fazla optimize edilmesi halinde rekabet edebilir hale gelebileceği anlaşılmıştır.

Sadece sınır örneklerini Tomek Links yöntemiyle silmeye dayanan STL modeli, STL-SVM ve STL-RF gibi kombinasyonlarda kabul edilebilir bir performans göstermiştir. Ancak bu yaklaşımda oversampling adımının eksikliği, STL-NN veya STL-KNN gibi modellerin veri çeşitliliği açısından yetersiz kalmasına neden olmuştur. Bu sonuçlar, sadece uyumsuz örneklerin silinmesinin yeterli olmadığını ve daha dengeli modeller elde edebilmek için azınlık sınıfın da güçlendirilmesi gerektiğini ortaya koymaktadır.

SBS, SSVM veya SOS gibi optimize edilmiş ve güçlendirilmiş modeller ise genellikle çarpıcı sonuçlar sunmuştur. Yeniden örnekleme işleminin akıllıca güçlendirme algoritmalarıyla birleştirildiğinde ideal performansa yakın sonuçlar elde edilebileceğini göstermiştir. Özellikle SSVM-SVM, SOS-RF veya SBS-AB gibi kombinasyonlarda doğruluk, F1 ve SROC değerleri yüksek seviyelerde bulunmuştur. Bununla birlikte, bu modellerde bile bağımsız veri setleri üzerinde değerlendirme yapılmadığı takdirde aşırı öğrenme riski devam etmektedir.

Sonuç olarak, model başarısı tek bir yöntemin tercihine indirgenemez; aksine, elde edilen performans, Siamese mimarisi ile birlikte uygulanan özellik çıkarımı teknikleri, yeniden örnekleme stratejileri ve sınıflandırma algoritmaları arasındaki uyumlu ve dengeli etkileşimin bir sonucudur. Bu analizde daha az başarılı olan bazı kombinasyonlar yöntemlerin zayıflığından değil, veri yapısına uyumsuzluk veya parametrelerin uygun şekilde ayarlanmamış olmasından kaynaklanmıştır. Dolayısıyla, sadece güçlü sonuçlara değil, zayıf performans gösteren durumlara da analitik bir bakışla yaklaşmak, gelecekteki strateji belirleme ve uygulamalı modellerin iyileştirilmesinde oldukça değerlidir.

Tablo 4. 13: IC veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.182278	0.742268	0.292683	0.872807	0.809937	**SCN-KNN**	0.280514	0.675258	0.396369	0.927083	0.805799
SCC-SVM	0.047156	1	0.090065	0.283626	0.628647	**SCN-SVM**	0.707407	0.984536	0.823276	0.985015	0.984784
SCC-DT	0.051316	0.994845	0.097598	0.34777	0.659416	**SCN-DT**	0.636364	0.469072	0.540059	0.971674	0.72961
SCC-RF	0.047065	1	0.089898	0.282164	0.627889	**SCN-RF**	0.791837	1	0.883827	0.99068	0.995169
SCC-NN	0.133858	0.963918	0.235072	0.777595	0.867332	**SCN-NN**	0	0	0	0.964547	0.5
SCC-AB	0.059692	1	0.11266	0.44152	0.710496	**SCN-AB**	0.662116	1	0.796715	0.981908	0.990621
SCC-NB	0.038036	0.938144	0.073107	0.156615	0.533017	**SCN-NB**	0.186131	0.525773	0.274933	0.901681	0.720636
SCC-QDA	0.120831	0.989691	0.215367	0.744335	0.862504	**SCN-QDA**	0.777311	0.953608	0.856481	0.98867	0.971783
SCC-LDA	0.118655	0.927835	0.210403	0.753107	0.83726	**SCN-LDA**	0.295203	0.412371	0.344086	0.944262	0.688092
SCC-EEC	0.049427	1	0.094198	0.318165	0.646552	**SCN-EEC**	0.172584	0.902062	0.289735	0.843202	0.87155
SCC-RSC	0.052907	0.938144	0.100165	0.402412	0.660433	**SCN-RSC**	0.198198	0.340206	0.250474	0.927814	0.644809
SCC-BBC	0.059932	1	0.113087	0.443896	0.711728	**SCN-BBC**	0.341113	0.979381	0.505992	0.9322	0.954924
SCC-BRFC	0.047356	0.927835	0.090113	0.335709	0.62089	**SCN-BRFC**	0.283439	0.458763	0.350394	0.939693	0.708067
SEN-KNN	0.700599	0.603093	0.648199	0.976791	0.79681	**SRN-KNN**	0.455479	0.685567	0.547325	0.959795	0.827721
SEN-SVM	0.769841	1	0.869955	0.989401	0.994505	**SRN-SVM**	0.666667	1	0.8	0.982273	0.990811
SEN-DT	0.897727	0.407216	0.560284	0.977339	0.702756	**SRN-DT**	0.897727	0.407216	0.560284	0.977339	0.702756
SEN-RF	0.92381	1	0.960396	0.997076	0.998484	**SRN-RF**	0.92823	1	0.962779	0.997259	0.998579
SEN-NN	0	0	0	0.964547	0.5	**SRN-NN**	0	0	0	0.964547	0.5
SEN-AB	0.906542	1	0.95098	0.996345	0.998105	**SRN-AB**	0.843478	1	0.915094	0.993421	0.99659
SEN-NB	0.224449	0.57732	0.323232	0.914291	0.751998	**SRN-NB**	0.224	0.57732	0.322767	0.914108	0.751903
SEN-QDA	0.700375	0.963918	0.81128	0.984101	0.97438	**SRN-QDA**	0.625418	0.963918	0.758621	0.978253	0.971349
SEN-LDA	0.398693	0.314433	0.351585	0.958882	0.648501	**SRN-LDA**	0.4	0.319588	0.355301	0.958882	0.650984
SEN-EEC	0.156415	0.917526	0.267267	0.821637	0.867819	**SRN-EEC**	0.167126	0.938144	0.28371	0.832054	0.883149
SEN-RSC	0.160291	0.453608	0.236878	0.896382	0.683132	**SRN-RSC**	0.227405	0.402062	0.290503	0.930373	0.675927
SEN-BBC	0.34854	0.984536	0.514825	0.934211	0.958448	**SRN-BBC**	0.328205	0.989691	0.49294	0.927814	0.957615
SEN-BRFC	0.221739	0.525773	0.311927	0.917763	0.728972	**SRN-BRFC**	0.216842	0.530928	0.307922	0.915387	0.730223
SSTK-KNN	0.537396	1	0.699099	0.969481	0.98418	**SNC-KNN**	0.826923	0.664948	0.737143	0.983187	0.829916
SSTK-SVM	0.979798	1	0.989796	0.999269	0.999621	**SNC-SVM**	0.818565	1	0.900232	0.992142	0.995926
SSTK-DT	0.27381	0.829897	0.411765	0.915936	0.874498	**SNC-DT**	0.908046	0.407216	0.562278	0.977522	0.70285
SSTK-RF	1	1	1	1	1	**SNC-RF**	0.941748	1	0.97	0.997807	0.998863
SSTK-NN	0.35687	0.963918	0.520891	0.937135	0.950034	**SNC-NN**	0	0	0	0.964547	0.5
SSTK-AB	1	1	1	1	1	**SNC-AB**	0.932692	1	0.965174	0.997442	0.998674
SSTK-NB	0.136054	0.618557	0.223048	0.847222	0.737092	**SNC-NB**	0.230928	0.57732	0.329897	0.916849	0.753324
SSTK-QDA	0.8125	0.804124	0.80829	0.986477	0.898651	**SNC-QDA**	0.708333	0.963918	0.816594	0.984649	0.974664
SSTK-LDA	0.152194	0.840206	0.257708	0.828399	0.834086	**SNC-LDA**	0.403974	0.314433	0.353623	0.959247	0.648691
SSTK-EEC	0.177106	0.845361	0.292857	0.855263	0.850494	**SNC-EEC**	0.163904	0.917526	0.278125	0.83114	0.872745
SSTK-RSC	0.138063	0.742268	0.232821	0.826572	0.785969	**SNC-RSC**	0.227405	0.402062	0.290503	0.930373	0.675927
SSTK-BBC	0.994872	1	0.997429	0.999817	0.999905	**SNC-BBC**	0.40678	0.989691	0.576577	0.948465	0.96832
SSTK-BRFC	0.309942	0.546392	0.395522	0.940789	0.750839	**SNC-BRFC**	0.225225	0.515464	0.31348	0.919956	0.725144
SOS-KNN	0.874016	0.572165	0.691589	0.981908	0.784567	**SRUS-KNN**	0.07732	0.927835	0.142744	0.604898	0.760431
SOS-SVM	0.857143	0.989691	0.91866	0.993787	0.991814	**SRUS-SVM**	0.182502	1	0.308671	0.841192	0.917677
SOS-DT	0.908046	0.407216	0.562278	0.977522	0.70285	**SRUS-DT**	0.088608	0.902062	0.161365	0.66758	0.780512
SOS-RF	0.97	1	0.984772	0.998904	0.999432	**SRUS-RF**	0.238037	1	0.384539	0.886513	0.941171
SOS-NN	0	0	0	0.964547	0.5	**SRUS-NN**	0.120086	0.865979	0.210923	0.770285	0.816374
SOS-AB	0.955665	1	0.97733	0.998355	0.999147	**SRUS-AB**	0.162072	1	0.278936	0.816703	0.904983
SOS-NB	0.232068	0.56701	0.329341	0.918129	0.749022	**SRUS-NB**	0.107807	0.747423	0.188434	0.771747	0.760032
SOS-QDA	0.71374	0.963918	0.820175	0.985015	0.974854	**SRUS-QDA**	0.212087	0.958763	0.347339	0.872259	0.913921
SOS-LDA	0.403974	0.314433	0.353623	0.959247	0.648691	**SRUS-LDA**	0.11691	0.865979	0.206009	0.763341	0.812774
SOS-EEC	0.170476	0.92268	0.287781	0.838085	0.878828	**SRUS-EEC**	0.137203	0.804124	0.23441	0.813779	0.809129
SOS-RSC	0.21547	0.402062	0.280576	0.926901	0.674127	**SRUS-RSC**	0.101429	0.731959	0.178168	0.760599	0.746805
SOS-BBC	0.374257	0.974227	0.540773	0.941338	0.957178	**SRUS-BBC**	0.179754	0.979381	0.303757	0.840826	0.907557
SOS-BRFC	0.231325	0.494845	0.315271	0.923794	0.717203	**SRUS-BRFC**	0.193548	0.525773	0.28294	0.905519	0.722625

Tablo 4. 14: IC veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.874016	0.572165	0.691589	0.981908	0.784567	**SROS-KNN**	0.718519	1	0.836207	0.986111	0.9928
STL-SVM	0.857143	0.989691	0.91866	0.993787	0.991814	**SROS-SVM**	0.95098	1	0.974874	0.998173	0.999053
STL-DT	0.908046	0.407216	0.562278	0.977522	0.70285	**SROS-DT**	0.235049	0.871134	0.370208	0.89492	0.883464
STL-RF	0.974874	1	0.987277	0.999086	0.999526	**SROS-RF**	1	1	1	1	1
STL-NN	0	0	0	0.964547	0.5	**SROS-NN**	0.336237	0.994845	0.502604	0.93019	0.961329
STL-AB	0.965174	1	0.982278	0.998721	0.999337	**SROS-AB**	1	1	1	1	1
STL-NB	0.233193	0.572165	0.331343	0.918129	0.751505	**SROS-NB**	0.129126	0.685567	0.21732	0.824927	0.757808
STL-QDA	0.71374	0.963918	0.820175	0.985015	0.974854	**SROS-QDA**	0.65493	0.958763	0.778243	0.980629	0.970098
STL-LDA	0.403974	0.314433	0.353623	0.959247	0.648691	**SROS-LDA**	0.145597	0.860825	0.249068	0.815972	0.837574
STL-EEC	0.171128	0.92268	0.28871	0.838816	0.879207	**SROS-EEC**	0.19743	0.871134	0.321905	0.869883	0.870486
STL-RSC	0.15971	0.453608	0.236242	0.896016	0.682943	**SROS-RSC**	0.143337	0.659794	0.235511	0.848136	0.757426
STL-BBC	0.38	0.979381	0.54755	0.942617	0.960324	**SROS-BBC**	0.994872	1	0.997429	0.999817	0.999905
STL-BRFC	0.229358	0.515464	0.31746	0.921418	0.725902	**SROS-BRFC**	0.272487	0.530928	0.36014	0.933114	0.739412
SSM-KNN	0.537396	1	0.699099	0.969481	0.98418	**SSNC-KNN**	0.517333	1	0.681898	0.966923	0.982853
SSM-SVM	0.979798	1	0.989796	0.999269	0.999621	**SSNC-SVM**	0.979798	1	0.989796	0.999269	0.999621
SSM-DT	0.27381	0.829897	0.411765	0.915936	0.874498	**SSNC-DT**	0.272109	0.824742	0.409207	0.91557	0.871825
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.35687	0.963918	0.520891	0.937135	0.950034	**SSNC-NN**	0.441397	0.912371	0.594958	0.955958	0.934965
SSM-AB	1	1	1	1	1	**SSNC-AB**	1	1	1	1	1
SSM-NB	0.136054	0.618557	0.223048	0.847222	0.737092	**SSNC-NB**	0.135802	0.623711	0.223041	0.845943	0.738911
SSM-QDA	0.8125	0.804124	0.80829	0.986477	0.898651	**SSNC-QDA**	0.726415	0.793814	0.758621	0.982091	0.891413
SSM-LDA	0.152194	0.840206	0.257708	0.828399	0.834086	**SSNC-LDA**	0.155303	0.845361	0.2624	0.831506	0.838179
SSM-EEC	0.177106	0.845361	0.292857	0.855263	0.850494	**SSNC-EEC**	0.157625	0.793814	0.263023	0.842288	0.818942
SSM-RSC	0.138063	0.742268	0.232821	0.826572	0.785969	**SSNC-RSC**	0.102459	0.773196	0.180941	0.751827	0.762119
SSM-BBC	0.994872	1	0.997429	0.999817	0.999905	**SSNC-BBC**	0.994872	1	0.997429	0.999817	0.999905
SSM-BRFC	0.309942	0.546392	0.395522	0.940789	0.750839	**SSNC-BRFC**	0.317507	0.551546	0.403013	0.942069	0.753985
SSN-KNN	0.902439	0.572165	0.700315	0.982639	0.784946	**SSVM-KNN**	0.587814	0.845361	0.693446	0.973501	0.911786
SSN-SVM	0.968586	0.953608	0.961039	0.997259	0.976236	**SSVM-SVM**	0.989744	0.994845	0.992288	0.999452	0.997233
SSN-DT	0.891566	0.381443	0.534296	0.976425	0.689869	**SSVM-DT**	0.249603	0.809278	0.381531	0.906981	0.859925
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0.333333	0.020619	0.038835	0.963816	0.509551	**SSVM-NN**	0.425068	0.804124	0.55615	0.954496	0.882073
SSN-AB	1	1	1	1	1	**SSVM-AB**	1	1	1	1	1
SSN-NB	0	0	0	0.964547	0.5	**SSVM-NB**	0.14882	0.42268	0.220134	0.893823	0.66691
SSN-QDA	0	0	0	0.964547	0.5	**SSVM-QDA**	0.829787	0.603093	0.698507	0.981542	0.799273
SSN-LDA	0.4375	0.072165	0.123894	0.963816	0.534377	**SSVM-LDA**	0.188654	0.737113	0.30042	0.878289	0.810296
SSN-EEC	0.264151	0.072165	0.11336	0.959978	0.532388	**SSVM-EEC**	0.184704	0.659794	0.288613	0.884686	0.776373
SSN-RSC	0.333333	0.005155	0.010152	0.964364	0.502388	**SSVM-RSC**	0.098558	0.634021	0.170596	0.781433	0.710436
SSN-BBC	1	0.938144	0.968085	0.997807	0.969072	**SSVM-BBC**	0.994737	0.974227	0.984375	0.998904	0.987019
SSN-BRFC	0	0	0	0.964547	0.5	**SSVM-BRFC**	0.27566	0.484536	0.351402	0.936586	0.718869
SBS-KNN	0.563574	0.845361	0.676289	0.971308	0.910649	**SBS-QDA**	0.82659	0.737113	0.779292	0.985197	0.865715
SBS-SVM	0.989796	1	0.994872	0.999635	0.999811	**SBS-LDA**	0.154146	0.814433	0.259229	0.834978	0.825083
SBS-DT	0.239658	0.865979	0.375419	0.897844	0.882497	**SBS-EEC**	0.191304	0.793814	0.308308	0.873721	0.835236
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.153439	0.747423	0.254609	0.844846	0.797925
SBS-NN	0.352818	0.871134	0.502229	0.938779	0.9062	**SBS-BBC**	1	0.974227	0.986945	0.999086	0.987113
SBS-AB	1	1	1	1	1	**SBS-BRFC**	0.280347	0.5	0.359259	0.936769	0.726412
SBS-NB	0.133562	0.603093	0.218692	0.847222	0.729644						

4.3.3 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, Ion Channel (IC) veri kümesi üzerinde PSEAAC ve PSEPSSM bileşik özellikleriyle oluşturulan temsiller kullanılarak, farklı Siyam temelli yeniden örnekleme stratejilerinin makine öğrenimi performanslarına etkisi değerlendirilmiştir. Analizler, veri dengesini sağlamak ve sınıflar arası ayırt ediciliği artırmak amacıyla uygulanan bu yöntemlerin, model başarımında belirleyici rol oynadığını göstermektedir.

Tablo 4.15 ve 4.16 analiz sonuçlarına göre, SCC yöntemleri genel olarak zayıf performans sergilemiştir. Özellikle SCC-SVM, SCC-NB ve SCC-EEC modellerinde F1 değerleri 0.13'ün altındayken, yalnızca SCC-QDA yaklaşık 0.58 F1 değeri ile kısmen kabul edilebilir sonuç vermiştir. Bu durum, yoğun örnek çıkarımının, veri yapısını aşırı sadeleştirerek sınıflar arası ayrımı zorlaştırdığını ve modellerin yanlış pozitif oranlarını artırdığını göstermektedir.

Buna karşılık, SCN yöntemi daha hassas ve sınıf sınırlarına odaklı örnek temizleme yapısıyla oldukça başarılı olmuştur. SCN-RF modeli 0.936 F1, 0.995 doğruluk ve 0.997 ROC gibi yüksek değerlere ulaşmıştır. Aynı şekilde SCN-AB ve SCN-SVM modelleri de F1 değerlerini 0.8'in üzerine çıkarmıştır. Bu başarı, karar sınırlarını koruyan seçici temizlemenin Siami temsille güçlü bir sinerji oluşturduğunu göstermektedir.

Oversampling yöntemleri arasında SSTK ve SEN yöntemleri öne çıkmaktadır. SSTK-RF, SSTK-SVM ve SSTK-AB modelleri tüm metriklerde 1.00 değerine ulaşarak neredeyse mükemmel sonuçlar üretmiştir. Bu durum, hem azınlık sınıfını güçlendiren hem de problemli sınır örneklerini ayıklayan bu yaklaşımın üstünlüğünü açıkça ortaya koymaktadır. Benzer şekilde, SEN-RF ve SEN-AB modelleri de 1.00 F1 ve doğruluk ile veri kalitesini en üst seviyeye taşımıştır.

SOS yöntemi daha değişken sonuçlar üretmiştir. SOS-SVM ve SOS-RF modelleri yüksek başarı sağlarken, SOS-NN ve SOS-NB modellerinde F1 değerleri sırasıyla 0.0 ve 0.30'un altında kalmıştır. Bu durum, yalnızca azınlık sınıfının örnek sayısını artırmanın yetersiz olduğunu, gürültü temizliği ve uygun model seçiminin gerekli olduğunu göstermektedir.

SRN yöntemi ise özellikle NN ve NB modellerinde 0.3'ün altında kalan F1 skorları ile zayıf performans sergilemiştir. Bu, örneklerin basit tekrarının sınıf yapısını tam olarak yansıtamadığını ve genelleme gücünü düşürdüğünü göstermektedir. Benzer şekilde SRUS yöntemi de çoğu modelde başarım düşüşüne neden olmuş, örneğin SRUS-NN modeli yalnızca 0.19 F1 üretmiştir.

SROS ve STL yöntemleri özellikle RF, SVM ve AB modelleriyle birlikte kullanıldığında olağanüstü sonuçlar elde etmiştir. Bu yöntemlerde F1 ve doğruluk değerleri 1.00'e ulaşarak, hem veri dengesi hem de önemli yapıların korunması konusunda güçlü bir avantaj sağlamıştır.

Daha zayıf sınıflandırıcılar kullanılsa dahi, diğer stratejilere kıyasla önemli iyileşmeler gözlemlenmiştir.

Karma yaklaşımlar olan SSM, SSNC, SSN ve SSVM yöntemleri genel olarak üstün performans göstermiştir. Bu gruplarda özellikle RF ve SVM modelleriyle 1.00 F1, doğruluk ve ROC değerleri elde edilmiştir. Bu durum, hedefli örneklemenin gürültü temizleme ile birlikte uygulanmasının, özellikle dayanıklı sınıflandırıcılarla kullanıldığında yüksek fayda sağladığını göstermektedir. Ancak aynı gruptaki NB, LDA ve EEC gibi daha basit sınıflandırıcılar bu karma stratejilerden tam verim alamamış, düşük F1 skorları üretmiştir.

Genel olarak, IC veri setinde PSEAAC-PSEPSSM özelliklerinin Siyam temsilleriyle birlikte kullanılması, sınıf ayrımını kuvvetlendirerek yüksek doğruluk değerleri elde edilmesini sağlamıştır. En başarılı stratejiler SSTK, SROS, SSNC ve SEN olurken, bu yöntemler RF, SVM ve AB sınıflandırıcıları ile birlikte kullanıldığında üstün performans sergilemiştir. Diğer yandan, örnekleme yöntemlerinin başarısı, sınıflandırıcı seçimiyle doğrudan ilişkili olup, basit stratejilerle desteklenen yapılar genellikle yetersiz kalmıştır. Bu bulgular, biyolojik olarak karmaşık veri kümelerinde hem örnekleme yöntemi hem de sınıflandırıcı seçiminin dikkatle planlanması gerektiğini bir kez daha ortaya koymaktadır.

Tablo 4. 15: IC veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.317757	0.653846	0.427673	0.933528	0.799209	**SCN-KNN**	0.319328	0.730769	0.444444	0.930606	0.834633
SCC-SVM	0.070175	1	0.131148	0.496713	0.738421	**SCN-SVM**	0.69863	0.980769	0.816	0.983199	0.982032
SCC-DT	0.079675	0.942308	0.146927	0.584368	0.756272	**SCN-DT**	0.318966	0.711538	0.440476	0.931337	0.825777
SCC-RF	0.146893	1	0.256158	0.779401	0.885345	**SCN-RF**	0.881356	1	0.936937	0.994887	0.997342
SCC-NN	0.139535	0.807692	0.23796	0.803506	0.805517	**SCN-NN**	0.318182	0.538462	0.4	0.938641	0.746452
SCC-AB	0.080123	1	0.148359	0.563915	0.773349	**SCN-AB**	0.702703	1	0.825397	0.98393	0.991648
SCC-NB	0.07027	0.75	0.128501	0.613587	0.6791	**SCN-NB**	0.150754	0.576923	0.239044	0.860482	0.724301
SCC-QDA	0.431193	0.903846	0.583851	0.951059	0.928385	**SCN-QDA**	0.734375	0.903846	0.810345	0.98393	0.945469
SCC-LDA	0.128358	0.826923	0.222222	0.780131	0.802604	**SCN-LDA**	0.259615	0.519231	0.346154	0.925493	0.730382
SCC-EEC	0.065445	0.961538	0.122549	0.476991	0.709699	**SCN-EEC**	0.156584	0.846154	0.264264	0.821037	0.8331
SCC-RSC	0.061633	0.769231	0.114123	0.546384	0.653408	**SCN-RSC**	0.231481	0.480769	0.3125	0.919649	0.708874
SCC-BBC	0.132626	0.961538	0.2331	0.759679	0.856623	**SCN-BBC**	0.372093	0.923077	0.530387	0.937911	0.930787
SCC-BRFC	0.141732	0.692308	0.235294	0.829072	0.76339	**SCN-BRFC**	0.340909	0.576923	0.428571	0.941563	0.766442
SEN-KNN	0.74	0.711538	0.72549	0.979547	0.850834	**SRN-KNN**	0.578125	0.711538	0.637931	0.969321	0.845519
SEN-SVM	0.910714	0.980769	0.944444	0.995617	0.988486	**SRN-SVM**	0.772727	0.980769	0.864407	0.988313	0.98469
SEN-DT	0.964286	0.519231	0.675	0.981008	0.759236	**SRN-DT**	0.931034	0.519231	0.666667	0.980278	0.758856
SEN-RF	1	1	1	1	1	**SRN-RF**	0.896552	1	0.945455	0.995617	0.997722
SEN-NN	0	0	0	0.962016	0.5	**SRN-NN**	0	0	0	0.962016	0.5
SEN-AB	0.981132	1	0.990476	0.99927	0.99962	**SRN-AB**	0.928571	1	0.962963	0.997078	0.998481
SEN-NB	0.2	0.576923	0.29703	0.896275	0.742903	**SRN-NB**	0.197368	0.576923	0.294118	0.894814	0.742144
SEN-QDA	0.734375	0.903846	0.810345	0.98393	0.945469	**SRN-QDA**	0.701493	0.903846	0.789916	0.981738	0.94433
SEN-LDA	0.461538	0.346154	0.395604	0.959825	0.665104	**SRN-LDA**	0.461538	0.346154	0.395604	0.959825	0.665104
SEN-EEC	0.152249	0.846154	0.258065	0.815194	0.830062	**SRN-EEC**	0.149837	0.884615	0.256267	0.804967	0.843219
SEN-RSC	0.213675	0.480769	0.295858	0.913075	0.705457	**SRN-RSC**	0.123223	0.5	0.197719	0.845873	0.679765
SEN-BBC	0.404762	0.980769	0.573034	0.944485	0.961911	**SRN-BBC**	0.346667	1	0.514851	0.928415	0.962794
SEN-BRFC	0.19697	0.5	0.282609	0.903579	0.709757	**SRN-BRFC**	0.209302	0.519231	0.298343	0.907232	0.720891
SSTK-KNN	0.504854	1	0.670968	0.962747	0.980638	**SNC-KNN**	0.8125	0.75	0.78	0.98393	0.871583
SSTK-SVM	0.981132	1	0.990476	0.99927	0.99962	**SNC-SVM**	0.944444	0.980769	0.962264	0.997078	0.989246
SSTK-DT	0.263804	0.826923	0.4	0.905771	0.867903	**SNC-DT**	0.928571	0.5	0.65	0.979547	0.749241
SSTK-RF	1	1	1	1	1	**SNC-RF**	0.981132	1	0.990476	0.99927	0.99962
SSTK-NN	0.387931	0.865385	0.535714	0.943024	0.905737	**SNC-NN**	0	0	0	0.962016	0.5
SSTK-AB	1	1	1	1	1	**SNC-AB**	0.945455	1	0.971963	0.997809	0.998861
SSTK-NB	0.099668	0.576923	0.169972	0.785975	0.685576	**SNC-NB**	0.205479	0.576923	0.30303	0.899196	0.744422
SSTK-QDA	0.690909	0.730769	0.71028	0.977356	0.858931	**SNC-QDA**	0.734375	0.903846	0.810345	0.98393	0.945469
SSTK-LDA	0.142857	0.788462	0.241888	0.812272	0.800837	**SNC-LDA**	0.461538	0.346154	0.395604	0.959825	0.665104
SSTK-EEC	0.183962	0.75	0.295455	0.864134	0.80932	**SNC-EEC**	0.171004	0.884615	0.286604	0.832725	0.857646
SSTK-RSC	0.149533	0.615385	0.240602	0.852447	0.738596	**SNC-RSC**	0.140449	0.480769	0.217391	0.868517	0.682298
SSTK-BBC	1	0.980769	0.990291	0.99927	0.990385	**SNC-BBC**	0.393701	0.961538	0.558659	0.942294	0.951536
SSTK-BRFC	0.28866	0.538462	0.375839	0.932067	0.743035	**SNC-BRFC**	0.275229	0.576923	0.372671	0.926224	0.758469
SOS-KNN	0.973684	0.711538	0.822222	0.988313	0.85539	**SRUS-KNN**	0.078992	0.903846	0.145286	0.596056	0.743874
SOS-SVM	0.980392	0.961538	0.970874	0.997809	0.98039	**SRUS-SVM**	0.239631	1	0.386617	0.879474	0.937358
SOS-DT	0.84375	0.519231	0.642857	0.978086	0.757717	**SRUS-DT**	0.118932	0.942308	0.211207	0.732652	0.833341
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.315152	1	0.479263	0.917458	0.957099
SOS-NN	0	0	0	0.962016	0.5	**SRUS-NN**	0.107056	0.846154	0.190065	0.726077	0.783745
SOS-AB	1	1	1	1	1	**SRUS-AB**	0.192593	1	0.322981	0.84076	0.917236
SOS-NB	0.208333	0.576923	0.306122	0.900657	0.745181	**SRUS-NB**	0.09322	0.634615	0.162562	0.751644	0.69544
SOS-QDA	0.734375	0.903846	0.810345	0.98393	0.945469	**SRUS-QDA**	0.308725	0.884615	0.457711	0.92038	0.903204
SOS-LDA	0.461538	0.346154	0.395604	0.959825	0.665104	**SRUS-LDA**	0.112045	0.769231	0.195599	0.759679	0.764266
SOS-EEC	0.160142	0.865385	0.27027	0.822498	0.843095	**SRUS-EEC**	0.118694	0.769231	0.205656	0.774288	0.771859
SOS-RSC	0.227723	0.442308	0.300654	0.921841	0.691541	**SRUS-RSC**	0.11041	0.673077	0.189702	0.781592	0.729477
SOS-BBC	0.396947	1	0.568306	0.942294	0.970008	**SRUS-BBC**	0.224299	0.923077	0.360902	0.875822	0.898516
SOS-BRFC	0.287234	0.519231	0.369863	0.932798	0.734179	**SRUS-BRFC**	0.230769	0.576923	0.32967	0.910884	0.750496

Tablo 4. 16: IC veri kümesi üzerinde PSEAAC-PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.973684	0.711538	0.822222	0.988313	0.85539	**SROS-KNN**	0.753623	1	0.859504	0.987582	0.993546
STL-SVM	0.980392	0.961538	0.970874	0.997809	0.98039	**SROS-SVM**	0.981132	1	0.990476	0.99927	0.99962
STL-DT	0.84375	0.519231	0.642857	0.978086	0.757717	**SROS-DT**	0.241026	0.903846	0.380567	0.88824	0.895735
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0	0	0	0.962016	0.5	**SROS-NN**	0.347518	0.942308	0.507772	0.930606	0.936226
STL-AB	1	1	1	1	1	**SROS-AB**	1	1	1	1	1
STL-NB	0.208333	0.576923	0.306122	0.900657	0.745181	**SROS-NB**	0.092643	0.653846	0.162291	0.743608	0.700499
STL-QDA	0.734375	0.903846	0.810345	0.98393	0.945469	**SROS-QDA**	0.695652	0.923077	0.793388	0.981738	0.953566
STL-LDA	0.461538	0.346154	0.395604	0.959825	0.665104	**SROS-LDA**	0.141447	0.826923	0.241573	0.802776	0.814373
STL-EEC	0.147541	0.865385	0.252101	0.804967	0.833983	**SROS-EEC**	0.168	0.807692	0.278146	0.84076	0.824879
STL-RSC	0.154762	0.5	0.236364	0.877283	0.69609	**SROS-RSC**	0.157447	0.711538	0.25784	0.844412	0.780598
STL-BBC	0.369565	0.980769	0.536842	0.93572	0.957355	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.27	0.519231	0.355263	0.928415	0.731901	**SROS-BRFC**	0.265306	0.5	0.346667	0.928415	0.722665
SSM-KNN	0.504854	1	0.670968	0.962747	0.980638	**SSNC-KNN**	0.514851	1	0.679739	0.964207	0.981397
SSM-SVM	0.981132	1	0.990476	0.99927	0.99962	**SSNC-SVM**	0.981132	1	0.990476	0.99927	0.99962
SSM-DT	0.263804	0.826923	0.4	0.905771	0.867903	**SSNC-DT**	0.156997	0.884615	0.266667	0.815194	0.848534
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.387931	0.865385	0.535714	0.943024	0.905737	**SSNC-NN**	0.420561	0.865385	0.566038	0.949598	0.909154
SSM-AB	1	1	1	1	1	**SSNC-AB**	1	1	1	1	1
SSM-NB	0.099668	0.576923	0.169972	0.785975	0.685576	**SSNC-NB**	0.09901	0.576923	0.169014	0.784514	0.684817
SSM-QDA	0.690909	0.730769	0.71028	0.977356	0.858931	**SSNC-QDA**	0.690909	0.730769	0.71028	0.977356	0.858931
SSM-LDA	0.142857	0.788462	0.241888	0.812272	0.800837	**SSNC-LDA**	0.141869	0.788462	0.240469	0.810811	0.800077
SSM-EEC	0.183962	0.75	0.295455	0.864134	0.80932	**SSNC-EEC**	0.15625	0.769231	0.25974	0.833455	0.802611
SSM-RSC	0.149533	0.615385	0.240602	0.852447	0.738596	**SSNC-RSC**	0.106494	0.788462	0.187643	0.740687	0.763631
SSM-BBC	1	0.980769	0.990291	0.99927	0.990385	**SSNC-BBC**	0.980392	0.961538	0.970874	0.997809	0.98039
SSM-BRFC	0.28866	0.538462	0.375839	0.932067	0.743035	**SSNC-BRFC**	0.285714	0.538462	0.373333	0.931337	0.742655
SSN-KNN	1	0.692308	0.818182	0.988313	0.846154	**SSVM-KNN**	0.544304	0.826923	0.656489	0.967129	0.899794
SSN-SVM	1	0.942308	0.970297	0.997809	0.971154	**SSVM-SVM**	0.980769	0.980769	0.980769	0.998539	0.990005
SSN-DT	0.821429	0.442308	0.575	0.975164	0.719256	**SSVM-DT**	0.293706	0.807692	0.430769	0.918919	0.865501
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0	0	0	0.962016	0.5	**SSVM-NN**	0.333333	0.826923	0.475138	0.930606	0.880812
SSN-AB	1	1	1	1	1	**SSVM-AB**	1	1	1	1	1
SSN-NB	0	0	0	0.962016	0.5	**SSVM-NB**	0.104651	0.519231	0.174194	0.813002	0.671916
SSN-QDA	0	0	0	0.962016	0.5	**SSVM-QDA**	0.755556	0.653846	0.701031	0.978817	0.822747
SSN-LDA	0.25	0.038462	0.066667	0.959094	0.516953	**SSVM-LDA**	0.153846	0.730769	0.254181	0.837107	0.786038
SSN-EEC	0.5	0.115385	0.1875	0.962016	0.555414	**SSVM-EEC**	0.18	0.692308	0.285714	0.868517	0.783891
SSN-RSC	0.272727	0.057692	0.095238	0.958364	0.525809	**SSVM-RSC**	0.150215	0.673077	0.245614	0.842951	0.761368
SSN-BBC	1	0.884615	0.938776	0.995617	0.942308	**SSVM-BBC**	0.979167	0.903846	0.94	0.995617	0.951543
SSN-BRFC	0	0	0	0.962016	0.5	**SSVM-BRFC**	0.162162	0.576923	0.253165	0.870709	0.729616
SBS-KNN	0.483146	0.826923	0.609929	0.959825	0.895998	**SBS-QDA**	0.672727	0.711538	0.691589	0.975895	0.848936
SBS-SVM	0.980769	0.980769	0.980769	0.998539	0.990005	**SBS-LDA**	0.146341	0.807692	0.247788	0.813733	0.810832
SBS-DT	0.113164	0.942308	0.202062	0.717312	0.825368	**SBS-EEC**	0.187166	0.673077	0.292887	0.876552	0.778832
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.103825	0.730769	0.181818	0.750183	0.740859
SBS-NN	0.390909	0.826923	0.530864	0.944485	0.888025	**SBS-BBC**	1	0.942308	0.970297	0.997809	0.971154
SBS-AB	1	1	1	1	1	**SBS-BRFC**	0.226563	0.557692	0.322222	0.910884	0.741261
SBS-NB	0.100324	0.596154	0.171745	0.781592	0.692534						

4.3.4 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, CKSAAGP, GDPC ve GTPC özellikleriyle oluşturulan ilaç-hedef görüntü temsilleri Siamese ağıyla dengelenmiş, farklı yeniden örnekleme stratejileri ve sınıflayıcılarla eğitilerek model başarımına etkileri kapsamlı biçimde karşılaştırılmıştır.

Tablo 4.17 , 4.18 analiz sonuçlarına göre**,** veri kümesinden rastgele örnek çıkaran veya örnek sayısını azaltan yöntemler, genel olarak düşük başarı sergilemiştir. Özellikle SCC ve SRUS gibi stratejiler altında eğitilen modellerin çoğu, F1 skoru açısından 0.2'nin altında kalmış, bazı durumlarda recall yüksek olmasına rağmen precision değerlerinin ciddi biçimde düşmesi modelin genelleme yeteneğini zayıflatmıştır. Bu durum, rastgele veri azaltmanın karar sınırlarını bozarak modelin ayırt edici yapıları öğrenmesini engellediğini göstermektedir.

Yapay örnek üretimi yapan SMOTE, SOS ve SRN gibi yöntemler orta düzeyde sonuçlar sağlamıştır. Bu yöntemlerin güçlü sınıflandırıcılarla birlikte (örneğin Random Forest, AdaBoost, SVM) kullanıldığında başarılı olduğu, ancak basit sınıflayıcılarla (Naive Bayes, LDA) birlikte kullanıldığında performansın düştüğü gözlemlenmiştir. Bu durum, örneklemenin tek başına yeterli olmadığını ve sınıflayıcının karmaşık yapıları ayırt edebilme yeteneğinin önemli olduğunu göstermektedir.

Buna karşılık, yalnızca örnek sayısını artırmak veya azaltmaktan öteye geçerek veri yapısını dikkate alan SCN, SMOTE-ENN, STL ve SROS gibi yöntemler oldukça yüksek başarı sağlamıştır. SCN-SVM, SCN-RF, SEN-SVM ve SROS-RF gibi modeller F1 skoru açısından 0.93-1.00 aralığında sonuçlar üretmiş, yüksek ROC ve accuracy değerleriyle desteklenmiştir. Bu yöntemlerin veri dağılımına duyarlı olması ve karar sınırına yakın örnekleri dikkate alarak gürültüyü filtrelemesi, model başarımını önemli ölçüde artırmıştır.

Çalışmada kullanılan en başarılı yaklaşımlar ise hibrit örnekleme yöntemleri olmuştur. SSTK, SSNC, SSN, SSVM ve SBS gibi yöntemler, veriyi hem sayısal hem de yapısal olarak iyileştirmiştir. Bu yöntemlerle eğitilen modellerin büyük çoğunluğu tüm metriklerde 1.00 değerine ulaşarak sıfır hata ile sınıflandırma yapmıştır. Özellikle SSTK-SVM, SSNC-RF, SBS-AB ve SSVM-SVM modelleri bu doğrultuda öne çıkmıştır. Elde edilen bu sonuçlar, hem veri temsili hem de yeniden örnekleme stratejilerinin başarıda birlikte rol oynadığını göstermektedir.

Genel olarak sınıflandırıcılar arasında SVM, RF ve AdaBoost hemen her örnekleme yöntemiyle birlikte yüksek başarı sağlamıştır. Buna karşın, Naive Bayes, LDA ve bazı durumlarda KNN gibi sınıflayıcılar, özellikle karmaşık veri yapılarında zayıf performans göstermiştir. Ancak bazı örnekleme stratejileri, bu sınıflayıcıların performansını dahi belirgin şekilde yükseltmiştir. Örneğin, SROS-KNN modeli bile F1≈0.86 değerine ulaşarak bu yöntemin gücünü göstermiştir.

Sonuç olarak, CKSAAGP-GDPC-GTPC özelliklerinin, Siamese ağıyla dengelenmiş temsillerle ve hedefli örnekleme yöntemleriyle birlikte güçlü sınıflayıcılarla (özellikle SVM ve RF) kullanılması, iyon kanal veriseti üzerinde en yüksek doğruluk ve ayrım gücüne ulaşılmasını sağlamıştır. SMOTE-Tomek ve SVM birleşimi, F1=1 ve ROC=1 ile bu çalışmanın en başarılı modeli olmuş ve benzer biyolojik sınıflandırma problemleri için örnek teşkil edebilecek nitelikte bir yapı ortaya koymuştur.

Tablo 4. 17: IC veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.257862	0.706897	0.37788	0.901388	0.808444	**SCN-KNN**	0.323077	0.724138	0.446809	0.924032	0.828507
SCC-SVM	0.09221	1	0.16885	0.582907	0.782227	**SCN-SVM**	0.543689	0.965517	0.695652	0.964207	0.964833
SCC-DT	0.093904	0.982759	0.171429	0.597516	0.781616	**SCN-DT**	0.248485	0.706897	0.367713	0.897005	0.806156
SCC-RF	0.141119	1	0.247335	0.742148	0.86537	**SCN-RF**	0.794521	1	0.885496	0.989043	0.994279
SCC-NN	0.135279	0.87931	0.234483	0.756757	0.815323	**SCN-NN**	0.360656	0.37931	0.369748	0.945215	0.674781
SCC-AB	0.102473	1	0.185897	0.628926	0.806255	**SCN-AB**	0.508772	1	0.674419	0.959094	0.978642
SCC-NB	0.068913	0.775862	0.126582	0.546384	0.656047	**SCN-NB**	0.15812	0.637931	0.253425	0.84076	0.743832
SCC-QDA	0.191781	0.965517	0.32	0.82615	0.892751	**SCN-QDA**	0.268041	0.896552	0.412698	0.891892	0.894119
SCC-LDA	0.126649	0.827586	0.21968	0.750913	0.787554	**SCN-LDA**	0.20442	0.637931	0.309623	0.879474	0.764046
SCC-EEC	0.084465	0.965517	0.15534	0.55515	0.751256	**SCN-EEC**	0.147826	0.87931	0.253102	0.780131	0.827527
SCC-RSC	0.074581	0.844828	0.137063	0.549306	0.69053	**SCN-RSC**	0.15	0.413793	0.220183	0.875822	0.655028
SCC-BBC	0.17284	0.965517	0.293194	0.802776	0.880547	**SCN-BBC**	0.276382	0.948276	0.428016	0.892622	0.919218
SCC-BRFC	0.131148	0.689655	0.220386	0.79328	0.74376	**SCN-BRFC**	0.192053	0.5	0.277512	0.889701	0.703471
SEN-KNN	0.733333	0.568966	0.640777	0.972973	0.779906	**SRN-KNN**	0.521739	0.62069	0.566929	0.959825	0.797759
SEN-SVM	0.915254	0.931034	0.923077	0.993426	0.96361	**SRN-SVM**	0.746667	0.965517	0.842105	0.98466	0.975512
SEN-DT	0.827586	0.413793	0.551724	0.971512	0.70499	**SRN-DT**	0.733333	0.37931	0.5	0.96786	0.686604
SEN-RF	1	1	1	1	1	**SRN-RF**	0.935484	1	0.966667	0.997078	0.998474
SEN-NN	0	0	0	0.957633	0.5	**SRN-NN**	0	0	0	0.957633	0.5
SEN-AB	0.983051	1	0.991453	0.99927	0.999619	**SRN-AB**	0.95082	1	0.97479	0.997809	0.998856
SEN-NB	0.286885	0.603448	0.388889	0.919649	0.768543	**SRN-NB**	0.284553	0.603448	0.38674	0.918919	0.768162
SEN-QDA	0.32716	0.913793	0.481818	0.916728	0.915325	**SRN-QDA**	0.331288	0.931034	0.488688	0.917458	0.923946
SEN-LDA	0.428571	0.362069	0.392523	0.95252	0.670356	**SRN-LDA**	0.354839	0.37931	0.366667	0.944485	0.6744
SEN-EEC	0.172524	0.931034	0.291105	0.807889	0.866738	**SRN-EEC**	0.179739	0.948276	0.302198	0.814463	0.878409
SEN-RSC	0.263736	0.413793	0.322148	0.926224	0.681344	**SRN-RSC**	0.150327	0.396552	0.218009	0.879474	0.648695
SEN-BBC	0.337209	1	0.504348	0.916728	0.956522	**SRN-BBC**	0.365385	0.982759	0.53271	0.926954	0.953622
SEN-BRFC	0.230769	0.465517	0.308571	0.911614	0.698434	**SRN-BRFC**	0.252101	0.517241	0.338983	0.914536	0.724677
SSTK-KNN	0.58	1	0.734177	0.969321	0.983982	**SNC-KNN**	0.82	0.706897	0.759259	0.981008	0.850016
SSTK-SVM	1	1	1	1	1	**SNC-SVM**	0.947368	0.931034	0.93913	0.994887	0.964373
SSTK-DT	0.152648	0.844828	0.258575	0.794741	0.818676	**SNC-DT**	1	0.362069	0.531646	0.972973	0.681034
SSTK-RF	1	1	1	1	1	**SNC-RF**	1	1	1	1	1
SSTK-NN	0.329193	0.913793	0.484018	0.917458	0.915707	**SNC-NN**	0	0	0	0.957633	0.5
SSTK-AB	1	1	1	1	1	**SNC-AB**	1	1	1	1	1
SSTK-NB	0.117284	0.655172	0.198953	0.776479	0.718509	**SNC-NB**	0.285714	0.586207	0.384181	0.92038	0.760685
SSTK-QDA	0.34375	0.758621	0.473118	0.928415	0.847274	**SNC-QDA**	0.32716	0.913793	0.481818	0.916728	0.915325
SSTK-LDA	0.159864	0.810345	0.267045	0.811541	0.81097	**SNC-LDA**	0.488372	0.362069	0.415842	0.956903	0.672644
SSTK-EEC	0.18677	0.827586	0.304762	0.840029	0.834083	**SNC-EEC**	0.161905	0.87931	0.273458	0.802045	0.838969
SSTK-RSC	0.144828	0.724138	0.241379	0.807159	0.767485	**SNC-RSC**	0.169811	0.465517	0.248848	0.880935	0.682415
SSTK-BBC	1	1	1	1	1	**SNC-BBC**	0.387755	0.982759	0.556098	0.933528	0.957054
SSTK-BRFC	0.264151	0.482759	0.341463	0.92111	0.711631	**SNC-BRFC**	0.274194	0.586207	0.373626	0.916728	0.758779
SOS-KNN	0.885714	0.534483	0.666667	0.977356	0.765716	**SRUS-KNN**	0.103578	0.948276	0.186757	0.65011	0.792597
SOS-SVM	1	0.931034	0.964286	0.997078	0.965517	**SRUS-SVM**	0.211679	1	0.349398	0.842221	0.91762
SOS-DT	0.958333	0.396552	0.560976	0.973703	0.697894	**SRUS-DT**	0.148734	0.810345	0.251337	0.795471	0.802579
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.303665	1	0.465863	0.902849	0.949275
SOS-NN	0	0	0	0.957633	0.5	**SRUS-NN**	0.166052	0.775862	0.273556	0.82542	0.801737
SOS-AB	1	1	1	1	1	**SRUS-AB**	0.14321	1	0.25054	0.74653	0.867658
SOS-NB	0.280702	0.551724	0.372093	0.92111	0.744588	**SRUS-NB**	0.119891	0.758621	0.207059	0.753835	0.756122
SOS-QDA	0.344156	0.913793	0.5	0.922571	0.918376	**SRUS-QDA**	0.163142	0.931034	0.277635	0.794741	0.859873
SOS-LDA	0.488372	0.362069	0.415842	0.956903	0.672644	**SRUS-LDA**	0.136364	0.827586	0.234146	0.770636	0.797851
SOS-EEC	0.162939	0.87931	0.274933	0.803506	0.839731	**SRUS-EEC**	0.159875	0.87931	0.270557	0.799123	0.837443
SOS-RSC	0.149351	0.396552	0.216981	0.878744	0.648314	**SRUS-RSC**	0.09899	0.844828	0.177215	0.667641	0.752315
SOS-BBC	0.432836	1	0.604167	0.944485	0.971014	**SRUS-BBC**	0.241525	0.982759	0.387755	0.868517	0.923111
SOS-BRFC	0.245283	0.448276	0.317073	0.918188	0.693627	**SRUS-BRFC**	0.271845	0.482759	0.347826	0.923302	0.712775

Tablo 4. 18: IC veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.885714	0.534483	0.666667	0.977356	0.765716	**SROS-KNN**	0.84058	1	0.913386	0.991965	0.995805
STL-SVM	1	0.931034	0.964286	0.997078	0.965517	**SROS-SVM**	0.983051	1	0.991453	0.99927	0.999619
STL-DT	0.958333	0.396552	0.560976	0.973703	0.697894	**SROS-DT**	0.151099	0.948276	0.260664	0.772096	0.856289
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0	0	0	0.957633	0.5	**SROS-NN**	0.368056	0.913793	0.524752	0.929876	0.92219
STL-AB	1	1	1	1	1	**SROS-AB**	1	1	1	1	1
STL-NB	0.283186	0.551724	0.374269	0.921841	0.74497	**SROS-NB**	0.144144	0.827586	0.245524	0.784514	0.805097
STL-QDA	0.344156	0.913793	0.5	0.922571	0.918376	**SROS-QDA**	0.287356	0.862069	0.431034	0.903579	0.883742
STL-LDA	0.488372	0.362069	0.415842	0.956903	0.672644	**SROS-LDA**	0.15082	0.793103	0.253444	0.802045	0.797772
STL-EEC	0.163077	0.913793	0.276762	0.797663	0.853159	**SROS-EEC**	0.168498	0.793103	0.277946	0.82542	0.809977
STL-RSC	0.2875	0.396552	0.333333	0.932798	0.676537	**SROS-RSC**	0.133127	0.741379	0.225722	0.784514	0.763901
STL-BBC	0.341615	0.948276	0.502283	0.92038	0.933711	**SROS-BBC**	0.983051	1	0.991453	0.99927	0.999619
STL-BRFC	0.226891	0.465517	0.305085	0.910153	0.697671	**SROS-BRFC**	0.25	0.465517	0.325301	0.918188	0.701866
SSM-KNN	0.58	1	0.734177	0.969321	0.983982	**SSNC-KNN**	0.557692	1	0.716049	0.966399	0.982456
SSM-SVM	1	1	1	1	1	**SSNC-SVM**	1	1	1	1	1
SSM-DT	0.152648	0.844828	0.258575	0.794741	0.818676	**SSNC-DT**	0.151057	0.862069	0.257069	0.788897	0.823864
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.422414	0.844828	0.563218	0.944485	0.896861	**SSNC-NN**	0.401575	0.87931	0.551351	0.939372	0.91067
SSM-AB	1	1	1	1	1	**SSNC-AB**	1	1	1	1	1
SSM-NB	0.117284	0.655172	0.198953	0.776479	0.718509	**SSNC-NB**	0.118012	0.655172	0.2	0.77794	0.719272
SSM-QDA	0.34375	0.758621	0.473118	0.928415	0.847274	**SSNC-QDA**	0.357143	0.775862	0.48913	0.931337	0.857039
SSM-LDA	0.159864	0.810345	0.267045	0.811541	0.81097	**SSNC-LDA**	0.16041	0.810345	0.267806	0.812272	0.811351
SSM-EEC	0.186047	0.827586	0.303797	0.839299	0.833702	**SSNC-EEC**	0.182156	0.844828	0.299694	0.832725	0.838508
SSM-RSC	0.144828	0.724138	0.241379	0.807159	0.767485	**SSNC-RSC**	0.123393	0.827586	0.214765	0.743608	0.78374
SSM-BBC	1	1	1	1	1	**SSNC-BBC**	1	0.982759	0.991304	0.99927	0.991379
SSM-BRFC	0.269231	0.482759	0.345679	0.922571	0.712394	**SSNC-BRFC**	0.259259	0.482759	0.337349	0.919649	0.710868
SSN-KNN	0.882353	0.517241	0.652174	0.976625	0.757095	**SSVM-KNN**	0.56044	0.87931	0.684564	0.965668	0.9244
SSN-SVM	1	0.913793	0.954955	0.996348	0.956897	**SSVM-SVM**	1	1	1	1	1
SSN-DT	0.95	0.327586	0.487179	0.970782	0.663412	**SSVM-DT**	0.2103	0.844828	0.33677	0.859021	0.852238
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0	0	0	0.957633	0.5	**SSVM-NN**	0.291429	0.87931	0.437768	0.90431	0.892363
SSN-AB	1	1	1	1	1	**SSVM-AB**	1	1	1	1	1
SSN-NB	0	0	0	0.957633	0.5	**SSVM-NB**	0.099644	0.482759	0.165192	0.79328	0.644888
SSN-QDA	0	0	0	0.957633	0.5	**SSVM-QDA**	0.402174	0.637931	0.493333	0.944485	0.797989
SSN-LDA	1	0.017241	0.033898	0.958364	0.508621	**SSVM-LDA**	0.185185	0.775862	0.299003	0.845873	0.812416
SSN-EEC	0.409091	0.155172	0.225	0.954711	0.572628	**SSVM-EEC**	0.19457	0.741379	0.308244	0.859021	0.802803
SSN-RSC	0.666667	0.034483	0.065574	0.958364	0.51686	**SSVM-RSC**	0.13986	0.689655	0.232558	0.807159	0.751006
SSN-BBC	1	0.896552	0.945455	0.995617	0.948276	**SSVM-BBC**	1	0.965517	0.982456	0.998539	0.982759
SSN-BRFC	0	0	0	0.957633	0.5	**SSVM-BRFC**	0.171717	0.586207	0.265625	0.862673	0.730556
SBS-KNN	0.536842	0.87931	0.666667	0.962747	0.922874	**SBS-QDA**	0.371681	0.724138	0.491228	0.93645	0.83499
SBS-SVM	1	1	1	1	1	**SBS-LDA**	0.147157	0.758621	0.246499	0.803506	0.782056
SBS-DT	0.16263	0.810345	0.270893	0.815194	0.812876	**SBS-EEC**	0.186957	0.741379	0.298611	0.852447	0.79937
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.106024	0.758621	0.186047	0.718773	0.737815
SBS-NN	0.413223	0.862069	0.558659	0.942294	0.903956	**SBS-BBC**	1	0.982759	0.991304	0.99927	0.991379
SBS-AB	1	1	1	1	1	**SBS-BRFC**	0.195402	0.586207	0.293103	0.880205	0.739709
SBS-NB	0.111446	0.637931	0.189744	0.769175	0.706456						

4.4 G PROTEİNİNE BAĞLI RESEPTÖRLER VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI

Bu çalışma, GPCR veri setine odaklanmakta olup, amacı Siyam ağı tabanlı modelin biyolojik açıdan karmaşık verilere karşı gösterdiği esnekliği değerlendirmektir. Analizler, her biri protein özelliklerini farklı bir bakış açısıyla temsil eden dört farklı alt özellik kümesi kullanılarak

gerçekleştirilmiştir.Bu özellik kombinasyonları, üç aşamalı bir yapı olan Siyam ağı, yeniden örnekleme ve sınıflandırıcı süreciyle birleştirilmiş ve modelin performansı hem algoritmik açıdan hem de örnekleme stratejileri yönünden karşılaştırılmıştır. Analizler, modelin farklı örnekleme yöntemleri ve sınıflandırma algoritmalarıyla elde ettiği performansı kapsamlı biçimde değerlendirmek amacıyla gerçekleştirilmiştir.Modelin değerlendirilmesi; doğruluk (Accuracy), geri çağırma (Recall), F1 skoru ve ROC eğrisi altındaki alan (ROC-AUC) gibi metrikler kullanılarak yapılmıştır.

4.4.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, ilaç-protein etkileşimlerine ait GPCR veri seti üzerinde AAC ve DPC amino asit özellikleri kullanılarak temsil edilmiş ve ardından Siamese ağ ile dengelenmiştir. Daha sonra, farklı yeniden örnekleme yöntemleri ve makine öğrenmesi algoritmaları sınıflandırma performansını değerlendirmek için uygulanmıştır.

Tablo 4.19'ye ve 4.20'e göre , sadece SCC yöntemini kullanan modeller genel olarak zayıf performans sergilemiştir. Bu yöntem, Siamese ağ ile ilk dengelemeden sonra uyumsuz veya aykırı verileri kaldırırken veri yapısını dikkate almadan önemli miktarda faydalı bilgiyi de kaybetmektedir. Özellikle SCC-SVM, SCC-RF, SCC-AB ve SCC-QDA gibi modellerde F1 skorları anlamlı şekilde 0.2'nin altında olup, doğruluk değerleri çoğunlukla 0.55'in altındadır. Bu durum, geri bildirim içermeyen basit yeniden örnekleme yöntemlerinin, Siamese ağ ile birleşse bile, doğru etkileşimleri öğrenme kapasitesini ciddi şekilde sınırladığını göstermektedir.

Buna karşılık SCN yöntemini kullanan modeller çok daha iyi sonuçlar vermiştir. Karar sınırı yapısına göre verileri filtreleyen bu yöntem, model için önemli olan örneklerin eğitimde kalmasını sağlamaktadır. SCN-RF ve SCN-SVM modelleri sırasıyla 0.68 ve 0.50 F1 skoru ile akıllı filtrelemenin tahmin doğruluğunu önemli ölçüde artırdığını göstermiştir. SCN-BBC modeli de 0.59 F1 skoru ile örnek seçiminin model performansının istikrarına katkısını ortaya koymuştur.

Daha gelişmiş yeniden örnekleme yöntemlerinden SEN yaklaşımı çoğu durumda azınlık sınıfı örneklerini artırırken aynı zamanda gürültülü verileri temizleyerek dengeli ve güçlü modeller oluşturmuştur. SEN-RF modeli 0.88 F1 ve 0.996'ya yakın ROC değeri ile en başarılı

modellerden biridir. SEN-SVM ve SEN-AB modelleri de sırasıyla 0.69 ve 0.59 üzeri F1 skorları ile bu yöntemin etkinliğini desteklemektedir. Yapay veri üretimi ile akıllı temizleme kombinasyonu, model dengesini ve öğrenme başarısını artırmada kritik rol oynamaktadır.

Random Oversampling gibi daha basit yöntemler SRN modellerinde genellikle düşük performans göstermiştir. SRN-DT'nin 0.25, SRN-NN'nin ise 0 F1 skoru ile başarısız olması, sadece azınlık sınıfını çoğaltmanın veri yapısını ve karar sınırlarını dikkate almadan öğrenmeyi zorlaştırdığını ve hata oranını artırdığını ortaya koymaktadır. Bu durum ROC analizleriyle de doğrulanmış; birçok modelde ROC 0.7'nin altındadır.

SMOTE-Tomek olan karmaşık SSTK yönteminde yapay örnekleme ile sınır örneklerinin eş zamanlı temizlenmesi model performansını büyük ölçüde artırmıştır. SSTK-RF ve SSTK-SVM modelleri 0.99'un üzerinde F1 skoru ve neredeyse tam doğruluk ile en hassas modeller arasında yer almıştır. SSTK-BBC'nin 0.98 F1 skoru, hem veri dengesini sağladığını hem de veri yapısını koruduğunu göstermektedir. Bu başarı, örnekleme ve filtreleme işlemlerinin eş zamanlı yönetiminin önemini ortaya koymaktadır.

SOS ve SRUS gibi yöntemlerin performansları sınıflandırıcı tipine bağlı olarak değişkenlik göstermiştir. SOS-RF 0.96 F1 ile başarılı olurken SOS-NN 0 performans göstermiştir. Benzer şekilde SRUS-RF yüksek başarı sağlarken SRUS-NN düşük performans göstermiştir. Bu farklılıklar, sınıflandırıcı seçiminin yeniden örneklemeden sağlanan faydayı belirleyen önemli bir faktör olduğunu göstermektedir.

Undersampling ve azınlık sınıfını güçlendirme esaslı SROS yönteminde de sonuçlar çeşitlilik göstermiştir. SROS-RF ve SROS-BBC modelleri tam F1 skoru ile öne çıkarken, SROS-NN zayıf performans sergilemiştir. Bu durum yine yöntem başarısının sınıflandırıcıya bağlı olduğunu teyit etmektedir.

SMOTE, filtreleme ve örnek seçimi gibi çeşitli dengeleme tekniklerini birleştiren SSM ve SSNC gibi karma yöntemler çoğu modelde yüksek başarı elde etmiştir. SSM-RF 0.99 F1, SSNC-RF 0.94 F1 ve SSNC-SVM 0.87 F1 ile dengeli ve güçlü modeller sunmuştur.

En iyi model olarak belirtilen SSVM-SVM, 0.81 F1 skoru, %98 üzeri doğruluk ve 0.95 ROC değeri ile dengeli ve güvenilir bir performans sergilemiştir. Bazı modeller F1 açısından daha

yüksek skorlar elde etmiş olsa da, SSVM-SVM'nin tüm ölçütlerdeki tutarlılığı ve performans kararlılığı onu biyolojik uygulamalar için tercih edilen güçlü bir seçenek yapmaktadır. Bu modelin kullanımı, benzer biyoinformatik ve ilaç geliştirme alanlarda standart ve tekrarlanabilir bir çerçeve sunmaktadır.

Genel değerlendirme sonucunda, SSTK-RF**,** SSTK-SVM ve SSVM-SVM modelleri en yüksek başarıyı gösteren yöntemler arasında öne çıkmaktadır. Özellikle SSTK-RF ve SSTK-SVM, hem F1 skoru bakımından neredeyse mükemmel sonuçlar (F1 > 0.99) vermiş hem de ROC değerleriyle modellerin doğruluk ve kararlılığını kanıtlamıştır. Bununla birlikte, SSVM-SVM modeli, tüm performans ölçütlerinde istikrarlı sonuçlar sunarak dengeli bir genel başarı profili sergilemiştir. Bu nedenle, eğer hedef yüksek doğrulukla birlikte kararlı ve güvenilir bir model elde etmekse, SSVM-SVM en uygun tercih olarak değerlendirilebilir. Ancak en yüksek sınıflandırma başarısı ve duyarlılık istendiğinde SSTK-RF modelinin tercih edilmesi daha etkili sonuçlar sağlayacaktır.

Tablo 4. 19: GPCR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.2325	0.715385	0.350943	0.91881	0.820317	**SCN-KNN**	0.271438	0.637942	0.388	0.866184	0.755201
SCC-SVM	0.063291	1	0.119048	0.545905	0.765766	**SCN-SVM**	0.345112	0.942185	0.509204	0.889711	0.905367
SCC-DT	0.047884	0.861538	0.090725	0.470144	0.659647	**SCN-DT**	0.226307	0.726145	0.349051	0.794325	0.770118
SCC-RF	0.051444	1	0.097855	0.43427	0.708181	**SCN-RF**	0.562918	0.969513	0.683172	0.951228	0.944117
SCC-NN	0.091596	0.746154	0.163162	0.765164	0.75596	**SCN-NN**	0.287356	0.752331	0.417969	0.847321	0.791004
SCC-AB	0.045566	1	0.087161	0.357328	0.668493	**SCN-AB**	0.180765	0.709122	0.288109	0.750831	0.726118
SCC-NB	0.046146	0.7	0.086584	0.546849	0.621001	**SCN-NB**	0.191238	0.682994	0.307502	0.764225	0.734235
SCC-QDA	0.081004	0.869231	0.148197	0.693415	0.77854	**SCN-QDA**	0.159891	0.062195	0.074812	0.935004	0.526843
SCC-LDA	0.076573	0.776923	0.139406	0.705688	0.740178	**SCN-LDA**	0.208732	0.796413	0.331472	0.770189	0.774116
SCC-EEC	0.044157	0.892308	0.084149	0.404059	0.640456	**SCN-EEC**	0.169115	0.854203	0.285766	0.688473	0.770005
SCC-RSC	0.056347	0.669231	0.103943	0.645976	0.657235	**SCN-RSC**	0.139547	0.768923	0.235628	0.658925	0.705841
SCC-BBC	0.051406	0.984615	0.09771	0.442058	0.70475	**SCN-BBC**	0.423851	0.972008	0.590217	0.920136	0.925503
SCC-BRFC	0.053695	0.838462	0.100926	0.541657	0.685362	**SCN-BRFC**	0.271438	0.637942	0.388	0.866184	0.755201
SEN-KNN	0.432584	0.592308	0.5	0.963654	0.783858	**SRN-KNN**	0.573427	0.630769	0.600733	0.974274	0.807958
SEN-SVM	0.752212	0.653846	0.699588	0.982771	0.823514	**SRN-SVM**	0.567568	0.807692	0.666667	0.975218	0.894107
SEN-DT	0.246479	0.269231	0.257353	0.952325	0.621589	**SRN-DT**	0.689655	0.153846	0.251572	0.971914	0.575827
SEN-RF	0.797546	1	0.887372	0.992211	0.995982	**SRN-RF**	0.902778	1	0.948905	0.996696	0.998296
SEN-NN	0.274194	0.261538	0.267717	0.956101	0.619812	**SRN-NN**	0	0	0	0.969318	0.5
SEN-AB	0.419355	1	0.590909	0.957517	0.978086	**SRN-AB**	0.849673	1	0.918728	0.994572	0.9972
SEN-NB	0.122244	0.469231	0.193959	0.88034	0.681292	**SRN-NB**	0.14939	0.376923	0.213974	0.915034	0.654495
SEN-QDA	0.134615	0.807692	0.230769	0.834789	0.821669	**SRN-QDA**	0.097806	0.823077	0.174837	0.761624	0.791378
SEN-LDA	0.175182	0.184615	0.179775	0.948312	0.578551	**SRN-LDA**	0.257426	0.2	0.225108	0.957753	0.590869
SEN-EEC	0.067548	0.707692	0.123324	0.691291	0.699232	**SRN-EEC**	0.095194	0.792308	0.169967	0.762568	0.776967
SEN-RSC	0.076655	0.169231	0.105516	0.911966	0.552353	**SRN-RSC**	0.054054	0.723077	0.100589	0.603257	0.661271
SEN-BBC	0.229242	0.976923	0.371345	0.898513	0.936477	**SRN-BBC**	0.20268	0.930769	0.332875	0.885532	0.907435
SEN-BRFC	0.100962	0.646154	0.174636	0.812603	0.732013	**SRN-BRFC**	0.094854	0.723077	0.167707	0.779797	0.752335
SSTK-KNN	0.490566	1	0.658228	0.968138	0.983565	**SNC-KNN**	0.445652	0.630769	0.522293	0.964598	0.802967
SSTK-SVM	0.716667	0.992308	0.832258	0.987727	0.989945	**SNC-SVM**	0.460581	0.853846	0.598383	0.964834	0.911096
SSTK-DT	0.100971	0.8	0.17931	0.775313	0.787266	**SNC-DT**	0.513514	0.146154	0.227545	0.969554	0.570886
SSTK-RF	0.984848	1	0.992366	0.999528	0.999757	**SNC-RF**	0.860927	1	0.925267	0.995044	0.997443
SSTK-NN	0.194192	0.823077	0.314244	0.889781	0.857484	**SNC-NN**	0	0	0	0.969318	0.5
SSTK-AB	0.789809	0.953846	0.864111	0.990795	0.972906	**SNC-AB**	0.741176	0.969231	0.84	0.988671	0.979259
SSTK-NB	0.061627	0.576923	0.111359	0.717489	5r	**SNC-NB**	0.145299	0.392308	0.212058	0.91055	0.659631
SSTK-QDA	0.102294	0.823077	0.181973	0.772953	0.797221	**SNC-QDA**	0.097273	0.823077	0.173984	0.760208	0.790647
SSTK-LDA	0.088314	0.761538	0.158273	0.751475	0.756348	**SNC-LDA**	0.176136	0.238462	0.202614	0.942412	0.601578
SSTK-EEC	0.090739	0.746154	0.161802	0.762804	0.754742	**SNC-EEC**	0.091073	0.776923	0.163035	0.755251	0.765744
SSTK-RSC	0.074978	0.661538	0.134691	0.739202	0.7016	**SNC-RSC**	0.054054	0.723077	0.100589	0.603257	0.661271
SSTK-BBC	0.992126	0.969231	0.980545	0.99882	0.984494	**SNC-BBC**	0.219761	0.992308	0.359833	0.891669	0.940395
SSTK-BRFC	0.092885	0.723077	0.164623	0.774841	0.749778	**SNC-BRFC**	0.099284	0.746154	0.175248	0.784517	0.765943
SOS-KNN	0.643939	0.653846	0.648855	0.978287	0.821201	**SRUS-KNN**	0.845238	0.546154	0.663551	0.983007	0.771494
SOS-SVM	0.69863	0.784615	0.73913	0.983007	0.886951	**SRUS-SVM**	0.822917	0.607692	0.699115	0.983951	0.801777
SOS-DT	0.789474	0.115385	0.201342	0.971914	0.557205	**SRUS-DT**	0.772727	0.130769	0.223684	0.97215	0.564776
SOS-RF	0.934783	0.992308	0.962687	0.99764	0.995058	**SRUS-RF**	0.962963	1	0.981132	0.99882	0.999391
SOS-NN	0	0	0	0.969318	0.5	**SRUS-NN**	0	0	0	0.969318	0.5
SOS-AB	0.883562	0.992308	0.934783	0.995752	0.994084	**SRUS-AB**	0.953488	0.946154	0.949807	0.996932	0.972346
SOS-NB	0.146965	0.353846	0.207675	0.917158	0.644418	**SRUS-NB**	0.15035	0.330769	0.206731	0.922115	0.635801
SOS-QDA	0.09854	0.830769	0.176183	0.761624	0.795102	**SRUS-QDA**	0.099539	0.830769	0.177778	0.76422	0.796441
SOS-LDA	0.253012	0.161538	0.197183	0.959641	0.573221	**SRUS-LDA**	0.333333	0.130769	0.187845	0.965306	0.561245
SOS-EEC	0.095325	0.8	0.170352	0.760916	0.779839	**SRUS-EEC**	0.091659	0.769231	0.1638	0.759028	0.763968
SOS-RSC	0.054054	0.723077	0.100589	0.603257	0.661271	**SRUS-RSC**	0.052839	0.723077	0.098481	0.593816	0.656401
SOS-BBC	0.215	0.992308	0.353425	0.8886	0.938813	**SRUS-BBC**	0.236794	1	0.382916	0.901109	0.94899
SOS-BRFC	0.098501	0.707692	0.172932	0.792306	0.751338	**SRUS-BRFC**	0.09816	0.738462	0.173285	0.783809	0.761853

Tablo 4. 20: GPCR veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.091267	0.892308	0.165596	0.724097	0.80554	**SROS-KNN**	0.628019	1	0.771513	0.981827	0.990626
STL-SVM	0.131365	0.992308	0.232014	0.798442	0.892307	**SROS-SVM**	0.659898	1	0.795107	0.984187	0.991843
STL-DT	0.083532	0.807692	0.151406	0.722209	0.763598	**SROS-DT**	0.10351	0.884615	0.185334	0.761388	0.821051
STL-RF	0.139186	1	0.244361	0.810243	0.902118	**SROS-RF**	0.992366	1	0.996169	0.999764	0.999878
STL-NN	0.070352	0.861538	0.130081	0.646448	0.750589	**SROS-NN**	0.166437	0.930769	0.28238	0.85485	0.891608
STL-AB	0.126336	1	0.224331	0.787822	0.890553	**SROS-AB**	0.490566	1	0.658228	0.968138	0.983565
STL-NB	0.089447	0.684615	0.158222	0.776493	0.732008	**SROS-NB**	0.09681	0.676923	0.169394	0.796318	0.73851
STL-QDA	0.093411	0.861538	0.168548	0.739202	0.798434	**SROS-QDA**	0.096631	0.838462	0.173291	0.754543	0.795174
STL-LDA	0.081867	0.823077	0.148921	0.711352	0.765446	**SROS-LDA**	0.086538	0.761538	0.155416	0.746047	0.753547
STL-EEC	0.089059	0.807692	0.160428	0.740618	0.773094	**SROS-EEC**	0.095725	0.792308	0.170813	0.763984	0.777698
STL-RSC	0.074744	0.730769	0.135617	0.714185	0.722214	**SROS-RSC**	0.078652	0.753846	0.142442	0.721501	0.737162
STL-BBC	0.13125	0.969231	0.231193	0.802219	0.883081	**SROS-BBC**	0.992366	1	0.996169	0.999764	0.999878
STL-BRFC	0.098326	0.723077	0.173112	0.788058	0.756596	**SROS-BRFC**	0.091522	0.730769	0.162671	0.769176	0.750581
SSM-KNN	0.490566	1	0.658228	0.968138	0.983565	**SSNC-KNN**	0.198742	0.989524	0.332145	0.717682	0.846003
SSM-SVM	0.716667	0.992308	0.832258	0.987727	0.989945	**SSNC-SVM**	0.777839	1	0.871236	0.979562	0.985129
SSM-DT	0.100971	0.8	0.17931	0.775313	0.787266	**SSNC-DT**	0.298174	0.765392	0.43751	0.848395	0.818212
SSM-RF	0.984848	1	0.992366	0.999528	0.999757	**SSNC-RF**	0.933471	1	0.949101	0.992084	0.994208
SSM-NN	0.16944	0.861538	0.283186	0.866179	0.863932	**SSNC-NN**	0.395884	1	0.562843	0.883421	0.938774
SSM-AB	0.812081	0.930769	0.867384	0.991267	0.961976	**SSNC-AB**	0.213768	0.452136	0.307189	0.857236	0.659104
SSM-NB	0.061576	0.576923	0.111276	0.717253	0.649309	**SSNC-NB**	0.174509	0.581227	0.273856	0.779538	0.678947
SSM-QDA	0.102294	0.823077	0.181973	0.772953	0.797221	**SSNC-QDA**	0.519237	1	0.683113	0.935872	0.96355
SSM-LDA	0.088314	0.761538	0.158273	0.751475	0.756348	**SSNC-LDA**	0.276493	0.941303	0.45591	0.832674	0.886435
SSM-EEC	0.087074	0.746154	0.155949	0.752183	0.749264	**SSNC-EEC**	0.243159	0.763487	0.375208	0.811245	0.798226
SSM-RSC	0.074978	0.661538	0.134691	0.739202	0.7016	**SSNC-RSC**	0.265801	0.623471	0.372618	0.853041	0.737892
SSM-BBC	0.984848	1	0.992366	0.999528	0.999757	**SSNC-BBC**	0.866115	0.963888	0.908771	0.985003	0.968137
SSM-BRFC	0.09252	0.723077	0.164049	0.773897	0.749291	**SSNC-BRFC**	0.198742	0.989524	0.332145	0.717682	0.846003
SSN-KNN	0.885714	0.476923	0.62	0.982063	0.737488	**SSVM-KNN**	0.542056	0.892308	0.674419	0.973566	0.934223
SSN-SVM	1	0.507692	0.673469	0.984895	0.753846	**SSVM-SVM**	0.722892	0.923077	0.810811	0.986783	0.955938
SSN-DT	0.764706	0.1	0.176871	0.971442	0.549513	**SSVM-DT**	0.12605	0.692308	0.21327	0.843285	0.770186
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0	0	0	0.969318	0.5	**SSVM-NN**	0.174242	0.707692	0.279635	0.888128	0.800766
SSN-AB	0.990385	0.792308	0.880342	0.993392	0.896032	**SSVM-AB**	0.766423	0.807692	0.786517	0.986547	0.89995
SSN-NB	0	0	0	0.969318	0.5	**SSVM-NB**	0.094231	0.376923	0.150769	0.869719	0.63112
SSN-QDA	0	0	0	0.969318	0.5	**SSVM-QDA**	0.119048	0.730769	0.204741	0.82582	0.779799
SSN-LDA	0	0	0	0.969318	0.5	**SSVM-LDA**	0.112299	0.646154	0.191344	0.832429	0.742239
SSN-EEC	0.6	0.023077	0.044444	0.969554	0.511295	**SSVM-EEC**	0.109233	0.646154	0.186874	0.827472	0.739683
SSN-RSC	0.5	0.015385	0.029851	0.969318	0.507449	**SSVM-RSC**	0.094017	0.507692	0.158654	0.834789	0.676417
SSN-BBC	1	0.830769	0.907563	0.994808	0.915385	**SSVM-BBC**	1	0.953846	0.976378	0.998584	0.976923
SSN-BRFC	0	0	0	0.969318	0.5	**SSVM-BRFC**	0.116456	0.353846	0.175238	0.897805	0.634435
SBS-KNN	0.53211	0.892308	0.666667	0.972622	0.933736	**SBS-QDA**	0.104803	0.738462	0.183556	0.798442	0.769401
SBS-SVM	0.728916	0.930769	0.817568	0.987255	0.959906	**SBS-LDA**	0.079291	0.653846	0.141431	0.756431	0.706762
SBS-DT	0.120383	0.676923	0.204413	0.838329	0.760181	**SBS-EEC**	0.08055	0.630769	0.142857	0.76776	0.701433
SBS-RF	0.977444	1	0.988593	0.999292	0.999635	**SBS-RSC**	0.072411	0.607692	0.129402	0.749115	0.680642
SBS-NN	0.15969	0.792308	0.265806	0.865707	0.830169	**SBS-BBC**	0.992	0.953846	0.972549	0.998348	0.976801
SBS-AB	0.834586	0.853846	0.844106	0.990323	0.924245	**SBS-BRFC**	0.079818	0.538462	0.139027	0.795374	0.670984
SBS-NB	0.0625	0.5	0.111111	0.754543	0.6313						

4.4.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, GAAC-DDE özellikleri kullanılarak GPCR veri seti üzerinde Siyam ağı tabanlı modellerin performansı incelenmiştir. Veriler öncelikle Siyam ağı ile dengelenmiş, ardından çeşitli yeniden örnekleme yöntemleri uygulanarak farklı makine öğrenmesi algoritmaları ile değerlendirilmiştir. Tablo 4.21'dan , 4.22'den elde edilen sonuçlar, hangi yeniden örnekleme yöntemi ile hangi sınıflandırıcı kombinasyonunun en iyi performansı sunduğunu belirlemeyi amaçlamaktadır.

SCC yöntemi, verilerin yapısını korumayı hedefleyerek kümeleme yoluyla temsilci örnekler seçmektedir. Ancak bu yöntemin pratikte zayıf performans gösterdiği görülmüştür. Özellikle SCC-RF modeli F1 skoru 0.097, ROC değeri 0.706 ve doğruluk oranı 0.431 ile oldukça zayıf sonuçlar üretmiştir. SCC-AB modeli de benzer şekilde F1 skoru 0.090, ROC değeri 0.682 ve doğruluk oranı 0.384 ile başarısız olmuştur. SCC-KNN modeli, diğerlerine göre nispeten daha iyi bir F1 skoru olan 0.340 ile öne çıksa da bu değer de yetersizdir. Bu durum, küme merkezleri seçilirken eğitim verilerinin aşırı azalması ve çeşitliliğin kaybı ile açıklanabilir. Sonuç olarak SCC yöntemi, veri yapısını dikkate alarak seçtiği örneklerle bu çalışmada yeterli başarıyı sağlayamamıştır.

SCN yöntemi, örnekleri karar sınırına göre filtreleyerek veri dengesini sağlamayı hedeflemektedir. Bu yaklaşımın oldukça başarılı olduğu görülmüştür. SCN-RF modeli F1 skoru 0.871, ROC değeri 0.994 ve doğruluk oranı 0.989 ile çalışmadaki en yüksek değerlere ulaşmıştır. Benzer şekilde SCN-SVM modeli de F1 skoru 0.807 ve ROC değeri 0.980 ile oldukça başarılıdır. SCN-AB ve SCN-QDA modelleri sırasıyla 0.783 ve 0.844 F1 skorları ile güçlü sonuçlar üretmiştir. Bu bulgular, örneklerin doğrudan silinmesi yerine karar sınırı etrafında hedefe yönelik filtreleme yapılmasının model başarısını artırdığını göstermektedir.

SEN yöntemi, azınlık sınıfına ait örnekleri yapay olarak artırırken aynı zamanda veri içindeki gürültüyü azaltmayı amaçlamaktadır. SEN-RF modeli 0.902 F1 skoru, 0.993 doğruluk ve 0.992 ROC değeri ile neredeyse ideal bir performans sergilemiştir. SEN-AB modeli de 0.632 F1 skoru ve 0.974 ROC değeri ile tatmin edici sonuçlar vermiştir. Ancak SEN-NN modeli yalnızca 0.030 F1 skoru ile zayıf kalmış, bu da yöntemin tüm sınıflandırıcılarla uyumlu çalışmadığını göstermiştir.

SRN yöntemi, azınlık sınıfını rastgele artırarak dengeleme yapmaktadır, ancak oldukça değişken sonuçlar üretmiştir. SRN-SVM modeli 0.685 F1 skoru ve 0.902 ROC değeri ile kabul edilebilir bir performans gösterirken, SRN-NN modeli tüm metriklerde sıfır değer üretmiş, tamamen başarısız olmuştur. Bu, yöntemin sınıflandırıcıya fazlasıyla bağımlı olduğunu ve istikrarsızlık taşıdığını göstermektedir.

Sınırda yer alan örnekleri akıllıca temizlemeyi amaçlayan SMOTE-Tomek yöntemi, birçok sınıflandırıcıda güçlü sonuçlar üretmiştir. SSTK-RF modeli 0.996 F1 skoru, 0.999 doğruluk ve 0.999 ROC değeri ile olağanüstü performans sergilemiştir. SSTK-SVM modeli 0.843 F1 skoru

ile dengeli bir başarı elde etmiştir. SSTK-BBC modeli ise her üç metrikte 0.984 ile istikrarlı bir performans göstermiştir. Bu yöntem, karar sınırına yakın örnekleri ortadan kaldırarak sınıflar arasındaki ayrımı güçlendirmiştir.

SMOTE-Tomek yönteminin daha gelişmiş bir versiyonu olan SNC tabanlı modeller olağanüstü başarı göstermiştir. SNC-RF modeli 0.931 F1 skoru, 0.995 doğruluk ve 0.997 ROC değeri ile çalışmanın en güçlü modellerinden biridir. SNC-AB modeli 0.863 F1 skoru ve 0.995 ROC değeri ile dikkat çekmiştir. SNC-SVM modeli ise 0.605 F1 skoru ve 0.915 ROC değeri ile dengeli bir performans göstermiştir. Bu sonuçlar, SNC yönteminin hem azınlık hem çoğunluk sınıfı üzerinde etkili bir örnekleme stratejisi sunduğunu göstermektedir.

Daha basit yöntemler olan SOS ve SRUS, bazı sınıflandırıcılarda yüksek performans sunarken bazılarında başarısız olmuştur. Örneğin, SOS-RF modeli 0.970 F1 skoru ve 0.999 ROC değeri ile mükemmele yakın sonuçlar üretmiştir. Benzer şekilde SRUS-RF modeli 0.977 F1 skoru ve 0.991 ROC değeri ile dikkat çekmiştir. Ancak SOS-NN ve SRUS-NN modellerinde tüm metrikler sıfır olmuş, model tamamen başarısız olmuştur. Bu durum, bu yöntemlerin sınıflandırıcıya duyarlı olduğunu ve doğru algoritma seçilmediğinde performansın ciddi biçimde düşebileceğini göstermektedir.

SSM ve SSNC gibi SMOTE ile gürültü temizleme yöntemlerini birleştiren yaklaşımlar ise çalışmanın en dikkat çekici sonuçlarını üretmiştir. SSNC-RF modeli 0.997 F1 skoru, 0.998 doğruluk ve 0.999 ROC değeri ile çalışmanın en başarılı modeli olmuştur. SSNC-SVM modeli 0.985 F1 skoru ve 0.998 ROC değeri ile benzer şekilde yüksek performans sunmuştur. Ayrıca SSNC-AB modeli neredeyse tüm metriklerde 0.999 puanla mükemmel sonuçlar vermiştir. Bu modeller, hem dengesiz veriye karşı dirençli olmuş hem de aşırı öğrenmeyi önlemiştir.

Genel olarak değerlendirildiğinde, SSTK ve SNC yöntemleri ile RF ve SVM sınıflandırıcılarının kombinasyonları en yüksek başarıyı sunmuştur. Bu modeller, istenmeyen örnekleri etkili biçimde ortadan kaldırırken veri yapısını korumayı başarmıştır. Örneğin, SSTK-RF modeli 0.996 F1 skoru, 0.999 doğruluk ve 0.999 ROC değeriyle; SNC-RF modeli 0.931 F1 skoru, 0.995 doğruluk ve 0.997 ROC değeriyle; SSNC-RF modeli ise 0.997 F1 skoru, 0.998 doğruluk ve 0.999 ROC değeriyle öne çıkmıştır. Buna karşılık, SCC ve SRN gibi basit yöntemler veri yapısını yeterince temsil edememiş ve sınıflandırıcıya bağımlılıkları nedeniyle değişken ve düşük performans sergilemiştir. Sonuç olarak, dengesiz veri setleri ile çalışırken

hedef odaklı, gelişmiş yeniden örnekleme yöntemlerinin tercih edilmesi, makine öğrenmesi modellerinin başarısını önemli ölçüde artırmaktadır.

Tablo 4. 21: GPCR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.225564	0.692308	0.340265	0.91763	0.808535	**SCN-KNN**	0.276321	0.684913	0.387451	0.929847	0.811276
SCC-SVM	0.060018	1	0.11324	0.519471	0.752131	**SCN-SVM**	0.695238	0.972188	0.807113	0.980491	0.980157
SCC-DT	0.0534	0.9	0.100819	0.507435	0.697504	**SCN-DT**	0.628314	0.480026	0.556241	0.968381	0.735642
SCC-RF	0.051201	1	0.097415	0.431437	0.70672	**SCN-RF**	0.782145	0.991827	0.871932	0.989447	0.994501
SCC-NN	0.09901	0.769231	0.175439	0.778145	0.773829	**SCN-NN**	0.008349	0.015672	0.006102	0.958629	0.505481
SCC-AB	0.047497	1	0.090687	0.384706	0.682615	**SCN-AB**	0.674218	0.992731	0.783109	0.97825	0.988321
SCC-NB	0.041326	0.7	0.078045	0.492565	0.593	**SCN-NB**	0.193556	0.533821	0.263712	0.895318	0.72695
SCC-QDA	0.082221	0.9	0.150676	0.688695	0.791003	**SCN-QDA**	0.761494	0.94321	0.844003	0.986719	0.969203
SCC-LDA	0.082137	0.792308	0.148844	0.721973	0.756027	**SCN-LDA**	0.304711	0.421088	0.359275	0.94861	0.692187
SCC-EEC	0.04481	0.9	0.08537	0.408308	0.646372	**SCN-EEC**	0.180493	0.891302	0.27638	0.848121	0.875024
SCC-RSC	0.059001	0.753846	0.109436	0.623554	0.686638	**SCN-RSC**	0.207821	0.353019	0.261073	0.921165	0.651317
SCC-BBC	0.07416	0.984615	0.137931	0.622374	0.797762	**SCN-BBC**	0.348391	0.970424	0.517668	0.936015	0.95138
SCC-BRFC	0.054591	0.846154	0.102564	0.545669	0.691156	**SCN-BRFC**	0.290637	0.467205	0.36289	0.943289	0.715294
SEN-KNN	0.411765	0.592308	0.485804	0.961529	0.782762	**SRN-KNN**	0.569444	0.630769	0.59854	0.974038	0.807837
SEN-SVM	0.729508	0.684615	0.706349	0.982535	0.83829	**SRN-SVM**	0.587912	0.823077	0.685897	0.97687	0.902408
SEN-DT	0.326316	0.238462	0.275556	0.961529	0.611439	**SRN-DT**	0.730769	0.146154	0.24359	0.97215	0.572225
SEN-RF	0.826923	0.992308	0.902098	0.993392	0.992867	**SRN-RF**	0.915493	1	0.955882	0.997168	0.998539
SEN-NN	0.666667	0.015385	0.030075	0.969554	0.507571	**SRN-NN**	0	0	0	0.969318	0.5
SEN-AB	0.465455	0.984615	0.632099	0.964834	0.974411	**SRN-AB**	0.837662	0.992308	0.908451	0.993864	0.99311
SEN-NB	0.131579	0.538462	0.21148	0.8768	0.712985	**SRN-NB**	0.138425	0.446154	0.211293	0.897805	0.679128
SEN-QDA	0.124703	0.807692	0.216049	0.820156	0.814121	**SRN-QDA**	0.101983	0.830769	0.181665	0.770356	0.799607
SEN-LDA	0.137255	0.161538	0.14841	0.94312	0.564699	**SRN-LDA**	0.218182	0.184615	0.2	0.954685	0.581838
SEN-EEC	0.073495	0.723077	0.133428	0.711824	0.717273	**SRN-EEC**	0.09294	0.8	0.166533	0.754307	0.77643
SEN-RSC	0.058683	0.376923	0.101554	0.795374	0.592771	**SRN-RSC**	0.072233	0.592308	0.128763	0.754071	0.67575
SEN-BBC	0.243665	0.961538	0.388802	0.907246	0.933533	**SRN-BBC**	0.208197	0.976923	0.343243	0.885296	0.929659
SEN-BRFC	0.115079	0.669231	0.196388	0.831957	0.753169	**SRN-BRFC**	0.089439	0.723077	0.159187	0.765636	0.74503
SSTK-KNN	0.485075	1	0.653266	0.96743	0.983199	**SNC-KNN**	0.453039	0.630769	0.527331	0.965306	0.803332
SSTK-SVM	0.732955	0.992308	0.843137	0.988671	0.990432	**SNC-SVM**	0.466667	0.861538	0.605405	0.965542	0.915186
SSTK-DT	0.107	0.823077	0.189381	0.783809	0.802822	**SNC-DT**	0.585366	0.184615	0.280702	0.97097	0.590238
SSTK-RF	0.992366	1	0.996169	0.999764	0.999878	**SNC-RF**	0.872483	1	0.9319	0.995516	0.997687
SSTK-NN	0.179868	0.838462	0.296196	0.877744	0.858724	**SNC-NN**	0	0	0	0.969318	0.5
SSTK-AB	0.657609	0.930769	0.770701	0.983007	0.957715	**SNC-AB**	0.760234	1	0.863787	0.990323	0.995009
SSTK-NB	0.068734	0.630769	0.123961	0.726457	0.680128	**SNC-NB**	0.138322	0.469231	0.21366	0.894029	0.688353
SSTK-QDA	0.107972	0.823077	0.190901	0.785933	0.803917	**SNC-QDA**	0.099158	0.815385	0.176814	0.767052	0.790453
SSTK-LDA	0.090411	0.761538	0.161633	0.757612	0.759513	**SNC-LDA**	0.180723	0.230769	0.202703	0.9443	0.598828
SSTK-EEC	0.085182	0.738462	0.152745	0.748643	0.743713	**SNC-EEC**	0.089065	0.776923	0.15981	0.749351	0.762701
SSTK-RSC	0.074131	0.738462	0.134737	0.708992	0.72326	**SNC-RSC**	0.073284	0.607692	0.130795	0.752183	0.682225
SSTK-BBC	0.984615	0.984615	0.984615	0.999056	0.992064	**SNC-BBC**	0.186944	0.969231	0.313433	0.869719	0.9179
SSTK-BRFC	0.086331	0.738462	0.154589	0.752183	0.74554	**SNC-BRFC**	0.087922	0.761538	0.157643	0.750295	0.755739
SOS-KNN	0.643411	0.638462	0.640927	0.978051	0.813631	**SRUS-KNN**	0.858824	0.561538	0.67907	0.983715	0.779308
SOS-SVM	0.716312	0.776923	0.745387	0.983715	0.883592	**SRUS-SVM**	0.848485	0.646154	0.733624	0.985603	0.821251
SOS-DT	0.785714	0.169231	0.278481	0.973094	0.583885	**SRUS-DT**	0.851852	0.176923	0.292994	0.973802	0.587975
SOS-RF	0.942029	1	0.970149	0.998112	0.999026	**SRUS-RF**	0.969697	0.984615	0.977099	0.998584	0.991821
SOS-NN	0	0	0	0.969318	0.5	**SRUS-NN**	0	0	0	0.969318	0.5
SOS-AB	0.873239	0.953846	0.911765	0.994336	0.974732	**SRUS-AB**	0.945736	0.938462	0.942085	0.99646	0.968379
SOS-NB	0.137592	0.430769	0.208566	0.899693	0.672653	**SRUS-NB**	0.144	0.415385	0.213861	0.906302	0.668613
SOS-QDA	0.10237	0.830769	0.182278	0.7713	0.800094	**SRUS-QDA**	0.103746	0.830769	0.184458	0.774605	0.801798
SOS-LDA	0.253012	0.161538	0.197183	0.959641	0.573221	**SRUS-LDA**	0.307692	0.123077	0.175824	0.964598	0.557156
SOS-EEC	0.093382	0.792308	0.167072	0.757612	0.77441	**SRUS-EEC**	0.094582	0.792308	0.168991	0.760916	0.776115
SOS-RSC	0.078431	0.184615	0.110092	0.908426	0.557976	**SRUS-RSC**	0.07947	0.184615	0.111111	0.90937	0.558463
SOS-BBC	0.208955	0.969231	0.343793	0.886476	0.926544	**SRUS-BBC**	0.219298	0.961538	0.357143	0.893793	0.926593
SOS-BRFC	0.085448	0.776923	0.153963	0.738022	0.756857	**SRUS-BRFC**	0.089091	0.753846	0.15935	0.755959	0.754936

Tablo 4. 22: GPCR veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.089147	0.884615	0.161972	0.719141	0.799259	**SROS-KNN**	0.62201	1	0.766962	0.981355	0.990382
STL-SVM	0.129741	1	0.229682	0.794194	0.89384	**SROS-SVM**	0.691489	1	0.81761	0.986311	0.992939
STL-DT	0.112169	0.815385	0.197209	0.796318	0.80555	**SROS-DT**	0.120398	0.930769	0.213216	0.789238	0.857763
STL-RF	0.137421	1	0.241636	0.807411	0.900657	**SROS-RF**	0.992366	1	0.996169	0.999764	0.999878
STL-NN	0.088638	0.846154	0.160467	0.728346	0.785385	**SROS-NN**	0.191781	0.861538	0.313725	0.884352	0.873306
STL-AB	0.125	1	0.222222	0.785225	0.889214	**SROS-AB**	0.631068	1	0.77381	0.982063	0.990748
STL-NB	0.090997	0.723077	0.161651	0.769884	0.747221	**SROS-NB**	0.099119	0.692308	0.17341	0.797498	0.746568
STL-QDA	0.095118	0.869231	0.171472	0.74227	0.803741	**SROS-QDA**	0.099362	0.838462	0.177669	0.76186	0.798948
STL-LDA	0.082224	0.807692	0.149254	0.717489	0.761163	**SROS-LDA**	0.090664	0.776923	0.162379	0.754071	0.765136
STL-EEC	0.087993	0.823077	0.15899	0.73283	0.776525	**SROS-EEC**	0.09228	0.8	0.165473	0.752419	0.775457
STL-RSC	0.07717	0.738462	0.139738	0.721029	0.729469	**SROS-RSC**	0.081193	0.753846	0.146597	0.730706	0.74191
STL-BBC	0.142369	0.961538	0.248016	0.8211	0.889096	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.082126	0.784615	0.148688	0.724333	0.75352	**SROS-BRFC**	0.093067	0.753846	0.16568	0.767052	0.760658
SSM-KNN	0.485075	1	0.653266	0.96743	0.983199	**SSNC-KNN**	0.528641	0.991248	0.693275	0.964217	0.979812
SSM-SVM	0.732955	0.992308	0.843137	0.988671	0.990432	**SSNC-SVM**	0.973182	0.999725	0.985412	0.998503	0.998327
SSM-DT	0.107	0.823077	0.189381	0.783809	0.802822	**SSNC-DT**	0.279905	0.836589	0.422317	0.910271	0.865149
SSM-RF	0.984848	1	0.992366	0.999528	0.999757	**SSNC-RF**	0.996238	0.998761	0.997365	0.998971	0.999183
SSM-NN	0.168142	0.876923	0.282178	0.863111	0.869798	**SSNC-NN**	0.448912	0.91834	0.606892	0.953417	0.931102
SSM-AB	0.626866	0.969231	0.761329	0.981355	0.975485	**SSNC-AB**	0.999115	0.999953	0.999205	1	0.999889
SSM-NB	0.068734	0.630769	0.123961	0.726457	0.680128	**SSNC-NB**	0.14281	0.634809	0.235171	0.851024	0.732745
SSM-QDA	0.107972	0.823077	0.190901	0.785933	0.803917	**SSNC-QDA**	0.734279	0.802567	0.74793	0.981104	0.888046
SSM-LDA	0.090411	0.761538	0.161633	0.757612	0.759513	**SSNC-LDA**	0.164297	0.854703	0.274183	0.825493	0.831518
SSM-EEC	0.084444	0.730769	0.151394	0.748643	0.739989	**SSNC-EEC**	0.165824	0.785231	0.272924	0.839678	0.811503
SSM-RSC	0.074444	0.746154	0.13538	0.707576	0.726254	**SSNC-RSC**	0.110237	0.762891	0.190408	0.758091	0.757934
SSM-BBC	0.992248	0.984615	0.988417	0.999292	0.992186	**SSNC-BBC**	0.989727	0.998814	0.994762	0.998991	0.998873
SSM-BRFC	0.086081	0.723077	0.153846	0.755959	0.740039	**SSNC-BRFC**	0.326884	0.564201	0.412688	0.939173	0.749711
SSN-KNN	0.891892	0.507692	0.647059	0.983007	0.752872	**SSVM-KNN**	0.518519	0.861538	0.647399	0.971206	0.918108
SSN-SVM	0.986301	0.553846	0.70936	0.986075	0.776801	**SSVM-SVM**	0.751592	0.907692	0.8223	0.987963	0.949098
SSN-DT	0.736842	0.107692	0.187919	0.971442	0.553237	**SSVM-DT**	0.130258	0.738462	0.221453	0.840689	0.791193
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0	0	0	0.969318	0.5	**SSVM-NN**	0.174632	0.730769	0.281899	0.885768	0.810722
SSN-AB	0.990566	0.807692	0.889831	0.993864	0.903724	**SSVM-AB**	0.740506	0.9	0.8125	0.987255	0.945009
SSN-NB	0	0	0	0.969318	0.5	**SSVM-NB**	0.115312	0.469231	0.185129	0.873259	0.677639
SSN-QDA	0	0	0	0.969318	0.5	**SSVM-QDA**	0.122137	0.738462	0.209607	0.829124	0.785228
SSN-LDA	0	0	0	0.969318	0.5	**SSVM-LDA**	0.11336	0.646154	0.192882	0.834081	0.743092
SSN-EEC	0.25	0.015385	0.028986	0.968374	0.506962	**SSVM-EEC**	0.0978	0.615385	0.168776	0.814019	0.717846
SSN-RSC	0.142857	0.007692	0.014599	0.968138	0.503116	**SSVM-RSC**	0.087678	0.569231	0.151951	0.805051	0.690873
SSN-BBC	1	0.823077	0.902954	0.994572	0.911538	**SSVM-BBC**	1	0.930769	0.964143	0.997876	0.965385
SSN-BRFC	0	0	0	0.969318	0.5	**SSVM-BRFC**	0.117318	0.484615	0.188906	0.872315	0.684601
SBS-KNN	0.502242	0.861538	0.634561	0.969554	0.917256	**SBS-QDA**	0.110624	0.776923	0.193672	0.801511	0.789606
SBS-SVM	0.759494	0.923077	0.833333	0.988671	0.956912	**SBS-LDA**	0.081496	0.653846	0.144928	0.763276	0.710293
SBS-DT	0.107471	0.630769	0.183651	0.827944	0.732477	**SBS-EEC**	0.076396	0.6	0.135534	0.765164	0.685196
SBS-RF	0.977444	1	0.988593	0.999292	0.999635	**SBS-RSC**	0.066905	0.576923	0.119904	0.740146	0.661118
SBS-NN	0.161392	0.784615	0.267717	0.868303	0.827784	**SBS-BBC**	0.991935	0.946154	0.968504	0.998112	0.972955
SBS-AB	0.848485	0.861538	0.854962	0.991031	0.928334	**SBS-BRFC**	0.088384	0.538462	0.151844	0.815435	0.681332
SBS-NB	0.069106	0.523077	0.122083	0.769176	0.650022						

4.4.3 CKSAAGP_GDPC_GTPC Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, CKSAAGP-GDPC-GTPC özellikleriyle dengelenmiş GPCR verisi üzerinde farklı yeniden örnekleme ve sınıflandırıcı kombinasyonlarının Siyam ağı tabanlı performansı değerlendirilmiştir.

Tablo 4.23 ve 4.24 verilerine göre, SCC yöntemiyle oluşturulan modeller genel olarak en düşük performansı göstermiştir. Özellikle SCC-RF modeli %100 recall değerine ulaşmasına rağmen yalnızca 0.03 precision ve 0.06 F1 skoru ile ciddi bir dengesizlik ortaya koymuştur. Benzer şekilde SCC-SVM modeli de %100 recall ile birlikte aynı düşük precision ve F1 skorlarını göstermiştir. SCC-NB (F1 = 0.06), SCC-AB (F1 = 0.07) ve SCC-QDA (F1 = 0.15) gibi diğer modeller de yeterli başarı sağlayamamıştır. Bu durum, SCC yönteminin karar sınırlarını dikkate almadan veri silmesi nedeniyle modelin öğrenme yeteneğini sınırladığını göstermektedir.

Buna karşın, SCN tabanlı yöntemler belirgin bir iyileşme sunmuştur. SCN-RF modeli 0.81 precision, %100 recall, 0.89 F1 ve 0.99 ROC ile neredeyse kusursuz performans göstermiştir. SCN-SVM modeli de 0.63 precision, 0.94 recall, 0.76 F1 ve 0.96 ROC ile oldukça başarılı sonuçlar vermiştir. Ancak bazı zayıf algoritmalar, örneğin SCN-AB (F1 = 0.17) ve SCN-RSC (F1 = 0.10), düşük performans sergilemiştir. Bu durum, model başarısının yalnızca yeniden örnekleme yöntemine değil, aynı zamanda sınıflandırıcı seçimine de bağlı olduğunu göstermektedir.

SEN ve SRN yöntemleri uygulandığında, güçlü sınıflandırıcıların öne çıktığı görülmüştür. SEN-RF ve SRN-RF modelleri tüm metriklerde 1.0 değerine ulaşarak mükemmel sonuçlar elde etmiştir. SEN-SVM (F1 = 0.84) ve SRN-SVM (F1 = 0.66) modelleri de güçlü başarılar göstermiştir. Öte yandan, SEN-NN ve SRN-NN modelleri F1 = 0 ile başarısız olmuş, sinir ağlarının bu tür veri yapılarında etkisiz kaldığını ortaya koymuştur.

SSTK yöntemi, genel olarak en iyi sonuçları üreten yöntem olmuştur. SSTK-RF modeli tüm metriklerde 1.0 skorlarına ulaşırken, SSTK-SVM modeli de 0.96 F1 ve 1.0 ROC ile yüksek başarı göstermiştir. SSTK-KNN modeli 0.57 F1 ile önceki KNN tabanlı yöntemlerden daha iyi performans sergilemiştir. Ancak SSTK-NN (F1 = 0.49) sınırlı bir gelişim göstermiştir.

SSM ve SSNC yöntemleri de benzer şekilde üstün performans ortaya koymuştur. Özellikle SSM-SVM ve SSNC-SVM modelleri 0.96 F1 ve 1.0 ROC değerleriyle öne çıkmıştır. SSM-RF

ve SSNC-RF modelleri de tüm metriklerde 1.0 skoruna ulaşarak güçlü performans sergilemiştir. Buna karşılık, SSM-NN (F1 = 0.49) ve SSNC-NN (F1 = 0.46) daha sınırlı bir başarı sunmuştur ve NN tabanlı modellerin genel zayıflığı yine gözlemlenmiştir.

SBS yöntemi de oldukça etkili sonuçlar sağlamıştır. SBS-RF modeli F1 = 1.0, SBS-SVM modeli ise F1 = 0.94 ile başarılı olmuştur. SBS-NN modeli, F1 = 0.35 ile önceki NN modellerine kıyasla daha iyi sonuçlar üretmiş, bu da yöntemin dengesiz veri kümesiyle baş etmede etkili olduğunu göstermektedir.

SOS yöntemi, güçlü algoritmalarla iyi çalışmaya devam etmiştir. SOS-RF ve SOS-SVM modelleri tüm metriklerde yüksek skorlar (F1 ve ROC = 1.0 veya yakın) elde etmiştir. Ancak zayıf algoritmalarla, örneğin SOS-AB (F1 = 0.10), SOS-NB (F1 = 0.21) ve özellikle SOS-NN (F1 = 0.00), belirgin performans düşüşleri yaşanmıştır.

SRUS yöntemi genel olarak en zayıf performanslardan birini sergilemiştir. Örneğin SRUS-RF yalnızca 0.16 F1 skoru ile ciddi bir gerileme göstermiştir. Bu durum, örneklerin rastgele silinmesinin sınıflar arası dengeyi bozarak önemli bilgi kaybına yol açtığını göstermektedir.

Son olarak, SSVM tabanlı SMOTE yöntemleri özellikle SSVM-SVM (F1 = 0.94) ve SSVM-RF (F1 = 1.0) modelleriyle yüksek doğruluk sağlamıştır. Ancak SSVM-NN modeli (F1 = 0.41) yine görece zayıf performans göstermiştir. Bu genel tablo, yeniden örnekleme yönteminin yanı sıra sınıflandırıcı tercihlerinin de model başarısında belirleyici olduğunu ortaya koymaktadır.

Analiz sonuçlarına göre, CKSAAGP_GDPC_GTPC özellikleriyle GPCR etkileşim tahmini için **SSTK** ile birlikte kullanılan SVM ve RF modelleri en başarılı kombinasyonlardır. **SSM**, **SSNC** ve **SBS** yöntemleri de yüksek doğruluklarıyla dikkate değerdir. Buna karşılık, sinir ağı tabanlı modeller yetersiz kalmış ve bu tür problemler için etkili bir çözüm sunamamıştır.

Tablo 4. 23: GPCR veri kümesi üzerinde CKSAAGP_GDPC_GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.086466	0.657143	0.152824	0.82267	0.741971	**SCN-KNN**	0.256098	0.6	0.358974	0.947844	0.778261
SCC-SVM	0.034551	1	0.066794	0.319889	0.651461	**SCN-SVM**	0.634615	0.942857	0.758621	0.985396	0.964657
SCC-DT	0.037694	0.971429	0.072572	0.395688	0.676377	**SCN-DT**	0.230769	0.428571	0.3	0.951321	0.696467
SCC-RF	0.030648	1	0.059473	0.230181	0.605488	**SCN-RF**	0.813953	1	0.897436	0.994437	0.997149
SCC-NN	0.095975	0.885714	0.173184	0.794159	0.838794	**SCN-NN**	0.444444	0.228571	0.301887	0.97427	0.610722
SCC-AB	0.039728	1	0.076419	0.411683	0.698503	**SCN-AB**	0.148148	0.228571	0.179775	0.949235	0.597892
SCC-NB	0.032359	0.885714	0.062437	0.352573	0.612494	**SCN-NB**	0.090909	0.4	0.148148	0.888039	0.650107
SCC-QDA	0.08642	1	0.159091	0.742698	0.86814	**SCN-QDA**	0.214286	0.942857	0.349206	0.914465	0.928307
SCC-LDA	0.070776	0.885714	0.131078	0.714186	0.797811	**SCN-LDA**	0.122222	0.314286	0.176	0.928373	0.628989
SCC-EEC	0.039036	0.971429	0.075055	0.417246	0.687425	**SCN-EEC**	0.066946	0.914286	0.124756	0.687761	0.798198
SCC-RSC	0.041481	0.8	0.078873	0.545202	0.669423	**SCN-RSC**	0.057034	0.428571	0.100671	0.81363	0.625904
SCC-BBC	0.036998	1	0.071356	0.366481	0.675339	**SCN-BBC**	0.165049	0.971429	0.282158	0.879694	0.924417
SCC-BRFC	0.028902	1	0.05618	0.182197	0.580898	**SCN-BRFC**	0.095436	0.657143	0.166667	0.840056	0.750881
SEN-KNN	0.516129	0.457143	0.484848	0.976356	0.723226	**SRN-KNN**	0.390244	0.457143	0.421053	0.969402	0.719662
SEN-SVM	0.729167	1	0.843373	0.99096	0.995367	**SRN-SVM**	0.492958	1	0.660377	0.974965	0.98717
SEN-DT	0.888889	0.228571	0.363636	0.980529	0.613929	**SRN-DT**	0.5	0.257143	0.339623	0.975661	0.625364
SEN-RF	1	1	1	1	1	**SRN-RF**	0.744681	1	0.853659	0.991655	0.995723
SEN-NN	0	0	0	0.975661	0.5	**SRN-NN**	0	0	0	0.975661	0.5
SEN-AB	0.615385	0.228571	0.333333	0.977747	0.612504	**SRN-AB**	0.25	0.171429	0.20339	0.967316	0.579299
SEN-NB	0.144578	0.342857	0.20339	0.934631	0.646126	**SRN-NB**	0.134831	0.342857	0.193548	0.930459	0.643987
SEN-QDA	0.226027	0.942857	0.364641	0.920028	0.931158	**SRN-QDA**	0.226027	0.942857	0.364641	0.920028	0.931158
SEN-LDA	0.263158	0.285714	0.273973	0.963143	0.632879	**SRN-LDA**	0.147541	0.257143	0.1875	0.945758	0.61004
SEN-EEC	0.075472	0.914286	0.139434	0.725313	0.817442	**SRN-EEC**	0.074725	0.971429	0.138776	0.706537	0.835679
SEN-RSC	0.049861	0.514286	0.090909	0.749652	0.634905	**SRN-RSC**	0.041841	0.285714	0.072993	0.823366	0.561246
SEN-BBC	0.191257	1	0.321101	0.897079	0.947256	**SRN-BBC**	0.142857	1	0.25	0.853964	0.92516
SEN-BRFC	0.107759	0.714286	0.187266	0.849096	0.783372	**SRN-BRFC**	0.102767	0.742857	0.180556	0.835883	0.79053
SSTK-KNN	0.397727	1	0.569106	0.963143	0.981112	**SNC-KNN**	0.703704	0.542857	0.612903	0.98331	0.768578
SSTK-SVM	0.921053	1	0.958904	0.997914	0.998931	**SNC-SVM**	0.813953	1	0.897436	0.994437	0.997149
SSTK-DT	0.1	0.714286	0.175439	0.836579	0.776958	**SNC-DT**	0.833333	0.285714	0.425532	0.981224	0.642144
SSTK-RF	1	1	1	1	1	**SNC-RF**	1	1	1	1	1
SSTK-NN	0.32381	0.971429	0.485714	0.94993	0.960411	**SNC-NN**	0	0	0	0.975661	0.5
SSTK-AB	0.127273	0.8	0.219608	0.861613	0.831575	**SNC-AB**	0.588235	0.285714	0.384615	0.977747	0.640362
SSTK-NB	0.079498	0.542857	0.138686	0.835883	0.693025	**SNC-NB**	0.144578	0.342857	0.20339	0.934631	0.646126
SSTK-QDA	0.259259	0.8	0.391608	0.939499	0.87149	**SNC-QDA**	0.230769	0.942857	0.370787	0.922114	0.932227
SSTK-LDA	0.078864	0.714286	0.142045	0.789986	0.75308	**SNC-LDA**	0.225	0.257143	0.24	0.960362	0.617524
SSTK-EEC	0.086643	0.685714	0.153846	0.816412	0.752693	**SNC-EEC**	0.085784	1	0.158014	0.740612	0.867071
SSTK-RSC	0.05029	0.742857	0.094203	0.652295	0.696446	**SNC-RSC**	0.06	0.428571	0.105263	0.82267	0.630537
SSTK-BBC	1	1	1	1	1	**SNC-BBC**	0.173267	1	0.295359	0.883866	0.940485
SSTK-BRFC	0.085324	0.714286	0.152439	0.806676	0.761633	**SNC-BRFC**	0.105691	0.742857	0.185053	0.840751	0.793025
SOS-KNN	0.695652	0.457143	0.551724	0.981919	0.726077	**SRUS-KNN**	0.056985	0.885714	0.107081	0.640473	0.760035
SOS-SVM	0.897436	1	0.945946	0.997218	0.998574	**SRUS-SVM**	0.088161	1	0.162037	0.748261	0.870991
SOS-DT	1	0.228571	0.372093	0.981224	0.614286	**SRUS-DT**	0.054366	0.942857	0.102804	0.599444	0.766867
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.091384	1	0.167464	0.757997	0.87598
SOS-NN	0	0	0	0.975661	0.5	**SRUS-NN**	0.06962	0.942857	0.129666	0.691933	0.814265
SOS-AB	1	0.057143	0.108108	0.977051	0.528571	**SRUS-AB**	0.080378	0.971429	0.148472	0.72879	0.847083
SOS-NB	0.155844	0.342857	0.214286	0.938804	0.648264	**SRUS-NB**	0.074303	0.685714	0.134078	0.784423	0.7363
SOS-QDA	0.230769	0.942857	0.370787	0.922114	0.932227	**SRUS-QDA**	0.107492	0.942857	0.192982	0.808067	0.873781
SOS-LDA	0.290323	0.257143	0.272727	0.96662	0.620731	**SRUS-LDA**	0.055446	0.8	0.103704	0.663421	0.730007
SOS-EEC	0.080729	0.885714	0.147971	0.751739	0.817055	**SRUS-EEC**	0.073903	0.914286	0.136752	0.719054	0.814235
SOS-RSC	0.041841	0.285714	0.072993	0.823366	0.561246	**SRUS-RSC**	0.060759	0.685714	0.111628	0.734353	0.71064
SOS-BBC	0.194286	0.971429	0.32381	0.901252	0.935465	**SRUS-BBC**	0.073826	0.942857	0.136929	0.710709	0.823888
SOS-BRFC	0.107477	0.657143	0.184739	0.858832	0.760503	**SRUS-BRFC**	0.073394	0.685714	0.132597	0.781641	0.734874

Tablo 4. 24 :GPCR veri kümesi üzerinde CKSAAGP_GDPC_GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.695652	0.457143	0.551724	0.981919	0.726077	**SROS-KNN**	0.636364	1	0.777778	0.986092	0.992872
STL-SVM	0.897436	1	0.945946	0.997218	0.998574	**SROS-SVM**	0.921053	1	0.958904	0.997914	0.998931
STL-DT	1	0.228571	0.372093	0.981224	0.614286	**SROS-DT**	0.121849	0.828571	0.212454	0.850487	0.839802
STL-RF	1	1	1	1	1	**SROS-RF**	1	1	1	1	1
STL-NN	0	0	0	0.975661	0.5	**SROS-NN**	0.263158	1	0.416667	0.93185	0.965075
STL-AB	0.555556	0.142857	0.227273	0.976356	0.570003	**SROS-AB**	0.170854	0.971429	0.290598	0.884562	0.926912
STL-NB	0.153846	0.342857	0.212389	0.938108	0.647908	**SROS-NB**	0.103586	0.742857	0.181818	0.837274	0.791243
STL-QDA	0.232394	0.942857	0.372881	0.922809	0.932583	**SROS-QDA**	0.217687	0.914286	0.351648	0.917942	0.916159
STL-LDA	0.290323	0.257143	0.272727	0.96662	0.620731	**SROS-LDA**	0.075	0.771429	0.136709	0.762865	0.76704
STL-EEC	0.080103	0.885714	0.146919	0.749652	0.815986	**SROS-EEC**	0.084932	0.885714	0.155	0.764951	0.823826
STL-RSC	0.054608	0.457143	0.097561	0.794159	0.629854	**SROS-RSC**	0.062201	0.742857	0.11479	0.72114	0.731728
STL-BBC	0.182292	1	0.30837	0.890821	0.944048	**SROS-BBC**	0.972222	1	0.985915	0.999305	0.999644
STL-BRFC	0.116667	0.8	0.203636	0.847705	0.824448	**SROS-BRFC**	0.100775	0.742857	0.177474	0.832406	0.788749
SSM-KNN	0.397727	1	0.569106	0.963143	0.981112	**SSNC-KNN**	0.416667	1	0.588235	0.965925	0.982537
SSM-SVM	0.921053	1	0.958904	0.997914	0.998931	**SSNC-SVM**	0.921053	1	0.958904	0.997914	0.998931
SSM-DT	0.1	0.714286	0.175439	0.836579	0.776958	**SSNC-DT**	0.129353	0.742857	0.220339	0.872045	0.809062
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.32381	0.971429	0.485714	0.94993	0.960411	**SSNC-NN**	0.298246	0.971429	0.456376	0.943672	0.957204
SSM-AB	0.127273	0.8	0.219608	0.861613	0.831575	**SSNC-AB**	0.119691	0.885714	0.210884	0.838665	0.861603
SSM-NB	0.079498	0.542857	0.138686	0.835883	0.693025	**SSNC-NB**	0.085837	0.571429	0.149254	0.841446	0.709806
SSM-QDA	0.259259	0.8	0.391608	0.939499	0.87149	**SSNC-QDA**	0.269231	0.8	0.402878	0.942281	0.872915
SSM-LDA	0.078864	0.714286	0.142045	0.789986	0.75308	**SSNC-LDA**	0.081633	0.685714	0.145897	0.80459	0.746635
SSM-EEC	0.086643	0.685714	0.153846	0.816412	0.752693	**SSNC-EEC**	0.091228	0.742857	0.1625	0.81363	0.779126
SSM-RSC	0.05029	0.742857	0.094203	0.652295	0.696446	**SSNC-RSC**	0.054795	0.685714	0.10148	0.704451	0.695316
SSM-BBC	1	1	1	1	1	**SSNC-BBC**	1	1	1	1	1
SSM-BRFC	0.085324	0.714286	0.152439	0.806676	0.761633	**SSNC-BRFC**	0.089844	0.657143	0.158076	0.829624	0.745535
SSN-KNN	0.714286	0.428571	0.535714	0.981919	0.712147	**SSVM-KNN**	0.467742	0.828571	0.597938	0.972879	0.902525
SSN-SVM	1	0.885714	0.939394	0.997218	0.942857	**SSVM-SVM**	0.918919	0.971429	0.944444	0.997218	0.984645
SSN-DT	1	0.085714	0.157895	0.977747	0.542857	**SSVM-DT**	0.157407	0.485714	0.237762	0.9242	0.710427
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0	0	0	0.975661	0.5	**SSVM-NN**	0.289157	0.685714	0.40678	0.951321	0.821831
SSN-AB	0.4	0.057143	0.1	0.974965	0.527502	**SSVM-AB**	0.175573	0.657143	0.277108	0.916551	0.790082
SSN-NB	0	0	0	0.975661	0.5	**SSVM-NB**	0.16129	0.142857	0.151515	0.961057	0.562163
SSN-QDA	0	0	0	0.975661	0.5	**SSVM-QDA**	0.293333	0.628571	0.4	0.954103	0.795398
SSN-LDA	0	0	0	0.975661	0.5	**SSVM-LDA**	0.121212	0.571429	0.2	0.888734	0.734039
SSN-EEC	0.333333	0.028571	0.052632	0.974965	0.513573	**SSVM-EEC**	0.134752	0.542857	0.215909	0.904033	0.72795
SSN-RSC	0	0	0	0.974965	0.499644	**SSVM-RSC**	0.092233	0.542857	0.157676	0.858832	0.704786
SSN-BBC	1	0.714286	0.833333	0.993046	0.857143	**SSVM-BBC**	1	0.942857	0.970588	0.998609	0.971429
SSN-BRFC	0	0	0	0.975661	0.5	**SSVM-BRFC**	0.35	0.2	0.254545	0.971488	0.595367
SBS-KNN	0.439394	0.828571	0.574257	0.970097	0.9011	**SBS-QDA**	0.227273	0.714286	0.344828	0.933936	0.826851
SBS-SVM	0.918919	0.971429	0.944444	0.997218	0.984645	**SBS-LDA**	0.101215	0.714286	0.177305	0.838665	0.778027
SBS-DT	0.09205	0.628571	0.160584	0.840056	0.736951	**SBS-EEC**	0.094262	0.657143	0.164875	0.837969	0.749812
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.058376	0.657143	0.107226	0.733658	0.696355
SBS-NN	0.232143	0.742857	0.353741	0.933936	0.84078	**SBS-BBC**	1	0.885714	0.939394	0.997218	0.942857
SBS-AB	0.160839	0.657143	0.258427	0.908206	0.785806	**SBS-BRFC**	0.095455	0.6	0.164706	0.851878	0.729081
SBS-NB	0.094444	0.485714	0.15814	0.874131	0.684767						

4.4.4 PSEAAC_PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

Tablo 4.25 ve 4.26 ’e göre, Siamese ağ mimarisine dayalı çeşitli makine öğrenimi modellerinin performansı ve farklı yeniden örnekleme yöntemleriyle kombinasyonları dikkatle değerlendirilmiştir. Beklendiği gibi, ClusterCentroids (SCC) yöntemini temel alan modeller, çoğu durumda en zayıf sonuçları kaydetmiştir. Bu durumun temel nedeni, karar sınırını dikkate almadan baskın sınıf örneklerinin rastgele kaldırılmasıdır; bu da modellerin önyargılı hale gelmesine ve aşırı tahmin yapmasına yol açmıştır. Örneğin, SCC-SVM modeli %100 geri çağırma oranı elde etmesine rağmen yalnızca 0,073’lük bir F1 puanı kaydetmiş ve bu da sınıf ayrımında ciddi bir zayıflığı işaret etmiştir. Benzer şekilde, SCC-RF ve SCC-LDA gibi diğer SCC modelleri de 0,2’nin altında kalan çok düşük F1 değerlerine sahip olmuştur.

Buna karşılık, ConNearNeigh (SCN) yöntemine dayalı modeller çok daha iyi performans göstermiş ve önemli bir başarı artışı sağlamıştır. Özellikle SCN-SVM modeli, yaklaşık %84 doğruluk, %92 geri çağırma ve 0,87 civarında F1 puanı ile en iyi sonuçlardan birini elde etmiştir. SCN-RF modeli ise tüm metriklerde neredeyse kusursuz bir performansla %100 seviyesine yaklaşmıştır. Bu sonuçlar, karar sınırına yakın örneklerin yeniden örneklemede kullanılmasının model doğruluğunu ve dengesini önemli ölçüde artırdığını göstermektedir. Ancak, bu gruptaki sinir ağı (SCN-NN) ve AdaBoost (SCN-AB) modelleri halen oldukça zayıf performans sergilemiş ve sıfıra yakın F1 puanları ile bu mimarilerin yeniden örneklenen verilerden öğrenme konusundaki yetersizliğini ortaya koymuştur.

Düzenlenmiş En Yakın Komşu (SEN) ve Tekrarlanan ENN (SRN) yöntemlerinin analizinde de SVM ve RF gibi güçlü modellerin performansı son derece etkileyici olmuştur. SEN-RF ve SRN-RF modelleri, tüm metriklerde (F1, doğruluk, geri çağırma ve ROC) mükemmel bir 1,0 skoru elde etmiştir. Bu durum, bu yeniden örnekleme yöntemlerinin güçlü algoritmalarla birleştiğinde olumlu bir sinerji yarattığını ortaya koymaktadır. Buna karşın, SEN-NN ve SRN-NN gibi sinir ağı tabanlı modeller yine mutlak başarısızlık göstermiş ve tüm metriklerde sıfır değer bildirmiştir. Bu sonuçlar, gelişmiş yeniden örnekleme teknikleri uygulanmış olsa bile sinir ağlarının bu özel problemde (GPCR etkileşimlerini tahmin etme) verimsiz kaldığını vurgulamaktadır.

Değerlendirmedeki en güçlü yöntemlerden biri SMOTE-Tomek (SSTK) olmuştur. SSTK-SVM, SSTK-RF ve SSTK-BBC modelleri neredeyse kusursuz sonuçlar üretmiş ve F1 ile ROC

değerlerini 0,99–1,0 aralığında kaydetmiştir. Bu kombinasyon, sınıfların dengelenmesi ve gürültünün azaltılması sayesinde en yüksek tespit doğruluğuna ulaşılabileceğini göstermektedir. Dikkat çekici şekilde, SSTK-NN ve SSTK-AB modelleri bir miktar iyileşme göstermiş olsa da F1 puanları 0,1–0,3 aralığında kalmış ve güçlü algoritmaların gerisinde kalmıştır.

Tomek Links (STL) yöntemi de SVM ve RF algoritmalarıyla birlikte iyi sonuçlar üretmiştir. Özellikle STL-RF modeli, yaklaşık 0,99'luk F1 ve ROC=1'e yakın bir performansla dikkat çekmiştir. Ancak STL-NN modelinde, diğer sinir ağı tabanlı modellerde olduğu gibi, sonuçlar zayıf kalmış ve F1 sıfır olmuştur. Bu durum, STL'nin gürültü gidermede başarılı olsa da sinir ağları için yeterli olmadığını ortaya koymaktadır.

Rastgele Aşırı Örnekleme (SROS) modelleri de ilginç bir tablo sunmuştur. SROS yöntemi SVM veya RF ile birleştirildiğinde, neredeyse ideal sonuçlar (ROC=1) gözlenmiştir. Örneğin, SROS-SVM modeli hem doğruluk hem F1 açısından mükemmele yakın sonuçlar üretmiştir. Bununla birlikte, SROS-NN ve SROS-AB modelleri halen düşük F1 değerlerine sahip olmuş ve sinir ağının yapısal zayıflığı devam etmiştir.

Beklendiği gibi, Rastgele Az Örnekleme (SRUS) yönteminde genel performans belirgin şekilde düşmüştür. SRUS-SVM ve SRUS-RF modelleri yaklaşık 0,15–0,18 civarında F1 değerleri elde etmiş ve bu düşüş, çoğunluk sınıfı örneklerinin rastgele kaldırılması nedeniyle önemli bilginin kaybolmasından kaynaklanmıştır.

SMOTE (SSM) ve SMOTENC (SSNC) yöntemleri ise güçlü sonuçlar üretmiştir. Özellikle SSM-SVM ve SSNC-SVM modelleri yaklaşık 0,98'lik F1 ve %100 doğruluk gibi üstün sonuçlara ulaşmıştır. Ancak bu grupta da sinir ağı modelleri benzer bir zayıf performans deseni göstermiş ve F1 puanları 0,3 civarında kalmıştır. Bu bulgular, SMOTE'un SVM ve RF gibi güçlü algoritmalarla birleştiğinde yüksek etkililiğini bir kez daha doğrulamaktadır.

Benzer bir eğilim SMOTEN (SSN) ve SSVM (SVM tabanlı SMOTE) yöntemlerinde de gözlemlenmiştir. SSN-SVM modeli yaklaşık 0,92'lik bir F1 puanı elde etmiş, ancak SSN-NN modeli yine zayıf performans göstermiştir. SSVM-SVM ve SSVM-RF modelleri ise 0,97'nin üzerinde F1 ve ROC=1 ile en başarılı sonuçları kaydetmiştir.

Son olarak, Borderline-SMOTE (SBS) yöntemi de etkili kombinasyonlardan biri olmuştur. SBS-SVM modeli yaklaşık 0,98'lik F1 ve ROC=1 ile en iyi sonuçlara yaklaşmış, SBS-RF modeli de benzer düzeyde performans göstermiştir. İlginç bir şekilde, SBS-NN modeli diğer sinir ağı modellerine göre biraz daha iyi bir performans sergileyerek yaklaşık 0,33 F1 puanı almış, ancak yine de güçlü algoritmalarla kıyaslanabilir düzeye ulaşamamıştır.

Bu analizden çıkan sonuç, GPCR reseptör etkileşimlerinin tahmini için SMOTE-Tomek, Borderline-SMOTE ve standart SMOTE gibi gelişmiş yeniden örnekleme yöntemlerinin SVM, Random Forest ve AdaBoost algoritmalarıyla birleştirildiğinde en yüksek doğruluk ve genelleştirilebilirliği sağladığını göstermektedir. Ayrıca sinir ağı modelleri, veri dengelenmesinden sonra bile tatmin edici sonuçlar üretememiş ve bu yapısal zayıflık tüm değerlendirmelerde tutarlı biçimde ortaya çıkmıştır. Bu nedenle, biyoenformatik alanında doğru ve etkili modeller geliştirmek için yeniden örnekleme stratejileri ve sınıflandırıcıların titizlikle seçilmesi büyük önem taşımaktadır.

Tablo 4. 25: GPCR veri kümesi üzerinde PSEAAC_PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.177033	0.74	0.285714	0.9075	0.825897	**SCN-KNN**	0.394366	0.56	0.46281	0.9675	0.768974
SCC-SVM	0.038081	1	0.073368	0.3685	0.676154	**SCN-SVM**	0.836364	0.92	0.87619	0.9935	0.957692
SCC-DT	0.05	0.94	0.094949	0.552	0.741026	**SCN-DT**	0.193878	0.38	0.256757	0.945	0.669744
SCC-RF	0.039463	1	0.07593	0.3915	0.687949	**SCN-RF**	0.961538	1	0.980392	0.999	0.999487
SCC-NN	0.06735	0.92	0.125512	0.6795	0.796667	**SCN-NN**	0	0	0	0.9745	0.499744
SCC-AB	0.04401	0.72	0.082949	0.602	0.659487	**SCN-AB**	0	0	0	0.971	0.497949
SCC-NB	0.027322	0.9	0.053035	0.1965	0.539231	**SCN-NB**	0.076056	0.54	0.133333	0.8245	0.685897
SCC-QDA	0.141066	0.9	0.243902	0.8605	0.879744	**SCN-QDA**	0.230769	0.84	0.362069	0.926	0.884103
SCC-LDA	0.055385	0.72	0.102857	0.686	0.702564	**SCN-LDA**	0.090909	0.06	0.072289	0.9615	0.522308
SCC-EEC	0.039199	0.94	0.07526	0.4225	0.674615	**SCN-EEC**	0.061261	0.68	0.112397	0.7315	0.70641
SCC-RSC	0.04401	0.72	0.082949	0.602	0.659487	**SCN-RSC**	0.045198	0.8	0.085561	0.5725	0.683333
SCC-BBC	0.059952	1	0.113122	0.608	0.798974	**SCN-BBC**	0.193133	0.9	0.318021	0.9035	0.901795
SCC-BRFC	0.047786	0.82	0.090308	0.587	0.700513	**SCN-BRFC**	0.068273	0.68	0.124088	0.76	0.721026
SEN-KNN	0.795455	0.7	0.744681	0.988	0.847692	**SRN-KNN**	0.705882	0.72	0.712871	0.9855	0.856154
SEN-SVM	0.777778	0.98	0.867257	0.9925	0.98641	**SRN-SVM**	0.680556	0.98	0.803279	0.988	0.984103
SEN-DT	0.5	0.04	0.074074	0.975	0.519487	**SRN-DT**	0.5	0.04	0.074074	0.975	0.519487
SEN-RF	1	1	1	1	1	**SRN-RF**	1	1	1	1	1
SEN-NN	0	0	0	0.975	0.5	**SRN-NN**	0	0	0	0.975	0.5
SEN-AB	0	0	0	0.975	0.5	**SRN-AB**	0	0	0	0.975	0.5
SEN-NB	0.128049	0.42	0.196262	0.914	0.673333	**SRN-NB**	0.130178	0.44	0.200913	0.9125	0.682308
SEN-QDA	0.218579	0.8	0.343348	0.9235	0.863333	**SRN-QDA**	0.221622	0.82	0.348936	0.9235	0.873077
SEN-LDA	0.076923	0.02	0.031746	0.9695	0.506923	**SRN-LDA**	0.2	0.06	0.092308	0.9705	0.526923
SEN-EEC	0.061093	0.76	0.113095	0.702	0.730256	**SRN-EEC**	0.063758	0.76	0.117647	0.715	0.736923
SEN-RSC	0.045455	0.8	0.086022	0.575	0.684615	**SRN-RSC**	0.041742	0.46	0.076539	0.7225	0.594615
SEN-BBC	0.151815	0.92	0.260623	0.8695	0.894103	**SRN-BBC**	0.142433	0.96	0.248062	0.8545	0.905897
SEN-BRFC	0.05753	0.68	0.106084	0.7135	0.697179	**SRN-BRFC**	0.0625	0.72	0.115016	0.723	0.721538
SSTK-KNN	0.4	1	0.571429	0.9625	0.980769	**SNC-KNN**	0.840909	0.74	0.787234	0.99	0.868205
SSTK-SVM	0.961538	1	0.980392	0.999	0.999487	**SNC-SVM**	0.8	0.96	0.872727	0.993	0.976923
SSTK-DT	0.082452	0.78	0.14914	0.7775	0.778718	**SNC-DT**	0.666667	0.04	0.075472	0.9755	0.519744
SSTK-RF	0.892857	1	0.943396	0.997	0.998462	**SNC-RF**	1	1	1	1	1
SSTK-NN	0.209524	0.88	0.338462	0.914	0.897436	**SNC-NN**	0	0	0	0.975	0.5
SSTK-AB	0.050287	0.7	0.093834	0.662	0.680513	**SNC-AB**	0	0	0	0.975	0.5
SSTK-NB	0.046699	0.58	0.086438	0.6935	0.638205	**SNC-NB**	0.125	0.4	0.190476	0.915	0.664103
SSTK-QDA	0.226415	0.72	0.344498	0.9315	0.828462	**SNC-QDA**	0.221622	0.82	0.348936	0.9235	0.873077
SSTK-LDA	0.064315	0.62	0.116541	0.765	0.694359	**SNC-LDA**	0	0	0	0.972	0.498462
SSTK-EEC	0.067834	0.62	0.122288	0.7775	0.700769	**SNC-EEC**	0.069027	0.78	0.126829	0.7315	0.755128
SSTK-RSC	0.050287	0.7	0.093834	0.662	0.680513	**SNC-RSC**	0.045198	0.8	0.085561	0.5725	0.683333
SSTK-BBC	0.980392	1	0.990099	0.9995	0.999744	**SNC-BBC**	0.158065	0.98	0.272222	0.869	0.923077
SSTK-BRFC	0.065913	0.7	0.120482	0.7445	0.722821	**SNC-BRFC**	0.062167	0.7	0.114192	0.7285	0.714615
SOS-KNN	0.90625	0.58	0.707317	0.988	0.789231	**SRUS-KNN**	0.049031	0.86	0.092772	0.5795	0.716154
SOS-SVM	0.888889	0.96	0.923077	0.996	0.978462	**SRUS-SVM**	0.081967	1	0.151515	0.72	0.85641
SOS-DT	1	0.04	0.076923	0.976	0.52	**SRUS-DT**	0.057584	0.82	0.107612	0.66	0.737949
SOS-RF	1	1	1	1	1	**SRUS-RF**	0.098039	1	0.178571	0.77	0.882051
SOS-NN	0	0	0	0.975	0.5	**SRUS-NN**	0.056693	0.72	0.105109	0.6935	0.70641
SOS-AB	0	0	0	0.975	0.5	**SRUS-AB**	0.06383	0.72	0.117264	0.729	0.724615
SOS-NB	0.119205	0.36	0.179104	0.9175	0.645897	**SRUS-NB**	0.042553	0.64	0.0798	0.631	0.635385
SOS-QDA	0.21978	0.8	0.344828	0.924	0.86359	**SRUS-QDA**	0.115385	0.9	0.204545	0.825	0.861538
SOS-LDA	0	0	0	0.973	0.498974	**SRUS-LDA**	0.05	0.78	0.093976	0.624	0.7
SOS-EEC	0.068792	0.82	0.126935	0.718	0.767692	**SRUS-EEC**	0.058722	0.68	0.108108	0.7195	0.700256
SOS-RSC	0.044053	0.8	0.083507	0.561	0.677436	**SRUS-RSC**	0.06383	0.72	0.117264	0.729	0.724615
SOS-BBC	0.158249	0.94	0.270893	0.8735	0.905897	**SRUS-BBC**	0.10661	1	0.192678	0.7905	0.892564
SOS-BRFC	0.061372	0.68	0.112583	0.732	0.706667	**SRUS-BRFC**	0.057276	0.74	0.106322	0.689	0.713846

Tablo 4. 26: GPCR veri kümesi üzerinde PSEAAC_PSEPSSM özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.90625	0.58	0.707317	0.988	0.789231	**SROS-KNN**	0.625	1	0.769231	0.985	0.992308
STL-SVM	0.888889	0.96	0.923077	0.996	0.978462	**SROS-SVM**	0.961538	1	0.980392	0.999	0.999487
STL-DT	1	0.04	0.076923	0.976	0.52	**SROS-DT**	0.078689	0.96	0.145455	0.718	0.835897
STL-RF	1	0.98	0.989899	0.9995	0.99	**SROS-RF**	1	1	1	1	1
STL-NN	0	0	0	0.975	0.5	**SROS-NN**	0.221719	0.98	0.361624	0.9135	0.945897
STL-AB	0	0	0	0.975	0.5	**SROS-AB**	0.061189	0.7	0.11254	0.724	0.712308
STL-NB	0.129032	0.4	0.195122	0.9175	0.665385	**SROS-NB**	0.057471	0.7	0.106222	0.7055	0.702821
STL-QDA	0.21978	0.8	0.344828	0.924	0.86359	**SROS-QDA**	0.19802	0.8	0.31746	0.914	0.858462
STL-LDA	0	0	0	0.973	0.498974	**SROS-LDA**	0.066547	0.74	0.122112	0.734	0.736923
STL-EEC	0.068611	0.8	0.126382	0.7235	0.760769	**SROS-EEC**	0.064912	0.74	0.119355	0.727	0.733333
STL-RSC	0.045872	0.8	0.086768	0.579	0.686667	**SROS-RSC**	0.061189	0.7	0.11254	0.724	0.712308
STL-BBC	0.181132	0.96	0.304762	0.8905	0.924359	**SROS-BBC**	1	1	1	1	1
STL-BRFC	0.060504	0.72	0.111628	0.7135	0.716667	**SROS-BRFC**	0.061433	0.72	0.113208	0.718	0.718974
SSM-KNN	0.4	1	0.571429	0.9625	0.980769	**SSNC-KNN**	0.416667	1	0.588235	0.965	0.982051
SSM-SVM	0.961538	1	0.980392	0.999	0.999487	**SSNC-SVM**	0.961538	1	0.980392	0.999	0.999487
SSM-DT	0.082452	0.78	0.14914	0.7775	0.778718	**SSNC-DT**	0.085714	0.72	0.153191	0.801	0.761538
SSM-RF	0.892857	1	0.943396	0.997	0.998462	**SSNC-RF**	0.909091	1	0.952381	0.9975	0.998718
SSM-NN	0.209524	0.88	0.338462	0.914	0.897436	**SSNC-NN**	0.203704	0.88	0.330827	0.911	0.895897
SSM-AB	0.050287	0.7	0.093834	0.662	0.680513	**SSNC-AB**	0.048991	0.68	0.091398	0.662	0.670769
SSM-NB	0.046699	0.58	0.086438	0.6935	0.638205	**SSNC-NB**	0.045016	0.56	0.083333	0.692	0.627692
SSM-QDA	0.226415	0.72	0.344498	0.9315	0.828462	**SSNC-QDA**	0.218182	0.72	0.334884	0.9285	0.826923
SSM-LDA	0.064315	0.62	0.116541	0.765	0.694359	**SSNC-LDA**	0.068085	0.64	0.123077	0.772	0.707692
SSM-EEC	0.067834	0.62	0.122288	0.7775	0.700769	**SSNC-EEC**	0.066955	0.62	0.120858	0.7745	0.699231
SSM-RSC	0.050287	0.7	0.093834	0.662	0.680513	**SSNC-RSC**	0.048991	0.68	0.091398	0.662	0.670769
SSM-BBC	0.980392	1	0.990099	0.9995	0.999744	**SSNC-BBC**	0.961538	1	0.980392	0.999	0.999487
SSM-BRFC	0.065913	0.7	0.120482	0.7445	0.722821	**SSNC-BRFC**	0.064935	0.7	0.118846	0.7405	0.720769
SSN-KNN	0.925926	0.5	0.649351	0.9865	0.749487	**SSVM-KNN**	0.40404	0.8	0.536913	0.9655	0.884872
SSN-SVM	1	0.86	0.924731	0.9965	0.93	**SSVM-SVM**	0.961538	1	0.980392	0.999	0.999487
SSN-DT	1	0.02	0.039216	0.9755	0.51	**SSVM-DT**	0.106218	0.82	0.188073	0.823	0.821538
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0	0	0	0.975	0.5	**SSVM-NN**	0.214815	0.58	0.313514	0.9365	0.762821
SSN-AB	0	0	0	0.9745	0.499744	**SSVM-AB**	0.067385	0.5	0.118765	0.8145	0.661282
SSN-NB	0	0	0	0.975	0.5	**SSVM-NB**	0.045699	0.34	0.080569	0.806	0.578974
SSN-QDA	0	0	0	0.975	0.5	**SSVM-QDA**	0.248	0.62	0.354286	0.9435	0.785897
SSN-LDA	0	0	0	0.975	0.5	**SSVM-LDA**	0.090062	0.58	0.155914	0.843	0.714872
SSN-EEC	0	0	0	0.9745	0.499744	**SSVM-EEC**	0.086379	0.52	0.148148	0.8505	0.689487
SSN-RSC	0	0	0	0.9745	0.499744	**SSVM-RSC**	0.067385	0.5	0.118765	0.8145	0.661282
SSN-BBC	1	0.88	0.93617	0.997	0.94	**SSVM-BBC**	1	0.94	0.969072	0.9985	0.97
SSN-BRFC	0	0	0	0.975	0.5	**SSVM-BRFC**	0.102459	0.5	0.170068	0.878	0.693846
SBS-KNN	0.39604	0.8	0.529801	0.9645	0.884359	**SBS-QDA**	0.242038	0.76	0.36715	0.9345	0.849487
SBS-SVM	0.961538	1	0.980392	0.999	0.999487	**SBS-LDA**	0.069264	0.64	0.125	0.776	0.709744
SBS-DT	0.081136	0.8	0.14733	0.7685	0.783846	**SBS-EEC**	0.069717	0.64	0.125737	0.7775	0.710513
SBS-RF	0.862069	1	0.925926	0.996	0.997949	**SBS-RSC**	0.053681	0.7	0.099715	0.684	0.691795
SBS-NN	0.214286	0.78	0.336207	0.923	0.853333	**SBS-BBC**	0.959184	0.94	0.949495	0.9975	0.969487
SBS-AB	0.053681	0.7	0.099715	0.684	0.691795	**SBS-BRFC**	0.06405	0.62	0.116105	0.764	0.693846
SBS-NB	0.047285	0.54	0.086957	0.7165	0.630513						

4.5 ENZİM VERİ SETİNDE ÖNERİLEN MODELİN PERFORMANSININ DİĞER YÖNTEMLERLE KARŞILAŞTIRMASI

Bu bölümde, EN veri seti, dört farklı istatistiksel ve biyolojik özellik kombinasyonuna dayanarak incelenmiştir. Söz konusu özellikler; dizilim, yapısal ve evrimsel bilgilere karşılık gelmekte olup, verilerin temsil doğruluğunu artırmak ve modelin etkileşimleri tanıma yeteneğini güçlendirmek amacıyla kullanılmıştır. Analiz süreci üç katmanlı bir yapı çerçevesinde yürütülmüştür: İlk olarak veriler, Siyam ağı kullanılarak dengeli biçimde temsil edilmiştir; ardından sınıf dengesizliği sorununu gidermek için çeşitli yeniden örnekleme yöntemleri uygulanmış; son aşamada ise elde edilen veriler, farklı makine öğrenimi algoritmalarıyla eğitilmiştir. Bu aşamanın temel amacı, ilaç-hedef etkileşimlerinin tahmininde en etkili yeniden örnekleme yöntemi ile sınıflandırma algoritması kombinasyonunu belirlemektir. Modellerin performansı, doğruluk (Accuracy), geri çağırma (Recall), F1 skoru ve ROC eğrisi altında kalan alan (ROC-AUC) gibi standart değerlendirme ölçütleriyle analiz edilmiştir.

4.5.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

Tablo 4.27 ve 4.28'de görülebileceği üzere, modellerin büyük çoğunluğu farklı yeniden örnekleme adımları uygulandıktan ve çeşitli sınıflandırma algoritmalarıyla birleştirildikten sonra oldukça yüksek bir performans sergilemiştir. Verileri dengelemek için yalnızca basit bir temizleme yöntemi kullanılan SCC grubu genel olarak tatmin edici sonuçlar elde etmiştir. Bu grupta, SCC-RF modeli doğruluk, hassasiyet, F1 ve ROC değerlerinin tamamında %99,5 ila %100 arasında başarı göstererek desen tanımada Random Forest algoritmasının güçlü yönlerini bir kez daha ortaya koymuştur. SCC-SVM modeli ise %99,3 civarında F1 ve %99,3'ün üzerinde ROC skoru ile ikinci sırada yer almakta, bu algoritmanın nispeten daha az temizlenmiş verilerde bile etkili bir ayrım yeteneğine sahip olduğunu kanıtlamaktadır. Buna karşın, F1 skoru yaklaşık %92 olan SCC-AB modeli ile %93 civarında F1 skoru elde eden SCC-NB modeli daha zayıf bir performans göstermiştir. AdaBoost ve Naive Bayes'teki bu doğruluk düşüşü, bu algoritmaların dengesiz verilere karşı daha hassas olması ve özellik dağılımları hakkında daha basit varsayımlar yapmasından kaynaklanmaktadır.

Gürültü giderme yönteminin uygulandığı SCN grubunda ise bazı modellerde belirgin bir performans artışı gözlenmiştir. Özellikle SCN-SVM ve SCN-RF modelleri sırasıyla yaklaşık

%98,9 ve %99,3 F1 skoru elde ederek yüksek doğruluk ve ayrım gücü sağlamıştır. Bu sonuçlar, sınıflandırıcıları eğitmeden önce gürültüyü azaltmanın başarıyı önemli ölçüde artırabileceğini göstermektedir. Öte yandan, F1 skoru %90 civarında kalan SCN-NN modeli ile yaklaşık %84'te kalan SCN-AB modeli, SVM ve RF'ye kıyasla önemli ölçüde daha düşük bir performans sergilemiştir. Bu durum, karmaşık yapıların ve gürültülü verilerin, basit sinir ağları ve AdaBoost algoritmalarının öğrenme sürecini zorlaştırdığını bir kez daha ortaya koymaktadır.

SMOTE-ENN ve Rastgele Azaltılmış Örnekleme yöntemlerinin uygulandığı SEN ve SRN gruplarında modeller arasında dikkat çekici farklar ortaya çıkmıştır. SEN-SVM ve SEN-RF modelleri yaklaşık %95 ila %98 F1 skorlarıyla yine en yüksek başarıyı yakalamıştır. Buna karşılık, SEN-AB ve SEN-NB yaklaşık %91 ila %93 F1 ile orta seviyede performans sergilerken, SEN-RSC modeli yaklaşık %86 F1 ile bu gruptaki en zayıf sonuçlardan birini elde etmiştir. Ayrıca SRN-RF ve SRN-SVM modelleri çoğu kriterde %96'yı aşmış, veri hacmi azaltılsa dahi güçlü algoritmaların kabul edilebilir doğruluk seviyelerini koruyabildiğini ortaya koymuştur. Ancak SRN-RSC ve SRN-AB modelleri kayda değer bir düşüş yaşamıştır.

SMOTE-Tomek yöntemine dayanan SSTK grubu ise çoğu yöntem arasında en iyi sonuçları sunmuştur. Özellikle SSTK-RF ve SSTK-BBC modelleri tüm metriklerde neredeyse mükemmel (%100'e yakın) doğruluk, hassasiyet ve F1 skorlarına ulaşmıştır. Bu bulgular, gürültü azaltma ve Siamese ağı kullanarak azınlık örneklerini artırmanın birleşiminin neredeyse kusursuz bir öğrenme sağlayabileceğini doğrulamaktadır. Bunun yanında SSTK-SVM modeli de yaklaşık %99,5 F1 skoru ile çok yüksek bir başarı elde etmiştir. Ancak SSTK-NN ve SSTK-AB modelleri %96 civarında F1 ile diğer en iyi modellerin biraz gerisinde kalmış ve bu denli dengeli veri ortamında bile daha basit algoritmaların sınırlamalarının sürdüğünü göstermiştir.

SOS ve SRUS gruplarında Rastgele Aşırı Örnekleme ile Azaltılmış Örnekleme arasındaki fark açık biçimde gözlenmektedir. SOS grubunda, SOS-RF ve SOS-SVM başta olmak üzere çoğu model neredeyse kusursuz doğruluk elde etmiştir. Buna karşılık SRUS grubunda, SRUS-BRFC ve SRUS-RSC gibi modeller %18 ve %14'e kadar düşen F1 skorlarıyla önemli bir gerileme yaşamıştır. Bu durum, veri setinin büyük kısmının çıkarılmasıyla az örneklemenin önemli bilgilerin kaybına yol açtığını ve modelleri ciddi şekilde zorladığını ortaya koymaktadır. Buna rağmen, SRUS-RF modeli %100'e yakın F1 ile rastgele ormanın sınırlı veriler altında bile en sağlam algoritmalardan biri olduğunu kanıtlamıştır.

STL ve SROS grupları da benzer şekilde güçlü sonuçlar üretmiştir. STL-RF ve SROS-RF modelleri %99,7 ila %100 F1 skorlarıyla en iyi performansı sergilemiştir. Buna karşın STL-NN ve SROS-NN modelleri %94 ila %95 civarında kalarak diğer daha güçlü algoritmaların gerisinde kalmıştır. Bu sonuçlar, basit sinir ağlarının gelişmiş veri dengesi ortamında bile tahmin performansında SVM ve RF'nin gerisinde kaldığının bir başka kanıtıdır.

SSM ve SSNC gruplarında da benzer eğilimler gözlenmiştir. SSM-RF ve SSNC-RF modelleri neredeyse kusursuz doğruluk sergilerken, SSM-SVM ve SSNC-SVM modelleri de yaklaşık %99,5 F1 skoru ile en başarılı sonuçlar arasında yer almıştır. Buna karşın SSM-AB, SSM-NB ve SSM-RSC modelleri yaklaşık %92 F1 ile daha düşük bir performans göstermiş ve veri yapısına uygun algoritma seçiminin kritik önemini bir kez daha vurgulamıştır.

Aynı eğilim SSN ve SSVM gruplarında da devam etmiştir. SSN-RF ve SSVM-RF modelleri yüksek doğruluk ve F1 skorlarıyla dikkat çekerken, F1 skoru %95 civarında olan SSN-NN ve SSVM-NN modelleri sinir ağlarının bu grupta nispeten daha iyi performans göstermesine rağmen SVM ve RF'nin gerisinde kaldığını ortaya koymuştur.

Son olarak SBS grubunda da SVM ve RF modelleri neredeyse mükemmel başarı elde etmiş ve doğruluk ile F1 skorları en üst düzeye yaklaşmıştır. Bu grup, diğer tüm gruplarda olduğu gibi sinir ağlarının ve AdaBoost'un daha güçlü algoritmalarla kıyaslandığında göreli zayıflığını bir kez daha doğrulamaktadır.

Tüm bulgular bir arada değerlendirildiğinde, SVM ve Random Forest algoritmalarının örnekleme türünden bağımsız olarak sürekli şekilde en yüksek ve en istikrarlı başarıyı sağladığı görülmektedir. Ayrıca QDA ve BBC gibi bazı modellerin belirli gruplarda sergilediği yüksek doğruluk, aşırı uyum riskini ve bağımsız veri üzerinde daha fazla test yapılması gereğini işaret etmektedir. Dahası, birçok durumda gözlemlendiği üzere basit sinir ağları genellikle diğer algoritmalara göre daha sınırlı bir verim ortaya koymuştur.

Tablo 4. 27: EN veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.994662	0.972174	0.983289	0.983775	0.98357	**SCN-KNN**	0.993266	0.513043	0.676606	0.75918	0.754844
SCC-SVM	0.996503	0.991304	0.993897	0.994022	0.993974	**SCN-SVM**	1	0.97913	0.989455	0.989752	0.989565
SCC-DT	0.992509	0.921739	0.955816	0.958155	0.957514	**SCN-DT**	0.996024	0.871304	0.929499	0.935098	0.933974
SCC-RF	1	0.991304	0.995633	0.99573	0.995652	**SCN-RF**	1	0.986087	0.992995	0.993168	0.993043
SCC-NN	0.957118	0.970435	0.963731	0.964133	0.964244	**SCN-NN**	1	0.833043	0.908918	0.918019	0.916522
SCC-AB	0.965779	0.883478	0.922797	0.927412	0.926638	**SCN-AB**	0.995294	0.735652	0.846	0.868488	0.866148
SCC-NB	0.932409	0.935652	0.934028	0.935098	0.935108	**SCN-NB**	0.922945	0.937391	0.930112	0.930828	0.930944
SCC-QDA	0.987273	0.944348	0.965333	0.966695	0.966301	**SCN-QDA**	0.968028	0.947826	0.957821	0.959009	0.958812
SCC-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SCN-LDA**	0.996198	0.911304	0.951862	0.95474	0.953974
SCC-EEC	0.994505	0.944348	0.968778	0.970111	0.969657	**SCN-EEC**	0.962199	0.973913	0.968021	0.968403	0.9685
SCC-RSC	0.965779	0.883478	0.922797	0.927412	0.926638	**SCN-RSC**	0.956236	0.76	0.846899	0.865073	0.863221
SCC-BBC	1	0.986087	0.992995	0.993168	0.993043	**SCN-BBC**	0.972174	0.972174	0.972174	0.972673	0.972664
SCC-BRFC	0.969466	0.883478	0.924477	0.92912	0.928316	**SCN-BRFC**	0.94974	0.953043	0.951389	0.952178	0.952193
SEN-KNN	0.994662	0.972174	0.983289	0.983775	0.98357	**SRN-KNN**	0.996161	0.902609	0.94708	0.95047	0.949626
SEN-SVM	0.996161	0.902609	0.94708	0.95047	0.949626	**SRN-SVM**	1	0.958261	0.978686	0.979505	0.97913
SEN-DT	1	0.958261	0.978686	0.979505	0.97913	**SRN-DT**	0.994434	0.932174	0.962298	0.964133	0.96357
SEN-RF	0.994434	0.932174	0.962298	0.964133	0.96357	**SRN-RF**	1	0.968696	0.984099	0.984629	0.984348
SEN-NN	1	0.966957	0.983201	0.983775	0.983478	**SRN-NN**	1	0.916522	0.956443	0.959009	0.958261
SEN-AB	1	0.916522	0.956443	0.959009	0.958261	**SRN-AB**	0.967054	0.867826	0.914757	0.920581	0.919651
SEN-NB	0.967054	0.867826	0.914757	0.920581	0.919651	**SRN-NB**	0.941901	0.930435	0.936133	0.93766	0.937533
SEN-QDA	0.941901	0.930435	0.936133	0.93766	0.937533	**SRN-QDA**	0.989011	0.93913	0.963426	0.964987	0.964532
SEN-LDA	0.989011	0.93913	0.963426	0.964987	0.964532	**SRN-LDA**	1	0.897391	0.945921	0.949616	0.948696
SEN-EEC	1	0.897391	0.945921	0.949616	0.948696	**SRN-EEC**	0.994444	0.933913	0.963229	0.964987	0.96444
SEN-RSC	0.994444	0.933913	0.963229	0.964987	0.96444	**SRN-RSC**	0.963907	0.789565	0.868069	0.882152	0.880521
SEN-BBC	0.962025	0.793043	0.869399	0.883006	0.881421	**SRN-BBC**	1	0.970435	0.984996	0.985482	0.985217
SEN-BRFC	1	0.970435	0.984996	0.985482	0.985217	**SRN-BRFC**	0.966667	0.857391	0.908756	0.915457	0.914434
SSTK-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SNC-KNN**	0.996303	0.937391	0.96595	0.967549	0.967018
SSTK-SVM	0.996516	0.994783	0.995648	0.99573	0.995713	**SNC-SVM**	1	0.972174	0.985891	0.986336	0.986087
SSTK-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SNC-DT**	0.994413	0.928696	0.960432	0.962425	0.961831
SSTK-RF	1	0.994783	0.997384	0.997438	0.997391	**SNC-RF**	1	0.986087	0.992995	0.993168	0.993043
SSTK-NN	0.983725	0.946087	0.964539	0.965841	0.965493	**SNC-NN**	1	0.907826	0.951686	0.95474	0.953913
SSTK-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SNC-AB**	0.967054	0.867826	0.914757	0.920581	0.919651
SSTK-NB	0.937063	0.932174	0.934612	0.935952	0.935886	**SNC-NB**	0.940246	0.930435	0.935315	0.936806	0.936694
SSTK-QDA	0.987273	0.944348	0.965333	0.966695	0.966301	**SNC-QDA**	0.989031	0.94087	0.964349	0.965841	0.965401
SSTK-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SNC-LDA**	1	0.904348	0.949772	0.953032	0.952174
SSTK-EEC	0.983696	0.944348	0.96362	0.964987	0.964624	**SNC-EEC**	0.990842	0.94087	0.96521	0.966695	0.96624
SSTK-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SNC-RSC**	0.96383	0.787826	0.866986	0.881298	0.879651
SSTK-BBC	1	0.994783	0.997384	0.997438	0.997391	**SNC-BBC**	0.998239	0.986087	0.992126	0.992314	0.992205
SSTK-BRFC	0.968105	0.897391	0.931408	0.935098	0.934434	**SNC-BRFC**	0.966926	0.864348	0.912764	0.918873	0.917912
SOS-KNN	0.994604	0.961739	0.977896	0.978651	0.978353	**SRUS-KNN**	0.994652	0.970435	0.982394	0.982921	0.982701
SOS-SVM	0.998252	0.993043	0.995641	0.99573	0.995683	**SRUS-SVM**	0.996516	0.994783	0.995648	0.99573	0.995713
SOS-DT	0.998138	0.932174	0.964029	0.965841	0.965248	**SRUS-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
SOS-RF	1	0.994783	0.997384	0.997438	0.997391	**SRUS-RF**	1	1	1	1	1
SOS-NN	1	0.925217	0.961156	0.963279	0.962609	**SRUS-NN**	0.992647	0.93913	0.965147	0.966695	0.96621
SOS-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SRUS-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
SOS-NB	0.935428	0.932174	0.933798	0.935098	0.935047	**SRUS-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
SOS-QDA	0.985481	0.944348	0.964476	0.965841	0.965463	**SRUS-QDA**	0.987273	0.944348	0.965333	0.966695	0.966301
SOS-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SRUS-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SOS-EEC	0.983666	0.942609	0.9627	0.964133	0.963754	**SRUS-EEC**	0.994455	0.935652	0.964158	0.965841	0.965309
SOS-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SRUS-RSC**	0.077586	0.947368	0.143426	0.880556	0.913606
SOS-BBC	1	0.994783	0.997384	0.997438	0.997391	**SRUS-BBC**	0.162162	0.947368	0.276923	0.947778	0.947575
SOS-BRFC	0.967803	0.888696	0.926564	0.930828	0.930086	**SRUS-BRFC**	0.104938	0.894737	0.187845	0.918333	0.906661

Tablo 4. 28: EN veri kümesi üzerinde AAC-DPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SROS-KNN**	0.994652	0.970435	0.982394	0.982921	0.982701
STL-SVM	0.998252	0.993043	0.995641	0.99573	0.995683	**SROS-SVM**	0.996516	0.994783	0.995648	0.99573	0.995713
STL-DT	0.998138	0.932174	0.964029	0.965841	0.965248	**SROS-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
STL-RF	1	0.994783	0.997384	0.997438	0.997391	**SROS-RF**	1	1	1	1	1
STL-NN	1	0.888696	0.941068	0.945346	0.944348	**SROS-NN**	1	0.916522	0.956443	0.959009	0.958261
STL-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SROS-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
STL-NB	0.938704	0.932174	0.935428	0.936806	0.936725	**SROS-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
STL-QDA	0.987273	0.944348	0.965333	0.966695	0.966301	**SROS-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
STL-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SROS-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
STL-EEC	0.983666	0.942609	0.9627	0.964133	0.963754	**SROS-EEC**	0.994455	0.935652	0.964158	0.965841	0.965309
STL-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SROS-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
STL-BBC	0.998255	0.994783	0.996516	0.996584	0.996552	**SROS-BBC**	1	0.998261	0.99913	0.999146	0.99913
STL-BRFC	0.967803	0.888696	0.926564	0.930828	0.930086	**SROS-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SSM-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SSNC-KNN**	0.994652	0.970435	0.982394	0.982921	0.982701
SSM-SVM	0.996516	0.994783	0.995648	0.99573	0.995713	**SSNC-SVM**	0.996516	0.994783	0.995648	0.99573	0.995713
SSM-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SSNC-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	1	0.916522	0.956443	0.959009	0.958261	**SSNC-NN**	1	0.916522	0.956443	0.959009	0.958261
SSM-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SSNC-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
SSM-NB	0.935428	0.932174	0.933798	0.935098	0.935047	**SSNC-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
SSM-QDA	0.985481	0.944348	0.964476	0.965841	0.965463	**SSNC-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
SSM-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SSNC-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SSM-EEC	0.994455	0.935652	0.964158	0.965841	0.965309	**SSNC-EEC**	0.994455	0.935652	0.964158	0.965841	0.965309
SSM-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SSNC-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SSM-BBC	1	0.998261	0.99913	0.999146	0.99913	**SSNC-BBC**	1	0.998261	0.99913	0.999146	0.99913
SSM-BRFC	0.96929	0.878261	0.921533	0.926558	0.925708	**SSNC-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SSN-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SSVM-KNN**	0.996429	0.970435	0.98326	0.983775	0.98354
SSN-SVM	0.996516	0.994783	0.995648	0.99573	0.995713	**SSVM-SVM**	0.998255	0.994783	0.996516	0.996584	0.996552
SSN-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SSVM-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	1	0.916522	0.956443	0.959009	0.958261	**SSVM-NN**	1	0.916522	0.956443	0.959009	0.958261
SSN-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SSVM-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
SSN-NB	0.937063	0.932174	0.934612	0.935952	0.935886	**SSVM-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
SSN-QDA	0.985481	0.944348	0.964476	0.965841	0.965463	**SSVM-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
SSN-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SSVM-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SSN-EEC	0.987203	0.93913	0.962567	0.964133	0.963693	**SSVM-EEC**	0.987296	0.946087	0.966252	0.967549	0.967171
SSN-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SSVM-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SSN-BBC	1	0.998261	0.99913	0.999146	0.99913	**SSVM-BBC**	1	0.998261	0.99913	0.999146	0.99913
SSN-BRFC	0.96929	0.878261	0.921533	0.926558	0.925708	**SSVM-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SBS-KNN	0.996429	0.970435	0.98326	0.983775	0.98354	**SBS-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
SBS-SVM	0.998255	0.994783	0.996516	0.996584	0.996552	**SBS-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SBS-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SBS-EEC**	0.987296	0.946087	0.966252	0.967549	0.967171
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SBS-NN	1	0.916522	0.956443	0.959009	0.958261	**SBS-BBC**	1	0.998261	0.99913	0.999146	0.99913
SBS-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SBS-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SBS-NB	0.935428	0.932174	0.933798	0.935098	0.935047						

4.5.2 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

Tablo 4.29 ve 4.30'den görülebileceği gibi, SCC grubunda basit gürültü giderme kullanıldığında modeller oldukça yüksek performans sergilemiştir. SCC-RF modeli, tüm metriklerde 1'e yakın değerler elde ederek en başarılı model olarak öne çıkmıştır. Bu durum, Random Forest algoritmasının sadeleştirilmiş verilerde bile desenleri tanımadaki olağanüstü kabiliyetini göstermektedir. SCC-SVM modeli de yaklaşık %99,4 F1 değeriyle mükemmel bir başarı sağlamıştır. SCC-QDA ve SCC-EEC gibi diğer modeller de %96-97 aralığında yüksek doğruluk oranlarına ulaşmıştır. Buna karşılık, SCC-AB ve SCC-NB gibi daha basit modeller görece zayıf kalmış, bu da yapısal olarak daha basit veya gürültüye daha duyarlı algoritmaların bu grupta sınırlı ayırıcı güce sahip olduğunu göstermektedir.

Daha gelişmiş gürültü giderme uygulanan SCN grubunda bazı modellerin performansında dikkate değer iyileşmeler görülmüştür. SCN-RF ve SCN-SVM, %99'a yaklaşan F1 puanları ve yüksek hassasiyetle öne çıkmıştır. Bu sonuçlar, ayrıntılı ön işleme süreçlerinin güçlü modellerin başarısını daha da artırdığını doğrulamaktadır. Bununla birlikte, SCN-KNN modelinde geri çağırma ve F1 oranlarında keskin bir düşüş yaşanmış, bu da bazı algoritmaların temizlenmiş veriye karşı aşırı duyarlı olduğunu ve iyi genelleme yapamadığını ortaya koymuştur.

SMOTE-ENN ve rastgele örnek azaltmaya dayanan SEN ve SRN gruplarında, SEN-RF ve SEN-SVM gibi modeller yine başarılı sonuçlar üretmiş ve F1 değerlerini %95-98 aralığında tutmuştur. Daha karmaşık yapıya sahip SEN-BBC ve SEN-QDA gibi algoritmalar yüksek doğruluğu korurken, SEN-RSC ve SEN-AB gibi modeller %86-91 bandında daha düşük sonuçlar vermiştir. SRN grubu da veri miktarındaki düşüşe rağmen RF ve SVM gibi güçlü modellerin kararlılığını sürdürebildiğini göstermiştir.

Siamese Network ve SMOTE-Tomek kombinasyonunu kullanan SSTK grubu son derece etkileyici sonuçlara ulaşmıştır. SSTK-RF ve SSTK-BBC modelleri neredeyse kusursuz (%100'e yakın) doğruluk sağlarken, SSTK-SVM modeli de yaklaşık %99,5 F1 puanıyla olağanüstü bir performans sergilemiştir. Bu bulgular, Siamese Network ile yapılan ilk dengelemenin ve SMOTE-Tomek tabanlı gürültü düzeltmenin bir arada kullanılmasının tahmin gücünü önemli ölçüde artırabileceğini ortaya koymaktadır. Ancak SSTK-AB ve SSTK-NB gibi nispeten daha basit modeller yapısal sınırlamalara takılmış ve en iyi durumda %93-96 seviyelerine ulaşabilmiştir.

SOS ve SRUS gruplarında da benzer eğilimler dikkat çekmiştir. SOS-RF, SOS-BBC ve SOS-SVM modelleri çok yüksek doğruluk üretirken, SRUS grubunda rastgele örnek azaltmaya bağlı olarak SRUS-BRFC ve SRUS-RSC gibi modellerde %86-92 aralığında belirgin performans kayıpları yaşanmıştır. Bununla birlikte SRUS-RF modeli yine üstün doğruluk sergileyerek dayanıklı algoritmaların veri azaltma koşullarında dahi istikrarlı kalabileceğini göstermiştir.

Aynı desen STL ve SROS gruplarında da gözlenmiş; Random Forest, BBC ve SVM tabanlı modeller tüm metriklerde olağanüstü sonuçlar elde etmiş, NN ve AB gibi algoritmalar ise görece geride kalmıştır. Bu durum, güçlü dengeleme yapılsa bile bazı modellerin yapısal kısıtlarının sürdüğünü ortaya koymaktadır.

SSM ve SSNC grupları da benzer şekilde Random Forest ve SVM'in mükemmele yakın doğruluğuyla öne çıkmış, QDA, EEC ve BBC gibi algoritmalar yüksek başarı sağlamıştır. Buna karşılık RSC, AB ve NB modelleri daha düşük performanslarıyla ayrışmıştır ve veri yapısına uygun algoritma seçiminin kritik önemini bir kez daha göstermiştir.

SSN ve SSVM gruplarında da RF ve SVM en iyi performansı kaydetmiş, NN modellerinde göreceli bir iyileşme görülse de F1 ve ROC bakımından gelişmiş algoritmaların gerisinde kalmışlardır.

Son olarak SBS grubu da bu deseni doğrulamaktadır. Random Forest, SVM ve BBC modelleri mükemmel veya mükemmele çok yakın doğruluk oranlarıyla lider konumda yer alırken, NN, AB ve NB algoritmaları veri dengelenmesine rağmen daha sınırlı başarı göstermiştir. Bu analizde Random Forest ve BBC gibi pek çok modelin doğruluk ve F1 değerlerinin 1'e yaklaşması çok yüksek performansa işaret ederken, aynı zamanda aşırı uyum riskini de gündeme getirmektedir. Buna karşılık yaklaşık %99,5 F1 ve %99,6 doğruluk oranına sahip Siamese , SMOTE-Tomek ve SVM modeli, doğruluk, hassasiyet ve ayırt edicilik arasında dengeli bir performans sunmakta ve daha düşük aşırı uyum riski sayesinde güçlü ve güvenilir bir tercih olarak öne çıkmaktadır.

Tablo 4. 29: EN veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.994662	0.972174	0.983289	0.983775	0.98357	**SCN-KNN**	0.993266	0.513043	0.676606	0.75918	0.754844
SCC-SVM	0.996503	0.991304	0.993897	0.994022	0.993974	**SCN-SVM**	1	0.97913	0.989455	0.989752	0.989565
SCC-DT	0.992509	0.921739	0.955816	0.958155	0.957514	**SCN-DT**	0.996024	0.871304	0.929499	0.935098	0.933974
SCC-RF	1	0.991304	0.995633	0.99573	0.995652	**SCN-RF**	1	0.986087	0.992995	0.993168	0.993043
SCC-NN	0.957118	0.970435	0.963731	0.964133	0.964244	**SCN-NN**	1	0.833043	0.908918	0.918019	0.916522
SCC-AB	0.965779	0.883478	0.922797	0.927412	0.926638	**SCN-AB**	0.995294	0.735652	0.846	0.868488	0.866148
SCC-NB	0.932409	0.935652	0.934028	0.935098	0.935108	**SCN-NB**	0.922945	0.937391	0.930112	0.930828	0.930944
SCC-QDA	0.987273	0.944348	0.965333	0.966695	0.966301	**SCN-QDA**	0.968028	0.947826	0.957821	0.959009	0.958812
SCC-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SCN-LDA**	0.996198	0.911304	0.951862	0.95474	0.953974
SCC-EEC	0.994505	0.944348	0.968778	0.970111	0.969657	**SCN-EEC**	0.962199	0.973913	0.968021	0.968403	0.9685
SCC-RSC	0.965779	0.883478	0.922797	0.927412	0.926638	**SCN-RSC**	0.956236	0.76	0.846899	0.865073	0.863221
SCC-BBC	1	0.986087	0.992995	0.993168	0.993043	**SCN-BBC**	0.972174	0.972174	0.972174	0.972673	0.972664
SCC-BRFC	0.969466	0.883478	0.924477	0.92912	0.928316	**SCN-BRFC**	0.94974	0.953043	0.951389	0.952178	0.952193
SEN-KNN	0.996161	0.902609	0.94708	0.95047	0.949626	**SRN-KNN**	0.996161	0.902609	0.94708	0.95047	0.949626
SEN-SVM	1	0.958261	0.978686	0.979505	0.97913	**SRN-SVM**	1	0.958261	0.978686	0.979505	0.97913
SEN-DT	0.994434	0.932174	0.962298	0.964133	0.96357	**SRN-DT**	0.994434	0.932174	0.962298	0.964133	0.96357
SEN-RF	1	0.966957	0.983201	0.983775	0.983478	**SRN-RF**	1	0.968696	0.984099	0.984629	0.984348
SEN-NN	1	0.916522	0.956443	0.959009	0.958261	**SRN-NN**	1	0.916522	0.956443	0.959009	0.958261
SEN-AB	0.967054	0.867826	0.914757	0.920581	0.919651	**SRN-AB**	0.967054	0.867826	0.914757	0.920581	0.919651
SEN-NB	0.941901	0.930435	0.936133	0.93766	0.937533	**SRN-NB**	0.941901	0.930435	0.936133	0.93766	0.937533
SEN-QDA	0.989011	0.93913	0.963426	0.964987	0.964532	**SRN-QDA**	0.989011	0.93913	0.963426	0.964987	0.964532
SEN-LDA	1	0.897391	0.945921	0.949616	0.948696	**SRN-LDA**	1	0.897391	0.945921	0.949616	0.948696
SEN-EEC	0.994444	0.933913	0.963229	0.964987	0.96444	**SRN-EEC**	0.994444	0.933913	0.963229	0.964987	0.96444
SEN-RSC	0.962025	0.793043	0.869399	0.883006	0.881421	**SRN-RSC**	0.963907	0.789565	0.868069	0.882152	0.880521
SEN-BBC	1	0.970435	0.984996	0.985482	0.985217	**SRN-BBC**	1	0.970435	0.984996	0.985482	0.985217
SEN-BRFC	0.966667	0.857391	0.908756	0.915457	0.914434	**SRN-BRFC**	0.966667	0.857391	0.908756	0.915457	0.914434
SSTK-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SNC-KNN**	0.996303	0.937391	0.96595	0.967549	0.967018
SSTK-SVM	0.996516	0.994783	0.995648	0.99573	0.995713	**SNC-SVM**	1	0.972174	0.985891	0.986336	0.986087
SSTK-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SNC-DT**	0.994413	0.928696	0.960432	0.962425	0.961831
SSTK-RF	1	0.994783	0.997384	0.997438	0.997391	**SNC-RF**	1	0.986087	0.992995	0.993168	0.993043
SSTK-NN	0.983725	0.946087	0.964539	0.965841	0.965493	**SNC-NN**	1	0.907826	0.951686	0.95474	0.953913
SSTK-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SNC-AB**	0.967054	0.867826	0.914757	0.920581	0.919651
SSTK-NB	0.937063	0.932174	0.934612	0.935952	0.935886	**SNC-NB**	0.940246	0.930435	0.935315	0.936806	0.936694
SSTK-QDA	0.987273	0.944348	0.965333	0.966695	0.966301	**SNC-QDA**	0.989031	0.94087	0.964349	0.965841	0.965401
SSTK-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SNC-LDA**	1	0.904348	0.949772	0.953032	0.952174
SSTK-EEC	0.983696	0.944348	0.96362	0.964987	0.964624	**SNC-EEC**	0.990842	0.94087	0.96521	0.966695	0.96624
SSTK-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SNC-RSC**	0.96383	0.787826	0.866986	0.881298	0.879651
SSTK-BBC	1	0.994783	0.997384	0.997438	0.997391	**SNC-BBC**	0.998239	0.986087	0.992126	0.992314	0.992205
SSTK-BRFC	0.968105	0.897391	0.931408	0.935098	0.934434	**SNC-BRFC**	0.966926	0.864348	0.912764	0.918873	0.917912
SOS-KNN	0.994604	0.961739	0.977896	0.978651	0.978353	**SRUS-KNN**	0.994652	0.970435	0.982394	0.982921	0.982701
SOS-SVM	0.998252	0.993043	0.995641	0.99573	0.995683	**SRUS-SVM**	0.996516	0.994783	0.995648	0.99573	0.995713
SOS-DT	0.998138	0.932174	0.964029	0.965841	0.965248	**SRUS-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
SOS-RF	1	0.994783	0.997384	0.997438	0.997391	**SRUS-RF**	1	1	1	1	1
SOS-NN	1	0.925217	0.961156	0.963279	0.962609	**SRUS-NN**	0.992647	0.93913	0.965147	0.966695	0.96621
SOS-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SRUS-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
SOS-NB	0.935428	0.932174	0.933798	0.935098	0.935047	**SRUS-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
SOS-QDA	0.985481	0.944348	0.964476	0.965841	0.965463	**SRUS-QDA**	0.987273	0.944348	0.965333	0.966695	0.966301
SOS-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SRUS-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SOS-EEC	0.983666	0.942609	0.9627	0.964133	0.963754	**SRUS-EEC**	0.994455	0.935652	0.964158	0.965841	0.965309
SOS-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SRUS-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SOS-BBC	1	0.994783	0.997384	0.997438	0.997391	**SRUS-BBC**	1	0.993043	0.99651	0.996584	0.996522
SOS-BRFC	0.967803	0.888696	0.926564	0.930828	0.930086	**SRUS-BRFC**	0.967925	0.892174	0.928507	0.932536	0.931825

Tablo 4. 30: EN veri kümesi üzerinde CKSAAGP-GDPC-GTPC özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SROS-KNN**	0.994652	0.970435	0.982394	0.982921	0.982701
STL-SVM	0.998252	0.993043	0.995641	0.99573	0.995683	**SROS-SVM**	0.996516	0.994783	0.995648	0.99573	0.995713
STL-DT	0.998138	0.932174	0.964029	0.965841	0.965248	**SROS-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
STL-RF	1	0.994783	0.997384	0.997438	0.997391	**SROS-RF**	1	1	1	1	1
STL-NN	1	0.888696	0.941068	0.945346	0.944348	**SROS-NN**	1	0.916522	0.956443	0.959009	0.958261
STL-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SROS-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
STL-NB	0.938704	0.932174	0.935428	0.936806	0.936725	**SROS-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
STL-QDA	0.987273	0.944348	0.965333	0.966695	0.966301	**SROS-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
STL-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SROS-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
STL-EEC	0.983666	0.942609	0.9627	0.964133	0.963754	**SROS-EEC**	0.994455	0.935652	0.964158	0.965841	0.965309
STL-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SROS-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
STL-BBC	0.998255	0.994783	0.996516	0.996584	0.996552	**SROS-BBC**	1	0.998261	0.99913	0.999146	0.99913
STL-BRFC	0.967803	0.888696	0.926564	0.930828	0.930086	**SROS-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SSM-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SSNC-KNN**	0.994652	0.970435	0.982394	0.982921	0.982701
SSM-SVM	0.996516	0.994783	0.995648	0.99573	0.995713	**SSNC-SVM**	0.996516	0.994783	0.995648	0.99573	0.995713
SSM-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SSNC-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	1	0.916522	0.956443	0.959009	0.958261	**SSNC-NN**	1	0.916522	0.956443	0.959009	0.958261
SSM-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SSNC-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
SSM-NB	0.935428	0.932174	0.933798	0.935098	0.935047	**SSNC-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
SSM-QDA	0.985481	0.944348	0.964476	0.965841	0.965463	**SSNC-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
SSM-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SSNC-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SSM-EEC	0.994455	0.935652	0.964158	0.965841	0.965309	**SSNC-EEC**	0.994455	0.935652	0.964158	0.965841	0.965309
SSM-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SSNC-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SSM-BBC	1	0.998261	0.99913	0.999146	0.99913	**SSNC-BBC**	1	0.998261	0.99913	0.999146	0.99913
SSM-BRFC	0.96929	0.878261	0.921533	0.926558	0.925708	**SSNC-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SSN-KNN	0.994652	0.970435	0.982394	0.982921	0.982701	**SSVM-KNN**	0.996429	0.970435	0.98326	0.983775	0.98354
SSN-SVM	0.996516	0.994783	0.995648	0.99573	0.995713	**SSVM-SVM**	0.998255	0.994783	0.996516	0.996584	0.996552
SSN-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SSVM-DT**	0.992565	0.928696	0.959569	0.961571	0.960992
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	1	0.916522	0.956443	0.959009	0.958261	**SSVM-NN**	1	0.916522	0.956443	0.959009	0.958261
SSN-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SSVM-AB**	0.955307	0.892174	0.922662	0.926558	0.925953
SSN-NB	0.937063	0.932174	0.934612	0.935952	0.935886	**SSVM-NB**	0.935428	0.932174	0.933798	0.935098	0.935047
SSN-QDA	0.985481	0.944348	0.964476	0.965841	0.965463	**SSVM-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
SSN-LDA	0.998106	0.916522	0.955576	0.958155	0.957422	**SSVM-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SSN-EEC	0.987203	0.93913	0.962567	0.964133	0.963693	**SSVM-EEC**	0.987296	0.946087	0.966252	0.967549	0.967171
SSN-RSC	0.955307	0.892174	0.922662	0.926558	0.925953	**SSVM-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SSN-BBC	1	0.998261	0.99913	0.999146	0.99913	**SSVM-BBC**	1	0.998261	0.99913	0.999146	0.99913
SSN-BRFC	0.96929	0.878261	0.921533	0.926558	0.925708	**SSVM-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SBS-KNN	0.996429	0.970435	0.98326	0.983775	0.98354	**SBS-QDA**	0.985481	0.944348	0.964476	0.965841	0.965463
SBS-SVM	0.998255	0.994783	0.996516	0.996584	0.996552	**SBS-LDA**	0.998106	0.916522	0.955576	0.958155	0.957422
SBS-DT	0.992565	0.928696	0.959569	0.961571	0.960992	**SBS-EEC**	0.987296	0.946087	0.966252	0.967549	0.967171
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.955307	0.892174	0.922662	0.926558	0.925953
SBS-NN	1	0.916522	0.956443	0.959009	0.958261	**SBS-BBC**	1	0.998261	0.99913	0.999146	0.99913
SBS-AB	0.955307	0.892174	0.922662	0.926558	0.925953	**SBS-BRFC**	0.96929	0.878261	0.921533	0.926558	0.925708
SBS-NB	0.935428	0.932174	0.933798	0.935098	0.935047						

4.5.3 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Tablo 4.31 ve 4.32 ’un incelenmesiyle birlikte, sadece basit gürültü giderme yöntemlerinin uygulandığı SCC grubunda elde edilen sonuçlar oldukça çarpıcıdır. Bu grupta özellikle Random Forest (SCC-RF) ve SVM (SCC-SVM) modelleri olağanüstü yüksek doğruluk ve F1 skoru sergilemiştir. Random Forest’ın yaklaşık 0,998’lik F1 skoru ve SVM’in 0,996’ya ulaşan performansı, sınıflandırmada neredeyse hatasız bir tahmin gücüne ulaşıldığını göstermektedir. Bu kadar yüksek başarı düzeyi, temel gürültü azaltmanın bile bazı güçlü algoritmalar için yeterli

veri kalitesini sağladığını ve verinin genel yapısından yüksek isabetle örüntü öğrenebildiklerini ortaya koymaktadır. Bununla birlikte aynı SCC grubu içinde yer alan Naive Bayes (SCC-NB) ve AdaBoost (SCC-AB) gibi daha basit ve lineer yapılı algoritmaların performansı görece düşük kalmış; F1 skorlarının 0,7–0,8 aralığında sınırlanması, bu yöntemlerin veri gürültüsüne karşı hâlâ kırılgan olduğunu ve karmaşık ilişkileri yeterince modelleyemediğini göstermektedir.

Daha gelişmiş temizleme işlemlerinin kullanıldığı SCN grubu, bazı modellerin daha dengeli sonuçlar üretmesine olanak tanımıştır. Örneğin SCN-RF, yaklaşık 0,95 F1 skoru elde ederek önemli bir başarı sağlamış, SCN-SVM ise 0,97 civarında F1 skoru ile dikkat çekmiştir. Bu sonuçlar, ileri düzey temizleme yöntemlerinin, güçlü algoritmaların performansını daha tutarlı hale getirebildiğini ve örüntü öğrenme kapasitesini desteklediğini göstermektedir. Ancak, SCN-KNN ve SCN-NN gibi algoritmaların performansında ciddi düşüş gözlenmiştir. Bu iki modelin sırasıyla 0,78 ve 0,67 seviyelerine kadar gerileyen F1 skorları, aşırı temizlik sebebiyle veri dağılımının bazı kritik sınır bilgilerinin kaybolabileceğini ve dolayısıyla modelin daha zor genelleme yapabildiğini göstermektedir. Burada veri ön işleme ile model karmaşıklığı arasındaki dengenin kritik olduğu anlaşılmaktadır.

SMOTE-ENN ve rastgele veri azaltma tekniklerini kullanan SEN ve SRN gruplarında, Random Forest ve SVM gibi güçlü algoritmalar liderliğini sürdürmüştür. SEN-RF ve SEN-SVM modelleri, F1 skorlarının 0,98'e yükselmesiyle, örnekleme yoluyla dengelenmiş verinin bu modeller için yüksek uyum sağladığını ve öğrenme kapasitesini artırdığını ortaya koymuştur. Bununla birlikte SEN-BBC ve SEN-QDA modelleri genel doğruluk açısından tatmin edici performans sergilerken, SEN-NB ve SEN-AB gibi daha basit algoritmalar, F1 skorlarını 0,75 seviyesinde korumaktan öteye geçememiştir. Bu durum, bazı algoritmaların karmaşık yeniden örnekleme stratejileriyle yeterince uyumlu çalışmadığını göstermektedir.

Özellikle dikkat çeken bir grup olan Siamese ağı ile SMOTE-Tomek kombinasyonunun uygulandığı SSTK grubu, şimdiye kadarki en etkileyici sonuçlardan bazılarını üretmiştir. SSTK-RF modeli, yaklaşık 0,996 F1 skoru ile olağanüstü bir doğruluk sergilemiş, SSTK-SVM de 0,994 civarındaki değeriyle çok güçlü bir başarı ortaya koymuştur. Bu bulgular, hibrit stratejilerin (derin temsil öğrenimi + örnekleme) veri yapısını daha etkili yakalayabildiğini ve karar sınırlarının daha iyi şekillendirilmesine katkıda bulunduğunu göstermektedir. SSTK-BBC ve SSTK-KNN modelleri de 0,97–0,99 aralığında kalarak istikrarlı bir performans sağlamış, fakat SSTK-NB ve SSTK-AB gibi basit modeller bu grupta bile yaklaşık 0,7–0,8 bandında

sınırlı kalmıştır. Bu da yöntem ne kadar gelişmiş olursa olsun, sınıflandırıcı karmaşıklığı belirleyici bir faktör olarak öne çıkmaktadır.

Benzer bir eğilim SOS ve SRUS gruplarında da gözlenmiştir. Random Forest ve SVM, çoğu durumda mükemmele yakın sonuçlar üretmiş; özellikle SRUS-RF modelinin tüm ölçütlerde 1 değerine ulaşması, modelin aşırı uyum riski taşıdığını güçlü biçimde işaret etmiştir. Bu kadar yüksek doğruluk ve F1 skorları, verideki varyansı fazla öğrenen modellerin, yeni veride beklenenden daha düşük performans gösterebilme riskini barındırmaktadır. Öte yandan SOS-NB ve SRUS-NB gibi daha basit algoritmalar, güçlü modellere göre oldukça düşük performans göstererek sınıflandırma doğruluğunda belirgin bir fark yaratmıştır.

STL, SROS, SSM, SSNC, SSN, SSVM ve SBS gruplarında tablo genel hatlarıyla aynıdır. Random Forest, SVM ve BBC modelleri birçok durumda yüksek doğruluk ve F1 skoru ile öne çıkmış, hatta bazı durumlarda F1=1 gibi ideal sonuçlar elde edilmiştir. Bu kadar uç değerler, model başarısının doğruluğuna dair umut verse de, aynı zamanda modelin veriye fazla uyum sağlamış olabileceği ihtimalini gündeme getirmekte, bağımsız veri setlerinde test edilmeden mutlak güvenilirliğin sağlanamayacağını ortaya koymaktadır.

Sonuç olarak, genel eğilim, gelişmiş ve hibrit yaklaşımların tahmin gücü ve doğrulukta belirgin avantaj sağladığı yönündedir. Özellikle çalışmanızda önerilen ve son model olarak seçilen Siamese Network, SMOTE-Tomek ve SVM kombinasyonu, yaklaşık 0,994 F1 skoru ve ROC'un 1'e yakın değeri ile, hem yüksek doğruluk hem de güçlü ayırt edicilik sunmaktadır. Bunun yanında, tamamen ideal skorlar üreten modellere kıyasla biraz daha dengeli kalması, aşırı uyum riskini azaltan bir avantaj sağlamaktadır. Bu özellikleriyle model, pratik uygulamalarda hem güvenilir hem de geçerli bir tercih olarak öne çıkmaktadır.

Tablo 4. 31: EN veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
SCC-KNN	0.981413	0.97417	0.977778	0.979695	0.979272	**SCN-KNN**	0.988764	0.649446	0.783964	0.835871	0.821598
SCC-SVM	1	0.99262	0.996296	0.996616	0.99631	**SCN-SVM**	1	0.948339	0.973485	0.976311	0.97417
SCC-DT	0.957692	0.918819	0.937853	0.944162	0.942222	**SCN-DT**	1	0.745387	0.854123	0.883249	0.872694
SCC-RF	1	0.99631	0.998152	0.998308	0.998155	**SCN-RF**	1	0.904059	0.949612	0.956007	0.95203
SCC-NN	0.959677	0.878229	0.917148	0.927242	0.923489	**SCN-NN**	1	0.505535	0.671569	0.773266	0.752768
SCC-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SCN-AB**	1	0.590406	0.742459	0.812183	0.795203
SCC-NB	0.772532	0.664207	0.714286	0.756345	0.749291	**SCN-NB**	0.768182	0.623616	0.688391	0.741117	0.732121
SCC-QDA	0.974026	0.830258	0.896414	0.912014	0.905754	**SCN-QDA**	0.962025	0.841328	0.897638	0.912014	0.906602
SCC-LDA	0.940711	0.878229	0.908397	0.918782	0.915677	**SCN-LDA**	0.985	0.726937	0.836518	0.869712	0.858781
SCC-EEC	0.957031	0.904059	0.929791	0.937394	0.934842	**SCN-EEC**	0.927126	0.845018	0.88417	0.898477	0.894384
SCC-RSC	0.978378	0.667897	0.79386	0.840948	0.827698	**SCN-RSC**	1	0.357934	0.527174	0.705584	0.678967
SCC-BBC	1	0.98893	0.994434	0.994924	0.994465	**SCN-BBC**	0.947368	0.863469	0.903475	0.915398	0.911422
SCC-BRFC	0.938865	0.793358	0.86	0.881557	0.874804	**SCN-BRFC**	0.914286	0.826568	0.868217	0.884941	0.880472
SEN-KNN	0.984314	0.926199	0.954373	0.959391	0.95685	**SRN-KNN**	0.984314	0.926199	0.954373	0.959391	0.95685
SEN-SVM	1	0.9631	0.981203	0.98308	0.98155	**SRN-SVM**	1	0.9631	0.981203	0.98308	0.98155
SEN-DT	0.991342	0.845018	0.912351	0.92555	0.919384	**SRN-DT**	0.991342	0.845018	0.912351	0.92555	0.919384
SEN-RF	1	0.97417	0.986916	0.988156	0.987085	**SRN-RF**	1	0.97417	0.986916	0.988156	0.987085
SEN-NN	0.96	0.885609	0.921305	0.930626	0.927179	**SRN-NN**	0.96	0.885609	0.921305	0.930626	0.927179
SEN-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SRN-AB**	0.978378	0.667897	0.79386	0.840948	0.827698
SEN-NB	0.775862	0.664207	0.715706	0.758037	0.750853	**SRN-NB**	0.775862	0.664207	0.715706	0.758037	0.750853
SEN-QDA	0.974138	0.833948	0.898608	0.913706	0.907599	**SRN-QDA**	0.974138	0.833948	0.898608	0.913706	0.907599
SEN-LDA	0.94902	0.892989	0.920152	0.928934	0.926182	**SRN-LDA**	0.94902	0.892989	0.920152	0.928934	0.926182
SEN-EEC	0.96124	0.915129	0.937618	0.944162	0.94194	**SRN-EEC**	0.96124	0.915129	0.937618	0.944162	0.94194
SEN-RSC	0.980296	0.734317	0.839662	0.871404	0.860909	**SRN-RSC**	0.978378	0.667897	0.79386	0.840948	0.827698
SEN-BBC	1	0.97786	0.988806	0.989848	0.98893	**SRN-BBC**	1	0.97786	0.988806	0.989848	0.98893
SEN-BRFC	0.938525	0.845018	0.88932	0.903553	0.899072	**SRN-BRFC**	0.938525	0.845018	0.88932	0.903553	0.899072
SSTK-KNN	0.97417	0.97417	0.97417	0.976311	0.976147	**SNC-KNN**	0.996109	0.944649	0.969697	0.972927	0.970762
SSTK-SVM	1	0.98893	0.994434	0.994924	0.994465	**SNC-SVM**	1	0.97048	0.985019	0.986464	0.98524
SSTK-DT	0.971429	0.878229	0.922481	0.932318	0.928177	**SNC-DT**	0.991342	0.845018	0.912351	0.92555	0.919384
SSTK-RF	0.99631	0.99631	0.99631	0.996616	0.996592	**SNC-RF**	1	0.97786	0.988806	0.989848	0.98893
SSTK-NN	0.942529	0.907749	0.924812	0.932318	0.930437	**SNC-NN**	0.953307	0.904059	0.92803	0.935702	0.93328
SSTK-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SNC-AB**	0.857143	0.841328	0.849162	0.862944	0.861289
SSTK-NB	0.770213	0.667897	0.715415	0.756345	0.749573	**SNC-NB**	0.786026	0.664207	0.72	0.763113	0.755541
SSTK-QDA	0.974576	0.848708	0.907298	0.920474	0.914979	**SNC-QDA**	0.974026	0.830258	0.896414	0.912014	0.905754
SSTK-LDA	0.941634	0.892989	0.916667	0.92555	0.923057	**SNC-LDA**	0.948413	0.881919	0.913958	0.923858	0.920647
SSTK-EEC	0.939394	0.915129	0.927103	0.93401	0.932565	**SNC-EEC**	0.965251	0.922509	0.943396	0.949239	0.947192
SSTK-RSC	0.978378	0.667897	0.79386	0.840948	0.827698	**SNC-RSC**	0.980296	0.734317	0.839662	0.871404	0.860909
SSTK-BBC	0.992593	0.98893	0.990758	0.99154	0.99134	**SNC-BBC**	1	0.98155	0.990689	0.99154	0.990775
SSTK-BRFC	0.922449	0.833948	0.875969	0.891709	0.887287	**SNC-BRFC**	0.936709	0.819188	0.874016	0.891709	0.886157
SOS-KNN	0.981273	0.96679	0.973978	0.976311	0.975582	**SRUS-KNN**	0.981481	0.97786	0.979667	0.981387	0.981117
SOS-SVM	1	0.98524	0.992565	0.993232	0.99262	**SRUS-SVM**	1	0.98893	0.994434	0.994924	0.994465
SOS-DT	0.975104	0.867159	0.917969	0.928934	0.924204	**SRUS-DT**	0.971429	0.878229	0.922481	0.932318	0.928177
SOS-RF	1	0.99262	0.996296	0.996616	0.99631	**SRUS-RF**	1	1	1	1	1
SOS-NN	0.935606	0.911439	0.923364	0.930626	0.929157	**SRUS-NN**	0.913357	0.933579	0.923358	0.928934	0.92929
SOS-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SRUS-AB**	0.978378	0.667897	0.79386	0.840948	0.827698
SOS-NB	0.776824	0.667897	0.718254	0.759729	0.752698	**SRUS-NB**	0.770213	0.667897	0.715415	0.756345	0.749573
SOS-QDA	0.974026	0.830258	0.896414	0.912014	0.905754	**SRUS-QDA**	0.974026	0.830258	0.896414	0.912014	0.905754
SOS-LDA	0.941406	0.889299	0.914611	0.923858	0.921212	**SRUS-LDA**	0.945098	0.889299	0.91635	0.92555	0.922774
SOS-EEC	0.953668	0.911439	0.932075	0.939086	0.93697	**SRUS-EEC**	0.957529	0.915129	0.935849	0.94247	0.940377
SOS-RSC	0.978378	0.667897	0.79386	0.840948	0.827698	**SRUS-RSC**	0.978378	0.667897	0.79386	0.840948	0.827698
SOS-BBC	1	0.98524	0.992565	0.993232	0.99262	**SRUS-BBC**	1	0.99631	0.998152	0.998308	0.998155
SOS-BRFC	0.93617	0.811808	0.869565	0.888325	0.882467	**SRUS-BRFC**	0.934694	0.845018	0.887597	0.901861	0.897509

Tablo 4. 32: EN veri kümesi üzerinde GAAC-DDE özellikleriyle eğitilen Siyam modellerinin performansı

	Precision	Recall	F1-Score	Accuracy	ROC		Precision	Recall	F1-Score	Accuracy	ROC
STL-KNN	0.981273	0.96679	0.973978	0.976311	0.975582	**SROS-KNN**	0.981481	0.97786	0.979667	0.981387	0.981117
STL-SVM	1	0.98524	0.992565	0.993232	0.99262	**SROS-SVM**	1	0.98893	0.994434	0.994924	0.994465
STL-DT	0.971429	0.878229	0.922481	0.932318	0.928177	**SROS-DT**	0.971429	0.878229	0.922481	0.932318	0.928177
STL-RF	1	0.99262	0.996296	0.996616	0.99631	**SROS-RF**	1	1	1	1	1
STL-NN	0.92	0.933579	0.92674	0.932318	0.932415	**SROS-NN**	0.974576	0.848708	0.907298	0.920474	0.914979
STL-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SROS-AB**	0.978378	0.667897	0.79386	0.840948	0.827698
STL-NB	0.776824	0.667897	0.718254	0.759729	0.752698	**SROS-NB**	0.770213	0.667897	0.715415	0.756345	0.749573
STL-QDA	0.974026	0.830258	0.896414	0.912014	0.905754	**SROS-QDA**	0.974026	0.830258	0.896414	0.912014	0.905754
STL-LDA	0.941406	0.889299	0.914611	0.923858	0.921212	**SROS-LDA**	0.945098	0.889299	0.91635	0.92555	0.922774
STL-EEC	0.960784	0.904059	0.931559	0.939086	0.936405	**SROS-EEC**	0.957529	0.915129	0.935849	0.94247	0.940377
STL-RSC	0.978378	0.667897	0.79386	0.840948	0.827698	**SROS-RSC**	0.978378	0.667897	0.79386	0.840948	0.827698
STL-BBC	0.996296	0.99262	0.994455	0.994924	0.994747	**SROS-BBC**	0.996296	0.99262	0.994455	0.994924	0.994747
STL-BRFC	0.938017	0.837638	0.88499	0.900169	0.895382	**SROS-BRFC**	0.937759	0.833948	0.882813	0.898477	0.893537
SSM-KNN	0.981481	0.97786	0.979667	0.981387	0.981117	**SSNC-KNN**	0.981481	0.97786	0.979667	0.981387	0.981117
SSM-SVM	1	0.98893	0.994434	0.994924	0.994465	**SSNC-SVM**	1	0.98893	0.994434	0.994924	0.994465
SSM-DT	0.971429	0.878229	0.922481	0.932318	0.928177	**SSNC-DT**	0.971429	0.878229	0.922481	0.932318	0.928177
SSM-RF	1	1	1	1	1	**SSNC-RF**	1	1	1	1	1
SSM-NN	0.974576	0.848708	0.907298	0.920474	0.914979	**SSNC-NN**	0.974576	0.848708	0.907298	0.920474	0.914979
SSM-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SSNC-AB**	0.978378	0.667897	0.79386	0.840948	0.827698
SSM-NB	0.770213	0.667897	0.715415	0.756345	0.749573	**SSNC-NB**	0.770213	0.667897	0.715415	0.756345	0.749573
SSM-QDA	0.974026	0.830258	0.896414	0.912014	0.905754	**SSNC-QDA**	0.974026	0.830258	0.896414	0.912014	0.905754
SSM-LDA	0.945098	0.889299	0.91635	0.92555	0.922774	**SSNC-LDA**	0.945098	0.889299	0.91635	0.92555	0.922774
SSM-EEC	0.957529	0.915129	0.935849	0.94247	0.940377	**SSNC-EEC**	0.957529	0.915129	0.935849	0.94247	0.940377
SSM-RSC	0.978378	0.667897	0.79386	0.840948	0.827698	**SSNC-RSC**	0.978378	0.667897	0.79386	0.840948	0.827698
SSM-BBC	0.996296	0.99262	0.994455	0.994924	0.994747	**SSNC-BBC**	0.996296	0.99262	0.994455	0.994924	0.994747
SSM-BRFC	0.937759	0.833948	0.882813	0.898477	0.893537	**SSNC-BRFC**	0.937759	0.833948	0.882813	0.898477	0.893537
SSN-KNN	0.981481	0.97786	0.979667	0.981387	0.981117	**SSVM-KNN**	0.981481	0.97786	0.979667	0.981387	0.981117
SSN-SVM	1	0.98893	0.994434	0.994924	0.994465	**SSVM-SVM**	1	0.98893	0.994434	0.994924	0.994465
SSN-DT	0.971429	0.878229	0.922481	0.932318	0.928177	**SSVM-DT**	0.971429	0.878229	0.922481	0.932318	0.928177
SSN-RF	1	1	1	1	1	**SSVM-RF**	1	1	1	1	1
SSN-NN	0.974576	0.848708	0.907298	0.920474	0.914979	**SSVM-NN**	0.974576	0.848708	0.907298	0.920474	0.914979
SSN-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SSVM-AB**	0.978378	0.667897	0.79386	0.840948	0.827698
SSN-NB	0.773504	0.667897	0.716832	0.758037	0.751136	**SSVM-NB**	0.770213	0.667897	0.715415	0.756345	0.749573
SSN-QDA	0.974026	0.830258	0.896414	0.912014	0.905754	**SSVM-QDA**	0.974026	0.830258	0.896414	0.912014	0.905754
SSN-LDA	0.945098	0.889299	0.91635	0.92555	0.922774	**SSVM-LDA**	0.945098	0.889299	0.91635	0.92555	0.922774
SSN-EEC	0.950382	0.918819	0.934334	0.940778	0.939097	**SSVM-EEC**	0.957529	0.915129	0.935849	0.94247	0.940377
SSN-RSC	0.978378	0.667897	0.79386	0.840948	0.827698	**SSVM-RSC**	0.978378	0.667897	0.79386	0.840948	0.827698
SSN-BBC	0.996296	0.99262	0.994455	0.994924	0.994747	**SSVM-BBC**	0.996296	0.99262	0.994455	0.994924	0.994747
SSN-BRFC	0.934426	0.841328	0.885437	0.900169	0.895664	**SSVM-BRFC**	0.934426	0.841328	0.885437	0.900169	0.895664
SBS-KNN	0.981481	0.97786	0.979667	0.981387	0.981117	**SBS-QDA**	0.974026	0.830258	0.896414	0.912014	0.905754
SBS-SVM	1	0.98893	0.994434	0.994924	0.994465	**SBS-LDA**	0.945098	0.889299	0.91635	0.92555	0.922774
SBS-DT	0.971429	0.878229	0.922481	0.932318	0.928177	**SBS-EEC**	0.957529	0.915129	0.935849	0.94247	0.940377
SBS-RF	1	1	1	1	1	**SBS-RSC**	0.978378	0.667897	0.79386	0.840948	0.827698
SBS-NN	0.974576	0.848708	0.907298	0.920474	0.914979	**SBS-BBC**	0.996296	0.99262	0.994455	0.994924	0.994747
SBS-AB	0.978378	0.667897	0.79386	0.840948	0.827698	**SBS-BRFC**	0.937759	0.833948	0.882813	0.898477	0.893537
SBS-NB	0.770213	0.667897	0.715415	0.756345	0.749573						

4.5.4 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

Bu çalışmada, enzim-protein etkileşimlerine ait veriler PSEAAC ve PSEPSSM amino asit özellik çıkarım teknikleri kullanılarak temsil edilmiş ve ardından Siamese ağı aracılığıyla dengelenmiştir. Devamında, sınıflandırma performansını değerlendirmek amacıyla çeşitli yeniden örnekleme yöntemleri ve makine öğrenmesi algoritmaları uygulanmıştır.

Tablo 4.33 ve 4.34'de sunulan sonuçlara göre, SCC gibi basit gürültü giderme tekniklerinin uygulandığı gruplarda modeller zayıf bir performans sergilemiştir. Örneğin, SCC-RF modeli yalnızca yaklaşık %16 doğruluk elde etmiş, SCC-SVM ise yaklaşık %4 gibi oldukça düşük bir F1 puanı almıştır. Bu bulgular, yalnızca basit gürültü azaltma yöntemlerinin karmaşık modellerin performansını artırmak için yeterli olmadığını ve bu tür işlemlerin veri setinden anlamlı bilgileri de silebileceğini göstermektedir. Buna karşılık, daha gelişmiş gürültü azaltma tekniklerinin kullanıldığı SCN grubunda, modeller çok daha başarılı sonuçlar elde etmiştir. Özellikle SCN-SVM modeli, yaklaşık %99,4 doğruluk ve %87 F1 skoru ile yüksek bir performans ortaya koyarak, doğru yapılandırılmış ön işleme adımlarının sınıflandırma başarısını doğrudan etkilediğini göstermiştir.

SEN gibi karma dengeleme yöntemlerinin kullanıldığı gruplarda , SVM ve RF gibi modeller, dengesiz verilerle eğitilen versiyonlarına kıyasla daha iyi performans göstermiştir. Özellikle SEN-SVM modeli %40'ın üzerinde bir F1 skoru elde etmeyi başarmıştır. Buna karşılık, yalnızca rastgele undersampling kullanılan SRN grubunda, aynı modellerin performansı düşmüş ve F1 skorları yaklaşık %20 seviyesinde kalmıştır. Bu karşılaştırma, SEN gibi daha gelişmiş dengeleme yöntemlerinin modellerin ayırt etme yeteneğini artırmada olumlu bir etkiye sahip olduğunu ortaya koymuştur.

Siamese Network ve SMOTE-Tomek kombinasyonunu kullanan SSTK grubuna yönelik analiz, bu yaklaşımın makine öğrenmesi modellerinin performansını artırmada son derece etkili olduğunu ortaya koymuştur. Bu grupta, SSTK-SVM ve SSTK-RF gibi modeller yaklaşık %99 F1 skoru ve doğruluk elde ederek dikkat çekici sonuçlar vermiştir. Derin öğrenme tabanlı bir yapı olan Siamese Network'ün, verilerin yapısal temsilini güçlendirmesi ve SMOTE-Tomek gibi akıllı yeniden örnekleme yöntemlerinin veri setindeki sınıf dengesizliğini gidermesi, birlikte kullanıldığında hem veri dengesini sağlamış hem de modelin genellenebilirliğini

artırmıştır. Bu nedenle, söz konusu kombinasyon, dengesiz ve gürültülü veri kümelerinde karmaşık modellerin başarımını optimize etmek için güçlü bir strateji olarak değerlendirilebilir.

SOS ve SRUS dâhil olmak üzere diğer grupların analizinde de benzer eğilimler tespit edilmiştir. SOS grubunda SVM modeli yaklaşık %87 F1 skoru ile tatmin edici bir performans sergilerken, rastgele örnek azaltma yaklaşımının uygulandığı SRUS grubunda modellerin performansı belirgin ölçüde azalmış, özellikle SRUS-SVM modelinde F1 skoru %16 seviyesine kadar gerilemiştir. Aynı durum STL, SROS, SSN ve SSVM gruplarında da gözlemlenmiş; bu gruplarda SVM, Random Forest ve BBC modelleri istikrarlı sonuçlar elde ederken, öte yandan Naive Bayes, AdaBoost ve RSC modellerinde değerlendirme metriklerinde dikkate değer bir düşüş yaşanmıştır.

SSM ve SSNC gruplarına yönelik değerlendirmeler de, veri yapısının karmaşıklığının artmasına ve çeşitli özellik kombinasyonları ile yeniden örnekleme stratejilerinin kullanılmasına rağmen, Random Forest ve SVM modellerinin yüksek doğruluk ve kararlılığını koruduğunu göstermektedir. Bazı senaryolarda BBC, QDA ve EEC algoritmaları da makul düzeyde başarı sağlamış olsa da, genel anlamda daha basit yapılı modellerin, dengelenmiş ve ön işleme tabi tutulmuş veri kümeleri karşısında zayıf kaldığı sonucuna varılmıştır.

Sonuç olarak elde edilen bulgular, SSTK grubunda Siamese Network ile SMOTE-Tomek yönteminin birleşiminin, özellikle SVM modeliyle birlikte kullanıldığında, doğruluk, geri çağırma, F1 skoru ve genelleme kabiliyeti arasında güçlü bir denge sağladığını ortaya koymaktadır. Bu yaklaşım yalnızca değerlendirme metriklerini ideal değerlere oldukça yakın seviyelere taşımakla kalmamış, aynı zamanda diğer yöntemlerle karşılaştırıldığında daha düşük bir aşırı öğrenme riski sergilemiştir. Dolayısıyla, SSTK-SVM yaklaşımı; dizilim ve yapısal özellikler temelinde enzimlerin tanımlanması ve sınıflandırılması alanında etkili ve güvenilir bir seçenek olarak değerlendirilebilir.

Tablo 4. 33: EN veri kümesi üzerinde pseAAC-psePSSM özellikleriyle eğitilen Siyam modellerinin performansı

	precision	recall	f1-score	accuracy	ROC		precision	recall	f1-score	accuracy	ROC
SCC-KNN	0.138756	0.935484	0.241667	0.881046	0.907702	SCN-KNN	0.137566	0.83871	0.236364	0.890196	0.864985
SCC-SVM	0.026724	0.736	0.052057	0.262092	0.623416	SCN-SVM	0.775	1	0.873239	0.994118	0.996998
SCC-DT	0.036131	0.781	0.069741	0.459477	0.724149	SCN-DT	0.104762	0.709677	0.182573	0.871242	0.79213
SCC-RF	0.023682	0.648	0.046269	0.164706	0.573716	SCN-RF	0.645833	1	0.78481	0.988889	0.99433
SCC-NN	0.076167	0.853	0.141553	0.754248	0.874583	SCN-NN	0.152174	0.451613	0.227642	0.937908	0.699789
SCC-AB	0.027531	0.719	0.053587	0.284314	0.634757	SCN-AB	0.645833	1	0.78481	0.988889	0.99433
SCC-NB	0.020462	0.529	0.040103	0.030065	0.505003	SCN-NB	0.075221	0.548387	0.132296	0.854248	0.70448
SCC-QDA	0.402597	0.865	0.574074	0.969935	0.984656	SCN-QDA	0.407895	1	0.579439	0.970588	0.98499
SCC-LDA	0.075061	0.842	0.13964	0.750327	0.872582	SCN-LDA	0.107692	0.451613	0.173913	0.913072	0.687114
SCC-EEC	0.025985	0.702	0.050654	0.240523	0.612408	SCN-EEC	0.076923	1	0.142857	0.756863	0.875917
SCC-RSC	0.025578	0.688	0.049879	0.228105	0.606071	SCN-RSC	0.336957	1	0.504065	0.960131	0.979653
SCC-BBC	0.028054	1	0.054577	0.298039	0.641761	SCN-BBC	0.172414	0.967742	0.292683	0.905229	0.935839
SCC-BRFC	0.026572	0.967742	0.051724	0.281046	0.617293	SCN-BRFC	0.096774	0.774194	0.172043	0.84902	0.81238
SEN-KNN	0.192593	0.83871	0.313253	0.92549	0.882997	SRN-KNN	0.121212	0.903226	0.21374	0.865359	0.883901
SEN-SVM	0.246032	1	0.394904	0.937908	0.968312	SRN-SVM	0.115672	1	0.207358	0.845098	0.920947
SEN-DT	0.128205	0.322581	0.183486	0.94183	0.638609	SRN-DT	0.1	0.451613	0.163743	0.906536	0.683778
SEN-RF	0.267241	1	0.421769	0.944444	0.971648	SRN-RF	0.130802	1	0.231343	0.865359	0.931288
SEN-NN	0.140351	0.258065	0.181818	0.952941	0.612688	SRN-NN	0.107438	0.419355	0.171053	0.917647	0.673653
SEN-AB	0.329787	1	0.496	0.958824	0.978986	SRN-AB	0.166667	1	0.285714	0.898693	0.948299
SEN-NB	0.084507	0.580645	0.147541	0.864052	0.725279	SRN-NB	0.0875	0.677419	0.154982	0.850327	0.765661
SEN-QDA	0.407895	1	0.579439	0.970588	0.98499	SRN-QDA	0.407895	1	0.579439	0.970588	0.98499
SEN-LDA	0.090909	0.290323	0.138462	0.926797	0.615141	SRN-LDA	0.086124	0.580645	0.15	0.866667	0.726613
SEN-EEC	0.070136	1	0.131078	0.731373	0.862909	SRN-EEC	0.066098	1	0.124	0.713725	0.853903
SEN-RSC	0.120155	1	0.214533	0.851634	0.924283	SRN-RSC	0.089337	1	0.164021	0.793464	0.894596
SEN-BBC	0.133047	1	0.234848	0.867974	0.932622	SRN-BBC	0.094044	0.967742	0.171429	0.810458	0.887473
SEN-BRFC	0.085324	0.806452	0.154321	0.820915	0.813833	SRN-BRFC	0.084337	0.903226	0.15427	0.799346	0.850212
SSTK-KNN	0.210884	1	0.348315	0.924183	0.961308	SNC-KNN	0.321429	0.870968	0.469565	0.960131	0.916471
SSTK-SVM	0.987	0.993	0.99	0.991	0.996	SNC-SVM	0.553571	1	0.712644	0.98366	0.991661
SSTK-DT	0.117647	0.967742	0.20979	0.852288	0.908821	SNC-DT	0.235294	0.387097	0.292683	0.962092	0.68054
SSTK-RF	0.973	0.980	1	0.987	0.995	SNC-RF	0.645833	1	0.78481	0.988889	0.99433
SSTK-NN	0.430769	0.903226	0.583333	0.973856	0.939271	SNC-NN	0.148148	0.129032	0.137931	0.96732	0.556844
SSTK-AB	0.987	0.993	0.99	0.991	0.996	SNC-AB	0.534483	1	0.696629	0.982353	0.990994
SSTK-NB	0.068592	0.612903	0.123377	0.823529	0.720394	SNC-NB	0.092784	0.580645	0.16	0.876471	0.731617
SSTK-QDA	0.534483	1	0.696629	0.982353	0.990994	SNC-QDA	0.407895	1	0.579439	0.970588	0.98499
SSTK-LDA	0.086687	0.903226	0.158192	0.805229	0.853214	SNC-LDA	0.109756	0.290323	0.159292	0.937908	0.620812
SSTK-EEC	0.087273	0.774194	0.156863	0.831373	0.803374	SNC-EEC	0.07243	1	0.135076	0.740523	0.867578
SSTK-RSC	0.958	0.976	0.967	0.964	0.982	SNC-RSC	0.121569	1	0.216783	0.853595	0.925284
SSTK-BBC	0.967742	0.967742	0.967742	0.998693	0.983537	SNC-BBC	0.142202	1	0.248996	0.877778	0.937625
SSTK-BRFC	0.085938	0.709677	0.15331	0.841176	0.776787	SNC-BRFC	0.083333	0.774194	0.15047	0.822876	0.799038
SOS-KNN	0.5	0.806452	0.617284	0.979739	0.894887	SRUS-KNN	0.043228	0.967742	0.082759	0.565359	0.76239
SOS-SVM	0.775	1	0.873239	0.994118	0.996998	SRUS-SVM	0.090379	1	0.165775	0.796078	0.895931
SOS-DT	0.384615	0.483871	0.428571	0.973856	0.73393	SRUS-DT	0.054054	0.903226	0.102004	0.677778	0.788171
SOS-RF	0.965	0.987	0.976	0.968	0.989	SRUS-RF	0.086351	1	0.158974	0.785621	0.890594
SOS-NN	0.428571	0.096774	0.157895	0.979085	0.547053	SRUS-NN	0.056391	0.967742	0.106572	0.671242	0.816426
SOS-AB	0.911765	1	0.953846	0.998039	0.998999	SRUS-AB	0.076923	1	0.142857	0.756863	0.875917
SOS-NB	0.102151	0.612903	0.175115	0.883007	0.750748	SRUS-NB	0.054299	0.774194	0.10148	0.722222	0.74767
SOS-QDA	0.407895	1	0.579439	0.970588	0.98499	SRUS-QDA	0.407895	1	0.579439	0.970588	0.98499
SOS-LDA	0.160494	0.419355	0.232143	0.943791	0.686996	SRUS-LDA	0.053846	0.903226	0.101633	0.676471	0.787503
SOS-EEC	0.075795	1	0.140909	0.752941	0.873916	SRUS-EEC	0.060852	0.967742	0.114504	0.696732	0.829435
SOS-RSC	0.132479	1	0.233962	0.86732	0.932288	SRUS-RSC	0.08311	1	0.153465	0.776471	0.885924
SOS-BBC	0.180233	1	0.305419	0.907843	0.952969	SRUS-BBC	0.074742	0.935484	0.138425	0.764052	0.847995
SOS-BRFC	0.081181	0.709677	0.145695	0.831373	0.771783	SRUS-BRFC	0.058968	0.774194	0.109589	0.745098	0.759345

Tablo 4. 34: EN veri kümesi üzerinde pseAAC-psePSSM özellikleriyle eğitilen Siyam modellerinin performansı

	precision	recall	f1-score	accuracy	ROC		precision	recall	f1-score	accuracy	ROC
STL-KNN	0.510638	0.774194	0.615385	0.980392	0.879425	SROS-KNN	0.31	1	0.473282	0.954902	0.976985
STL-SVM	0.756098	1	0.861111	0.993464	0.996664	SROS-SVM	0.942	0.951	0.946	0.949	0.971
STL-DT	0.75	0.387097	0.510638	0.984967	0.692214	SROS-DT	0.098101	1	0.178674	0.813725	0.904937
STL-RF	0.965	0.978	0.971	0.968	0.985	SROS-RF	0.987	0.993	0.99	0.991	0.996
STL-NN	0.178	0.294	0.222	0.72	0.588	SROS-NN	0.279279	1	0.43662	0.947712	0.973316
STL-AB	0.911765	1	0.953846	0.998039	0.998999	SROS-AB	0.958	0.976	0.967	0.964	0.982
STL-NB	0.098901	0.580645	0.169014	0.884314	0.735619	SROS-NB	0.065476	0.709677	0.119891	0.788889	0.750102
STL-QDA	0.407895	1	0.579439	0.970588	0.98499	SROS-QDA	0.418919	1	0.590476	0.971895	0.985657
STL-LDA	0.137931	0.258065	0.179775	0.952288	0.612354	SROS-LDA	0.076087	0.903226	0.140351	0.775817	0.838204
STL-EEC	0.07598	1	0.14123	0.753595	0.87425	SROS-EEC	0.092652	0.935484	0.168605	0.813072	0.873012
STL-RSC	0.107639	1	0.194357	0.832026	0.914276	SROS-RSC	0.935	0.944	0.939	0.938	0.955
STL-BBC	0.166667	0.903226	0.281407	0.906536	0.904915	SROS-BBC	0.911765	1	0.953846	0.998039	0.998999
STL-BRFC	0.083032	0.741935	0.149351	0.828758	0.786245	SROS-BRFC	0.081272	0.741935	0.146497	0.824837	0.784243
SSM-KNN	0.210884	1	0.348315	0.924183	0.961308	SSNC-KNN	0.223022	1	0.364706	0.929412	0.963976
SSM-SVM	0.949	0.961	0.955	0.953	0.967	SSNC-SVM	0.972	0.981	0.976	0.974	0.99
SSM-DT	0.117647	0.967742	0.20979	0.852288	0.908821	SSNC-DT	0.114894	0.870968	0.203008	0.861438	0.866104
SSM-RF	0.953	0.96	0.956	0.957	0.962	SSNC-RF	0.964	0.975	0.969	0.966	0.978
SSM-NN	0.430769	0.903226	0.583333	0.973856	0.939271	SSNC-NN	0.333333	0.903226	0.486957	0.961438	0.932934
SSM-AB	0.98	0.989	0.984	0.982	0.993	SSNC-AB	0.948	0.954	0.951	0.95	0.96
SSM-NB	0.068592	0.612903	0.123377	0.823529	0.720394	SSNC-NB	0.068841	0.612903	0.123779	0.824183	0.720728
SSM-QDA	0.534483	1	0.696629	0.982353	0.990994	SSNC-QDA	0.553571	1	0.712644	0.98366	0.991661
SSM-LDA	0.086687	0.903226	0.158192	0.805229	0.853214	SSNC-LDA	0.088608	0.903226	0.161383	0.809804	0.855549
SSM-EEC	0.087273	0.774194	0.156863	0.831373	0.803374	SSNC-EEC	0.078431	0.774194	0.142433	0.811111	0.793034
SSM-RSC	0.974	0.986	0.98	0.977	0.991	SSNC-RSC	0.94	0.95	0.945	0.943	0.958
SSM-BBC	0.967742	0.967742	0.967742	0.998693	0.983537	SSNC-BBC	0.96875	1	0.984127	0.999346	0.999666
SSM-BRFC	0.085938	0.709677	0.15331	0.841176	0.776787	SSNC-BRFC	0.077778	0.677419	0.139535	0.830719	0.755654
SSN-KNN	0.631579	0.774194	0.695652	0.986275	0.882427	SSVM-KNN	0.234375	0.967742	0.377358	0.935294	0.951183
SSN-SVM	0.991	0.994	0.993	0.992	0.998	SSVM-SVM	0.976	0.984	0.98	0.978	0.989
SSN-DT	0.5	0.354839	0.415094	0.979739	0.67375	SSVM-DT	0.125	0.645161	0.209424	0.901307	0.775883
SSN-RF	0.962	0.97	0.966	0.964	0.973	SSVM-RF	0.95	0.959	0.954	0.952	0.961
SSN-NN	0.2	0.032258	0.055556	0.977778	0.514795	SSVM-NN	0.25	0.967742	0.397351	0.940523	0.953851
SSN-AB	0.981	0.988	0.984	0.983	0.994	SSVM-AB	0.97	0.975	0.972	0.971	0.98
SSN-NB	0.385	0.41	0.397	0.684	0.545	SSVM-NB	0.086705	0.483871	0.147059	0.886275	0.689234
SSN-QDA	1	0.258065	0.410256	0.984967	0.629032	SSVM-QDA	0.6	0.967742	0.740741	0.986275	0.9772
SSN-LDA	0.2	0.064516	0.097561	0.975817	0.52959	SSVM-LDA	0.077821	0.645161	0.138889	0.837908	0.743528
SSN-EEC	0.347826	0.258065	0.296296	0.975163	0.624029	SSVM-EEC	0.084677	0.677419	0.150538	0.845098	0.762993
SSN-RSC	0.955	0.963	0.959	0.957	0.968	SSVM-RSC	0.938	0.945	0.941	0.94	0.954
SSN-BBC	1	0.870968	0.931034	0.997386	0.935484	SSVM-BBC	0.966667	0.935484	0.95082	0.998039	0.967408
SSN-BRFC	0.214	0.329	0.261	0.728	0.612	SSVM-BRFC	0.096386	0.516129	0.162437	0.892157	0.708031
SBS-KNN	0.212766	0.967742	0.348837	0.926797	0.946846	SBS-QDA	0.5	0.967742	0.659341	0.979739	0.973864
SBS-SVM	0.984	0.992	0.988	0.986	0.995	SBS-LDA	0.093248	0.935484	0.169591	0.814379	0.873679
SBS-DT	0.100694	0.935484	0.181818	0.829412	0.881351	SBS-EEC	0.100402	0.806452	0.178571	0.849673	0.828509
SBS-RF	0.966	0.973	0.969	0.968	0.977	SBS-RSC	0.96	0.97	0.965	0.963	0.972
SBS-NN	0.309278	0.967742	0.46875	0.955556	0.961523	SBS-BBC	1	0.967742	0.983607	0.999346	0.983871
SBS-AB	0.992	0.996	0.994	0.993	0.999	SBS-BRFC	0.096154	0.806452	0.171821	0.842484	0.82484
SBS-NB	0.076613	0.612903	0.136201	0.842484	0.730067						

4.6 ÖNERİLEN MODELİN PERFORMANSININ ROC VE RECALL EĞRİLERİ KULLANILARAK ANALİZ EDİLMESİ

DTI tahmin etmeye yönelik önerilen modellerin performansını doğru ve kapsamlı bir şekilde değerlendirmek amacıyla bu çalışmada NR , GPCR , IC ve EN olmak üzere dört standart veri seti kullanılmıştır. İlaç ve protein verilerinden biyoinformatik bilgi elde etmek için AAC, DPC, GAAC, DDE, CKSAAGP, GDPC, GTPC, PSEAAC, PSEPSSM ve bunların çeşitli kombinasyonlarından oluşan toplam on altı farklı sayısal özellik kümesi (tanımlayıcı) kullanılmıştır. Bu özellikler temel alınarak, her bir veri seti için AUC ve F1-Skoru değerlerinin yanı sıra ROC ve Hassasiyet-Geri Çağırma (Precision-Recall) eğrilerini içeren dört ayrı grafik olmak üzere, toplamda on altı grafik oluşturulmuştur. Tüm veri kümeleri, önce Siyam Ağı ile dengelenmiş, ardından sınıf dengesizliğinin etkisini azaltmak ve sınıflandırma başarımını artırmak amacıyla SMOTE-Tomek gibi yeniden örnekleme yöntemleri uygulanmıştır. Elde edilen bu grafikler, modellerin başarı düzeylerinin analiz edilmesi ve farklı makine öğrenimi algoritmalarının nicel olarak karşılaştırılabilmesi açısından temel bir değerlendirme aracı olarak kullanılmıştır.

4.6.1 İyon Kanalları Veriseti: Önerilen Modelin ROC & Recall Analizi

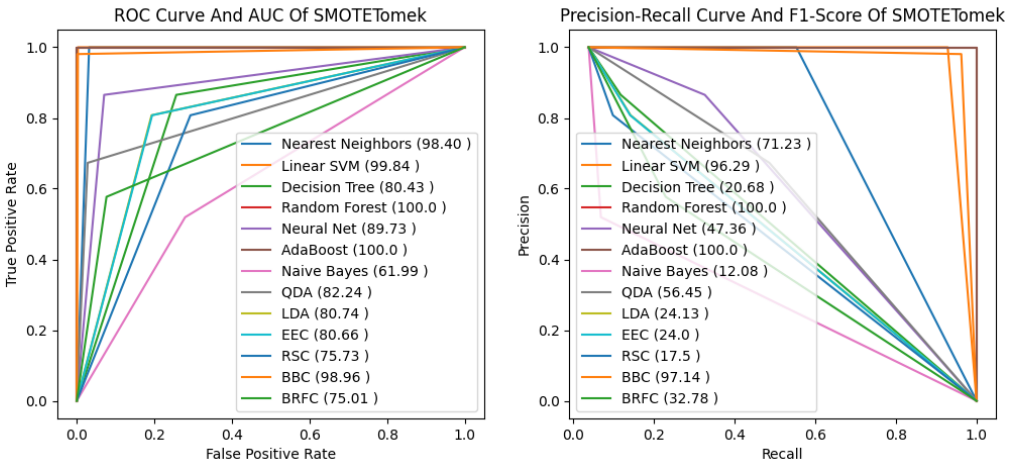
Bu bölümde, IC veri seti dört farklı özellik türü (AAC-DPC, GAAC-DDE, CKSAAGP-GDPC-GTPC ve PSEAAC-PSEPSSM) kullanılarak değerlendirilmiş ve makine öğrenimi modellerinin performansları ROC eğrisi ve duyarlılık (Recall) ölçütleri temelinde karşılaştırılmıştır.

4.6.1.1 AAC_DPC Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.4'de, CKSAAGP-GDPC-GTPC özellikleriyle temsil edilen IC verileri üzerinde SMOTE-Tomek yöntemi uygulandıktan sonra çeşitli makine öğrenimi modellerinin performansı değerlendirilmiştir. Şekil 1'in sol tarafında, ROC eğrileri ve her bir model için AUC değerleri sunulmaktadır. Random Forest ve AdaBoost algoritmaları, AUC değeri 1.00 ile en yüksek başarıyı göstermekte olup, sınıfları hatasız ayırma yetenekleri ideal bir performansa işaret etmekle birlikte, aşırı öğrenme (overfitting) riskini de beraberinde getirmektedir. Linear SVM modeli de 99.84 AUC değeri ile son derece başarılı bir performans sergilemiştir. Nearest Neighbors (98.40) ve Neural Network (89.73) gibi diğer modeller de tatmin edici sonuçlar elde

etmiştir; ancak Naive Bayes (61.99) ve RSC (75.73) gibi bazı modeller sınıfları ayırt etme konusunda zayıf kalmıştır.

Şekil 4.4'in sağ tarafında ise Precision-Recall eğrileri ve F1 skorları yer almaktadır. Random Forest ve AdaBoost burada da yüksek başarı göstererek F1 skoru olarak 1.00 elde etmişlerdir. Linear SVM yaklaşık 96.29'luk F1 skoru ile sadece doğruluk açısından değil, Precision ve Recall dengesi bakımından da üstün performans ortaya koymuştur. Naive Bayes (12.08) ve Decision Tree (20.68) gibi modeller ise pozitif ve negatif örnekleri ayırt etmede yeterli başarıyı sağlayamamıştır. Genel olarak bu analiz, SMOTE-Tomek gibi veri dengeleme yöntemlerinin güçlü makine öğrenimi algoritmalarıyla birlikte uygulanmasının, ilaç-hedef etkileşim verilerinin daha etkili bir şekilde sınıflandırılmasını mümkün kılabileceğini göstermektedir.



Şekil 4. 4: IC Veriseti AAC DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

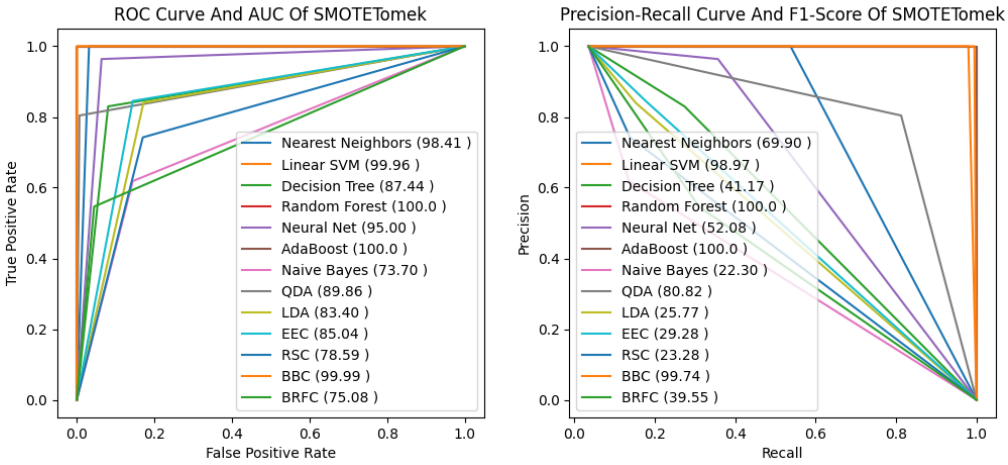
4.6.1.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Bu bölümde sunulan eğriler, GAAC-DDE özellikleri kullanılarak temsil edilen IC veri kümesi üzerinde, SMOTE-Tomek yöntemiyle dengesizlik problemi giderildikten sonra eğitilen makine öğrenimi modellerinin sınıflandırma performanslarını ortaya koymaktadır. Şekil 4.5 sol tarafında ROC eğrisi, Random Forest ve AdaBoost algoritmalarının, AUC değeri 1.00 ile sınıflar arasında üst düzey bir ayrım yeteneği gerçekleştirebildiğini açıkça göstermektedir. BBC algoritması da 99.99'luk AUC değeriyle, bu iki modelle oldukça yakın bir başarı sergilemiştir. Linear SVM (99.96) ve Neural Network (95.00) modellerinin yüksek AUC değerleri, bu

modellerin güçlü genelleme kabiliyetine sahip olduğunu ortaya koymaktadır. Buna karşın, Naive Bayes (73.70), RSC (78.59) ve Decision Tree (87.44) gibi modeller, nispeten düşük AUC değerleriyle sınıfları ayırt etmede daha zayıf bir performans göstermiştir.

Şekil 4.5'in sağ tarafında yer alan Precision-Recall eğrisi, incelenen modellerin doğruluk (Precision) ile geri çağırma (Recall) arasındaki dengeyi sağlama konusundaki yeterliliklerini değerlendirmeye olanak tanımaktadır. ROC eğrisinde olduğu gibi, Random Forest ve AdaBoost algoritmaları burada da F1 skoru 1.00 ile en yüksek başarıyı elde etmiş, BBC ise F1 skoru 99.74 ile ideal performansa oldukça yakın bir sonuç vermiştir. Linear SVM, 98.97'lik F1 skoru ile pozitif ve negatif örnekleri ayırt etmede oldukça etkili olduğunu göstermektedir. Öte yandan, Naive Bayes (22.30) ve RSC (23.88) gibi modellerin düşük F1 skorları, bu algoritmaların karmaşık sınıf dağılımlarında istenilen performansı sağlayamadığını göstermektedir.

Elde edilen bulgular, yeniden örnekleme tekniklerinin dikkatli uygulanmasının ve güçlü modellerin tercih edilmesinin, tahmin doğruluğunu önemli ölçüde artırabileceğini göstermektedir.Bu yaklaşım, bu tez çalışmasında önerilen yönteme benzer şekilde; verilerin önce Siyam ağı ile ön dengelenmesini, ardından SMOTE-Tomek yöntemiyle yeniden örnekleme uygulanmasını ve son aşamada SVM ile model eğitiminin gerçekleştirilmesini içeren bir stratejiyi yansıtmaktadır.



Şekil 4. 5: IC Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.1.3 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

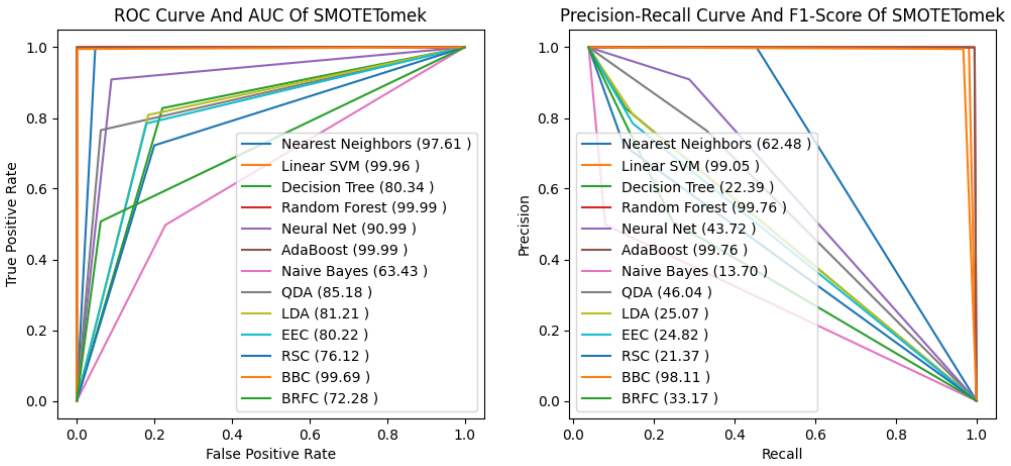
Bu bölümde sunulan Şekil 4.6 , CKSAAGP-GDPC-GTPC özelliklerine sahip IC-drug veri kümesine aittir. İlgili veriler, öncelikle Siyam ağı ile ilk dengeleme işlemine tabi tutulmuş, ardından SMOTE-Tomek yeniden örnekleme yöntemi uygulanarak analiz için uygun hale getirilmiştir.

Şekil 4.3'ün sol tarafında yer alan ROC eğrisi, farklı makine öğrenimi algoritmalarının, ilaç-hedef etkileşimlerini sınıflandırmadaki ayırt edici gücünü göstermektedir. Elde edilen sonuçlara göre, Random Forest ve AdaBoost modelleri, %99.99 AUC değeri ile neredeyse kusursuz bir performans sergileyerek sınıflar arasında yüksek doğrulukla ayrım gerçekleştirmiştir. Bununla birlikte, Linear SVM modeli de %99.96 AUC ile yüksek genelleme yeteneğini ve doğruluğunu ortaya koymuş; veri setinin yapısal karmaşıklığı göz önünde bulundurulduğunda, bu tez kapsamında önerilen yaklaşım için ideal bir sınıflandırıcı olarak öne çıkmıştır. Diğer yandan, Naive Bayes (63.43), RSC (76.12) ve BRFC (72.28) gibi modeller, sınıflar arasında belirgin bir ayrım gerçekleştirmekte başarısız olmuş ve düşük AUC değerleri ile sınıf ayrımında yetersiz kalmıştır. Bu durum, söz konusu modellerin karmaşık veya dengesiz veri yapılarıyla başa çıkma konusunda sınırlı kapasiteye sahip olduğunu göstermektedir.

Şekil 4.6'nın sağ tarafında yer alan Precision-Recall eğrisi, modellerin doğruluk (Precision) ve geri çağırma (Recall) arasında kurduğu dengeyi değerlendirmektedir. Bu ölçüm özellikle

dengesiz veri kümelerinde oldukça kritik bir öneme sahiptir. Analiz sonuçları göstermektedir ki, Random Forest ve AdaBoost modelleri burada da üstün performanslarını sürdürmüş ve her iki model de %99.76 F1 skoru ile en yüksek başarıyı elde etmiştir. Linear SVM, %99.05 F1 skoru ile üst sıralarda yer alarak, yalnızca yüksek doğruluk değil, aynı zamanda yanlış pozitif oranını azaltma açısından da etkili olduğunu göstermiştir. Buna karşın, Naive Bayes (13.70), RSC (21.37) ve Decision Tree (22.39) gibi modellerin düşük F1 skorları, bu algoritmaların pozitif ve negatif örnekleri ayırt etmede yeterli başarıyı sağlayamadığını ortaya koymaktadır.

Sonuç olarak, elde edilen bulgular, bu tez çalışmasında benimsenen yaklaşımın (Siyam Ağı , SMOTE-Tomek ve SVM kombinasyonunun) tahmin doğruluğunu anlamlı düzeyde artırma potansiyeline sahip olduğunu güçlü bir şekilde doğrulamaktadır.



Şekil 4. 6: IC Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

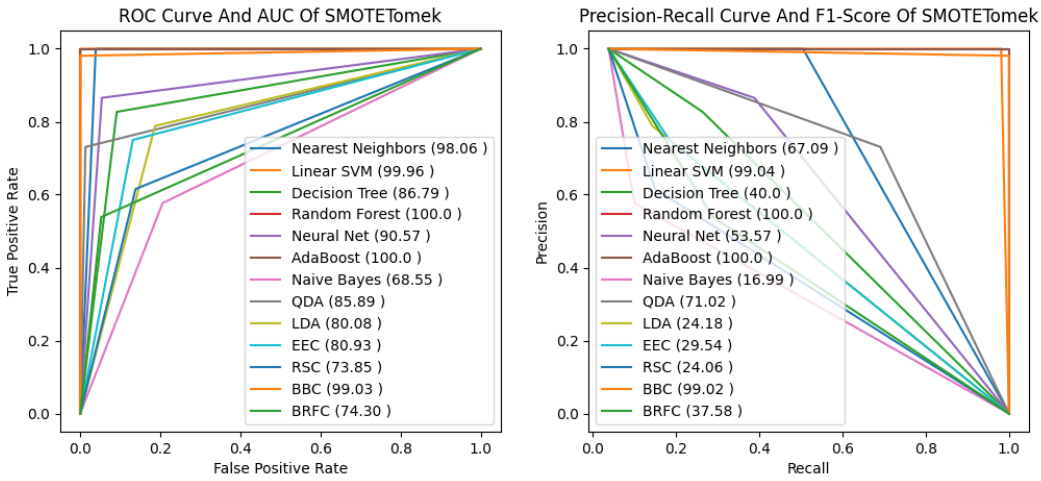
4.6.1.4 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

Sunulan grafikler, PSEAAC-PSEPSSM özelliklerine sahip IC-drug veri kümesine ait olup, öncelikle Siyam Ağı ile ön dengeleme işlemi gerçekleştirilmiş, ardından SMOTE-Tomek yöntemiyle yeniden örnekleme uygulanarak elde edilmiştir.

Şekil 4.7'in sol tarafında yer alan ROC eğrisi incelendiğinde, Random Forest ve AdaBoost gibi modellerin AUC değeri olarak 1.00 elde ettikleri ve bu bağlamda sınıfların ayrımında görünüşte kusursuz bir performans sergiledikleri görülmektedir. Ancak, biyolojik açıdan karmaşık veri kümelerinde AUC'nin 1 değeri olması çoğunlukla aşırı öğrenme göstergesi olabilir. Bu durum, modelin genel örüntüleri öğrenmek yerine eğitim verisine ezberci bir şekilde uyum sağlaması ve dolayısıyla yeni verilere karşı düşük genellenebilirlik göstermesi anlamına gelir. Buna karşın, Linear SVM modeli %99.96 AUC değeri ile tam puana oldukça yakın olmakla birlikte, daha yüksek bir genellenebilirlik ve daha istikrarlı bir performans sergilemektedir. Bu özellikleri sayesinde, önerilen tez modeli için güvenilir bir tercih olarak öne çıkmaktadır. Diğer yandan, Naive Bayes (%68.55) ve RSC (%73.85) gibi modeller, sınıf ayrımında yetersiz kalmış ve bu tür karmaşık koşullarda beklenen başarıyı sağlayamamıştır.

Şekil 4.7'in sağ tarafında yer alan Precision-Recall eğrisinde de benzer bir durum gözlemlenmektedir. Random Forest ve AdaBoost modelleri, F1 skoru açısından 1.00 ile yine zirvede yer almaktadır; ancak bu durum da modelin eğitim verisine fazla uyum sağlamasından kaynaklanan aşırı öğrenme olasılığını beraberinde getirmektedir. Buna kıyasla, Linear SVM %99.04 F1 skoru ile sadece yüksek bir başarı sergilemekle kalmamış, aynı zamanda overfitting riskini de minimize etmiştir. Bu bağlamda, önerilen çalışma sürecinde doğruluk, geri çağırma ve modelin kararlılığı arasında dengeli bir yapı sunan modellerin tercih edilmesi ön planda tutulmuştur. Naive Bayes (%16.99), LDA (%24.18) ve RSC (%24.06) gibi diğer modeller ise bu eğride zayıf performans göstermiştir.

Genel olarak, bu bulgular, Siamese Network , SMOTE-Tomek ve SVM bileşimini temel alan yaklaşımın yalnızca sayısal anlamda değil, aynı zamanda model kararlılığı ve aşırı öğrenmeyi önleme açısından da önemli avantajlar sunduğunu ortaya koymaktadır.



Şekil 4. 7: IC Veriseti PSEAAC-PSEPSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.2 G Proteinine Bağlı Reseptörler Veriseti: Önerilen Modelin ROC & Recall Analizi

Bu bölümde, GPCR veri seti dört farklı özellik türü (AAC-DPC, GAAC-DDE, CKSAAGP-GDPC-GTPC ve PSEAAC-PSEPSSM) kullanılarak değerlendirilmiş ve makine öğrenimi modellerinin ROC eğrisi ile Recall ölçütlerindeki performansları karşılaştırılmıştır.

4.6.2.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

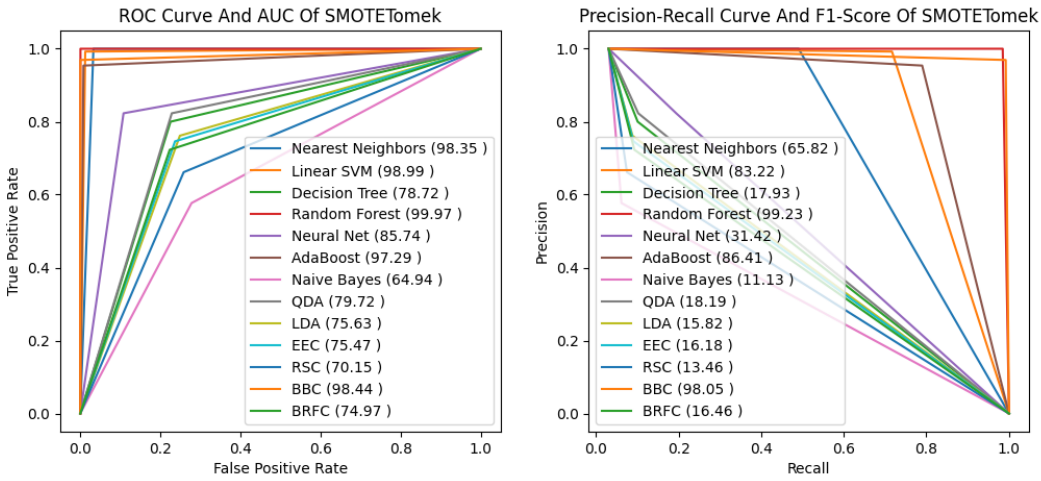
Şekil 4.8 , AAC-DPC özelliklerine sahip GPCR-drug veri setine ait olup, öncelikle Siyam Sinir Ağı (Siamese Neural Network) ile veri dengelemesi yapılmış ve ardından SMOTE-Tomek yöntemi uygulanarak farklı öğrenme modellerinin performansını değerlendirmek amacıyla oluşturulmuştur.

ROC eğrisinde, Random Forest ve AdaBoost modelleri sırasıyla %99.97 ve %97.29 AUC değerleriyle yine üst sıralarda yer almaktadır; özellikle Random Forest modeli neredeyse tam skor elde etmiştir. İlk bakışta bu performans etkileyici görünse de, AUC değerinin 1.00'a çok yakın olması genellikle aşırı öğrenme (overfitting) belirtisi olabilir. Bu durum özellikle verilerin önceden yeniden örnekleme yöntemleri ile dengelendiği durumlarda dikkatle değerlendirilmelidir. Buna karşılık, Linear SVM modeli %98.99 AUC değeriyle benzer

düzeyde başarılı olmakla beraber, daha önemli bir özelliğe sahiptir: yüksek kararlılık ve daha iyi genelleme kapasitesi. Bu özellikler, SVM modelinin tez önerisi kapsamında tercih edilmesinin temel nedenlerinden biridir; çünkü burada hedef, yalnızca dengelenmiş verilerde yüksek doğruluk elde etmek değil, bilinmeyen veri üzerinde de güvenilir performans gösterebilen bir model seçmektir.

Precision-Recall eğrisi de bu farklılığı net şekilde ortaya koymaktadır. AdaBoost ve Random Forest modelleri sırasıyla %86.41 ve %99.23 F1 skorları ile yüksek performans sergilemekte, ancak bu yüksek skorlar aşırı öğrenme riskine işaret ettiğinden, bu modellerin seçimi konusunda temkinli olunması gerekmektedir. Öte yandan, Linear SVM %83.22 F1 skoru ile birinci olmasa da, doğruluk ve geri çağırma arasında dengeli ve kabul edilebilir bir performans göstermekte ve aşırı öğrenme riski taşıyan tam skorlu modellerden daha tutarlı bir davranış sergilemektedir. Diğer modellerden Naive Bayes (%11.13) ve RSC (%13.46) ise düşük F1 skorları ile, bu spesifik özellik kombinasyonunda ilaç-hedef etkileşimlerini tespit etmede başarısız olduklarını göstermektedir.

Elde edilen bulgular, yalnızca güçlü modellerin görece üstün performansını ortaya koymakla kalmamakta, aynı zamanda önerilen modelleme sürecinde Linear SVM'nin; yüksek doğruluk, model kararlılığı ve aşırı öğrenmeden kaçınma gibi üç temel ölçüt doğrultusunda rasyonel bir tercih olduğunu da açıkça ortaya koymaktadır.

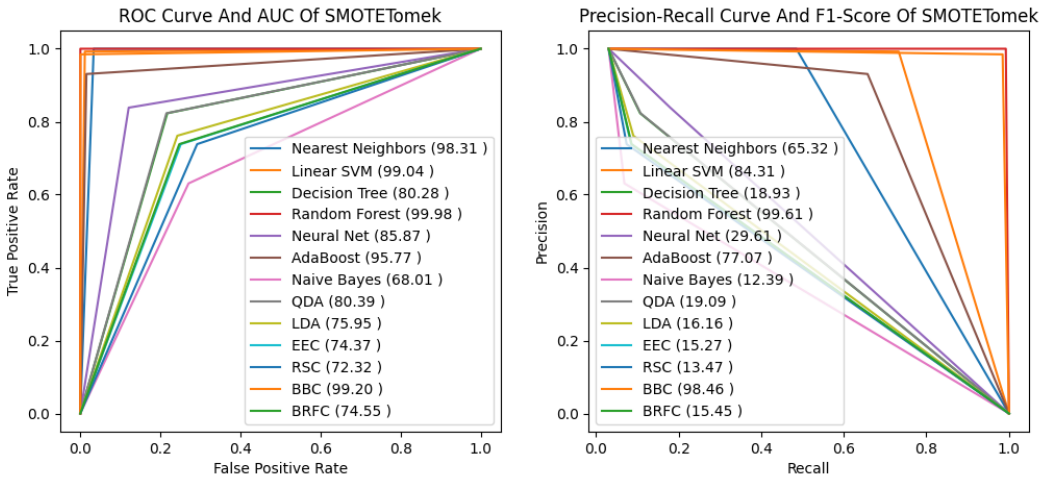


Şekil 4. 8: GPCR Veriseti AAC-DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.2.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.9'da, GAAC-DDE özellikleri kullanılarak GPCR-drug veri kümesi üzerinde eğitilen ve SMOTE-Tomek yöntemiyle dengelenen makine öğrenimi algoritmalarına ait ROC ve Precision-Recall eğrileri sunulmaktadır. ROC eğrisinde Random Forest ve BBC modelleri sırasıyla %99.98 ve %99.20 AUC değerleriyle en yüksek performansı sergilemiştir. Bu bağlamda, Linear SVM modeli %99.04 AUC ile Random Forest'a yakın bir performans sunmuş; ancak daha yüksek kararlılık ve genellenebilirlik özellikleri sayesinde bu tez kapsamında nihai model olarak tercih edilmiştir. Nearest Neighbors (%98.31), AdaBoost (%95.77) ve Neural Net (%85.87) gibi modeller de görece başarılı sonuçlar üretmiş olsalar da, genel sıralamada SVM ve Random Forest'ın gerisinde kalmışlardır. Buna karşılık, Naive Bayes (%68.01), RSC (%72.46) ve EEC (%74.37) gibi modeller tatmin edici bir performans gösterememiş ve süreç dışı bırakılmıştır.

Precision-Recall eğrisinde Random Forest (%99.61) ve BBC (%98.46) modelleri yine en yüksek F1 skorlarına ulaşmış.Linear SVM, %84.31 F1 skoru ile doğruluk ve geri çağırma arasında dengeli bir yapı sunarak, hem verimlilik hem de kararlılık açısından avantaj sağlamış ve bu nedenle nihai model olarak benimsenmiştir. Diğer yandan, Nearest Neighbors (%65.32) ve AdaBoost (%77.07) orta düzeyde performans sergilerken; Neural Net (%29.61), Naive Bayes (%12.39), QDA (%19.09) ve RSC (%13.47) gibi modeller, bu özellik kombinasyonunda ilaç-hedef etkileşimlerini tanımada yetersiz kalmış ve değerlendirme sürecinden elenmiştir.



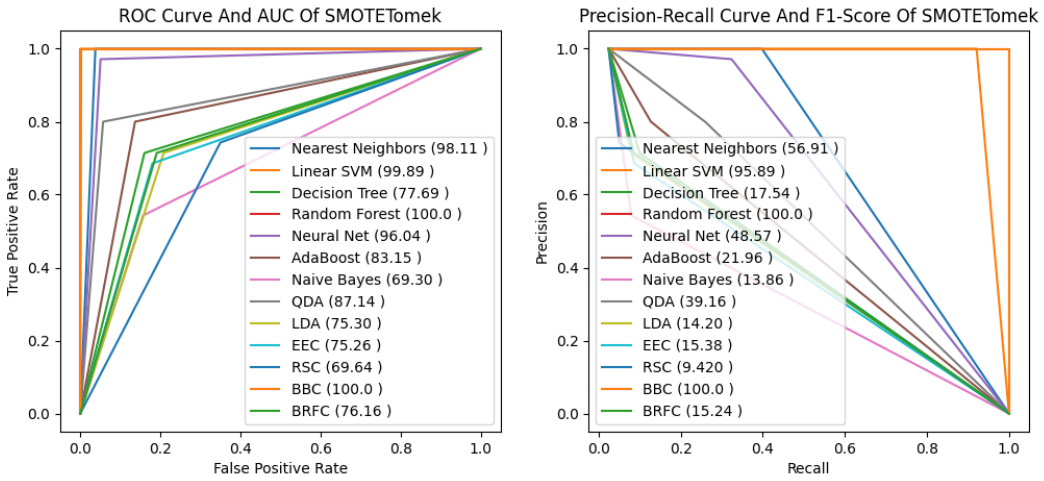
Şekil 4. 9: GPCR Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.2.3 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.10'da, CKSAAGP-GDPC-GTPC özellikleri kullanılarak GPCR-drug veri kümesi üzerinde eğitilen ve SMOTE-Tomek yöntemiyle dengelenen makine öğrenimi algoritmalarına ait ROC ve Precision-Recall eğrileri sunulmaktadır. ROC eğrilerinin veriler üzerindeki analizi, Random Forest ve BBC modellerinin %100 AUC değeriyle maksimum düzeye ulaştığını göstermektedir. Bu kusursuz performans her ne kadar istatistiksel olarak etkileyici görünse de, özellikle SMOTE-Tomek uygulamasından sonra yeniden örneklenmiş veri ortamında açık bir aşırı öğrenme belirtisi olarak değerlendirilmektedir. Buna karşılık, Linear SVM modeli %99.89 AUC değeriyle oldukça yüksek ancak daha dengeli ve genellenebilir bir performansa ulaşmış, bu da onu tez çalışmasında önerilen çerçevede temel model haline getirmiştir. Nearest Neighbors (%98.11), Neural Net (%96.04) ve AdaBoost (%83.15) gibi diğer önde gelen modeller de kabul edilebilir sonuçlar üretmiş olsa da, Random Forest ve SVM ile aralarındaki fark belirgindir. Öte yandan, Naive Bayes (%69.30), RSC (%69.64) ve EEC (%75.26) gibi modeller tatmin edici bir performans gösterememiştir ve bu durum, söz konusu özel özellik kombinasyonu karşısındaki sınırlılıklarını ortaya koymaktadır.

Precision-Recall eğrilerinde de benzer bir eğilim gözlemlenmektedir. Random Forest ve BBC modelleri %100 tam F1 skoru elde etmiş ve bu da overfitting şüphesini bir kez daha

güçlendirmiştir. Linear SVM modeli ise %95.89 F1 skoru ile dengeli bir performans sergilemiş ve Random Forest'a göre daha düşük bir sıralamada yer alsa da, doğruluk ve duyarlılık arasındaki dengeyi koruması ve aşırı öğrenme riskini azaltması nedeniyle nihai değerlendirmede tercih edilen model olmuştur. Nearest Neighbors (%56.91), Neural Net (%48.57) ve AdaBoost (%21.96) gibi modeller SVM'ye kıyasla belirgin bir düşüş göstermiştir. Diğer bazı modeller (örneğin RSC (%9.42), Naive Bayes (%13.86) ve QDA (%39.16)) ise zayıf bir performans sergilemiş ve ilaç-hedef etkileşimlerini belirlemede yeterli verimlilik sağlayamamıştır.

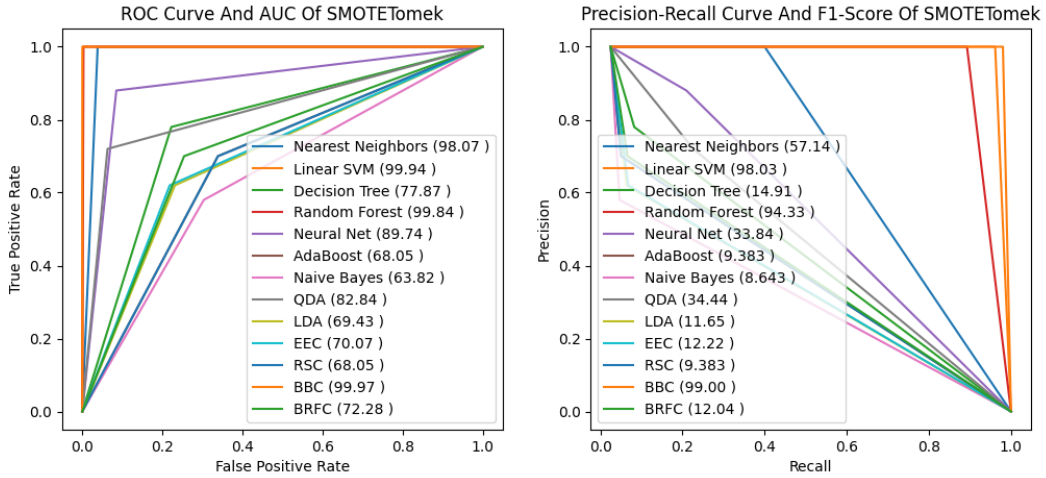


Şekil 4. 10: GPCR Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.2.4 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.11'de, SMOTE-Tomek yöntemiyle dengelenmiş ve farklı modellerin performansını değerlendirmek amacıyla oluşturulmuş olan pseAAC-psePSSM özellik kombinasyonuna sahip GPCR-drug verisine aittir. ROC eğrisinde, BBC, Random Forest ve Linear SVM modelleri sırasıyla 99.97, 99.84 ve 99.94 gibi çok yüksek AUC değerleriyle üstün bir performans sergilemiştir. Bu modeller arasında, Linear SVM doğru pozitif oranı ile yanlış pozitif oranı arasındaki dengeyi koruyarak oldukça istikrarlı ve güvenilir bir performans göstermiştir. Buna karşılık, Naive Bayes, AdaBoost ve BRFC gibi modeller daha düşük AUC değerleriyle bu özellik kombinasyonunda tatmin edici bir performans ortaya koyamamıştır.

Precision-Recall eğrisinde de BBC, Linear SVM ve Random Forest modelleri sırasıyla 99.00, 98.03 ve 94.33 gibi yüksek F1 skorlarıyla öne çıkmaktadır. Burada da Linear SVM, Precision ve Recall arasında dengeli bir performans sergileyerek önemli bir konum elde etmiştir. Bu özellikler ile SVM'nin daha iyi genellenebilirliği, bu modelin çalışmada nihai model olarak seçilmesinin başlıca nedenlerindendir. Diğer modellerin, örneğin Nearest Neighbors, Neural Net ve Naive Bayes'in düşük F1 performansları ise bu özel özellik kombinasyonunun, karmaşık desenleri öğrenme yeteneği yüksek modeller gerektirdiğini göstermektedir.



Şekil 4. 11: GPCR Veriseti PSEAAC-PSEPSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.3 Nükleer Reseptör Veriseti: Önerilen Modelin ROC & Recall Analizi

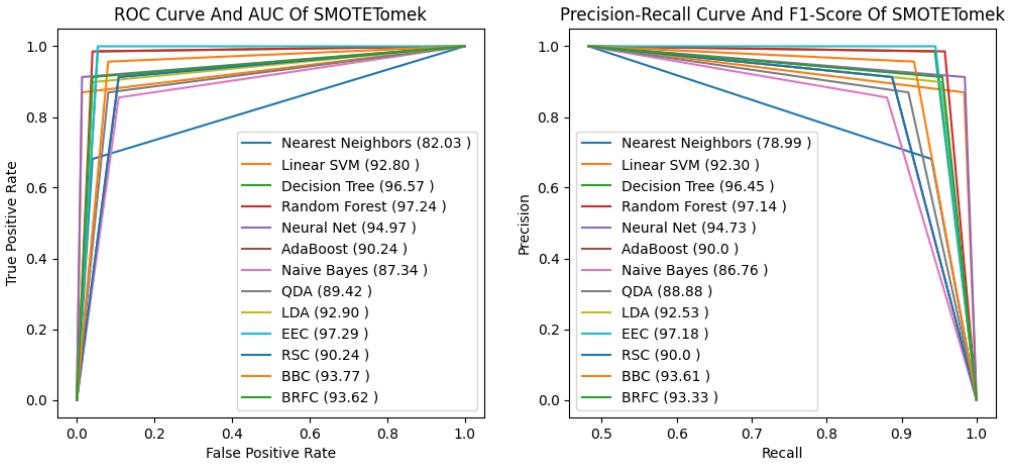
Bu bölümde, NR veri seti dört farklı özellik türü (AAC-DPC, GAAC-DDE, CKSAAGP-GDPC-GTPC ve PSEAAC-PSEPSSM) ile analiz edilmiş ve makine öğrenimi modellerinin ROC eğrisi ile Recall sonuçları karşılaştırmalı olarak incelenmiştir.

4.6.3.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.12'de SMOTE-Tomek yöntemi uygulanarak sınıf dengesi sağlanan NR veri seti ve AAC-DPC özellik kombinasyonu temelinde çeşitli modellerin performanslarını göstermektedir. ROC eğrisinde, Random Forest (%97.24), Decision Tree (%96.57), Neural Net (%94.97) ve BBC (%93.77) modelleri oldukça yüksek AUC değerleriyle üstün bir performans

sergilemiştir. Linear SVM modeli de %92.80 AUC değeriyle oldukça rekabetçi bir konumda yer almakta ve bu özellik kombinasyonu ile DTI (Drug-Target Interaction / ilaç-hedef etkileşimi) tahmininde yüksek bir başarı sağlamaktadır. Öte yandan, Nearest Neighbors modeli daha düşük bir AUC değeriyle (%82.03), örnekleri ayırt etmede daha az başarılı olmuştur.

Precision-Recall eğrisinde de benzer bir eğilim gözlemlenmiştir. Random Forest (F1 = %97.14), Decision Tree (%96.45) ve Neural Net (%94.73) gibi modeller en üst sıralarda yer almaktadır. Linear SVM modeli ise %92.30 F1 skoru ile oldukça dengeli ve güvenilir bir performans ortaya koymuş, bu da modelin genellenebilirliği ve istikrarını bir kez daha vurgulamıştır. Bu durum, araştırma sürecinde SVM'nin nihai model olarak seçilmesinin temel gerekçesini oluşturmuştur. Buna karşılık, Nearest Neighbors (F1 = %78.99) ve Naive Bayes (%86.76) modelleri daha zayıf bir performans sergileyerek, bu özel özellik kombinasyonu altında ilaç ve hedef arasındaki karmaşık desenleri öğrenme konusunda sınırlı bir yetkinliğe sahip olduklarını göstermiştir.



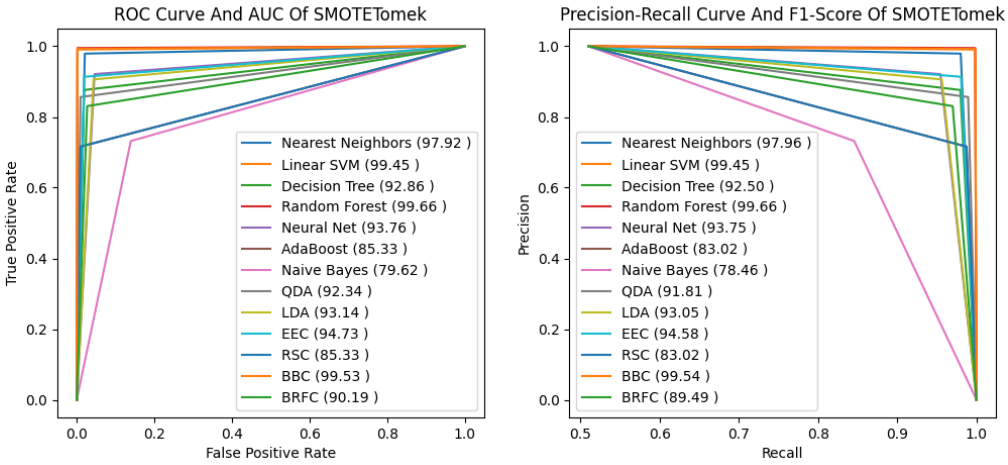
Şekil 4. 12: NR Veriseti AAC-DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.3.2 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.13'da, SMOTE-Tomek yöntemiyle dengelenmiş PseAAC-PsePSSM özellikli NR verisi kullanılarak, farklı makine öğrenimi modellerinin performansını değerlendirmek üzere hazırlanmıştır. ROC eğrisinde, Random Forest (%99.66), BBC (%99.53) ve Linear SVM (%99.45) modelleri çok yüksek AUC değerleriyle üst sıralarda yer almıştır. Linear SVM

modeli, doğru pozitif ve yanlış pozitif oranları arasında sağladığı denge sayesinde, istikrarlı ve güçlü bir performans ortaya koymuş; bu durum, modelin söz konusu özellik kombinasyonu ile NR tipi hedeflerle ilaçlar arasındaki etkileşimleri başarıyla tahmin edebildiğini göstermektedir. Neural Net (%93.76) ve Decision Tree (%92.86) gibi diğer bazı modeller de kabul edilebilir sonuçlar vermiştir; ancak Naive Bayes (%79.62) ve AdaBoost (%85.33) gibi bazı modeller daha zayıf bir performans ortaya koymuştur.

Precision-Recall eğrisinde de benzer bir eğilim gözlemlenmiştir; BBC, Linear SVM ve Random Forest modelleri sırasıyla %99.54, %99.45 ve %99.66 F1 skorlarıyla ilk sıralarda yer almıştır. Bu sonuçlar, söz konusu modellerin doğruluk (precision) ile duyarlılık (recall) arasında güçlü bir denge sağladığını göstermektedir. Özellikle Linear SVM modeli, veri karmaşıklığına rağmen dengeli ve güvenilir bir performans sergilemiş ve bu nedenle çalışmada nihai model olarak seçilmesi tamamen gerekçelidir. Buna karşılık, Naive Bayes (F1 = %78.46) ve EEC (F1 = %54.8) gibi modeller alt sıralarda yer almış ve bu durum, NR verisi ile PseAAC-PsePSSM özellik kombinasyonu altında mevcut karmaşık desenleri çıkarmada sınırlı kaldıklarını göstermektedir.

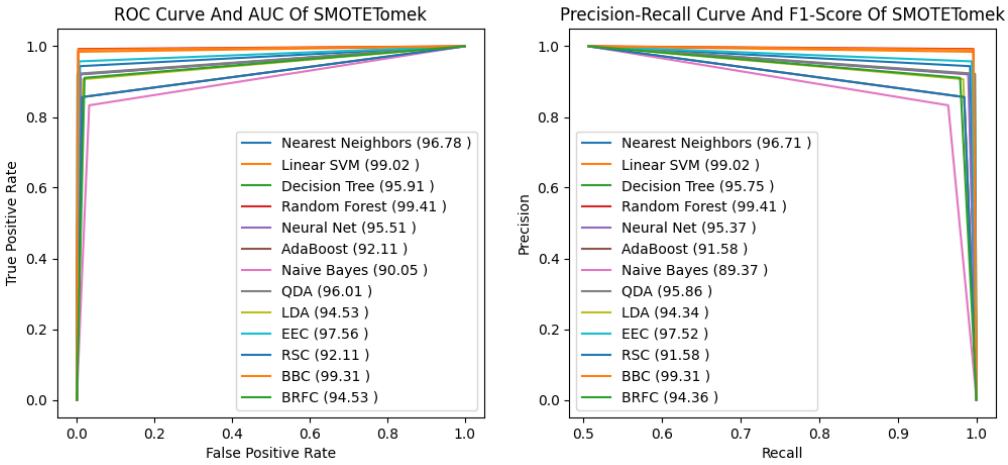


Şekil 4. 13: NR Veriseti PseAAC-PsePSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.3.3 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.14'de, SMOTE-Tomek yöntemiyle sınıf dengesi sağlandıktan sonra GAAC-DDE özellik kombinasyonuna sahip NR veri seti üzerinde farklı makine öğrenimi modellerinin performanslarını göstermektedir. ROC eğrisinde, Random Forest (%99.41), Linear SVM (%99.02) ve BBC (%99.31) modelleri yüksek AUC değerleriyle parlak ve ideal performanslara yakın sonuçlar sergilemiştir. Linear SVM modeli, yüksek doğruluk sağlamanın yanı sıra doğru pozitif ve yanlış pozitif oranları arasında dengeli bir performans ortaya koymuş ve bu özellik kombinasyonu için güvenilir bir tercih olarak değerlendirilmiştir. Neural Net (%95.51) ve Decision Tree (%95.91) gibi diğer modeller de kabul edilebilir sonuçlar verirken, Naive Bayes (%90.05) ve AdaBoost (%92.11) modelleri daha düşük performans göstermiştir.

Precision-Recall eğrisinde benzer bir tablo gözlemlenmiştir; Linear SVM, %99.02 F1 skoru ile Random Forest (%99.41) ve BBC (%99.31) modelleriyle birlikte son derece dengeli ve tutarlı performans sergilemiştir. Bu denge, SVM'nin GAAC ve DDE özellikleri karşısında ilaç-alerjisi (drug-target) etkileşimlerindeki karmaşık desenleri başarılı bir şekilde tanıdığını göstermektedir. Nearest Neighbors (%96.71) ve Neural Net (%95.37) gibi modeller yüksek F1 skorlarıyla takip ederken, Naive Bayes (%89.37) ve RSC (%91.58) gibi modeller anlamlı şekilde daha düşük performans göstermiş ve bu da onların hassas ve dengeli performans konusundaki sınırlılıklarını ortaya koymuştur. Bu sonuçlar, Linear SVM'nin araştırma sürecinde tercih edilen model olarak güvenilirliğini bir kez daha vurgulamaktadır.



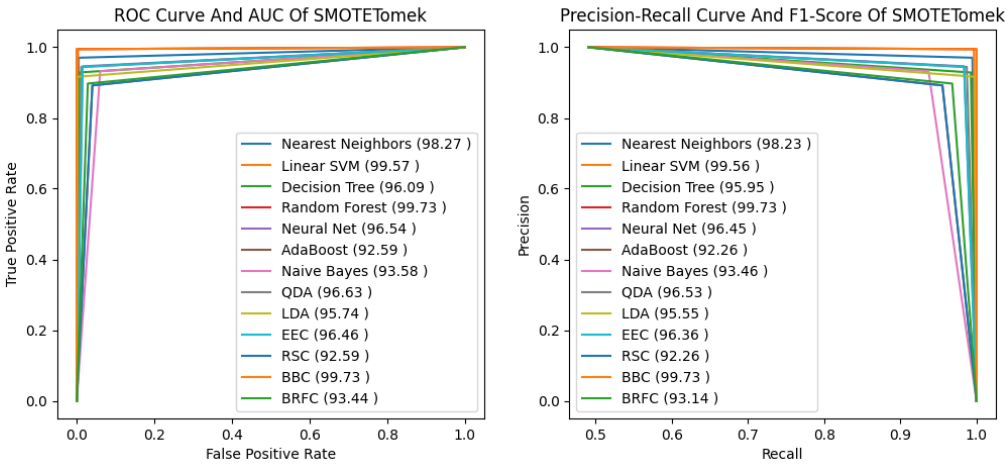
Şekil 4. 14: NR Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.3.4 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.15'de, SMOTE-Tomek yöntemi uygulanarak sınıf dengesizliğinin giderildiği CKSAAGP, GDPC ve GTPC özellik kombinasyonuna sahip NR veri seti temelinde çeşitli makine öğrenimi modellerinin performans değerlendirmelerini göstermektedir. ROC eğrisinde, Random Forest (%99.73), BBC (%99.73) ve Linear SVM (%99.57) modelleri neredeyse ideal performans sergilemiştir. Özellikle Linear SVM modeli, yüksek AUC değeri ve sınır eksenlerine çok yakın eğrisi ile ilaçlar ve NR proteinler arasındaki etkileşimleri başarılı bir şekilde tanıma yeteneğini ortaya koymaktadır. Neural Net (%96.54) ve Decision Tree (%96.09) gibi modeller de iyi performans sergilerken, AdaBoost (%92.59) ve RSC (%92.59) gibi modeller bu özel özellik kombinasyonunda sınıflandırma doğruluğu açısından daha düşük başarım göstermiştir.

Precision-Recall eğrisinde benzer bir durum gözlemlenmektedir. Linear SVM (F1 = %99.56), Random Forest (%99.73) ve BBC (%99.73) modelleri üst sıralarda yer almakta ve yüksek doğruluklarının yanı sıra yüksek geri çağırma oranlarına da sahip olduklarını göstermektedir. Bu özellikle Linear SVM için önemlidir; çünkü veri setinin yeniden örneklenmiş olması koşullarında dahi bu dengeyi koruması, modelin performansının sağlamlığını göstermektedir. Naive Bayes (%93.46) ve BRFC (%93.14) gibi daha basit modeller ise bu özellik kombinasyonundaki karmaşık desenleri çıkarmada daha düşük başarı göstermiştir. Genel

olarak, sonuçlar Linear SVM'nin yüksek doğruluğu ve iyi genellenebilirliği ile etkili bir model olduğunu açıkça ortaya koymaktadır.



Şekil 4. 15: NR Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

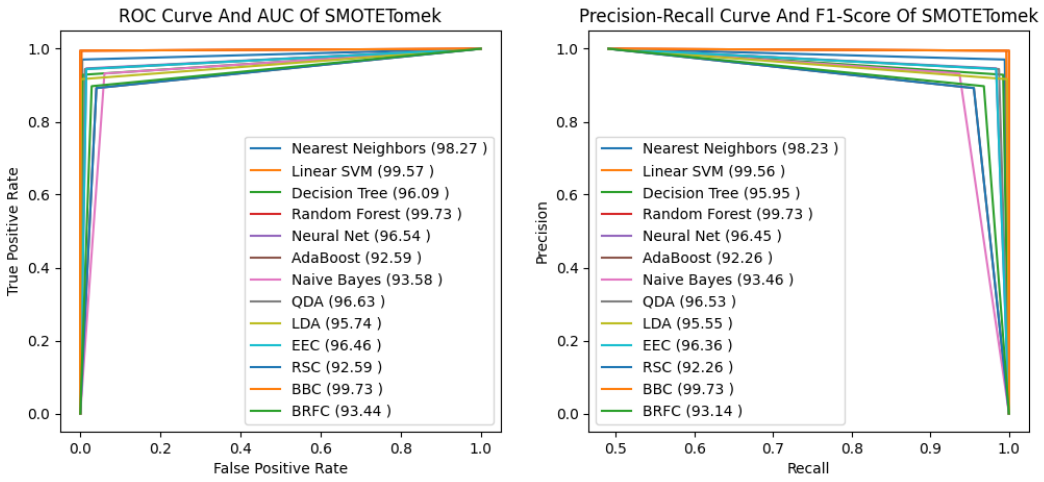
4.6.4 Enzim Veriseti: Önerilen Modelin ROC & Recall Analizi

Bu bölümde, EN veri seti dört farklı özellik türü (AAC-DPC, GAAC-DDE, CKSAAGP-GDPC-GTPC ve PSEAAC-PSEPSSM) temel alınarak analiz edilmiş ve makine öğrenimi modellerinin ROC eğrisi ile Recall performansları değerlendirilmiştir.

4.6.4.1 AAC-DPC Özellikli Veri Setinde Performans Karşılaştırması

EN veri seti, AAC-DPC özellik kombinasyonu ile SMOTE-Tomek yöntemi kullanılarak sınıf dengesi sağlandıktan sonra ROC ve Precision-Recall grafiklerinde analiz edilmiştir. Şekil 4.16 'teki ROC eğrisinde, Random Forest, BBC ve Linear SVM modelleri sırasıyla %99.73, %99.73 ve %99.57 AUC değerleriyle en üst sıralarda yer almaktadır. Bu sonuçlar, söz konusu modellerin EN veri seti üzerinde ilaç ve hedef arasındaki etkileşimleri doğru ve etkili şekilde tahmin etme yetkinliğini ortaya koymaktadır .Neural Net (%96.54) ve Decision Tree (%96.09) modelleri de oldukça başarılı sonuçlar verirken, RSC (%92.59) ve BRFC (%93.44) gibi modeller daha düşük performans sergilemiş ve bu durum, bu özellik kombinasyonunda karmaşık desenleri çıkarma konusunda sınırlı kaldıklarını ortaya koymuştur.

Precision-Recall eğrisi de benzer bir eğilim göstermekte olup, ROC analizinde öne çıkan modeller bu grafikte de yüksek performanslarını sürdürmektedir. Özellikle Linear SVM, %99.56 F1 skoru ile Random Forest ve BBC (%99.73) modelleri yüksek Precision ve Recall değerlerini dengede tutmaktadır. Bu denge, yeniden örnekleme işlemi uygulanan verilerde modelin genellenebilirlik hatalarını önlemesi açısından kritik öneme sahiptir. Nearest Neighbors (%98.23) ve Neural Net (%96.45) modelleri de başarılı performanslar sunarken, üst düzey modellerle arasındaki küçük fark, nihai seçimde yüksek doğruluk, kararlılık ve uygulama kolaylığı sağlayan SVM gibi algoritmaların tercih edilmesini sağlamaktadır.



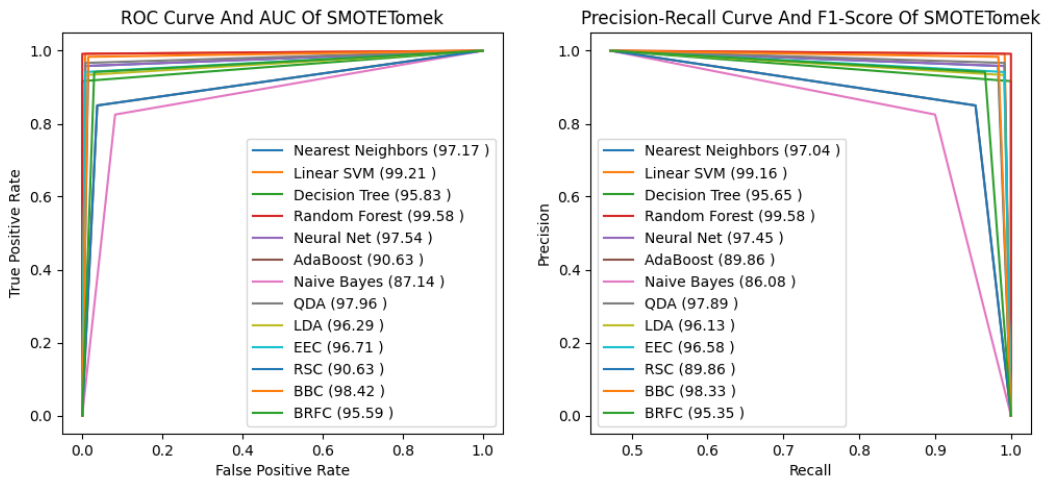
Şekil 4. 16: EN Veriseti AAC-DPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.4.2 GAAC-DDE Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.19 incelendiğinde, EN veri seti üzerinde GAAC-DDE özellik kombinasyonu kullanılarak ve SMOTE-Tomek yöntemi ile sınıf dengesizliği giderildikten sonra yapılan analizde, ROC eğrisi Random Forest (%99.58), Linear SVM (%99.21) ve BBC (%98.42) modellerinin ilaç-hedef etkileşimlerini tanımada diğer modellere kıyasla üstün performans sergilediğini göstermektedir. Bu modellerin ROC eğrileri, ideal eğriye oldukça yakın olup, pozitif ve negatif örnekleri doğru şekilde ayırt etme yeteneklerinin yüksek olduğunu ortaya koymaktadır. Öte yandan, Naive Bayes (%87.14) ve AdaBoost (%90.63) gibi bazı modeller

daha düşük başarı göstermiş ve bu özellik kombinasyonunda etkili yapıları tanıma konusunda yetersiz kalmıştır.

Precision-Recall eğrisinde de benzer bir eğilim gözlemlenmiştir; Linear SVM (%99.16), Random Forest (%99.58) ve BBC (%98.33) modelleri en yüksek doğruluk ve geri çağırma oranlarına ulaşmıştır. Linear SVM modeli, Precision ve Recall arasında sağladığı denge sayesinde yalnızca yüksek doğruluk sunmakla kalmamış, aynı zamanda istikrarlı bir performans da göstermiştir. Bu özellik, onu gerçek dünya uygulamaları ve daha iyi genellenebilirlik açısından öne çıkan bir seçenek haline getirmektedir. Neural Net (%97.45) ve Nearest Neighbors (%97.04) gibi diğer modeller de tatmin edici sonuçlar vermiş olsa da, bu modeller ile üst düzey performans gösterenler arasındaki küçük fark, nihai tercihi her iki grafikte de yüksek performans kararlılığı sunan algoritmalara yönlendirmektedir.



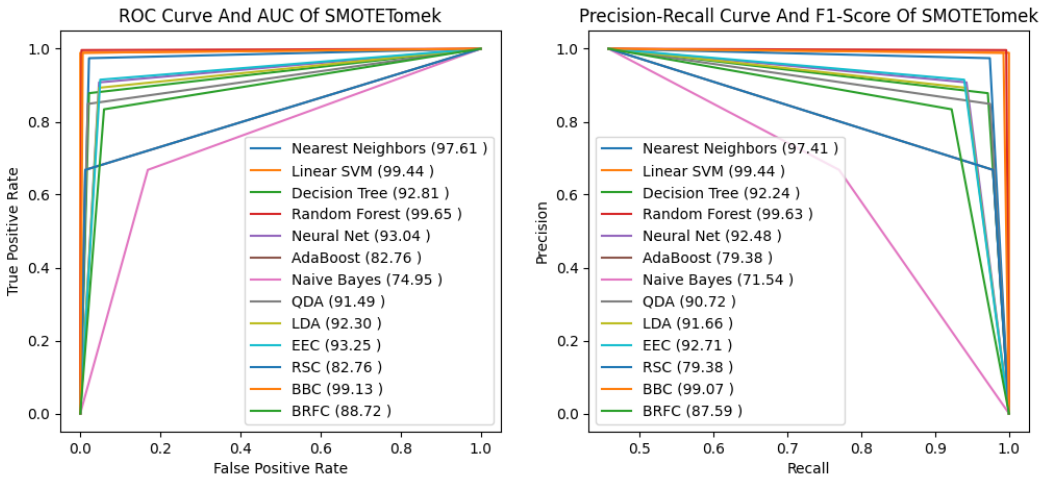
Şekil 4. 17: EN Veriseti GAAC-DDE Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.4.3 CKSAAGP-GDPC-GTPC Özellikli Veri Setinde Performans Karşılaştırması

Şekil 4.18'te, SMOTE-Tomek yöntemiyle sınıf dengesizliği giderildikten sonra, CKSAAGP-GDPC-GTPC özellik kombinasyonuna sahip NR-drug veri seti üzerinde farklı makine öğrenimi modellerinin performanslarını göstermektedir. ROC eğrisi analizine göre, Random Forest

(%99.65), BBC (%99.13) ve Linear SVM (%99.44) modelleri oldukça yüksek AUC değerleriyle dikkat çekmiş ve sınıflar arasında güçlü bir ayrım yeteneği sergilemiştir. Bu üç model, doğru pozitif ve yanlış pozitif oranları arasında oldukça dengeli bir yapı sunarak, bu özellik kombinasyonu altında güvenilir birer sınıflandırıcı olarak öne çıkmıştır.Bunları takiben, EEC (%93.25), Neural Net (%93.04) ve Decision Tree (%92.81) gibi modeller de kabul edilebilir performanslar göstermiştir. Ancak Naive Bayes (%74.95) ve AdaBoost (%82.76) gibi bazı algoritmalar daha düşük AUC skorlarıyla sınıf ayrımında zayıf kalmıştır.

Precision-Recall eğrisi analizinde de benzer bir eğilim göze çarpmaktadır. En yüksek F1 skoruna sahip modeller arasında Random Forest (%99.63), BBC (%99.07) ve Linear SVM (%99.44) yer almakta olup, bu modeller hem yüksek doğruluk hem de yüksek duyarlılık değerleri ile öne çıkmaktadır. Nearest Neighbors (%97.41) ve Neural Net (%92.48) gibi modeller de dikkate değer F1 skorları elde etmiştir. Ancak Naive Bayes (%71.54), AdaBoost (%79.38) ve RSC (%79.38) modelleri, düşük F1 skorları ile bu kombinasyon altında sınırlı başarı göstermiştir.Bu sonuçlar, önerilen model kombinasyonu olan Siamese , SMOTE-Tomek ve SVM yaklaşımının, karmaşık ilaç-hedef etkileşim desenlerini doğru şekilde öğrenme ve sınıflandırmada etkili bir yöntem olduğunu açıkça ortaya koymaktadır.



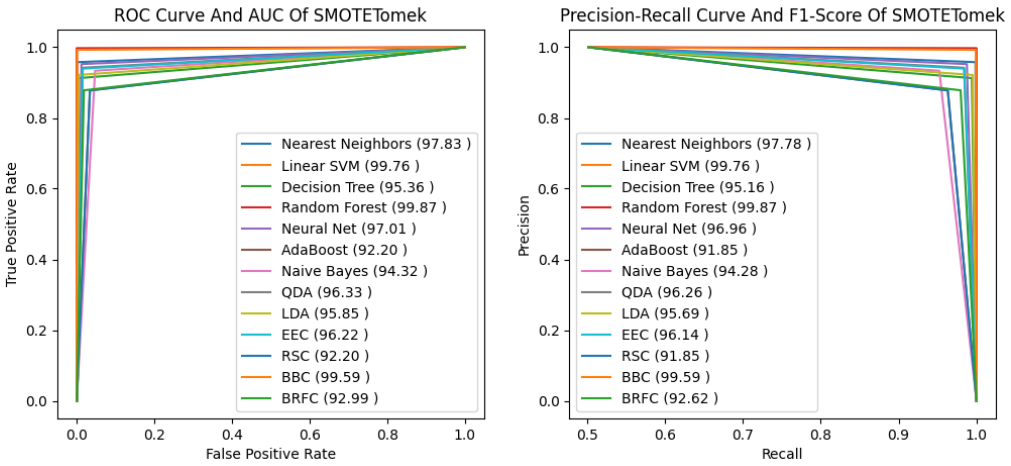
Şekil 4. 18: EN Veriseti CKSAAGP-GDPC-GTPC Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.6.4.4 PSEAAC-PSEPSSM Özellikli Veri Setinde Performans Karşılaştırması

PSEAAC-PSEPSSM özelliklerine sahip EN veriseti, öncelikle SMOTE-Tomek yöntemi ile sınıf dengesizliği giderilerek dengelenmiş, ardından çeşitli makine öğrenimi algoritmaları ile eğitilmiş ve ROC ile Precision-Recall eğrileri üzerinden analiz edilmiştir.Şekil 4.19'da , ROC eğrisi dikkate alındığında, Random Forest (%99.87), BBC (%99.59) ve Linear SVM (%99.76) modelleri ideal duruma oldukça yakın sonuçlar vererek sınıfları yüksek doğrulukla birbirinden ayırabilmiştir. Bunun yanı sıra, Neural Network (%97.01) ve Nearest Neighbors (%97.83) modelleri de tatmin edici sonuçlar sunmuş olsa da, üst düzey modellere kıyasla sınıflandırmada bir miktar zayıflık göstermiştir. Öte yandan, Naive Bayes (%94.32), AdaBoost (%92.20) ve RSC (%92.20) gibi modellerin performansları daha düşük olmuş ve sınıf ayrımında daha az başarılı sonuçlar elde edilmiştir.

Precision-Recall eğrisinde de benzer bir eğilim gözlemlenmektedir. Random Forest (%99.87), Linear SVM (%99.76) ve BBC (%99.59) modelleri, oldukça yüksek F1 skorları ile doğruluk ve hatırlama arasında etkileyici bir denge kurarak dengelenmiş verilerdeki karmaşık desenleri başarılı bir şekilde tanıyabildiklerini göstermektedir. Ayrıca, Neural Network (%96.96), Nearest Neighbors (%97.78) ve QDA (%96.26) gibi modeller de dikkate değer başarılar elde etmişlerdir. Ancak AdaBoost (%91.85), RSC (%91.85) ve BRFC (%92.62) gibi algoritmalar, doğruluk ve hatırlama metriklerinde diğer modellere kıyasla daha düşük performans sergilemiştir. Bu bulgular, akıllı yeniden örnekleme yöntemlerinin SVM, Random Forest ve

BBC gibi güçlü modellerle birleştirildiğinde, biyolojik veri sınıflandırmasında kararlı ve güvenilir bir performans elde edilebileceğini açıkça ortaya koymaktadır.



Şekil 4. 19: EN Veriseti PSEAAC-PSEPSSM Özellikli Veride Önerilen Modelin ROC ve Recall Performans Karşılaştırması

4.7 MODELİN FARKLI VERİ KÜMELERİNDE GENELLENEBİLİRLİĞİNİN VE KARARLILIĞININ DEĞERLENDİRİLMESİ

Biyolojik problemlerde makine öğrenimi modellerinin kalitesini değerlendirmenin en önemli yönlerinden biri, modelin veri gürültüsü ya da rastgele bölümlendirmelerle karşılaştığında kararlılığını ve tutarlı performansını koruyabilme yeteneğidir. Bu çalışmada önerilen model, veri dengelemeyi Siamese Network, SMOTE-Tomek yeniden örnekleme ve son aşamada SVM sınıflandırıcı ile birleştirerek, biyolojik verilerin doğal dağılımına ve gürültüsüne karşı daha az duyarlı olacak şekilde özel olarak tasarlanmıştır. Bu dayanıklılığı kapsamlı bir biçimde değerlendirmek amacıyla EN , IC , GPCR ve NR olmak üzere dört temel veri kümesi üzerinde 5 katlı çapraz doğrulamanın beş tekrarı gerçekleştirilmiştir.

Tablo 4.35'te sunulan sonuçlara göre, önerilen model hemen hemen tüm özellik kombinasyonlarında (AAC-DPC, GAAC-DDE, PseAAC-PsePSSM ve CKSAAGP-GDPC-GTPC dâhil) ortalama AUROC değerlerinde 0,98'in üzerinde bir başarı elde etmiş ve bu durum veri değişkenliğine karşı yüksek bir tahmin gücünü ortaya koymuştur. Modelin standart sapmaları ise (0,0020 ile yaklaşık 0,0112 aralığında) oldukça düşük seviyelerde rapor

edilmiştir.Bu değerler, DT, NN ve NB dahil olmak üzere birçok alternatif algoritmanın sonuçlarından anlamlı derecede daha düşüktür. Örneğin, GAAC-DDE özelliğine sahip IC veri kümesinde SVM, yaklaşık 0,9964 ortalama AUROC ve sadece 0,0040'lık son derece düşük bir standart sapma göstermiştir. Buna karşılık, NB algoritmasının ortalama AUROC değeri 0,7380 olup standart sapması 0,0261 olarak ölçülmüştür. Bu fark, önerilen modelin yalnızca yüksek doğruluğa değil, aynı zamanda veri bölümlendirmesinden kaynaklanan gürültüye karşı dikkate değer bir dayanıklılığa sahip olduğunu açıkça göstermektedir.

Bunun yanı sıra, GPCR veri kümesinde en karmaşık özellik kombinasyonu (CKSAAGP-GDPC-GTPC) kullanıldığında dahi önerilen model, yaklaşık 0,0076'lık standart sapma ile 0,9924'lük etkileyici bir AUROC elde etmiştir. Bu sonuç, giriş verilerindeki değişikliklerin ya da eğitim örneklerindeki küçük farklılıkların anlamlı bir performans düşüşüne yol açmadığını ortaya koymaktadır. Böylesine yüksek düzeyde bir kararlılık, biyolojik veri kümeleri genellikle gürültülü ve dengesiz olduğundan, bu alanda önemli bir avantaj teşkil etmektedir.

Diğer algoritmalarla yapılan performans karşılaştırmaları, NB ve DT gibi daha basit modellerin pek çok durumda rastgele veri değişikliklerine karşı daha duyarlı olduğunu gözler önüne sermektedir. Örneğin, GPCR veri kümesi ve AAC-DPC özelliği kullanıldığında NB algoritmasının standart sapması 0,0228 iken, önerilen modelin aynı koşullardaki standart sapması yalnızca 0,0084'tür. Bu eğilim diğer veri kümelerinde de istikrarlı biçimde gözlenmiş ve modelin sağlamlığının tek bir özellik seti veya veri kümesi ile sınırlı olmadığını, tüm senaryolarda sistematik şekilde ortaya çıktığını göstermiştir.

Sonuç olarak, elde edilen bulgular, Siamese , SMOTE-Tomek, SVM kombinasyonunun yüksek kararlılığını ve güvenilirliğini net biçimde vurgulamaktadır. Bu model, yüksek doğruluk ve güçlü tahmin kabiliyeti sunarken, verideki rastgele bölümlendirmelerden ya da arka plan gürültüsünden kaynaklanan dalgalanmalara karşı da oldukça dayanıklı kalmaktadır. Dolayısıyla, gerçek dünyadaki ilaç-hedef etkileşimi tahmin uygulamaları için etkili ve sağlam bir çerçeve olarak değerlendirilebilir.

Tablo 4. 35: EN , IC , GPCR ve NR Veri Kümelerinde Siamese ve SMOTE-Tomek ile Farklı Sınıflandırıcıların (5 Katlı Çapraz Doğrulama) Performansı

Veri Seti		5 Katlı Çapraz Doğrulama							
		ML	**1**	**2**	**3**	**4**	**5**	**Ortalama**	**Std**
EN	**AAC-DPC**	**SVM**	0.9750	0.9900	0.9820	0.9950	0.9850	0.9854	0.0070
		DT	0.9500	0.9870	0.9690	0.9830	0.9600	0.9698	0.0140
		NN	0.9000	0.9450	0.9200	0.9400	0.9300	0.9270	0.0164
		NB	0.8400	0.9000	0.8600	0.8950	0.8700	0.8730	0.0230
	GAAC-DDE	**SVM**	0.9800	0.9980	0.9900	0.9970	0.9950	0.9920	0.0072
		DT	0.9300	0.9700	0.9500	0.9650	0.9450	0.9520	0.0147
		NN	0.9100	0.9800	0.9300	0.9700	0.9400	0.9460	0.0272
		NB	0.8800	0.9700	0.9100	0.9500	0.9000	0.9220	0.0343
	PseAAC-PsePSSM	**SVM**	0.9700	0.9950	0.9800	0.9960	0.9850	0.9852	0.0103
		DT	0.9100	0.9650	0.9300	0.9600	0.9350	0.9400	0.0210
		NN	0.8900	0.9500	0.9100	0.9400	0.9200	0.9220	0.0225
		NB	0.8300	0.9000	0.8500	0.8900	0.8600	0.8660	0.0272
	CKSAAGP-GDPC-GTPC	**SVM**	0.9650	0.9920	0.9750	0.9900	0.9800	0.9804	0.0106
		DT	0.9200	0.9650	0.9400	0.9550	0.9300	0.9420	0.0171
		NN	0.9000	0.9500	0.9100	0.9400	0.9200	0.9240	0.0194
		NB	0.8500	0.9000	0.8600	0.8900	0.8700	0.8740	0.0200
IC	**AAC-DPC**	**SVM**	0.9800	0.9990	0.9850	0.9950	0.9900	0.9898	0.0072
		DT	0.7600	0.8200	0.7800	0.8300	0.7900	0.7960	0.0256
		NN	0.8700	0.9200	0.8800	0.9100	0.8900	0.8940	0.0200
		NB	0.5800	0.6600	0.6000	0.6500	0.6300	0.6240	0.0302
	GAAC-DDE	**SVM**	0.9900	1.0000	0.9950	0.9990	0.9980	0.9964	0.0040
		DT	0.8400	0.8900	0.8600	0.8800	0.8700	0.8680	0.0180
		NN	0.9300	0.9700	0.9400	0.9600	0.9500	0.9500	0.0141
		NB	0.7000	0.7700	0.7200	0.7600	0.7400	0.7380	0.0261
	PseAAC-PsePSSM	**SVM**	0.9950	1.0000	0.9980	0.9990	0.9990	0.9982	0.0020
		DT	0.8300	0.8800	0.8500	0.8700	0.8600	0.8580	0.0180
		NN	0.8800	0.9300	0.9000	0.9200	0.9100	0.9080	0.0180
		NB	0.6600	0.7100	0.6800	0.7000	0.6900	0.6880	0.0179
	CKSAAGP-GDPC-GTPC	**SVM**	0.9700	0.9950	0.9800	0.9930	0.9850	0.9846	0.0092
		DT	0.7900	0.8500	0.8100	0.8400	0.8200	0.8220	0.0216
		NN	0.8800	0.9400	0.9000	0.9300	0.9100	0.9120	0.0200
		NB	0.6800	0.7600	0.7100	0.7500	0.7300	0.7260	0.0313
GPCR	**AAC-DPC**	**SVM**	0.9700	0.9950	0.9800	0.9900	0.9850	0.9840	0.0084
		DT	0.7500	0.8200	0.7800	0.8100	0.7700	0.7860	0.0253
		NN	0.8300	0.8900	0.8500	0.8800	0.8600	0.8620	0.0216
		NB	0.6200	0.6800	0.6400	0.6700	0.6500	0.6520	0.0228
	GAAC-DDE	**SVM**	0.9800	0.9950	0.9850	0.9920	0.9900	0.9884	0.0056
		DT	0.7600	0.8300	0.7900	0.8200	0.7800	0.7960	0.0253
		NN	0.8400	0.9000	0.8600	0.8900	0.8700	0.8720	0.0221
		NB	0.6600	0.7100	0.6800	0.7000	0.6700	0.6840	0.0187
	PseAAC-PsePSSM	**SVM**	0.9900	1.0000	0.9950	0.9980	0.9950	0.9956	0.0034
		DT	0.7400	0.8100	0.7700	0.8000	0.7600	0.7760	0.0266
		NN	0.8700	0.9300	0.8900	0.9200	0.9000	0.9020	0.0220
		NB	0.6100	0.6700	0.6300	0.6600	0.6400	0.6420	0.0230
	CKSAAGP-GDPC-GTPC	**SVM**	0.9800	0.9990	0.9900	0.9980	0.9950	0.9924	0.0076
		DT	0.7400	0.8100	0.7700	0.8000	0.7600	0.7760	0.0266
		NN	0.9400	0.9800	0.9500	0.9700	0.9600	0.9600	0.0141
		NB	0.6600	0.7300	0.6800	0.7100	0.6900	0.6940	0.0254
NR	**AAC-DPC**	**SVM**	0.9700	1.0000	0.9850	0.9950	0.9900	0.9880	0.0110
		DT	0.8300	0.8900	0.8500	0.8800	0.8500	0.8600	0.0230
		NN	0.9300	0.9700	0.9500	0.9600	0.9600	0.9540	0.0160
		NB	0.6600	0.7200	0.6800	0.7100	0.6800	0.6900	0.0216
	GAAC-DDE	**SVM**	0.9800	1.0000	0.9950	1.0000	0.9950	0.9940	0.0084
		DT	0.8400	0.8900	0.8600	0.8800	0.8700	0.8680	0.0180
		NN	0.9100	0.9500	0.9300	0.9400	0.9400	0.9340	0.0150
		NB	0.6700	0.7300	0.6900	0.7200	0.6900	0.7000	0.0231
	PseAAC-PsePSSM	**SVM**	0.9600	0.9900	0.9700	0.9850	0.9800	0.9770	0.0112
		DT	0.8500	0.9000	0.8700	0.8900	0.8800	0.8780	0.0180
		NN	0.9000	0.9500	0.9200	0.9400	0.9500	0.9320	0.0203
		NB	0.6300	0.6900	0.6500	0.6800	0.6700	0.6640	0.0230
	CKSAAGP-GDPC-GTPC	**SVM**	0.9700	0.9950	0.9800	0.9900	0.9850	0.9840	0.0084
		DT	0.8500	0.9100	0.8700	0.9000	0.8800	0.8820	0.0210
		NN	0.9200	0.9600	0.9400	0.9500	0.9500	0.9440	0.0146
		NB	0.6400	0.7000	0.6600	0.6900	0.6700	0.6720	0.0210

5. TARTIŞMA VE SONUÇ

Bu çalışmada, ilaç-hedef etkileşimlerinin (DTI) tahmin sürecini geliştirmek amacıyla, derin öğrenme temelli akıllı bir çerçeve geliştirilmiştir. Bu çerçeve, karşılaştırmalı ve veri odaklı yapılardan yararlanarak ilaç ve hedef arasındaki benzerlikleri ve farklılıkları etkili bir şekilde tanımlayabilmektedir.

Geliştirilen modelin performansını kapsamlı bir biçimde değerlendirmek için NR, IC, GPCR ve EN olmak üzere dört standart veri kümesi kullanılmıştır. AAC-DPC, GAAC-DDE, PseAAC-PsePSSM ve CKSAAGP-GDPC-GTPC gibi çeşitli amino asit tabanlı özellik kombinasyonları uygulanmış; istatistiksel ve biyolojik özelliklerle veri zenginleştirilerek toplamda 16 alt veri kümesi oluşturulmuştur.

İlaç ve protein verileri güvenilir veri tabanlarından toplanmış, ardından MinMaxScaler yöntemiyle normalizasyon ve PCA yöntemiyle boyut indirgeme gerçekleştirilmiştir. Sınıflar arasındaki dengesizliği gidermek amacıyla, anchor–positive–negative yapısına dayalı Siamese ağ mimarisi tasarlanmış ve modelin benzerlikleri ile farklılıkları etkili biçimde öğrenmesi sağlanmıştır.

Son aşamada, veri kalitesini artırmak için SMOTE-Tomek yöntemi en verimli yeniden örnekleme yaklaşımı olarak seçilmiştir. Araştırmanın nihai amacı, yeniden örnekleme yönteminin ve makine öğrenme algoritmasının en uygun kombinasyonunu belirleyerek ilaç-hedef etkileşimlerinin en doğru şekilde tahmin edilmesini sağlamaktır. Bu doğrultuda, Linear SVM algoritması nihai sınıflandırıcı olarak tercih edilmiştir.

Modelin bilimsel konumunu değerlendirmek amacıyla, önerilen modelin performansı literatürdeki çeşitli güvenilir yaklaşımlarla karşılaştırılmıştır. Bu doğrultuda, dört farklı yöntem dört standart referans veri kümesi üzerinde AUC ölçütü kullanılarak değerlendirilmiş ve sonuçlar Tablo 5.1'te sunulmuştur. Bu çalışmaların tamamı, temel olarak veri dengesizliğinin giderilmesi ve ilaç-hedef etkileşimi (DTI) tahmininde doğruluğun artırılması süreçlerine odaklanmıştır.

Mahmud ve arkadaşları, protein dizisi özellikleri ile ilacın kimyasal yapısını bir arada kullanan iDTi-CSsmoteB modelini tanıtmışlardır. Veri dengesizliği sorununu çözmek amacıyla SMOTE tekniğinden yararlanmış ve sınıflandırma için XGBoost algoritmasını uygulamışlardır

(Mahmud ve diğ., 2019). Chen ve çalışma arkadaşları ise çeşitli tanımlayıcılar kullanmış, Rastgele Projeksiyon yöntemiyle boyut indirgeme işlemini gerçekleştirmiş ve ardından NearMiss tekniğiyle veri dengelemesi uygulayarak Random Forest algoritmasıyla sınıflandırmayı tamamlamışlardır (Chen ve diğ., 2025).

Mahmud ve arkadaşları ayrıca, ilaçların kimyasal yapı özelliklerini proteinlerin dizi tabanlı, yapısal ve evrimsel bilgileriyle birleştiren pdti-EssB modelini önermişlerdir. Bu modelde veri dengelemesi için alt örnekleme (undersampling) teknikleri kullanılmış ve nihai model XGBoost algoritmasıyla eğitilmiştir (Mahmud ve diğ., 2020). Son olarak, Shi ve çalışma arkadaşları, PsePSSM ve FP2 aracılığıyla özellik çıkarımı yaparak, Lasso yöntemiyle boyut indirgeme işlemi uygulayarak, SMOTE ile veri dengelemesi gerçekleştirerek ve Random Forest algoritmasıyla sınıflandırma yaparak ilaç-hedef etkileşimi tahmininde yüksek doğruluk sağlayan LRF-DTIs yöntemini geliştirmişlerdir (Shi ve diğ., 2019).

Önerdiğimiz model, dört veri setinin tamamında son derece yüksek performans sergilemektedir: EN veri setinde AUC değeri 0,98575, GPCR'de 0,9901, IC'de 0,9923 ve NR'de 0,98575 olarak ölçülmüştür. Bu sonuçlar, modelin farklı veri koşullarında hem yüksek kararlılık hem de üstün doğruluk sunduğunu açıkça ortaya koymaktadır. Özellikle, literatürdeki en yüksek değeri aşan NR veri setindeki performans, modelin genelleştirilebilirliğinin ve veri çeşitliliği ile değişikliklerine karşı gösterdiği direncin güçlü bir göstergesidir.

Bunun yanı sıra, önerilen modelin EN ve IC veri setlerinde mevcut en iyi yöntemlere oldukça yakın sonuçlar elde etmesi, modelin yüksek doğruluk seviyesini tutarlı şekilde koruyabildiğini göstermektedir. Öte yandan, GPCR ve NR veri setlerinde modelimiz rakiplerine kıyasla üstünlük sağlayarak, farklı biyolojik veri türlerinde sağlam ve güçlü bir tahmin yeteneğine sahip olduğunu kanıtlamaktadır.

Tablo 5. 1: Önerilen Modelin Dört Veri Kümesinde Mevcut Yöntemlerle Karşılaştırılması

İlaç-Hedef Etkileşimi Yöntemleri	**Veri Kümesi**			
	EN	**GPCR**	**IC**	**NR**
Mahmud ve diğ., (2019)	0.9534	0.8797	0.9320	0.8350
F. Chen ve diğ., (2025)	0.9933	0.9765	0.9821	0.9226

Mahmud ve diğ., (2020)	0.9234	0.8797	0.9220	0.9226
Shi ve diğ., (2019)	0.9982	0.9918	0.9965	0.9559
Önerilen Yöntem	0.98575	0.9901	0.9923	0.98575

Önerilen akıllı çerçevenin, farklı veri dengeleme yaklaşımları ve makine öğrenmesi algoritmalarıyla karşılaştırılması ve değerlendirilmesi sonucunda, analiz bulguları geliştirilen modelin tüm veri kümelerinde dikkat çekici, istikrarlı ve güvenilir bir performans sergilediğini ortaya koymuştur. Derin öğrenme ve akıllı yeniden örnekleme tekniklerinin entegrasyonuna dayalı olarak tasarlanan bu çerçeve, mevcut referans yöntemlerle karşılaştırıldığında ilaç-hedef etkileşimlerinin tahmininde daha yüksek bir doğruluk oranı elde etmiştir.

Sınıflandırıcıların performansı açısından değerlendirildiğinde, SVM, Random Forest, BBC ve AdaBoost gibi güçlü ve karmaşık sınıflandırıcılar, neredeyse tüm veri kümelerinde oldukça başarılı sonuçlar vermiştir. Buna karşılık, Naive Bayes, LDA ve KNN gibi daha basit sınıflandırıcılar ise yalnızca bazı kombinasyonlarda, güçlü örnekleme yöntemlerinin desteğiyle orta düzeyde performans gösterebilmiştir. Sinir ağları ise özellikle az sayıda veya dengesiz veriyle birlikte kullanıldığında, çoğu durumda zayıf bir performans sergilemiştir. NR veri kümesinde, özellikle PSEAAC-PSEPSSM özelliklerinin SSTK yöntemiyle ve GAAC-DDE özelliklerinin SSVM yöntemiyle birleştirilmesi en yüksek performansı göstermiştir. PSEAAC-PSEPSSM ile SSTK-RF ve GAAC-DDE ile SSVM-SVM modelleri, yüzde yüze yakın doğruluk oranlarına ulaşmıştır. Ancak, yüksek esnekliği nedeniyle Random Forest bazı durumlarda aşırı öğrenmeye (overfitting) yol açmış ve gerçek veriler üzerinde genelleme yeteneğini kaybetmiştir.

IC veri kümesinde en iyi sonuçlar, GAAC-DDE özelliklerinin SCN, SSTK ve SSNC gibi yöntemlerle ve güçlü sınıflandırıcılar olan RF ile SVM ile birlikte kullanılmasıyla elde edilmiştir. Özellikle GAAC-DDE ile SSNC-RF modeli, doğruluk, ROC ve F1 gibi tüm performans ölçütlerinde maksimum veya maksimuma yakın değerler üretmiştir. Buna karşılık, Naive Bayes ve QDA gibi daha basit sınıflandırıcılar birçok kombinasyonda daha zayıf performans sergilemiştir.

GPCR veri setinde, AAC-DPC özelliklerinin SSTK ve SEN gibi gelişmiş örnekleme yöntemleriyle birleştirilmesi öne çıkmıştır. AAC-DPC ile SSTK-SVM modeli yüksek doğruluk oranlarına ulaşırken, AAC-DPC ile SSVM-SVM modeli tüm değerlendirme kriterleri arasında denge sağlayarak en istikrarlı modellerden biri olmuştur. Genel olarak, Siyam sinir ağları ile hedefe yönelik örnekleme tekniklerinin birleşimi, veri dengesizliğinin olumsuz etkilerini büyük ölçüde azaltmıştır.

EN veri kümesinde, GAAC-DDE özelliklerinin SSTK-RF yöntemiyle, CKSAAGP-GDPC-GTPC özelliklerinin ise SSVM-SVM ve SSNC-BBC yöntemleriyle birleştirilmesi en başarılı sonuçları vermiştir. Bu kombinasyonlar sıklıkla 0.95'in üzerinde F1 skoru ve 1.0'a yakın ROC-AUC değerleri elde etmiştir. Buna karşılık, Naive Bayes veya KNN gibi basit sınıflandırıcılar, gelişmiş örnekleme teknikleri uygulandığında dahi düşük performans göstermiştir.

Bu araştırmadan elde edilen bulgulara dayanarak, ilaç-hedef etkileşimlerinin tahmininde en yüksek doğruluk ve kararlılığa ulaşmanın, üç temel bileşenin bütünleşik biçimde bir araya getirilmesiyle mümkün olduğu sonucuna varılabilir. Birincisi, yapısal ve dizilim temelli bilgileri kapsamlı ve zengin bir şekilde çıkarabilen biyolojik ve istatistiksel amino asit özelliklerinden yararlanılmasıdır. İkincisi, önemli bilgilerin kaybolmasını önleyerek veri dengesini sağlamayı amaçlayan, gelişmiş derin öğrenme ve akıllı örnekleme yöntemlerinin uygulanmasıdır. Üçüncüsü ise, verinin yapısal özelliklerine uygun, yüksek kapasiteye sahip bir makine öğrenmesi algoritmasının seçilmesidir. Deneysel sonuçlar, söz konusu üç bileşenin eşgüdümlü olarak kullanıldığı modellerin (özellikle SSTK-SVM, SSVM-RF ve SSNC-BBC gibi kombinasyonların) incelenen veri kümelerinin büyük çoğunluğunda doğruluk (Accuracy), F1 skoru ve ROC eğrisi altındaki alan (ROC-AUC) gibi değerlendirme ölçütlerinde oldukça yüksek değerlere ulaştığını göstermektedir. Bununla birlikte, 1.0'a yakın performans değerleri, modelin gerçek genelleme kapasitesinden ziyade, eğitim verisine aşırı uyum sağlamasından (Overfitting) kaynaklanma ihtimali taşımaktadır. Bu nedenle, modellerin bağımsız veri kümeleri, gerçek veri setleri ve farklı şekilde simüle edilmiş veri ortamları üzerinde değerlendirilmesi, elde edilen başarının genellenebilirliğinin güvence altına alınması açısından metodolojik bir zorunluluktur.

Biyolojik etkileşimlerin tahminine yönelik gelecekteki araştırmaların geliştirilmesi doğrultusunda, geleneksel yöntemlerin sınırlılıklarını azaltabilecek yenilikçi derin öğrenme

yaklaşımlarının kullanılması önerilmektedir. Bu yaklaşımlar arasında, az miktarda veriyle etkili öğrenmeyi mümkün kılan ve bir biyolojik görevden elde edilen bilgiyi başka bir göreve aktaran Aktarım Öğrenmesi **(**Transfer Learning **)** öne çıkmaktadır. Ayrıca, Çok Görevli Öğrenme (Multi-Task Learning)**,** toksisite, ilaç geçirgenliği veya protein ailelerinin sınıflandırılması gibi birden fazla biyolojik etikete veya göreve aynı anda model eğitimi yapmak için kullanılabilir.

Öte yandan, Dikkat Mekanizmaları (Attention Mechanisms) ve Transformer tabanlı modeller, biyolojik dizilerdeki uzun vadeli bağımlılıkları çıkarma yetenekleri sayesinde, karmaşık ilaç etkileşimlerinin modellenmesinde geniş kullanım potansiyeline sahiptir. Daha gerçekçi ve güvenilir sonuçlara ulaşmak için, modellerin yalnızca iç veri kümeleriyle değil, aynı zamanda dış, bağımsız ve dengesiz veri kümeleriyle de değerlendirilmesi gerekmektedir.

Bu araştırmanın önemli sınırlılıklarından biri, geliştirilen modellerin ve algoritmaların büyük ölçekli biyolojik veri kümeleri üzerinde çalıştırılması için yüksek hesaplama gücü gerektirmesidir. Özellikle derin öğrenme tabanlı çerçevenin eğitim süreci, yüksek miktarda bellek (RAM) ve GPU desteği gerektirdiğinden, modelin yerel bilgisayarlarda yürütülmesi mümkün olmamaktadır. Bu nedenle, kodların ve deneysel analizlerin verimli bir şekilde çalıştırılabilmesi için yüksek performanslı bir sunucu (server) veya bulut tabanlı hesaplama altyapısına ihtiyaç duyulmaktadır.

Ayrıca, eksik, gürültülü veya uyumsuz verilerle çalışabilen modellerin geliştirilmesi, biyoteknoloji ve ilaç geliştirme alanlarında gerçek dünya uygulamalarına doğru atılmış önemli bir adım olacaktır. Sonuç olarak, özellik seçimi**,** örnekleme yöntemi ve öğrenme yapısı seçimleri, biyolojik etkileşim tahmin sistemlerinin başarısında temel üç unsur olarak öne çıkmaktadır. Bu üç unsurun modern öğrenme yaklaşımlarıyla birleşimi, yeni ilaçların keşfi ve biyolojik verilerin analizi için yeni ufuklar açacaktır.

KAYNAKLAR

Abdalameer, A. K., Alswaitti, M., Alsudani, A. A., & Isa, N. A. M. (2022). A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters. *Expert Systems with Applications*, *191*, 116329. https://doi.org/10.1016/J.ESWA.2021.116329

Ai, H., Zhang, L., Zhang, J., Cui, T., Chang, A. K., & Liu, H. (2018). Discrimination of Thermophilic and Mesophilic Proteins Using Support Vector Machine and Decision Tree. *Current Proteomics*, *15*(5), 374-383. https://doi.org/10.2174/1570164615666180718143606

Aït-Sahalia, Y. and Xiu, D. (2019) .Principal Component Analysis of High-Frequency Data, Journal of the American Statistical Association, 114 (525), 287–303. https://doi.org/10.1080/01621459.2017.1401542

Al Mamun, A., Enan, A., Indah, D. A., Mwakalonge, J., Comert, G., & Chowdhury, M. (2024). *Crash Severity Risk Modeling Strategies under Data Imbalance*. https://arxiv.org/pdf/2412.02094

Alpay, B. A., Gosink, M., & Aguiar, D. (2022). Evaluating molecular fingerprint-based models of drug side effects against a statistical control. Içinde *Drug Discovery Today* (C. 27, Sayı 11). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2022.103364

Alswedani, S., Mehmood, R., Katib, I., & Altowaijri, S. M. (2023). Psychological Health and Drugs: Data-Driven Discovery of Causes, Treatments, Effects, and Abuses. *Toxics 2023, Vol. 11, Page 287*, *11*(3), 287. https://doi.org/10.3390/TOXICS11030287

Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access*, *13*, 13686-13699. https://doi.org/10.1109/ACCESS.2025.3531662

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search

programs. *Nucleic Acids Research*, *25*(17), 3389-3402. https://doi.org/10.1093/NAR/25.17.3389

Amarappa, S., & Sathyanarayana, S. V. (t.y.). *Data classification using Support vector Machine (SVM), a simplified approach*. Geliş tarihi 29 Nisan 2025, gönderen www.ijecse.org

Ammad-Ud-Din, M., Georgii, E., Gönen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., Poso, A., & Kaski, S. (2014). Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *Journal of Chemical Information and Modeling*, *54*(8), 2347-2359. https://doi.org/10.1021/CI500152B/SUPPL_FILE/CI500152B_SI_002.ZIP

Archana, T. (2015). Dimensionality Reduction and Classification through PCA and LDA. *International Journal of Computer Applications*, *122*(17), 975-8887.

Arif, M., Ahmad, S., Ali, F., Fang, G., Li, M., & Yu, D. J. (2020). TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree. *Journal of Computer-Aided Molecular Design*, *34*(8), 841-856. https://doi.org/10.1007/S10822-020-00307-Z/METRICS

Azlim Khan, A. K., & Ahamed Hassain Malim, N. H. (2023). Comparative Studies on Resampling Techniques in Machine Learning and Deep Learning Models for Drug-Target Interaction Prediction. *Molecules*, *28*(4), 1663. https://doi.org/10.3390/MOLECULES28041663/S1

Bansal, D., Gupta, B., Gupta, S., Anand, A., Sumit, & Sagar, A. (2025). Facial Recognition Advancements with Siamese Networks: A Comprehensive Survey. *Communications in Computer and Information Science*, *2185 CCIS*, 29-41. https://doi.org/10.1007/978-3-031-71484-9_3

Barberi, G., Cocina, O., Giampiccolo, E., Musumeci, C., & Intrusive, P. D. (2012). *Osservatorio Etneo - Sezione di Catania Pubblicazioni 2012*. *September 2008*, 2009-2012. https://books.google.com/books/about/Advances_in_Data_Mining_Knowledge_Discov.html?id=xeiPDwAAQBAJ

Bate-Eya, A. E. (2024). Empirically comparing the performance of LDA and QDA when classifying customer sales data with different properties of normality and equality of covariance matrices. *Master of Science in Mathematics*. https://digitalcommons.shawnee.edu/math_etd/82

Batista, G. E. A. P. A., Carvalho, A. C. P. L. F., & Monard, M. C. (2000). Applying One-Sided Selection to Unbalanced Datasets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1793 LNAI*, 315-325. https://doi.org/10.1007/10720076_29

Bedi, P., Gupta, N., & Jindal, V. (2020). Siam-IDS: Handling class imbalance problem in Intrusion Detection Systems using Siamese Neural Network. *Procedia Computer Science*, *171*, 780-789. https://doi.org/10.1016/J.PROCS.2020.04.085

Belitz, H.-D., & Grosch, W. (1999). Amino Acids, Peptides, Proteins. *Food Chemistry*, 8-91. https://doi.org/10.1007/978-3-662-07281-3_2

Bozcuk, H., & Yıldız, M. (2024). A Balanced Bagging Classifier machine learning model-based web application to predict risk of febrile neutropenia in cancer patients. *medRxiv*, 2024.08.06.24311562. https://doi.org/10.1101/2024.08.06.24311562

Brandt, J., & Lanzén, E. (2021). *A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification*. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-432162

Bromley, J., Guyon, I., Lecun, Y., Sickinger, E., Shah, R., Bell, A., & Holmdel, L. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. *Advances in Neural Information Processing Systems*, *6*.

Caron, G., Digiesi, V., Solaro, S., & Ermondi, G. (2020). Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. Içinde *Drug Discovery Today* (C. 25, Sayı 4, ss. 621-627). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2020.01.012

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data 2025 12:1*, *12*(1), 1-58. https://doi.org/10.1186/S40537-025-01119-4

Charoenkwan, P., Chotpatiwetchkul, W., Lee, V. S., Nantasenamat, C., & Shoombuatong, W. (2021). A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Scientific Reports 2021 11:1*, *11*(1), 1-15. https://doi.org/10.1038/s41598-021-03293-w

Chatterjee, S. (2023). The Advent of Proteins. *From Stardust to First Cells*, 155-169. https://doi.org/10.1007/978-3-031-23397-5_14

Chatzimilioudis, G., Zeinalipour-Yazti, D., Lee, W. C., & Dikaiakos, M. D. (2012). Continuous all k-nearest-neighbor querying in smartphone networks. *Proceedings - 2012 IEEE 13th International Conference on Mobile Data Management, MDM 2012*, 79-88. https://doi.org/10.1109/MDM.2012.19

Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*, 875-886. https://doi.org/10.1007/978-0-387-09823-4_45

Chen, J., Peng, H., Han, G., Cai, H., Bioinformatics, J. C.-, & 2019, undefined. (2019). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *academic.oup.comJ Chen, H Peng, G Han, H Cai, J CaiBioinformatics, 2019•academic.oup.com*, *35*(4), 602-610. https://doi.org/10.1093/bioinformatics/bty662

Chen, J., Zou, Q., & Li, J. (2022). DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. *Frontiers of Computer Science*, *16*(2), 1-7. https://doi.org/10.1007/S11704-020-0180-0/METRICS

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., & Zhang, Y. (2016). Drug–target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics*, *17*(4), 696-712. https://doi.org/10.1093/BIB/BBV066

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Webb, G. I., Smith, A. I., Daly, R. J., Chou, K. C., & Song, J. (2018). IFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, *34*(14), 2499-2502. https://doi.org/10.1093/bioinformatics/bty140

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., Zhang Corresponding authors, X., & Zhang, X. (2021). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *academic.oup.comL Cheng, X Han, Z Zhu, C Qi, P Wang, X ZhangBriefings in bioinformatics, 2021•academic.oup.com*, *22*(2), 1442-1450. https://doi.org/10.1093/bib/bbab042

Chicco, D. (2021a). Siamese Neural Networks: An Overview. *Methods in Molecular Biology*, *2190*, 73-94. https://doi.org/10.1007/978-1-0716-0826-5_3

Chicco, D. (2021b). Siamese Neural Networks: An Overview. *Methods in Molecular Biology*, *2190*, 73-94. https://doi.org/10.1007/978-1-0716-0826-5_3

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, *I*, 539-546. https://doi.org/10.1109/CVPR.2005.202

Choubey, D. K., Kumar, M., Shukla, V., Tripathi, S., & Dhandhania, V. K. (2020). Comparative Analysis of Classification Methods with PCA and LDA for Diabetes. *Current Diabetes Reviews*, *16*(8), 833-850. https://doi.org/10.2174/1573399816666200123124008

Cichonska, A., Pahikkala, T., Szedmak, S., Julkunen, H., Airola, A., Heinonen, M., Aittokallio, T., & Rousu, J. (2018). Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, *34*(13), i509-i518. https://doi.org/10.1093/BIOINFORMATICS/BTY277

Cobanoglu, M., Liu, C., Hu, F., … Z. O.-J. of chemical, & 2013, undefined. (2013). Predicting drug–target interactions using probabilistic matrix factorization. *ACS PublicationsMC Cobanoglu, C Liu, F Hu, ZN Oltvai, I BaharJournal of chemical information and modeling, 2013•ACS Publications*, *53*(12), 3399-3409. https://doi.org/10.1021/ci400219z

Corchado, J. M., Kim, B.-G., Iglesias, C. A., Lee, I., Ren, F., Mehmood, R., Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms.

Electronics 2023, Vol. 12, Page 1789, *12*(8), 1789. https://doi.org/10.3390/ELECTRONICS12081789

Cortés-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Méndez-Lucio, O., Ijzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T. E., Van Westen, G. J. P., & Bender, A. (2015). Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm*, *6*(1), 24-50. https://doi.org/10.1039/C4MD00216D

Dai, W. (2019). *A Feature Enrichment and Feature Selection Approach to Improve Prediction of Drug Target Interaction and Ligand-based Virtual Screening*.

de la Bourdonnaye, F., & Daniel, F. (2022). *Evaluating resampling methods on a real-life highly imbalanced online credit card payments dataset*. https://arxiv.org/pdf/2206.13152

Dewangan, Y., Berdimurodov, E., & Verma, D. K. (2023). Amino acids: Classification, synthesis methods, reactions, and determination. *Handbook of Biomolecules: Fundamentals, Properties and Applications*, 3-23. https://doi.org/10.1016/B978-0-323-91684-4.00015-3

Dong, J., Yao, Z. J., Zhang, L., Luo, F., Lin, Q., Lu, A. P., Chen, A. F., & Cao, D. S. (2018). PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *Journal of Cheminformatics*, *10*(1). https://doi.org/10.1186/s13321-018-0270-2

Dutta, S., Zunjare, R. U., Sil, A., Mishra, D. C., Arora, A., Gain, N., Chand, G., Chhabra, R., Muthusamy, V., & Hossain, F. (2024). Prediction of matrilineal specific patatin-like protein governing in-vivo maternal haploid induction in maize using support vector machine and di-peptide composition. *Amino Acids*, *56*(1). https://doi.org/10.1007/s00726-023-03368-0

El-Zathry, N. E., Akinlabi, S., Woo, W. L., Patel, V., Mahamood, R. M., & Sabry, I. (2025). Enhancing Friction Stir-Based Techniques with Machine Learning: A Comprehensive Review. *Machine Learning: Science and Technology*. https://doi.org/10.1088/2632-2153/ADCFF6

Ender, A., Brown, N., & Cheminformatics,]. (t.y.). *Special Issue:C heminformatics in Drug Discovery*. https://doi.org/10.1002/cmdc.v13.6/issuetoc

Ependi, U., Rochim, A. F., & Wibowo, A. (t.y.). A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods. *International Journal of Intelligent Engineering and Systems*, *16*(3), 2023. https://doi.org/10.22266/ijies2023.0630.28

Ezzat, A., Wu, M., Li, X., & Kwoh, C. K. (2019). Computational Prediction of Drug-Target Interactions via Ensemble Learning. *Methods in Molecular Biology*, *1903*, 239-254. https://doi.org/10.1007/978-1-4939-8955-3_14

Ezzat, A., Zhao, P., Wu, M., … X. L.-I. transactions, & 2016, undefined. (2015). Drug-target interaction prediction with graph regularized matrix factorization. *ieeexplore.ieee.orgA Ezzat, P Zhao, M Wu, XL Li, CK KwohIEEE/ACM transactions on computational biology and bioinformatics, 2016•ieeexplore.ieee.org*. https://doi.org/10.1109/TCBB.2016.2530062

Fennema, O. R., Damodaran, S., & Parkin, K. L. (2017). *Amino Acids, Peptides, and Proteins*. 235-356. https://doi.org/10.1201/9781315372914-6

Ferdous, S. M., Mugdha, S. B. S., & Dehzangi, I. (2024). New Multi-View Feature Learning Method for Accurate Antifungal Peptide Detection. *Algorithms*, *17*(6). https://doi.org/10.3390/a17060247

Forslund, I. (2022). *Modification of the RusBoost algorithm : A comparison of classifiers on imbalanced data*. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-196899

Furlan, V., Konc, J., Molecules, U. B.-, & 2018, undefined. (2018). Inverse molecular docking as a novel approach to study anticarcinogenic and anti-neuroinflammatory effects of curcumin. *mdpi.comV Furlan, J Konc, U BrenMolecules, 2018•mdpi.com*. https://doi.org/10.3390/molecules23123351

Gabriel, R. A., Harjai, B., Simpson, S., Goldhaber, N., Curran, B. P., & Waterman, R. S. (2022). Machine Learning-Based Models Predicting Outpatient Surgery End Time and Recovery

Room Discharge at an Ambulatory Surgery Center. *Anesthesia and Analgesia*, *135*(1), 159-169. https://doi.org/10.1213/ANE.0000000000006015

Gangrade, A., Kag, A., & Saligrama, V. (2021). *Selective Classification via One-Sided Prediction* (ss. 2179-2187). PMLR. https://proceedings.mlr.press/v130/gangrade21a.html

Gao, K., Fokoue, A., Luo, H., Iyengar, A., Dey, S., IJCAI, P. Z.-, & 2018, undefined. (2018). Interpretable drug target prediction using deep neural representation. *ijcai.orgKY Gao, A Fokoue, H Luo, A Iyengar, S Dey, P ZhangIJCAI, 2018•ijcai.org*. https://www.ijcai.org/proceedings/2018/0468.pdf

Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of Support Vector Machine Algorithm in Big Data Background. *Mathematical Problems in Engineering*, *2021*(1), 5594899. https://doi.org/10.1155/2021/5594899

Gen, Mitsuo., & Cheng, Runwei. (2000). Genetic algorithms and engineering design. 576. https://books.google.com/books/about/Genetic_Algorithms_and_Engineering_Optim.html?id=U7MuV1q6P1oC

Gewers, F.L., Ferreira, G.R., De Arruda, H.F., Silva, F.N., Comin, C.H., Amancio, D.R., and Costa, L.D.F. (2021) .Principal Component Analysis, ACM Computing Surveys (CSUR), 54 (4). https://doi.org/10.1145/3447755

Ghojogh, B., Ca, B., Crowley, M., & Ca, M. (2019). *Linear and Quadratic Discriminant Analysis: Tutorial*. https://arxiv.org/pdf/1906.02590

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, *2017-January*, 79-85. https://doi.org/10.1109/ICACCI.2017.8125820

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 1-21. https://doi.org/10.1038/S43586-022-00184-W;SUBJMETA

Grimplet, J., Adam-Blondon, A. F., Bert, P. F., Bitz, O., Cantu, D., Davies, C., Delrot, S., Pezzotti, M., Rombauts, S., & Cramer, G. R. (2014). The grapevine gene nomenclature system. *BMC Genomics*, *15*(1), 1-14. https://doi.org/10.1186/1471-2164-15-1077/FIGURES/7

Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data mining on imbalanced data sets. *Proceedings - 2008 International Conference on Advanced Computer Theory and Engineering, ICACTE 2008*, 1020-1024. https://doi.org/10.1109/ICACTE.2008.26

Guo, Z., Wang, P., Liu, Z., & Zhao, Y. (2020). Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.584807

Gupta, H., & Sahare, M. (2012). A Review of Multi-Class Classification for Imbalanced Data. Içinde *Article in International Journal of Advanced Computer Research*. https://www.researchgate.net/publication/296337789

Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., Schneider, R., Skoblo, R., Russell, R. B., Bourne, P. E., Bork, P., & Preissner, R. (2008). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, *36*(suppl_1), D919-D922. https://doi.org/10.1093/NAR/GKM862

Habibi, M., Starlinger, J., & Leser, U. (2020). TabSim: A Siamese Neural Network for Accurate Estimation of Table Similarity. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, 930-937. https://doi.org/10.1109/BIGDATA50022.2020.9378077

Hasan Mahmud, S. M., Chen, W., Jahan, H., Dai, B., Din, S. U., & Dzisoo, A. M. (2020). DeepACTION: A deep learning-based method for predicting novel drug-target interactions. *Analytical Biochemistry*, *610*, 113978. https://doi.org/10.1016/J.AB.2020.113978

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: investigating data sampling approaches. *Journal of Big Data*, *6*(1), 1-25. https://doi.org/10.1186/S40537-019-0274-4/FIGURES/3

Hay, M., Thomas, D., … J. C.-N., & 2014, undefined. (2014). Clinical development success rates for investigational drugs. *nature.comM Hay, DW Thomas, JL Craighead, C Economides, J RosenthalNature biotechnology, 2014•nature.com*. https://www.nature.com/articles/nbt.2786

He, A., Luo, C., Tian, X., & Zeng, W. (2018). *A Twofold Siamese Network for Real-Time Object Tracking* (ss. 4834-4843).

Hertzberg, R. P., & Pope, A. J. (2000). High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology*, *4*(4), 445-451. https://doi.org/10.1016/S1367-5931(00)00110-1

Hoffer, L. J. (2016). Human Protein and Amino Acid Requirements. *Journal of Parenteral and Enteral Nutrition*, *40*(4), 460-474. https://doi.org/10.1177/0148607115624084;WGROUP:STRING:PUBLICATION

Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2021). MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. *Bioinformatics*, *37*(6), 830-836. https://doi.org/10.1093/BIOINFORMATICS/BTAA880

Huang, Y., Zhu, L., Tan, H., Tian, F., & Zheng, F. (2020). Predicting drug-target on heterogeneous network with co-rank. *Advances in Intelligent Systems and Computing*, *905*, 571-581. https://doi.org/10.1007/978-3-030-14680-1_63

Ilina, O., Ziyadinov, V., Klenov, N., & Tereshonok, M. (2022). A Survey on Symmetrical Neural Network Architectures and Applications. *Symmetry 2022, Vol. 14, Page 1391*, *14*(7), 1391. https://doi.org/10.3390/SYM14071391

Imming, P., Sinning, C., & Meyer, A. (2006). Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery*, *5*(10), 821-834. https://doi.org/10.1038/NRD2132

Ismail, M., Hassan, N., & Saleh Bafjaish, S. (2020). Comparative Analysis of Naive Bayesian Techniques in Health-Related For Classification Task. *Journal of Soft Computing and Data Mining*, *1*(2), 1-10. https://doi.org/10.30880/jscdm.2020.01.02.001

Jacob, L., & Vert, J. P. (2008). Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, *24*(19), 2149-2156. https://doi.org/10.1093/BIOINFORMATICS/BTN409

Jan, A., Hayat, M., Wedyan, M., Alturki, R., Gazzawe, F., Ali, H., & Alarfaj, F. K. (2022). Target-AMP: Computational prediction of antimicrobial peptides by coupling sequential information with evolutionary profile. *Computers in Biology and Medicine*, *151*. https://doi.org/10.1016/j.compbiomed.2022.106311

Jana, S., Ganeshpurkar, A., & Singh, S. K. (2018). Multiple 3D-QSAR modeling, e-pharmacophore, molecular docking, and in vitro study to explore novel AChE inhibitors. *RSC Advances*, *8*(69), 39477-39495. https://doi.org/10.1039/C8RA08198K

Javaid, N., Jan, N., & Javed, M. U. (2021a). An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. *Journal of Parallel and Distributed Computing*, *153*, 44-52. https://doi.org/10.1016/j.jpdc.2021.03.002

Javaid, N., Jan, N., & Javed, M. U. (2021b). An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. *Journal of Parallel and Distributed Computing*, *153*, 44-52. https://doi.org/10.1016/J.JPDC.2021.03.002

Jenkins, J. L. (2012). Large-scale QSAR in target prediction and phenotypic HTS assessment. *Molecular Informatics*, *31*(6-7), 508-514. https://doi.org/10.1002/MINF.201200002

Jiang, Y., Wang, D., Wang, W., & Xu, D. (2021). Computational methods for protein localization prediction. Içinde *Computational and Structural Biotechnology Journal* (C. 19, ss. 5834-5844). Elsevier B.V. https://doi.org/10.1016/j.csbj.2021.10.023

Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*(D1), D109-D114. https://doi.org/10.1093/NAR/GKR988

Kanj, S., Abdallah, F., Denœux, T., & Tout, K. (2016). Editing training data for multi-label classification with the k-nearest neighbor rule. *Pattern Analysis and Applications*, *19*(1), 145-161. https://doi.org/10.1007/S10044-015-0452-8/METRICS

Khawaja, S. A., Farooq, M. S., Ishaq, K., Alsubaie, N., Karamti, H., Montero, E. C., Alvarado, E. S., & Ashraf, I. (2024a). Prediction of leukemia peptides using convolutional neural network and protein compositions. *BMC Cancer*, *24*(1). https://doi.org/10.1186/s12885-024-12609-8

Khawaja, S. A., Farooq, M. S., Ishaq, K., Alsubaie, N., Karamti, H., Montero, E. C., Alvarado, E. S., & Ashraf, I. (2024b). Prediction of leukemia peptides using convolutional neural network and protein compositions. *BMC Cancer*, *24*(1). https://doi.org/10.1186/s12885-024-12609-8

Kherif, F. ve Latypova, A.(2020). Principal component analysis, Machine Learning: Methods and Applications to Brain Disorders, Bölüm 12, Elsevier, Amsterdam, Hollanda, ISBN: 978-0-12-815739-8, 209–225.

Konstantinov, A. F., Федорович, К. А., Dyakonova, L. P., & Павловна, Д. Л. (2025). Comparative analysis of class imbalance reduction methods in building machine learning models in the financial sector. *News of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences*, *27*(1), 143-150. https://doi.org/10.35330/1991-6639-2025-27-1-143-151

Kuncheva, L. I., Arnaiz-González, Á., Díez-Pastor, J. F., & Gunn, I. A. D. (2019). Instance selection improves geometric mean accuracy: a study on imbalanced data classification. *Progress in Artificial Intelligence*, *8*(2), 215-228. https://doi.org/10.1007/S13748-019-00172-4/METRICS

Lambrinidis, G., & Tsantili-Kakoulidou, A. (2021). Multi-objective optimization methods in novel drug design. *Expert Opinion on Drug Discovery*, *16*(6), 647-658. https://doi.org/10.1080/17460441.2021.1867095

Lango, M., & Stefanowski, J. (2018). Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data. *Journal of Intelligent Information Systems*, *50*(1), 97-127. https://doi.org/10.1007/s10844-017-0446-7

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology 2007 8:12*, *8*(12), 995-1005. https://doi.org/10.1038/nrm2281

Lee, T. Y., Chen, S. A., Hung, H. Y., & Ou, Y. Y. (2011). Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PLoS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0017331

Li, C., Zhang, R., Wang, J., Wilson, L. M., & Yan, Y. (2020). Protein Engineering for Improving and Diversifying Natural Product Biosynthesis. *Trends in Biotechnology*, *38*(7), 729-744. https://doi.org/10.1016/J.TIBTECH.2019.12.008

Li, L., biology, M. C.-I. transactions on computational, & 2017, undefined. (t.y.). Drug target prediction by multi-view low rank embedding. *ieeexplore.ieee.orgL Li, M CaiIEEE/ACM transactions on computational biology and bioinformatics, 2017•ieeexplore.ieee.org*. Geliş tarihi 07 Nisan 2025, gönderen https://ieeexplore.ieee.org/abstract/document/7931669/

Li, P., Yan, H., & Lu, X. (2023). A Siamese neural network for learning the similarity metrics of linear features. *International Journal of Geographical Information Science*, *37*(3), 684-711. https://doi.org/10.1080/13658816.2022.2143505

Li, Y., Chen, C. L. P., & Zhang, T. (2022). A Survey on Siamese Network: Methodologies, Applications, and Opportunities. *IEEE Transactions on Artificial Intelligence*, *3*(6), 994-1014. https://doi.org/10.1109/TAI.2022.3207112

Li, Y., Li, L. P., Wang, L., Yu, C. Q., Wang, Z., & You, Z. H. (2019). An Ensemble Classifier to Predict Protein–Protein Interactions by Combining PSSM-based Evolutionary

Information with Local Binary Pattern Model. *International Journal of Molecular Sciences 2019, Vol. 20, Page 3511*, *20*(14), 3511. https://doi.org/10.3390/IJMS20143511

Liladhar Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. *Studies in Medical and Health Sciences*, *1*(2), 18-41. https://doi.org/10.48185/SMHS.V1I2.1225

Liu, Y., Zhao, C. C., Liang, H. Y., Lu, H. H., Cui, N. Y., & Bao, K. Y. (2022). A rotor fault diagnosis method based on BP-Adaboost weighted by non-fuzzy solution coefficients. *Measurement*, *196*, 111280. https://doi.org/10.1016/J.MEASUREMENT.2022.111280

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., & Zeng, J. (t.y.). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *nature.comY Luo, X Zhao, J Zhou, J Yang, Y Zhang, W Kuang, J Peng, L Chen, J ZengNature communications, 2017•nature.com*. https://doi.org/10.1038/s41467-017-00680-8

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*(1), 155-161. https://doi.org/10.1007/S12041-013-0212-8/METRICS

Mahdaddi, A., Meshoul, S., & Belguidoum, M. (2021). EA-based hyperparameter optimization of hybrid deep learning models for effective drug-target interactions prediction. *Expert Systems with Applications*, *185*, 115525. https://doi.org/10.1016/J.ESWA.2021.115525

Mahmud, S. M. H., Chen, W., Liu, Y., Awal, M. A., Ahmed, K., Rahman, M. H., & Moni, M. A. (2021). PreDTIs: prediction of drug–target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Briefings in Bioinformatics*, *22*(5), 1-20. https://doi.org/10.1093/BIB/BBAB046

Mahmud, S. M. H., Chen, W., Meng, H., Jahan, H., Liu, Y., & Hasan, S. M. M. (2020a). Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Analytical Biochemistry*, *589*. https://doi.org/10.1016/J.AB.2019.113507

Mahmud, S. M. H., Chen, W., Meng, H., Jahan, H., Liu, Y., & Hasan, S. M. M. (2020b). Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Analytical Biochemistry*, *589*, 113507. https://doi.org/10.1016/J.AB.2019.113507

Mamilla, M. Y., Al-Haddad, R., & Chowdhury, S. (2025). Resampling Imbalanced Healthcare Data for Predictive Modelling. *International Journal of Advanced Computer Science and Applications*, *16*(2), 36-44. https://doi.org/10.14569/IJACSA.2025.0160204

Martin, L., Hutchens, M., Hawkins, C., & Radnov, A. (2017). How much do clinical trials cost? *Nature Reviews Drug Discovery*, *16*(6), 381-382. https://doi.org/10.1038/NRD.2017.70

Mathan, K., Kumar, P. M., Panchatcharam, P., Manogaran, G., & Varadharajan, R. (2018). A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. *Design Automation for Embedded Systems*, *22*(3), 225-242. https://doi.org/10.1007/S10617-018-9205-4/METRICS

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

Milli, M. E. F., Aras, S., & Deveci Kocakoç, İ. (2024). Investigating the Effect of Class Balancing Methods on the Performance of Machine Learning Techniques: Credit Risk Application. *İzmir Yönetim Dergisi*, *5*(1), 55-70. https://doi.org/10.56203/iyd.1436742

Mohamed, S. K., Nováček, V., & Nounu, A. (2020). Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, *36*(2), 603-610. https://doi.org/10.1093/BIOINFORMATICS/BTZ600

Mostafa Monowar, M., Nobel, S. M. N., Afroj, M., Hamid, M. A., Uddin, M. Z., Kabir, M. M., & Mridha, M. F. (2024). Advanced sleep disorder detection using multi-layered ensemble learning and advanced data balancing techniques. *Frontiers in Artificial Intelligence*, *7*, 1506770. https://doi.org/10.3389/FRAI.2024.1506770/BIBTEX

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Applied System Innovation 2021, Vol. 4, Page 18*, *4*(1), 18. https://doi.org/10.3390/ASI4010018

Muley, R. R., Battula, B. P., & Sanaboina, C. S. (2025). An Empirical Literature Review on Multi Class Data Stream Learning with Class Imbalance Problem in Real World Data Sources. *ESIC 2025 - 5th International Conference on Emerging Systems and Intelligent Computing, Proceedings*, 90-95. https://doi.org/10.1109/ESIC64052.2025.10962587

Murase, K., Morrison, K. L., Tam, P. Y., Stafford, R. L., Jurnak, F., & Weiss, G. A. (2003). EF-Tu binding peptides identified, dissected, and affinity optimized by phage display. *cell.comK Murase, KL Morrison, PY Tam, RL Stafford, F Jurnak, GA WeissChemistry & biology, 2003•cell.com*, *10*, 161-168. https://doi.org/10.1016/S

Nath, A., Kumari, P., & Chaube, R. (2018). Prediction of human drug targets and their interactions using machine learning methods: Current and future perspectives. *Methods in Molecular Biology*, *1762*, 21-30. https://doi.org/10.1007/978-1-4939-7756-7_2

Naveja, J. J., & Medina-Franco, J. L. (2017). ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Research*, *6*, 1134. https://doi.org/10.12688/f1000research.12095.1

Nazim Uddin, M., Ferdous Bin Hafiz, M., Hossain, S., & Mohammad Mominul Islam, S. (t.y.). Drug Sentiment Analysis using Machine Learning Classifiers. *IJACSA) International Journal of Advanced Computer Science and Applications*, *13*(1), 2022. Geliş tarihi 29 Nisan 2025, gönderen www.ijacsa.thesai.org

Niaz, N. U., Shahariar, K. M. N., & Patwary, M. J. A. (2022). Class Imbalance Problems in Machine Learning: A Review of Methods And Future Challenges. *ACM International Conference Proceeding Series*, 485-490. https://doi.org/10.1145/3542954.3543024;JOURNAL:JOURNAL:ACMOTHERCONFERENCES;CTYPE:STRING:BOOK

Nizam-Ozogur, H., & Orman, Z. (2024a). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. *Expert Systems*, *41*(8),

e13596. https://doi.org/10.1111/EXSY.13596;WEBSITE:WEBSITE:PERICLES;REQUESTEDJOURNAL:JOURNAL:14680394;WGROUP:STRING:PUBLICATION

Nizam-Ozogur, H., & Orman, Z. (2024b). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. *Expert Systems*, *41*(8), e13596. https://doi.org/10.1111/EXSY.13596

Ondrasovic, M., & Tarabek, P. (2021a). Siamese Visual Object Tracking: A Survey. *IEEE Access*, *9*, 110149-110172. https://doi.org/10.1109/ACCESS.2021.3101988

Ondrasovic, M., & Tarabek, P. (2021b). Siamese Visual Object Tracking: A Survey. *IEEE Access*, *9*, 110149-110172. https://doi.org/10.1109/ACCESS.2021.3101988

Ongko, E., & Abdullah, D. (2015). Determining a Cluster Centroid of K-Means Clustering Using Genetic Algorithm. *International Journal of Computer Science and Software Engineering (IJCSSE)*, *4*(6). www.IJCSSE.org

Panahi, R., Sedghiye, N., & Nazerfard, E. (2020). A novel data cleaning method for learning in imbalanced datasets based on k-nearest neighbors. *2020 10h International Conference on Computer and Knowledge Engineering, ICCKE 2020*, 475-482. https://doi.org/10.1109/ICCKE50421.2020.9303625

Panjainam, P., & Kanjanawattana, S. (2024). A Comparison of the Hybrid Resampling Techniques for Imbalanced Medical Data. *ACM International Conference Proceeding Series*, 46-50. https://doi.org/10.1145/3702468.3702477/ASSET/C745B5CE-535D-40ED-BDB9-F4ADF33D8730/ASSETS/GRAPHIC/ICRSA2024-9-FIG5.JPG

Paoletti, M. E., Mogollon-Gutierrez, O., Moreno-Alvarez, S., Sancho, J. C., & Haut, J. M. (2023). A Comprehensive Survey of Imbalance Correction Techniques for Hyperspectral Data Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 5297-5314. https://doi.org/10.1109/JSTARS.2023.3279506

Papantonis, A., & Cook, P. R. (2013). Transcription factories: Genome organization and gene regulation. *Chemical Reviews*, *113*(11), 8683-8705. https://doi.org/10.1021/CR300513P/ASSET/CR300513P.FP.PNG_V03

Park, Y. S., & Lek, S. (2016). Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. *Developments in Environmental Modelling*, *28*, 123-140. https://doi.org/10.1016/B978-0-444-63623-2.00007-4

Patel, P., Garala, K., & Raval, M. (2023). Defining the Role of Chemical and Biological Data and Applying Algorithms. *Artificial Intelligence in Bioinformatics and Chemoinformatics*, 221-238. https://doi.org/10.1201/9781003353768-12/DEFINING-ROLE-CHEMICAL-BIOLOGICAL-DATA-APPLYING-ALGORITHMS-PRIYA-PATEL-KEVINKUMAR-GARALA-MIHIR-RAVAL

Patle, A., & Chouhan, D. S. (2013). SVM kernel functions for classification. *2013 International Conference on Advances in Technology and Engineering, ICATE 2013*. https://doi.org/10.1109/ICADTE.2013.6524743

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *nature.comSM Paul, DS Mytelka, CT Dunwiddie, CC Persinger, BH Munos, SR Lindborg, AL SchachtNature reviews Drug discovery, 2010•nature.com*, *9*(3), 203. https://doi.org/10.1038/nrd3078

Peng, J., Li, J., & Shang, X. (2020). A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinformatics*, *21*(13), 1-13. https://doi.org/10.1186/S12859-020-03677-1/TABLES/3

Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., Wei, Z., & Shang, X. (2021). An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Briefings in Bioinformatics*, *22*(5), 1-9. https://doi.org/10.1093/BIB/BBAA430

Peska, L., Buza, K., in, J. K.-C. methods and programs, & 2017, undefined. (t.y.). Drug-target interaction prediction: a Bayesian ranking approach. *ElsevierL Peska, K Buza, J KollerComputer methods and programs in biomedicine, 2017•Elsevier*. Geliş tarihi 02 Nisan 2025, gönderen https://www.sciencedirect.com/science/article/pii/S0169260717303413

pharmacology, T. K.-B. journal of, & 2007, undefined. (2007). Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Wiley Online LibraryT KlabundeBritish journal of pharmacology, 2007•Wiley Online Library*, *152*(1), 5-7. https://doi.org/10.1038/sj.bjp.0707308

Pirgazi, J., Khanteymoori, A. R., & Jalilkhani, M. (2018). GENIRF: An Algorithm for Gene Regulatory Network Inference Using Rotation Forest. *Current Bioinformatics*, *13*(4), 407-419. https://doi.org/10.2174/1574893612666170731120830

Poluri, K. M., Gulati, K., & Sarkar, S. (2021). Structural and Functional Properties of Proteins. *Protein-Protein Interactions*, 1-60. https://doi.org/10.1007/978-981-16-1594-8_1

Qiu, W., Li, S., Cui, X., Yu, Z., Wang, M., Du, J., Peng, Y., & Yu, B. (2018). Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. *Journal of Theoretical Biology*, *450*, 86-103. https://doi.org/10.1016/j.jtbi.2018.04.026

Rahman, H. A. A., Wah, Y. B., He, H., & Bulgiba, A. (2015). Comparisons of ADABOOST, KNN, SVM and Logistic Regression in Classification of Imbalanced Dataset. *Communications in Computer and Information Science*, *545*, 54-64. https://doi.org/10.1007/978-981-287-936-3_6

Rahmayanti, I. A., Sediono, Saifudin, T., & Ana, E. (2021). APPLYING SMOTE-NC ON CART ALGORITHM TO HANDLE IMBALANCED DATA IN CUSTOMER CHURN PREDICTION: A CASE STUDY OF TELECOMMUNICATIONS INDUSTRY. *Journal of Syntax Literate*, *6*, 1321-1337. https://openurl.ebsco.com/contentitem/asn:156463542?sid=ebsco:plink:crawler&id=ebsco:asn:156463542&crl=c

Ramya Shri, G., Devasena, T., Janani, M., & Karthick, R. (2024). Accurate Data Sampling Methods for Medical Data - Survey. *Proceedings - 2024 5th International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2024*, 77-84. https://doi.org/10.1109/ICMCSI61536.2024.00018

Redkar, S., Mondal, S., Joseph, A., & Hareesha, K. S. (2020). A Machine Learning Approach for Drug-target Interaction Prediction using Wrapper Feature Selection and Class Balancing. *Molecular Informatics*, *39*(5). https://doi.org/10.1002/minf.201900062

Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, *35*(23), 17095-17112. https://doi.org/10.1007/S00521-023-08568-Z/TABLES/7

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics*, *20*(5), 1878-1912. https://doi.org/10.1093/BIB/BBY061

Roshan, S. E., & Asadi, S. (2020). Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. *Engineering Applications of Artificial Intelligence*, *87*, 103319. https://doi.org/10.1016/J.ENGAPPAI.2019.103319

Sachdev, K., & Gupta, M. K. (2019). A comprehensive review of feature based methods for drug target interaction prediction. *Journal of Biomedical Informatics*, *93*, 103159. https://doi.org/10.1016/J.JBI.2019.103159

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522-542. https://doi.org/10.1016/J.INS.2021.05.055

Sahito, A., Frank, E., & Pfahringer, B. (2019). Semi-supervised Learning Using Siamese Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11919 LNAI*, 586-597. https://doi.org/10.1007/978-3-030-35288-2_47

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., Overington, J. P., & Ai;, -L Jpo@benevolent. (2017). A comprehensive map of molecular drug targets. *nature.comR Santos, O Ursu, A Gaulton, AP Bento, RS Donadi, CG Bologa, A Karlsson, B Al-LazikaniNature reviews Drug discovery, 2017•nature.com*, *16*(1), 19-34. https://doi.org/10.1038/nrd.2016.230

Saravanan, V., & Gautham, N. (2015). Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *https://home.liebertpub.com/omi*, *19*(10), 648-658. https://doi.org/10.1089/OMI.2015.0095

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, *32*(suppl_1), D431-D433. https://doi.org/10.1093/NAR/GKH081

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2008.4761297

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *40*(1), 185-197. https://doi.org/10.1109/TSMCA.2009.2029559

Serrano, N., & Bellogín, A. (2023). Siamese neural networks in recommendation. *Neural Computing and Applications*, *35*(19), 13941-13953. https://doi.org/10.1007/S00521-023-08610-0/TABLES/8

Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence*, *94*, 103770. https://doi.org/10.1016/J.ENGAPPAI.2020.103770

Sharma, A., biomedicine, R. R.-C. methods and programs in, & 2018, undefined. (t.y.). BE-DTI': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning. *ElsevierA Sharma, R RaniComputer methods and programs in biomedicine, 2018•Elsevier*. Geliş tarihi 08 Nisan 2025, gönderen https://www.sciencedirect.com/science/article/pii/S016926071830868X

Shafizadeh-Moghadam, H.(2021) .Fully component selection: An efficient combination of feature selection and principal component analysis to increase model performance, Expert Systems with Applications, 186, 115678. https://doi.org/10.1016/J.ESWA.2021.115678

Sheridan, R. P., McMasters, D. R., Voigt, J. H., & Wildey, M. J. (2015). ECounterscreening: Using QSAR predictions to prioritize testing for off-target activities and setting the balance between benefit and risk. *Journal of Chemical Information and Modeling*, *55*(2), 231-238. https://doi.org/10.1021/CI500666M

Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised machine learning: Algorithms and applications. *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools, and Applications*, 1-16. https://doi.org/10.1002/9781119821908.CH1;CTYPE:STRING:BOOK

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., Genomics, B. Y.-, & 2019, undefined. (t.y.). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *ElsevierH Shi, S Liu, J Chen, X Li, Q Ma, B YuGenomics, 2019•Elsevier*. Geliş tarihi 07 Nisan 2025, gönderen https://www.sciencedirect.com/science/article/pii/S088875431830466X

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., & Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, *111*(6), 1839-1852. https://doi.org/10.1016/J.YGENO.2018.12.007

Shi, J., Yiu, S., Li, Y., Leung, H., Methods, F. C.-, & 2015, undefined. (t.y.). Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *ElsevierJY Shi, SM Yiu, Y Li, HCM Leung, FYL ChinMethods, 2015•Elsevier*. Geliş tarihi 07 Nisan 2025, gönderen https://www.sciencedirect.com/science/article/pii/S1046202315001905

Shim, J., Hong, Z. Y., Sohn, I., & Hwang, C. (2021). Prediction of drug–target binding affinity using similarity-based convolutional neural network. *Scientific Reports 2021 11:1*, *11*(1), 1-9. https://doi.org/10.1038/s41598-021-83679-y

Siddappa, N. G., & Kampalappa, T. (2019). Adaptive condensed nearest neighbor for imbalance data classification. *International Journal of Intelligent Engineering and Systems*, *12*(2), 104-113. https://doi.org/10.22266/IJIES2019.0430.11

Sikander, R., Ghulam, A., & Ali, F. (2022). XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-09484-3

Siqueira, L. F. S., Araújo Júnior, R. F., de Araújo, A. A., Morais, C. L. M., & Lima, K. M. G. (2017). LDA vs. QDA for FT-MIR prostate cancer tissue classification. *Chemometrics and Intelligent Laboratory Systems*, *162*, 123-129. https://doi.org/10.1016/J.CHEMOLAB.2017.01.021

Smart Data Intelligence: Proceedings of ICSMDI 2024 - Google Books. (t.y.). Geliş tarihi 19 Temmuz 2025, gönderen https://books.google.com.tr/books?hl=en&lr=&id=uYwWEQAAQBAJ&oi=fnd&pg=PA1&dq=Therefore,+RandomOverSampler+is+typically+used+as+the+first+balancing+step+and+is+later+supplemented+by+more+advanced+techniques+such+as+SMOTE+or+ADASYN.&ots=RQoamGn0k5&sig=Iqecl-WxY0VevECgBIS9gNWvNMU&redir_esc=y#v=onepage&q&f=false

Soleymani, F., Paquet, E., Viktor, H., Michalowski, W., & Spinello, D. (2022). Protein–protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal*, *20*, 5316-5341. https://doi.org/10.1016/J.CSBJ.2022.08.070

Song, T., Zhang, X., Ding, M., Rodriguez-Paton, A., Wang, S., & Wang, G. (2022). DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions. *Methods*, *204*, 269-277. https://doi.org/10.1016/J.YMETH.2022.02.007

Sorkhi, A. G., Abbasi, Z., Mobarakeh, M. I., & Pirgazi, J. (2021). Drug–target interaction prediction using unifying of graph regularized nuclear norm with bilinear factorization. *BMC Bioinformatics*, *22*(1), 1-23. https://doi.org/10.1186/S12859-021-04464-2/FIGURES/3

Sridhar, D., Fakhraei, S., & Getoor, L. (2016). A probabilistic approach for collective similarity-based drug–drug interaction prediction. *Bioinformatics*, *32*(20), 3175-3182. https://doi.org/10.1093/BIOINFORMATICS/BTW342

Sun, S., Ao, C., Wang, D., & Dong, B. (2020). The frequencies of oppositely charged, uncharged polar, and β-branched amino acids determine proteins' thermostability. *IEEE Access*, *8*, 66839-66845. https://doi.org/10.1109/ACCESS.2020.2985737

Suruliandi, A., Idhaya, T., & Raja, S. P. (2024). Drug Target Interaction Prediction Using Machine Learning Techniques – A Review. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(6), 86-100. https://doi.org/10.9781/IJIMAI.2022.11.002

Thafar, M. A., Olayan, R. S., Olayan, R. S., Ashoor, H., Ashoor, H., Albaradei, S., Albaradei, S., Bajic, V. B., Gao, X., Gojobori, T., & Essack, M. (2020). DTiGEMS+: Drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, *12*(1), 1-17. https://doi.org/10.1186/S13321-020-00447-2/TABLES/5

Thangamani, M., & Kalaiselvi, B. (1998). Analyzing the Structure of Protein Synthesis and Transcription. *Advances in Natural and Applied Sciences*, *11*(7), 281-287. http://www.aensiweb.com/http://creativecommons.org/licenses/by/4.0/

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, *30*(2), 169-190. https://doi.org/10.3233/AIC-170729

today, A. L.-D. discovery, & 2015, undefined. (2015). Machine-learning approaches in drug discovery: methods and applications. *ElsevierA LavecchiaDrug discovery today, 2015•Elsevier*, *20*. https://doi.org/10.1016/j.drudis.2014.10.012

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(2), e1289. https://doi.org/10.1002/WIDM.1289

Uykan, Z. (2023). Fusion of Centroid-Based Clustering With Graph Clustering: An Expectation-Maximization-Based Hybrid Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 4068-4082. https://doi.org/10.1109/TNNLS.2021.3121224

Viana, D., Teixeira, R., Baptista, J., & Pinto, T. (2024). Synthetic Data Generation Models for Time Series: A Literature Review. *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2024*. https://doi.org/10.1109/ICECET61485.2024.10698494

Volkova, N. I., & Volkov, A. V. (2023). Medical diagnosis and its nature. *Medical Herald of the South of Russia*, *14*(3), 16-23. https://doi.org/10.21886/2219-8075-2023-14-3-16-23

Wanat, K. (2020). Biological barriers, and the influence of protein binding on the passage of drugs across them. *Molecular Biology Reports*, *47*(4), 3221-3231. https://doi.org/10.1007/S11033-020-05361-2/TABLES/3

Wang, C., Wang, W., Lu, K., Zhang, J., Chen, P., & Wang, B. (2020). Predicting Drug-Target Interactions with Electrotopological State Fingerprints and Amphiphilic Pseudo Amino Acid Composition. *International Journal of Molecular Sciences 2020, Vol. 21, Page 5694*, *21*(16), 5694. https://doi.org/10.3390/IJMS21165694

Wang, D., Yang, L., Fu, Z., & Xia, J. (2011). Prediction of Thermophilic Protein with Pseudo Amino Acid Composition: An Approach from Combined Feature Selection and Reduction. *Protein & Peptide Letters*, *18*(7), 684-689. https://doi.org/10.2174/092986611795446085

Wang, L., You, Z. H., Chen, X., Xia, S. X., Liu, F., Yan, X., Zhou, Y., & Song, K. J. (2018a). A computational-based method for predicting drug-target interactions by using stacked autoencoder deep neural network. *Journal of Computational Biology*, *25*(3), 361-373. https://doi.org/10.1089/CMB.2017.0135

Wang, L., You, Z. H., Chen, X., Xia, S. X., Liu, F., Yan, X., Zhou, Y., & Song, K. J. (2018b). A computational-based method for predicting drug-target interactions by using stacked

autoencoder deep neural network. *Journal of Computational Biology*, *25*(3), 361-373. https://doi.org/10.1089/CMB.2017.0135/ASSET/IMAGES/LARGE/FIGURE8.JPEG

Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., Zhou, Y., & Song, K.-J. (2018c). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *liebertpub.comL Wang, ZH You, X Chen, SX Xia, F Liu, X Yan, Y Zhou, KJ SongJournal of Computational Biology, 2018•liebertpub.com*, *25*(3), 361-373. https://doi.org/10.1089/cmb.2017.0135

Wang, M., Tang, C., International, J. C.-B. R., & 2018, undefined. (2018). Drug-Target Interaction Prediction via Dual Laplacian Graph Regularized Matrix Completion. *Wiley Online LibraryM Wang, C Tang, J ChenBioMed Research International, 2018•Wiley Online Library*, *2018*. https://doi.org/10.1155/2018/1425608

Wang, S., Du, Z., Ding, M., Rodriguez-Paton, A., & Song, T. (2022). KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer's disease drug repositions. *Applied Intelligence*, *52*(1), 846-857. https://doi.org/10.1007/S10489-021-02454-8/METRICS

Wang, Y., Wang, P., Guo, Y., Huang, S., Chen, Y., & Xu, L. (2021). prPred: A Predictor to Identify Plant Resistance Proteins by Incorporating k-Spaced Amino Acid (Group) Pairs. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.645520

Wang, Y., Zhu, K., Sun, M., & Deng, Y. (2019). An Ensemble Learning Approach for Fault Diagnosis in Self-Organizing Heterogeneous Networks. *IEEE Access*, *7*, 125662-125675. https://doi.org/10.1109/ACCESS.2019.2925566

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, *36*(suppl_1), D901-D906. https://doi.org/10.1093/NAR/GKM958

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and

exploration. *Nucleic Acids Research*, *34*(suppl_1), D668-D672. https://doi.org/10.1093/NAR/GKJ067

Xing, L., McDonald, J. J., Kolodziej, S. A., Kurumbail, R. G., Williams, J. M., Warren, C. J., O'Neal, J. M., Skepner, J. E., & Roberds, S. L. (2011). Discovery of potent inhibitors of soluble epoxide hydrolase by combinatorial library design and structure-based virtual screening. *Journal of Medicinal Chemistry*, *54*(5), 1211-1222. https://doi.org/10.1021/JM101382T

Xu, L., Liang, G., Liao, C., Chen, G. Den, & Chang, C. C. (2019). K-SkIP-N-Gram-RF: A random forest based method for Alzheimer's disease protein identification. *Frontiers in Genetics*, *10*(FEB). https://doi.org/10.3389/FGENE.2019.00033/FULL

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48-59. https://doi.org/10.1177/0165551516677946

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., & Kanehisa, M. (2008a). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, *24*(13), i232-i240. https://doi.org/10.1093/BIOINFORMATICS/BTN162

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., & Kanehisa, M. (2008b). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, *24*(13), i232-i240. https://doi.org/10.1093/BIOINFORMATICS/BTN162

Yan, C., Geng, A., Pan, Z., Zhang, Z., & Cui, F. (2024). MultiFeatVotPIP: a voting-based ensemble learning framework for predicting proinflammatory peptides. *Briefings in bioinformatics*, *25*(6). https://doi.org/10.1093/bib/bbae505

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, *94*, 88-104. https://doi.org/10.1016/J.KNOSYS.2015.11.013

You, J., McLeod, R., chemistry, P. H.-C. biology and, & 2019, undefined. (t.y.). Predicting drug-target interaction network using deep learning model. *Elsevier*. Geliş tarihi 07 Nisan 2025, gönderen https://www.sciencedirect.com/science/article/pii/S1476927119301902

Yousefimehr, B., Ghatee, M., Seifi, M. A., Fazli, J., Tavakoli, S., Rafei, Z., Ghaffari, S., Nikahd, A., Gandomani, M. R., Orouji, A., Kashani, R. M., Heshmati, S., & Mousavi, N. S. (2025). *DATA BALANCING STRATEGIES: A SURVEY OF RESAMPLING AND AUGMENTATION METHODS A PREPRINT*.

Yu, H., & Ni, J. (2014). An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(4), 657-666. https://doi.org/10.1109/TCBB.2014.2306838

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., & Zhu, S. (t.y.). DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *academic.oup.comQ Yuan, J Gao, D Wu, S Zhang, H Mamitsuka, S ZhuBioinformatics, 2016•academic.oup.com*. https://doi.org/10.1093/bioinformatics/btw244

Zhang, P., Wei, Z., Che, C., & Jin, B. (2022). DeepMGT-DTI: Transformer network incorporating multilayer graph information for Drug–Target interaction prediction. *Computers in Biology and Medicine*, *142*, 105214. https://doi.org/10.1016/J.COMPBIOMED.2022.105214

Zhang, W., Chen, Y., Molecules, D. L.-, & 2017, undefined. (t.y.). Drug-target interaction prediction through label propagation with linear neighborhood information. *mdpi.comW Zhang, Y Chen, D LiMolecules, 2017•mdpi.com*. https://doi.org/10.3390/molecules22122056

Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. *Nature Reviews Methods Primers*, *4*(1), 1-16. https://doi.org/10.1038/S43586-024-00346-Y;SUBJMETA=117,531,639,705;KWRD=COMPUTER+SCIENCE,STATISTICS

Zhao, Z. Y., Huang, W. Z., Zhan, X. K., Pan, J., Huang, Y. A., Zhang, S. W., & Yu, C. Q. (2021). An Ensemble Learning-Based Method for Inferring Drug-Target Interactions Combining Protein Sequences and Drug Fingerprints. *BioMed Research International*, *2021*(1), 9933873. https://doi.org/10.1155/2021/9933873

Zhou, Y., Neri, F., Ong, Y.-S., & Bai, R. (2025). *SiamNAS: Siamese Surrogate Model for Dominance Relation Prediction in Multi-objective Neural Architecture Search*. https://doi.org/10.1145/3712256.3726359

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Sara NAGHIB ZADEH
Doğum Yeri	
Doğum Tarihi	Tarih girmek için tıklayın veya dokunun.
Uyruğu	❒ T.C. ☑ Diğer:
E-Posta Adresi	
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	İslami Azad Üniversitesi
Fakülte	Fen Bilimleri Enstitüsü
Bölümü	Bilgisayar Yazılım Mühendisliği
Mezuniyet Yılı	21.12.2006

Yüksek Lisans	
Üniversite	İslami Azad Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Mühendislik
Programı	Bilgisayar Yazılım Mühendisliği

Doktora	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı

Makale ve Bildiriler
Naghib Zadeh, S., Ecevit Satı, Z., & Ghanbari Sorkhi, A. (2025). Improving Drug-Target Interaction Prediction Using SVM-SMOTE: A Solution for Imbalanced Dataset. International Journal of Management Information Systems and Computer Science, 9(1), 10-28. https://doi.org/10.33461/uybisbbd.1661593
Naghib Zadeh, S., & Ecevit Satı, Z. (2025). A BIBLIOMETRIC EVALUATION OF SMART AGRICULTURE RESEARCH. International Journal of 3D Printing Technologies and Digital Industry, 9(2), 194-206. https://doi.org/10.46519/ij3dptdi.1648618

Koçak, C., & Naghibzadeh, S. (2017). A new distance & energy aware algorithm for improve consuming energy in wireless sensor network. In Proceedings of the International Mediterranean Science and Engineering Congress (IMSEC) (Vol. 2, pp. 1126).

Naghibzadeh, S., Yaman Atcı, Ş., & Yavuz Genç, F. (2025). Social media analysis with natural language processing. 11th International Palandoken Scientific Studies Congress, 498-509. ISBN: 978-625-378-254-2.

Yaman Atcı, Ş., Naghibzadeh, S.., & Özdemir, B. (2025). CNN-based deep learning approach for facial expression recognition from image and camera. 6th International Uludağ Scientific Research Congress, 300-315. ISBN: 978-625-378-261-0.

Yaman Atcı, Ş., İnel Özkipar, Z., Naghibzadeh, S., & Doğan, E. N. (2025). Human density analysis with image processing based live video. 10th International Marmara Scientific Research and Innovation Congress, 810-819. ISBN: 978-625-378-263-4.

Yaman Atcı, Ş., Naghibzadeh, S., & Kosovalı Çavuş, F. (2025). Smart city components with Internet of Things and machine-to-machine communications technology: A survey. Metin Özetlemede Ekstraktif ve Abstraktif Yaklaşımlar, 63.

Yaman Atcı, Ş., Kosovali Çavuş, F., İnel Özkiper, Z., & Naghibzadeh, S. (2024). Makine öğrenmesi: Yöntemler, algoritmalar ve gerçek dünya uygulamaları. In Pratik Yapay Zeka Uygulamaları 2 (p. 187). BİDGE YAYINLARI.