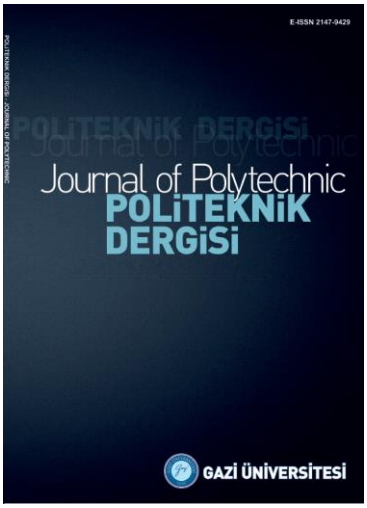


POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

RNA m⁶A modifikasyon bölgelerinin sınıflandırılması için öznitelik çıkarma ve boyut azaltma yöntemlerinin karşılaştırılması

A Comparison for feature extraction and dimension reduction methods for classification of RNA m⁶A modification sites

***Yazar(lar) (Author(s))**: Batuhan NURAY[1], Volkan ALTUNTAŞ[2]*

ORCID[1]: 0009-0008-1345-3363

ORCID[2]: 0000-0003-3144-8724

To cite to this article: Batuhan N. ve Volkan A., "RNA m⁶A Modifikasyon Bölgelerinin Sınıflandırılması için Öznitelik Çıkarma ve Boyut Azaltma Yöntemlerinin Karşılaştırılması", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Batuhan N. ve Volkan A., "RNA m⁶A Modifikasyon Bölgelerinin Sınıflandırılması için Öznitelik Çıkarma ve Boyut Azaltma Yöntemlerinin Karşılaştırılması", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1511303

RNA m^{6}A Modifikasyon Bölgelerinin Sınıflandırılması için Öznitelik Çıkarma ve Boyut Azaltma Yöntemlerinin Karşılaştırılması

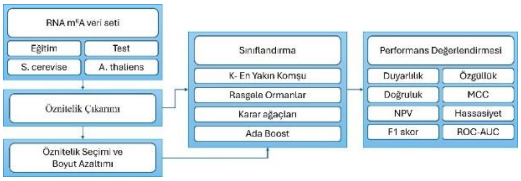
A Comparison for Feature Extraction and Dimension Reduction Methods for Classification of RNA m^{6}A Modification Sites

Önemli noktalar (Highlights)

- *N6-metiladenozin modifikasyonları için 35 öznitelik çıkarma, 9 boyut azaltma ve 4 sınıflandırma algoritmalarının performanslarının karşılaştırılması. (Comparison of the performance of 35 feature extraction, 9 dimensionality reduction and 4 classification algorithms for the N6-methlyadenosine modifications.)*

Grafik Özet (Graphical Abstract)

Bu çalışmada RNA'da gerçekleşen N6-metiladenozin modifikasyon bölgelerinin tespiti için yapılacak biyoenformatik çalışmalarda kullanılan farklı öznitelik çıkarma, öznitelik seçme ve boyut düşürme algoritmalarının performansları karşılaştırılarak incelenmiştir. (In this study, performances of various feature extraction, feature selection and dimensionality reduction algorithms used in bioinformatics studies to identify N6-methlyadenosine modification sites in RNA were compared and examined.)



Şekil. Geliştirilen modelin akış şeması. **/Figure.** Flowchart of the developed model.

Amaç (Aim)

Bu çalışmada araştırmacılara algoritma seçimi konusunda bir kaynak sağlamak amaçlanmıştır. (In this study, it was aimed to provide researchers with a resource for algorithm selection.)

Tasarım ve Yöntem (Design & Methodology)

N6-metiladenozin modifikasyonunun tespiti için 2 veri seti, 35 öznitelik çıkarma, 9 boyut azaltma ve 4 sınıflandırma algoritmasının performansları karşılaştırılmıştır. (The performance of 2 datasets, 35 feature extraction algorithm, 9 dimension reduction and 4 classification algorithms were compared for N6-methlyadenosine modification.)

Özgünlük (Originality)

Literatürde yer alan farklı öznitelik seçme ve boyut azaltma algoritmalarının N6-metiladenozin modifikasyonlarını tespiti için performanslarının karşılaştırmalı olarak incelendiği bir çalışma yoktur. (There is no study in the literature that examines the performance of different feature selection and dimensionality reduction algorithms comparatively for the detection of N6-methlyadenosine modifications.)

Bulgular (Findings)

PS4 öznitelik çıkarma algoritması, elastik net boyut düşürme algoritmasının RF sınıflandırma algoritması ile yüksek doğruluk gösterdiği görülmüştür. (It was observed that the PS4 feature extraction algorithm, elastic net dimension reduction algorithm showed high accuracy with the RF classification algorithm.)

Sonuç (Conclusion)

Elastik net-RF algoritmalarının birlikte kullanılması, m^{6}a bölgeleri için yüksek performans verdiği görülmüştür. (The combination of PS4-Elastic net-RF algorithms was found to give high performance for m^{6}a regions.)

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

RNA m^{6}A Modifikasyon Bölgelerinin Sınıflandırılması için Öznitelik Çıkarma ve Boyut Azaltma Yöntemlerinin Karşılaştırılması

Araştırma Makalesi / Research Article

Batuhan NURAY[1*], Volkan ALTUNTAŞ[2]

[1]Mühendislik ve Doğa Bilimleri Fakültesi, Biyomühendislik Bölümü, Bursa Teknik Üniversitesi, Türkiye
[2]Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Müh. Bölümü, Bursa Teknik Üniversitesi, Türkiye
(Geliş/Received : 05.07.2024 ; Kabul/Accepted : 01.09.2024 ; Erken Görünüm/Early View : 10.09.2024)

ÖZ

Bu çalışmada RNA'da sıklıkla meydana gelen N6-metiladenozin (m^{6}A) modifikasyon bölgelerinin belirlenmesi ve gelecekte yapılacak çalışmalar için farklı öznitelik çıkarıcılar, öznitelik seçiciler ve boyut düşürme algoritmalarının, K-en yakın komşu, Adaboost, Rasgele ormanlar ve Karar ağaçları sınıflandırma algoritmaları kullanılarak performanslarının karşılaştırılması amaçlanmıştır. 35 farklı öznitelik çıkarma algoritması, 9 farklı boyut azaltma ve öznitelik seçici algoritma ve 4 farklı sınıflandırma algoritması kullanılarak algoritmaların m^{6}A modifikasyon bölgelerinin tanımlamasındaki performansları değerlendirilmiştir. Yapılan çalışmanın sonunda nükleotid setlerinin gen dizisi içindeki birlikte görülme sıklığını dikkate alarak öznitelik çıkarımı yapan PS öznitelik çıkarma algoritması, Elastik net boyut azaltma algoritması ve Rastgele ormanlar sınıflandırma algoritmasının birlikte kullanılmasının m^{6}A modifikasyon bölgelerinin belirlenmesinde yüksek performans gösterdiği görülmüştür.

Anahtar Kelimeler: N6-metiladenozin(m^{6}A) bölgeleri, Öznitelik Çıkarımı, Öznitelik Seçimi, K-En Yakın Komşu, Adaboost, Rasgele Ormanlar, Karar Ağaçları.

A Comparison for Feature Extraction and Dimension Reduction Methods for Classification of RNA m^{6}A Modification Sites

ABSTRACT

In this study, the aim was to identify N6-methyladenosine (m^{6}A) modification sites that frequently occur in RNA and to compare the performance of different feature extractors, feature selectors, and dimension reduction algorithms using K-nearest neighbor, AdaBoost, Random Forest, and Decision Tree classification algorithms for future studies. The performance of the algorithms in identifying m6A modification sites was evaluated using 35 different feature extraction algorithms, 9 different dimension reduction and feature selection algorithms, and 4 different classification algorithms. At the end of the study, it was observed that the combination of the PS feature extraction algorithm, which considers the co-occurrence frequency of nucleotide sets within the gene sequence, the Elastic Net dimension reduction algorithm, and the Random Forest classification algorithm showed high performance in the identification of m6A modification sites.

Keywords: Solar air collector, conical spring, fuzzy logic, modeling, outlet temperature, thermal efficiency.

1. GİRİŞ (INTRODUCTION)

Genetik ve genomik alanında yapılan çalışmalar ilerledikçe, hücrelerin içerdiği genetik materyalin kontrol, düzenleme, bilgi aktarımı gibi farklı mekanizmalarını karmaşıklığı anlaşılmıştır. Biyolojinin temel prensiplerinden olan santral dogma modelinde DNA'dan dışarı bilgi aktarımı RNA aracılığıyla gerçekleşir. RNA'nın yapısındaki modifikasyonlar hücre içi sinyal iletiminde kritik rol oynar ve gen ifadesinin hassas bir şekilde düzenlenmesini sağlar. Günümüze kadar yapılan çalışmalarda RNA'nın 150'den fazla modifikasyonu keşfedilmiştir ve modifikasyon türleri arasında bulunan N6-metiladenozin (m^{6}A) en yaygın ve fazla görülen modifikasyon türüdür. m^{6}A modifikasyonları temel olarak insanlarda (H. sapiens) DRACH dizi motifleri (D = A, G veya U; R = A veya G; H = A,C veya U), Saccharomyces cerevisiae' de (S. cerevise) RGAC (R = A veya G) dizi motifinde ve Arabidopsis thaliana' da (A. Thaliana) RRACH (R = A veya G, H = A, C veya U) dizi motifi içinde gerçekleşir ve dur kodonları (UAA, UGA ve UAG) etrafında yaygın olarak görülür [1], [2], [3], [4]. m^{6}A'nın varlığı mRNA üzerinde m^{6}A bölgeleri oluşturan yazıcı enzim (METTL3-METTL14 metiltransferaz) kompleksi, m^{6}A'yı bağlayan okuyucu proteinler, m^{6}A'yı kaldıran silici enzimler tarafından RNA üzerinde gerçekleşen m^{6}A modifikasyonu dinamik olarak düzenlenir [1]. Yapılan çalışmalarda RNA'da gerçekleşen m^{6}A modifikasyonunun RNA'nın lokalizasyonu ve bozunması, alternatif splicing, hücre farklılaşması ve yeniden programlama, sirkadyen saatin düzenlenmesi gibi farklı biyolojik işlevlerle ilişkili olduğu keşfedilmiştir [5]. RNA'da geçekleşen m^{6}A modifikasyon bölgeleri prokaryotlar, ökaryotlar ve virüsler gibi canlı organizmaların farklı kategorilerinde

*Sorumlu Yazar (Corresponding Author)
e-posta : 23435020006@ogr.btu.edu.tr

Çizelge 1. Bulanık mantık kural çizelgesundan örneklemler (Samples from thefuzzy rule table)

Veri Seti	m^6A bölgelerinin sayısı	m^6A olmayan bölgelerinin sayısı	Toplam sayı	Dizinin uzunluğu	Dizi Motifi
S. cerevise	1307	1307	2614	51	GAC
A. thaliens	394	394	788	25	RRACH

bulunur [6]. m^6A modifikasyonunun RNA'da yaygın olarak görülmesi, farklı canlı kategorilerinde bulunması ve birçok önemli biyolojik süreçle ilişkilendirilmesi sebebiyle yaygın ve en iyi çalışılmış RNA modifikasyonlarından biridir [7]. m^6A modifikasyonu kanser, obezite, enfeksiyon ve otoimmün hastalıklar, nöral ve metabolik bozukluklar, viral enfeksiyon gibi insan hastalıklarını da etkiler [8],[9]. m^6A bölgeleri iki farklı yolla tespit tanımlanabilir. m^6A dizileme (m^6A-seq), çapraz bağlama immünopresipitasyonu, metillenmiş RNA immünopresipitasyonu (MeRIP) gibi çok sayıda deneysel yöntem vardır. Kullanılan bu deneysel yöntemler, m^6A bölgelerini tanımlamada önemli araçlardır. m^6A bölgelerini tanımlamada deneysel yöntemler pahalı ve zaman alıcı olduğu için uygulanabilir değillerdir. Deneysel yöntemlerin uygulanması pahalıdır ve uzun süre aldığından, büyük veri kümelerinde uygulanabilirliği düşüktür. Bu sebeple m^6A bölgelerinin tanımlanmasında biyoenformatik, makine öğrenmesi teknikleri gibi yenilikçi yöntemler önemli araçlar olarak ön plana çıkmıştır [10]. Bu yöntemler, RNA dizilerinde potansiyel m6A bölgelerinin belirleyerek deneysel çalışmaların maliyetini ve zamanını azaltabilir, ayrıca geniş çapta biyolojik süreçlerin anlaşılmasına katkı sağlayabilir. Bu bağlamda, m6A bölgelerini tahmin etmek için veri analitiği, makine öğrenimi ve hesaplamalı biyoloji gibi alanlardan gelen yöntemlerin kullanılması gerekmektedir. Ancak, bu yöntemlerin hassasiyeti, doğruluğu ve genel performansı üzerinde çalışmalar devam etmektedir. Bu problemin çözümü, hücresel düzeyde m^6A modifikasyonlarının rolünü ve etkisini daha iyi anlamamıza ve potansiyel olarak yeni tedavi stratejilerinin geliştirilmesine yol açabilir. Bu çabalar, biyolojik süreçlerin derinlemesine anlaşılmasına ve hastalıkların daha etkili bir şekilde tedavi edilmesine olanak sağlayabilir.

Bu makalede yapılan çalışmalar, m^6A modifikasyon bölgelerinin belirlenmesi için farklı öznitelik çıkarıcılar, öznitelik seçiciler ve boyut düşürme algoritmalarının, K-en yakın komşu, Adaboost, Rasgele Ormanlar ve Karar Ağaçları sınıflandırma algoritmaları kullanılarak karşılaştırılması amaçlanmıştır. Bu karşılaştırmanın sonuçları, gelecekte m^6A modifikasyon bölgelerinin tespit edilmesi ve sınıflandırılması için yapılacak araştırmalar için S. Cerevise ve A. Thaliens veri setlerinde kullanımı ve elde edilen sonuçların derlenmesi açısından önemli bir kaynak ve yardımcı çalışma niteliğinde olacaktır.

2. MATERYAL VE METOD (MATERIAL and METHOD)

2.1 Veri Seti (Data Set)

Tahmin modeli oluşturmada kullanılan veri setlerinin kalitesi, oluşturulan modelin etkili ve doğru eğitilmesi için yapının temelini oluşturur. Bu çalışmada, Saccharomyces cerevisiae (S. cerevise) ve Arabidopsis thaliana (A. Thaliana) veri setleri oluşturulan modelin performansını değerlendirmek için kullanılmıştır. Kullanılan veri setleri yapılarında pozitif ve negatif örnekler içermektedir. S. Cerevise eğitim veri seti, A. Thaliana bağımsız test veri seti olarak kullanılmıştır. S. Cerevise veri seti, GAC motifini içermektedir ve eksik nükleotidler dizinin ayna görüntüsüyle tamamlanmaktadır. S. Cerevise veri seti 1:1 oranında pozitif ve negatif diziler içermektedir ve her bir dizi 51 nükleotid uzunluğundadır. A. Thaliana veri seti RRACH motifini içermektedir. Yüksek benzerliğe sahip dizilerin sayısını azaltmak ve veri setinden çıkartmak için CD-HIT aracı kullanılarak %60'tan fazla benzerliğe sahip diziler belirlenmiş ve veri setinden çıkartılarak A. Thaliana veri seti hazırlanmıştır. A. Thaliana veri seti 1:1 oranında pozitif ve negatif diziler içermektedir ve her bir dizi 25 nükleotid uzunluğundadır. Çalışmada kullanılan veri setleri Zhang ve arkadaşlarının yaptığı çalışmadan elde edilmiştir [2]. Veri seti 10.03.2024 tarihinde https://github.com/QUST-AIBBDRC/StackRAM adresinden indirilmiştir. Kullanılan veri kümelerinin özellikleri Çizelge 1'de gösterilmiştir.

2.2 Öznitelik Çıkarma Algoritmaları (Feature Extraction Algorithms)

Öznitelik çıkarma, makine öğrenimi algoritmalarındaki en önemli ve temel adımdır. Öznitelik çıkarma algoritmaları veri setindeki gen dizilerini ifade edecek aritmetik değerler elde ederler. Gen dizilerinin sayısal veriler olarak ifade edilmesi biyoenformatik ve bilgisayar biliminde kullanılan istatistiksel modeller sayısal veriler üzerinde çalışır ve oluşturulan bu veriler ile eğitilirler. Öznitelikler, hesaplanan ilk veri kümesine dayanarak çıkarılır ve bu özelliklerin tekrarlanmayan, tanımlayıcı ve veri setini genelleme ve öğrenme aşamalarını doğru şekilde yönlendirecek nitelikte olması gerekmektedir [11]. Yapılan bu çalışmada literatürde yaygın olarak kullanılan ve iyi performans vereceği düşünülen algoritmalar seçilmiştir. Çizelge 2'de belirtilen farklı öznitelik çıkarma yöntemleri m^6A bölgelerini tanımlamadaki performansını değerlendirmek için kullanılmıştır.

Çizelge 2. Kullanılan öznitelik çıkarma algoritmaları [12],[13],[14],[15],[16]. (The feature extraction algorithms used.)

Öznitelik çıkarma yöntemi	Açıklama
Kmer (Tip1 ve Tip 2)	Gen dizilerindeki komşu nükleik asitlerin ham ve normalleştirilmiş alt dizilerinin görülme frekanslarının dikkate alan öznitelik çıkarma yöntemi.
Mismatch	Gen dizileri arasındaki eşleşmeyen baz çiftlerinin dağılımı ve frekanslarını dikkate alan öznitelik çıkarma yöntemi.
Subsequence	Gen dizilerinde yer alan alt dizilerin görülme frekanslarını dikkate alan öznitelik çıkarma yöntemi.
NAC	Gen dizilerinde yer alan A,U,C,G nükleik asitlerinin bileşim oranlarını dikkate alan öznitelik çıkarma yöntemi.
ANF	Nükleotidlerin gen dizileri içindeki kümülatif frekanslarını dikkate alan öznitelik çıkarma yöntemi.
ENAC	NAC yönteminin geliştirilmesiyle bileşim oranlarını daha fazla detay içeren nükleik asit oranlarını dikkate alan öznitelik çıkarma yöntemi.
Binary	Nükleik asitlerin 0 ve 1 ile temsil edildiği öznitelik çıkarma yöntemi.
PS (2, 3 ve 4)	İkili, üçlü ve dörtlü nükleotid setlerinin gen dizisi içindeki birlikte görülme sıklığını dikkate alan öznitelik çıkarma yöntemi.
CKSNAP (Tip1 ve Tip 2)	Gen dizisinin belirli bir aralığında ham ve normalleştirilmiş baz çiftlerinin bulunma frekanslarını dikkate alan öznitelik çıkarma yöntemi.
NCP	Nükleotidlerin kimyasal özelliklerini (polarite, hidrofobiklik vb.) dikkate alındığı öznitelik çıkarma yöntemi.
ASDC	Gen dizisindeki boşlukları atlayarak çift amino asit frekanslarını inceleyen öznitelik çıkarma yöntemi.
DBE	Di-nükleotitlerin 0 ve 1 ile temsil edildiği öznitelik çıkarma yöntemi.
LPDF	Gen dizisi içindeki di-nükleotitlerin spesifik pozisyonlarda görülme sıklıklarını dikkate alan öznitelik çıkarma yöntemi.
DPCP (Tip1 ve Tip 2)	Di-nükleotidlerin fiziksel ve kimyasal karakteristiklerini dikkate alan öznitelik çıkarma yöntemi.
MMI	Di-nükleotit ve tri-nükleotit özelliklerine dayalı olarak paylaşılan karşılıklı bilgiyi hesaplayan öznitelik çıkarma yöntemi.
Z eğrisi (9, 12, 36, 48 ve 144bit)	Ribonükleotidler, di-ribonükleotitler ve tri-nükleotitlerin gen içindeki konumlarına ve frekansların dikkate alan öznitelik çıkarma yöntemi.
NMBroto	Gen dizisindeki bazların spesifik uzaklıklarda birlikte görülme frekansını dikkate alan öznitelik çıkarma yöntemi.
Moran	Gen dizisinde yer alan bazların korelasyonunu dikkate alan öznitelik çıkarma yöntemi.
Geary	Gen dizisinde yer alan bazlar arasında farklılıkları ölçen istatistiksel bir öznitelik çıkarma yöntemi.
DAC	Di-nükleotidlerin otokorelasyonlarını dikkate alan öznitelik çıkarma yöntemi.
DCC	Di-nükleotidlerin çapraz korelasyonlarını dikkate alan öznitelik çıkarma yöntemi.
DACC	Di-nükleotidlerin oto ve çapraz korelasyonlarını dikkate alan öznitelik çıkarma yöntemi.

DCC	Di-nükleotidlerin çapraz korelasyonlarını dikkate alan öznitelik çıkarma yöntemi.
DACC	Di-nükleotidlerin oto ve çapraz korelasyonlarını dikkate alan öznitelik çıkarma yöntemi.
PseDNC	Gen dizisinde yer alan di-nükleotidlerin pseudo bileşimini dikkate alan öznitelik çıkarma yöntemi.
PseKNC	Gen dizisinde yer alan k uzunluğundaki nükleotidlerin pseudo bileşimini dikkate alan öznitelik çıkarma yöntemi.
PCPseDNC	Gen dizisinde yer alan di-nükleotidlerin pseudo bileşimini paralel korelasyon kullanarak hesaplayan öznitelik çıkarma yöntemi.
SCPseDNC	Gen dizisinde yer alan di-nükleotidlerin pseudo bileşimini seri korelasyon kullanarak hesaplayan öznitelik çıkarma yöntemi.

2.3 Öznitelik Seçme ve Boyut Azaltma Algoritmaları (Feature Selection and Dimension Reduction Algorithms)

Öznitelik çıkarma algoritmalarının oluşturduğu öznitelik vektörleri yüksek boyutlara sahiptir ve sahip oldukları yüksek boyutlar ağır hesaplama sorunlarına yol açmaktadır. Boyut azaltma algoritmaları, öznitelik vektörlerinin boyutlarını azaltma, destekleyici, hayati ve gerekli bir tekniktir. Öznitelik vektörlerini harekete geçiren önemli değişkenleri etkileyerek yüksek boyutlu bir veri kümesini daha düşük boyutlu bir veri kümesine dönüştürerek veri kümelerini kırpar, daha az boyutla temsil edecek şekilde günceller ve örneklendirir [17]. Öznitelik çıkarma algoritmaları çok sayıda tekrarlayan ve ilgili olmayan öznitelikler oluştururlar. Bu sebeple çıkarılan öznitelikler arasında en iyi özelliklerin alt kümelerini seçen bir yöntem uygulamak çok önemlidir. Öznitelik seçme algoritmaları, oluşturulan modellerin performansını iyileştirmek için makine öğrenim yöntemlerinde önemli bir ön işlem haline gelmiştir. Yerel olarak doğrusal gömme (LLE), spektral gömme (SE), temel bileşen analizi (PCA), ekstra ağaçlar (ET), negatif olmayan matris çarpanlarına ayırma (NMF) makine öğrenim algoritmalarında yaygın olarak kullanılan farklı öznitelik seçme algoritmaları vardır. Bu çalışmada literatürde yaygın olarak kullanılan ve iyi performans vereceği düşünülen Yerel olarak doğrusal gömme (LLE), spektral gömme (SE), temel bileşen analizi (PCA), ekstra ağaçlar (ET), negatif olmayan matris çarpanlarına ayırma (NMF), tekil değer ayrışımı (SVD), ısomap (Imap), mutual bilgi (MM) ve elastik net algoritmaları kullanılmıştır. Kullanılan algoritmaların özellikleri Çizelge 3'de anlatılmıştır.

2.4 K-En Yakın Komşu (KNN) (K-Nearest Neighbour (KNN))

K-en yakın komşu algoritması, biyoenformatik çalışmalarda ve sınıflandırma problemlerinde yaygın olarak kullanılan temel ve kolay uygulanabilen bir algoritmik sınıflandırma yöntemdir [25],[26]. K-en yakın komşu algoritması, denetimli bir öğrenme mekanizmasıdır ve yeni örnek sorgusunun sonucu ortak en yakın komşu grubuna göre sınıflandırılır. KNN algoritması, yeni sorgu örneğinin tahmin değeri olarak yerellik sınıflandırılmasını kullanır. Bu algoritmanın amacı, özniteliklere ve eğitim örneklerine dayalı olarak yeni bir nesneyi sınıflandırmaktır. Seçilen özellikler modüle girdi olarak verilir ve sorgu noktasına en yakın olan K değeri seçilir [27]. Sorgu örneği ile tüm eğitim örnekleri arasındaki mesafe Denkleme 1'e göre hesaplanır [11]. Mesafeler k minimum olacak şekilde sıralanır ve en yakın komşular belirlenir. Sınıflandırılan örneğinin tahmin değeri olarak en yakın komşuların kategorisinin basit çoğunluğu kullanılır [27].

$$\boldsymbol{d}\,(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sum_{i=1}^{n} \sqrt{(\boldsymbol{x_{i1}} - \boldsymbol{x_{i20}})^2} \qquad (1)$$

2.5 AdaBoost (AdaBoost)

Adaboost algoritması, bilim insanı Freund tarafından önerilen, makine öğrenmesi alanında yüksek performansa sahip, en başarılı ve yaygın olarak kullanılan algoritmalardan biridir. Sınıflandırma ve regresyon problemlerinde kullanılmak üzere geliştirilmiştir. Adaboost algoritması, zayıf sınıflandırıcıların bir araya toplanmasıyla güçlü bir sınıflandırıcı oluşturur [28]. Adaboost algoritması, güçlü bir sınıflandırıcı oluşturmak için aşağıda belirtilen 6 hesaplama adımından oluşur. Bu hesaplama adımları:

1. Başlangıç ağırlıklarının belirlenmesi: Her bir eğitim örneğine eşit ağırlıklar atanır ve sınıflandırma sürecine katkısı belirlenir.

2. Zayıf sınıflandırıcıların eğitimi: Algoritma belirli sayıda bir dizi sınıflandırıcıyı (yaygın olarak karar ağaçları kullanılmaktadır) eğitir ve performanslarına göre ağırlık değerleri atanır.

3. Hata oranı hesaplanması: Her bir sınıflandırıcı için hata oranı, yanlış sınıflandırılmış örneklerin toplam ağırlıklarının, tüm örneklerin toplam ağırlıklarına oranı olarak tanımlanır.

4. Zayıf sınıflandırıcıların ağırlıklarının güncellenmesi: Hesaplanan hata oranlarına bağlı olarak, zayıf sınıflandırıcıların ağırlıkları güncellenir ve düşük hata oranına sahip sınıflandırıcılar yüksek, düşük hata oranına sahip sınıflandırıcılar düşük ağırlıklar atanır.

5. Örnek ağırlıklarının güncellenmesi: Yanlış sınıflandırılan örneklerin ağırlıkları artırılarak, doğru sınıflandırılan örneklerin ağırlıkları azaltılır. Bu yapılan işlem sonraki sınıflandırıcıların yanlış sınıflandırılan örneklere daha fazla odaklanmasını sağlar.

6. Sonuçların birleştirilmesi: Zayıf sınıflandırıcılardan elde edilen sonuçlar ağırlıklarına göre birleştirilir ve güçlü bir sınıflandırıcı oluşturmak için kullanılır.

2.6 Karar Ağaçları (Decision Tree DT)

Çizelge 3. Kullanılan öznitelik seçme ve boyut azaltma algoritmaları [2],[17],[18],[19],[20],[21],[22],[23],[24] (Feature selection and dimension reduction algorithms used.)

Kullanılan yöntemler	Boyut azaltma metodu özellikleri
LLE	LLE, yüksek boyutlu verilerden daha düşük boyutlu bir alana yerel doğrusal ilişkileri koruyarak veriyi daha düşük bir boyutla temsil eder.
SE	SE, graf teorisine dayalı bir boyut indirgeme tekniğidir. Genellikle veri noktaları arasındaki benzerlikleri temsil eden bir komşuluk grafı oluşturur ve bu gafın laplace matrisini özdeğerleri ve özvektörlerini kullanarak veriyi daha düşük bir boyutla temsil eder.
PCA	PCA, yüksek boyutlu verilerin ana varyans yönlerini belirleyerek boyutları azaltan bir yöntemdir. Verilerin varyansını en iyi temsil eden ortogonal bileşenleri seçer ve veriyi daha düşük bir boyutla temsil eder.
ET	ET, Karar ağaçlarının rastgele örnekler ve özellikler kullanılarak oluşturulduğu bir topluluk yöntemidir. Rasgele ormanlara benzer ancak bölme noktaları rastgele seçilir. Yüksek varyanslı ve karmaşık veri setlerinde güçlü performans sergiler.
NMF	NMF, Bir matrisi iki düşük boyutlu negatif olmayan matrisin çarpımı olarak ayırır.
SVD	Bir matrisi üç farklı matrisin çarpımı olarak ifade eden bir lineer cebir tekniğidir. Veri boyutunu azaltma, gürültü temizleme ve özellik çıkarımı gibi çeşitli uygulamalarda kullanılır. Yaygın olarak yüksek boyutlu verilerde önemli bilgileri koruyarak veri boyutunu azaltmak için etkili bir yöntemdir.
Imap	Imap, veri noktaları arasında geodezik mesafeleri kullanarak yüksek boyutlu verileri düşük boyutlu bir alana küçülterek boyut azaltmayı hedefler. Imap manifold öğrenme yöntemlerinden biridir ve veinin altta yatan geometrik yapısını korumayı amaçlar.
MM	İki değişken arasındaki bağımlılığı ölçen bir istatistiksel metottur. MM'de değişkenler arasındaki bilgi paylaşımı değerlendirilir ve iki değişkenin ortak bilgisini belirler ve değişkenlerin bağımsız olup olmadığına karar verir.
Elastik net	Elastik net, L1 (Lasso) ve L2 (Ridge) düzenlileştirmelerini birleştiren bir regresyon modelidir. Elastik net, Lasso'nun seyrek yapı avantajlarını ve Ridge'in küçük katsayılar üreterek aşırı uyumu önleme avantajlarını birleştirir.

Karar ağaçları, bir ağacın yapısına benzer şekilde kök, dallar ve yapraklardan oluşur. Yapıda kökler düğümü, dallar ve yapraklar yapının dallarını temsil eder. Her bir iç düğümde bir öznitelik test edilir, testin sonucu dallarda, sınıf etiketleri yapraklarda gösterilir. Kök düğüm ağacın en üstünde yer alır ve diğer yapılır bu düğümden oluşur. Karar ağacı, her düğümün bir özelliği (özniteliği) gösterdiği, her bağlantının (dal) bir kararı (kuralı) temsil ettiği ve her yaprağın bir sonucu (kategorik veya sürekli bir değer) gösterdiği yapıdır [29]. Karar ağaçları insan düşünce biçimini taklit ettiği için verilerin anlamlandırılması ve yorumlanmasında başarılı bir yöntemdir. Karar ağaçlarının amacı tüm veri seti için bu yapıyı oluşturarak, yapıda yer alan her bir yaprakta tek bir sonuca ulaşmaktır. Karar ağaçlarının anlaşılması kolay yapısı, karmaşık karar süreçlerini görselleştirmek ve açıklamak için ideal bir yöntem sunar. Bu yapısı, sayesinde farklı özniteliklerin nasıl etkileşime girdiği ve etkileşimlerinin sonuçlara etkisi net bir şekilde ortaya konulabilmektedir.

2.7 Rasgele Ormanlar (Random Forest RF)

Rasgele ormanlar, her bir ağacın en sık görülen sınıfa oy vermesiyle girdilere sınıf atanmasını sağlayan, sınıflandırma ağaçlarından oluşan bir topluluktur. Rasgele ağaçlar algoritması, her düğümde en iyi değişkenle yerine en iyi rastgele alt kümelerden birini seçerek dallandırma işlemi yapar ve genelleme hatasını azaltır. Ağaçların çeşitliliğini artırmak için eğitim verilerinin farklı ağaçlar oluşturmak için farklı alt

kümelerden büyümesini sağlamak amacıyla bootstrap aggregating (bagging) yöntemlerini kullanır. Bagging, orijinal veri kümesinin rastgele bir sonraki alt küme oluşturulurken seçilen verilerin silinmeden yeniden örneklenmesi ile yapılan bir eğitim veri seti oluşturma tekniğidir. Bagging kullanılarak oluşturulan her bir alt küme, her bir ağacı büyütmek için kullanılan eğitim veri setinin belirli bir oranını içerir. Eğitim alt kümesinde bulunmayan veriler, "out-of-bag" (OOB) olarak adlandırılan başka bir alt kümenin parçası olarak ağaca dahil edilir. Her bir ağaç için, bootstrap yöntemiyle seçilmeyen elemanlardan farklı bir OOB alt kümesi oluşturulur. OOB'ye dahil olan elemanlar, ağacın sınıflandırma performansını değerlendirmek için kullanılabilir. Yanlış sınıflandırmaların toplam OOB elemanlarına oranı, genelleme hatasının tarafsız bir tahminini sağlar ve bu oran, özellik seçimi için kullanılabilir. Rastgele ormanlar, en iyi bölünmeyi seçmek için bir elemanın diğer sınıflara göre saflığını ölçen Gini indeksini kullanır ve bu sayede belirli bir özellik kombinasyonu kullanılarak, bir karar ağacı maksimum derinliğe kadar (budama yapılmadan) büyütülür [30].

2.8 Performans Değerlendirmesi (Performance Evaluation)

Yapılan çalışmada oluşturulan modelin istatistiksel tahmin yeteneğini belirlemek için bağımsız veri setleri ve jackknife testi kullanılmıştır. Jackknife yöntemi yapılan çalışmalarda oluşturulan modelin doğruluğunu test etmek için kullanımı yaygın bir şekilde artan bir yöntemdir [2]. Jackknife yönteminde kullanılan veri setinde bulunan elemanlardan 1 tanesi eğitim setinden ayrılır ve modelin eğitimi diğer veriler üzerinden yapılır. Eğitilen model, eğitim grubundan ayrılan 1 veri noktası üzerinden test edilir, bu işlem veri setindeki bütün veriler için tekrarlanır ve her bir eleman tahmin edilmeye çalışılır [31].

Sınıflandırıcıların tahmin yeteneklerini ve performanslarının birbirleriyle adil bir şekilde karşılaştırmak için duyarlılık (Sn), özgüllük (Sp), doğruluk (ACC), Mathew korelasyon sayısı (MCC), negatif öngörü değeri (NPV), hassasiyet ve F1 skor ölçütleri kullanılmıştır [32].

$$Sn = \frac{TP}{TP+FN} \quad (2)$$

$$Sp = \frac{TN}{FP+TN} \quad (3)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$MCC = \frac{TP*TN-FP*FB}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

$$F1\ skor = \frac{2*TP}{2*TP+FP+FN} \quad (6)$$

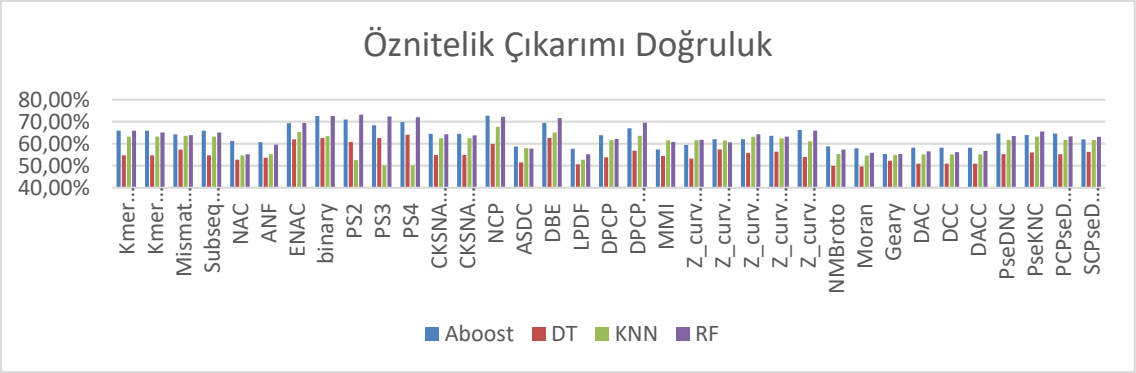
Denklem 2, 3, 4, 5, 6'da kullanılan, TP m^6A olan dizilerin doğru bir şekilde m^6A olan diziler olarak tanındığı sayı, TN m^6A olmayan dizilerin doğru bir şekilde m^6A olmayan diziler olarak tanındığı sayı, FP m^6A olmayan dizilerin m^6A olan diziler olarak tanındığı sayı, FN m^6A olan dizilerin m^6A olmayan diziler olarak tanındığı sayıyı temsil eder. Sn ve Sp değeri oluşturulan modelin örnekleri pozitif ve negatif olarak doğru tahmin etme kabiliyetini temsil eder. ACC, MCC, kesinlik ve duyarlılık değerlerinin harmonik ortalamasından elde edilen F1 skor değeri [33], TP ve FP arasındaki ilişkiyi gösteren ROC eğrisi ve ROC eğrisinin altında kalan alanı temsil eden AUC değeri oluşturulan modelin performansını değerlendirmek ve farklı sınıflandırıcılar kullanılarak oluşturulan modellerin performanslarının birbiriyle karşılaştırılmasında kullanılmıştır.

3. SONUÇLAR VE TARTIŞMA (RESULTS AND DISCUSSION)

3.1 Farklı Öznitelik Çıkarma Yöntemlerinin Karşılaştırılması (Comparison of Different Feature Extraction Methods)

Gen dizilerinden, m^6A bölgelerinin tanımlanmasındaki önemli zorluklardan biri bilgi açısından zengin özniteliklerin çıkarılmasıdır. Farklı öznitelik çıkarma yöntemleri, kullanılan veri setinin doğası, yapılan çalışmanın amaçları ve özellikleri gereği farklı davranışlar gösterebilmektedirler. Çalışılan problem ve veri setinde, kullanılan öznitelik çıkarıcıların performansı geliştirilen modelin performansını doğrudan etkilemektedir.

Çizelge 2. 'de belirtilen çeşitli öznitelik çıkarma yöntemleri, m^6A bölgelerini tahmin etme yeteneklerini ve performanslarını karşılaştırmak için incelenmiştir. Kullanılan öznitelik çıkarma yöntemlerinin oluşturduğu, öznitelik vektörleri K-en yakın komşu, Adaboost, rasgele ormanlar ve karar ağaçları sınıflandırıcılarına girdi olarak verilmiştir. Öznitelik çıkarma yöntemlerinin performansları karşılaştırılırken, S. Cerevise veri seti ve jackknife metodu kullanılarak KNN, Adaboost, RF ve DT sınıflandırıcılarında sınıflandırma işlemi yapılmıştır. Yapılan sınıflandırma sonunda elde edilen doğruluk değerleri Şekil 1.'de gösterilmiştir.



Şekil 1. Farklı öznitelik çıkarma yöntemlerinin farklı sınıflandırıcılardaki doğruluk değerleri. (Accuracy values of different feature extraction methods in different classifiers.)

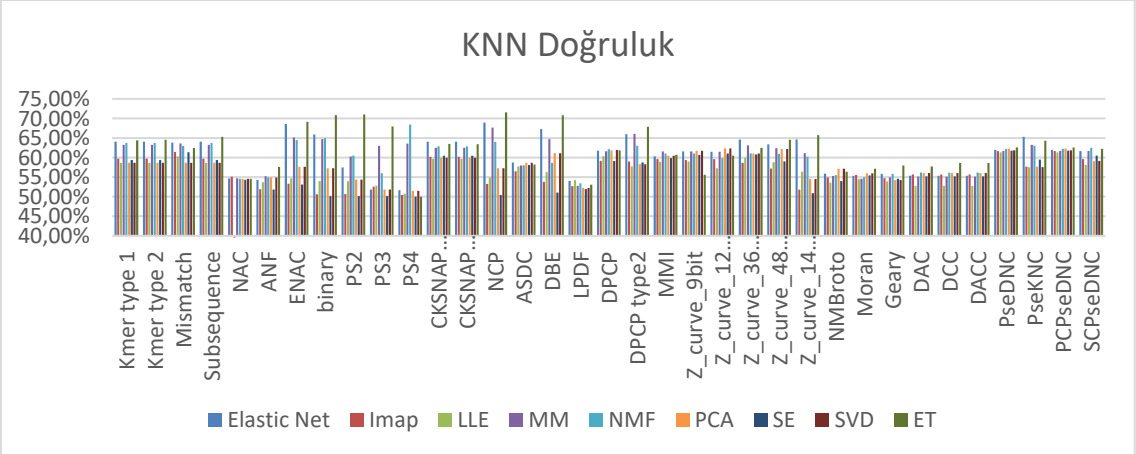
Şekil 1.'de farklı öznitelik çıkarma yöntemlerinin değişik sınıflandırıcılar ile m^{6}A bölgelerinin tanımlamada elde ettiği doğruluk değerleri gösterilmiştir. KNN, Adaboost, RF ve DT sınıflandırıcılarında elde edilen ortalama doğruluk değerleri sırasıyla %59.67, %63.6, %63.47 ve %64.12'dir. Sınıflandırıcılarda farklı öznitelik çıkarıcılar kullanılarak elde edilen en yüksek doğruluk değerleri KNN-NCP %67.67, Adaboost-NCP %72.72, RF-PS2 %73.22 ve DT-PS4 %64.12'dir.

Elde edilen sonuçlar, farklı sınıflandırıcılar ile farklı öznitelik çıkarma yöntemlerinin m^{6}A bölgelerinin belirlenmesindeki performanslarını göstermektedir. NCP öznitelik çıkarma algoritması KNN ve Adaboost sınıflandırıcılarında, PS öznitelik çıkarma algoritmaları RF ve DT sınıflandırıcılarında elde ettikleri performansla m^{6}A bölgelerinin belirlenmesinde etkili öznitelik çıkarım yöntemi olduğunu göstermektedir.

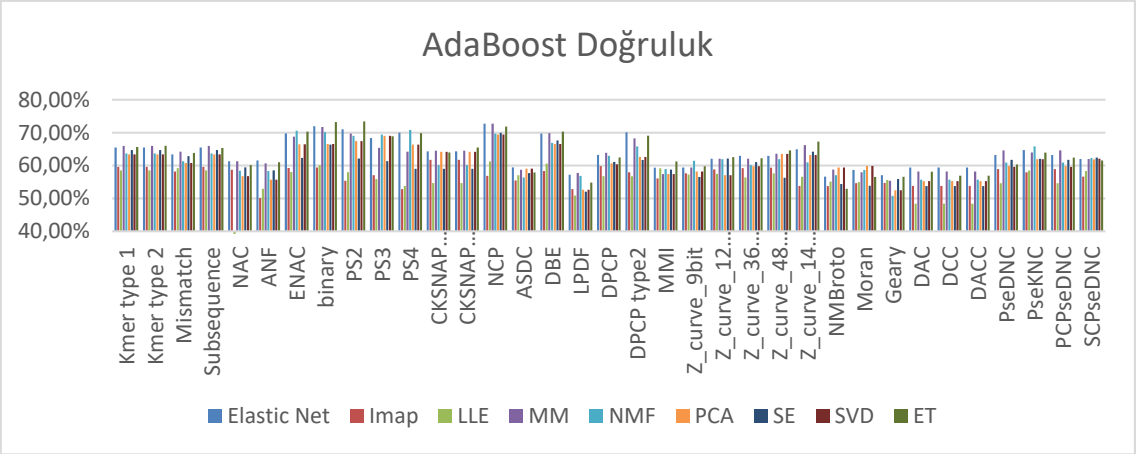
3.2 Farklı Öznitelik Seçim ve Boyut Azaltma Algoritmalarının Karşılaştırılması (Comparison of Different Feature Selection and Dimension Reduction Algorithms)

Oluşturulan öznitelik çıkarma yöntemleriyle oluşturulan vektörlerin içeriklerindeki ilgili olmayan bilgilerin çıkarılması ve en iyi temsil özelliklerine sahip değerlerin belirlenmesi için boyut küçültme ve öznitelik seçme algoritmaları kullanılmaktadır. LLE, SE, PCA, ET, NMF, SVD, Imap, MM ve Elastic net algoritmaları S. Cerevise veri seti'nin Çizelge 2.'de yer alan öznitelik çıkarma algoritmaları ile oluşturulan vektörlerine uygulanmıştır. Boyut küçültme ve seçim işlemine tabi tutulmuş vektörler jackknife metodu kullanılarak KNN, Adaboost, RF ve DT sınıflandırıcılarına girdi olarak verilmiştir.

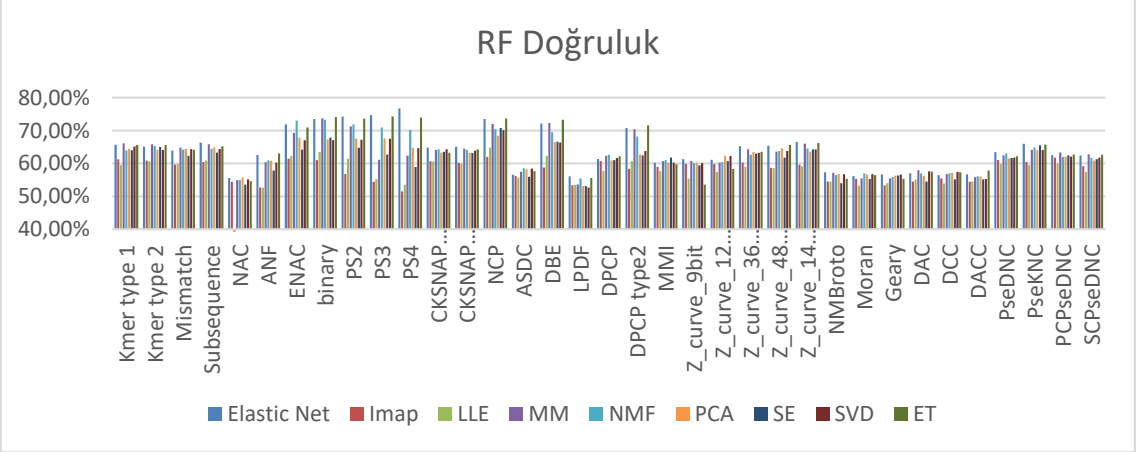
Elde edilen performans değerleri boyut küçültme ve seçim işlemine tabi tutulmamış orijinal verilerin performans değerleri ile karşılaştırılmıştır. Farklı sınıflandırıcı ve boyut küçültme algoritmaları kullanılarak elde edilen doğruluk değerleri Şekil 2., Şekil 3. Şekil 4. ve Şekil 5.'de verilmiştir.



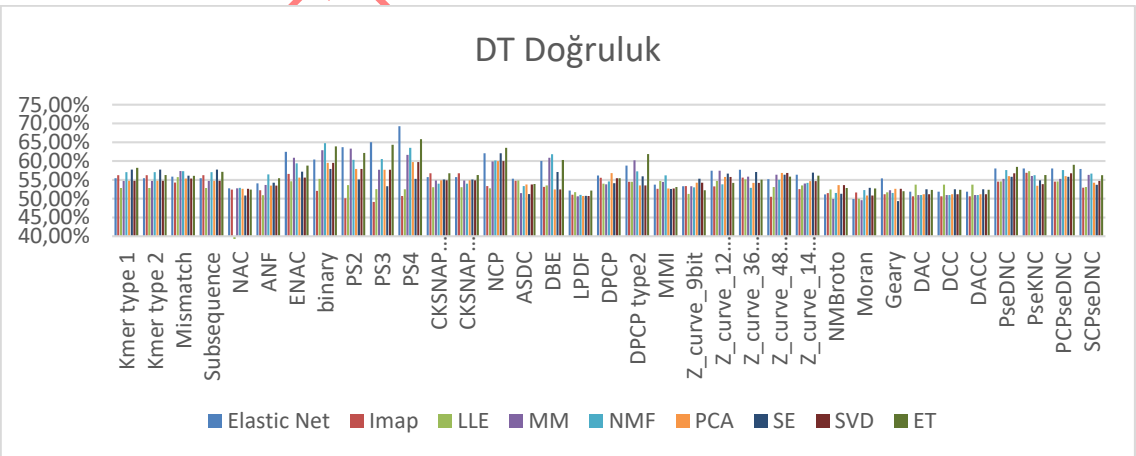
Şekil 2. KNN sınıflandırıcısı kullanılarak orijinal verilerin boyut küçültme algoritmaları kullanılarak sınıflandırılması ile elde edilmiş doğruluk değerleri. (Accuracy values obtained by classifying the original data using the KNN classifier using dimension reduction algorithms.)



Şekil 3. Adaboost sınıflandırıcısı kullanılarak orijinal verilerin boyut küçültme algoritmaları kullanılarak sınıflandırılması ile elde edilmiş doğruluk değerleri. (Accuracy values obtained by classifying the original data using the Adaboost classifier using dimension reduction algorithms.)



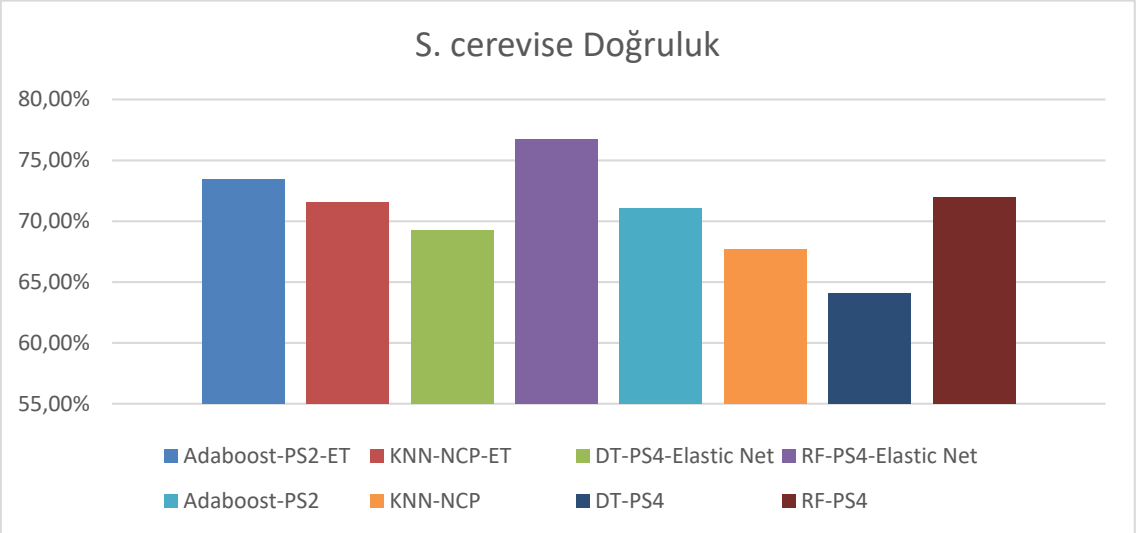
Şekil 4. RF sınıflandırıcısı kullanılarak orijinal verilerin boyut küçültme algoritmaları kullanılarak sınıflandırılması ile elde edilmiş doğruluk değerleri. (Accuracy values obtained by classifying the original data using the RF classifier using dimension reduction algorithms.)



Şekil 5. DT sınıflandırıcısı kullanılarak orijinal verilerin boyut küçültme algoritmaları kullanılarak sınıflandırılması ile elde edilmiş doğruluk değerleri. (Accuracy values obtained by classifying the original data using the DT classifier using dimension reduction algorithms.)

Uygulanan boyut küçültme algoritmaları oluşturulan farklı sınıflandırma algoritmaları ile oluşturulan modellerin doğruluk performanslarını pozitif ve negatif olarak etkilemişlerdir. KNN sınıflandırıcısında ET, elastik net, MM ve NMF algoritmaları ortalama doğruluk performansını sırasıyla %2.50, %1.06, %1.06 ve %0.68 arttırmıştır. LLE, SE, Imap, SVD ve PCA algoritmaları ortalama doğruluk performansını sırasıyla %4.65, %3.23, %3.12, %1.82 ve %1.80 azaltmıştır. Adaboost sınıflandırıcısında elastik net algoritması ortalama

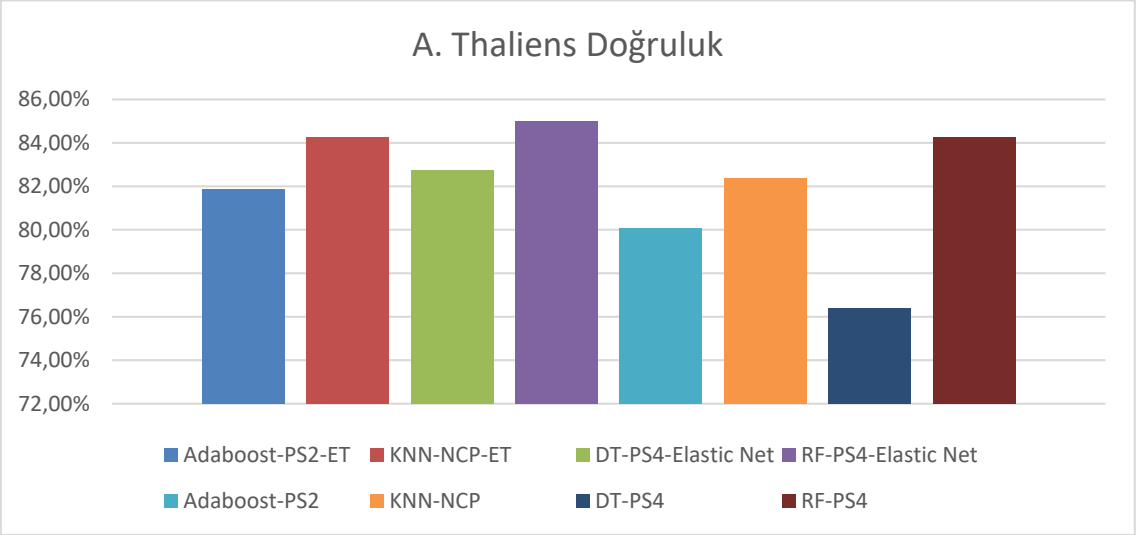
doğruluk performansını %0.11 arttırmıştır. LLE, Imap, SE, SVD, PCA ve ET algoritmaları ortalama doğruluk performansını sırasıyla %9.11, %6.58, %3.43, %2.52, %2.50 ve %0.25 azaltmıştır. RF sınıflandırıcısında elastik net ve ET algoritmaları ortalama doğruluk performansını sırasıyla %0.68 ve %0.16 arttırmıştır. LLE, Imap, SE, SVD ve PCA algoritmaları ortalama doğruluk performansını sırasıyla %7.07, %5.4, %2.50, %1.40 ve %1.52 azaltmıştır. DT sınıflandırıcısında elastik net, ET ve NMF algoritmaları ortalama doğruluk performansını sırasıyla %1.13, %1.09 ve %0.30 arttırmıştır. LLE, Imap, SE, SVD, PCA ve MM algoritmaları ortalama doğruluk performansını sırasıyla %3.58, %2.25, %0.66, %0.88, %0.91 ve %0.11 azaltmıştır. Elastik net ve ET öznitelik çıkarma algoritmaları hepsinde iyi çalışan bir "herkese uyan tek" yöntem olduğunu göstermiştir. Boyut düşürme algoritmalarının kullanımı sonucunda öznitelik çıkarma algoritmalarının sınıflandırılmasında, PS2 boyut azaltma algoritması, Adaboost sınıflandırıcısı ile ET boyut azaltma algoritması ile kullanıldığında doğruluğu %2.37, NCP boyut azaltma algoritması, KNN sınıflandırıcısı ile ET boyut azaltma algoritması ile kullanıldığında doğruluğu %3.86, PS4 boyut azaltma algoritması, DT sınıflandırıcısı ile Elastik net boyut azaltma algoritması ile kullanıldığında doğruluğu %5.16, PS4 boyut azaltma algoritması, RT sınıflandırıcısı ile Elastik net boyut azaltma algoritması ile kullanıldığında doğruluğu %4.74 arttırdığı gözlemlenmiştir. En yüksek doğruluk değeri PS4 öznitelik çıkarma algoritması, Elastik net boyut azaltma algoritması ve RF sınıflandırıcı kullanılarak %76.74 olarak elde edilmiştir. Farklı sınıflandırıcılar için en yüksek performansı veren boyut küçültme algoritmaları ve orijinal verilerin doğruluk değerleri Şekil 6.'de verilmiştir.



Şekil 6. Farklı sınıflandırıcılar için en yüksek performansı veren boyut küçültme algoritmaları ve orijinal verilerin doğruluk değerleri . (Highest performing dimension reduction algorithms for different classifiers and accuracy values of original data.)

3.3 Farklı Çapraz Doğrulama Veri Setlerinin Yüksek Performans Veren Öznitelik Çıkarıcı ve Boyut Azaltma Yöntemleri ile Test Edilmesi (Testing Different Cross-Validation Data Sets with High Performance Feature Extractor and Dimension Reduction Methods)

S. cerevise veri seti kullanılarak yapılan öznitelik çıkarma algoritmalarının ve boyut azalta algoritmalarının performanslarının incelenmesi ile NCP, PS2 ve PS4 öznitelik çıkarma algoritmalarının ET ve Elastik net boyut azaltma algoritmaları ile kullanıldığında yüksek doğruluk gösterdiği görülmüştür. Elde edilen bu sonuçların bağımsız veri seti ile test edilmesi ve validasyonunu sağlamak için S. cerevise veri seti ile aynı özelliklere sahip A. Thaliens veri seti kullanılmıştır. A. Thaliens bağımsız veri seti, NCP, PS2 ve PS4 öznitelik çıkarma algoritmaları ile öznitelik vektörleri oluşturulmuştur. Oluşturulan öznitelik vektörleri elastik net ve ET boyut azaltma algoritmaları ile boyut azaltımı yapılmıştır. Elde edilen öznitelik vektörleri KNN, Adaboost, DT, RF sınıflandırıcıları kullanılarak performans değerlendirilmesi yapılmıştır. Elde edilen doğruluk değerleri Şekil 7.'de gösterilmiştir.



Şekil 7. A. Thaliens NCP, PS2 ve PS4 özellik çıkarma algoritmaları ile elde edilen özellik vektörleri ile elastik ağ ve ET boyut indirgeme algoritması ile işlenerek üretilen özellik vektörünün farklı sınıflandırıcılar ile elde edilen doğruluk değerleri. (Accuracy values obtained with different classifiers of the feature vectors obtained by Thaliens NCP, PS2 and PS4 feature extraction algorithms and the feature vector produced by processing by the elastic net and ET dimension reduction algorithm.)Farklı sınıflandırıcılar için en yüksek performansı veren boyut küçültme algoritmaları ve orijinal verilerin doğruluk değerleri . (Highest performing dimension reduction algorithms for different classifiers and accuracy values of original data.)

A. Thaliens, bağımsız veri seti kullanılarak yapılan testler, S. cerevise veri seti kullanılarak elde edilen sonuçlar ile paralel çıkmıştır. A. Thaliens bağımsız test veri seti kullanılarak yapılan testte en yüksek doğruluk değeri S. cerevise eğitim veri setinde de yüksek performans veren PS4 öznitelik çıkarma algoritması, Elastik net boyut azaltma algoritması ve RF sınıflandırıcı ile %85.03 olarak elde edilmiştir. PS4 öznitelik çıkama algoritması, Elastik net boyut azaltma algoritmasının RF sınıflandırıcı ile kullanılmasının RNA nükleotit dizilerinde bulunan m^6A bölgelerinin tanımlanmasında yüksek performans verdiği görülmüştür.

4. SONUÇ (CONCLUSION)

Bu makale, RNA nükleotit dizilerini kullanarak, RNA'da gerçekleşen m^6A modifikasyon bölgelerinin tanımlanması ve tahmin edilmesi için yapılan çalışmalarda kullanılabilecek farklı öznitelik çıkarma algoritmaları, boyut azaltma algoritmaları ve sınıflandırma algoritmaları incelenmiştir. İncelenen öznitelik çıkarma ve boyut azaltma algoritmalarının KNN, Adaboost, DT ve RF sınıflandırıcıları ile performansları incelenmiştir ve birbirleri ile karşılaştırılmıştır. Yapılan karşılaştırmalarda elde edilen performans değerleri ve model doğruluklar ile m^6A bölgelerinin tanımlanması için geliştirilen modellerde PS4 öznitelik çıkarma algoritmasının, bu öznitelik çıkarma algoritması ile elastik net boyut azaltma algoritmasının RF sınıflandırma algoritması ile kullanımı m^6A bölgelerinin tanımlamak için etkili ve yüksek doğruluğa sahip olduğu görülmüştür. m^6A bölgelerinin tanımlamak için PS4 öznitelik çıkarma algoritması, bu öznitelik çıkarma algoritması ile Elastik net boyut azaltma algoritmasının RF sınıflandırma algoritması ile kullanımı A. Thaliens bağımsız veri seti kullanılarak da elde edilen sonuçlar doğrulanmıştır.

Sonuç olarak yapılan çalışmada m6A modifikasyon bölgelerinin tespiti için kullanılabilecek öznitelik çıkarma ve boyut azaltma algoritmalarının, farklı sınıflandırıcılar kullanılarak performansları karşılaştırılmıştır ve bu alanda çalışma yapacak araştırmacılara bir ön çalışma niteliğinde, algoritma geliştiricilerine farklı yöntemlerin kapsamlı bir değerlendirmesini ve uygun yöntemlerin seçilmesinde yararlı olacağını umuyoruz. Gelecekteki çalışmalarda, bu algoritmaların performansını daha da iyileştirmek amacıyla optimizasyon ve makine öğrenimi algoritmalarının uygulanması planlanmaktadır. Bu sayede, özellikle büyük veri kümeleri üzerinde daha etkili ve hızlı sonuçlar elde edilmesi mümkün olacaktır.

KISALTMALAR (SYMBOLS AND ABBREVIATIONS)

m^6A: N6-metiladenozin

S. cerevise: Saccharomyces cerevisiae

A. Thaliana: Arabidopsis thaliana

Kmer: K-mer Sayımı

Mismatch: Uyumsuzluk profili

Subsequence: Alt dizi profili

NAC: Nükleik asit kompozisyonu

ANF: Birikimli nükleotid frekansı

ENAC: Gelişmiş nükleik asit kompozisyonu

Binary: İkili kodlama

PS: Pozisyona özgü skorlama matrisi

CKSNAP: k-Aralıklı Nükleotid Çiftlerinin Kompozisyonu

NCP: Nükleotid Kimyasal Özelliği

ASDC: Dinükleotid Korelasyonlarının Ortalama Toplamı

DBE: Dinükleotid Tabanlı Kodlama

LPDF: Yerel Pozisyona Özgü Dinükleotid Frekansı

DPCP: Dinükleotid Fizikokimyasal Özelliği

MMI: Pozisyonlar Arası Karşılıklı Bilgi

NMBroto: Normalleştirilmiş Moreau-Broto Otokorelasyonu

Moran: Moran'ın Otokorelasyonu

Geary: Geary'nin Otokorelasyonu

DAC: Dinükleotid Tabanlı Otokovaryans

DCC: Dinükleotid Tabanlı Çapraz Kovaryans

DACC: Dinükleotid Tabanlı Otomatik Çapraz Kovaryans

PseDNC: Pseudo Dinükleotid Kompozisyonu

PseKNC: Pseudo K-tuple Nükleotid Kompozisyonu

PCPseDNC: Pozisyona Özgü Pseudo Dinükleotid Kompozisyonu

SCPseDNC: Seri Korelasyon Pseudo Dinükleotid Kompozisyonu

LLE: Yerel olarak doğrusal gömme

SE: Spektral gömme

PCA: Temel bileşen analizi

ET: Ekstra ağaçlar

NMF: Negatif olmayan matris çarpanlarına ayırma

SVD: Tekil değer ayrışımı

Imap: İzometrik haritalama

MM: Mutual bilgi

KNN: K-en yakın komşu

Sn: Duyarlılık

Sp: Özgüllük

ACC: Doğruluk

MCC: Mathew korelasyon sayısı

NPV: Negatif öngörü değeri

CD-HIT: Yüksek kimlikli kümeleme ile toleranslı veri tabanı

Bagging : Bootstrap aggregating

OOB : Out-of-bag

DT: Karar ağaçları

RF: Rastgele ormanlar

TEŞEKKÜR (ACKNOWLEDGEMENT)

Yazarlar, Bursa Teknik Üniversitesi Yüksek Başarımlı Hesaplama Laboratuvarı'na, TÜBİTAK ULAKBİM Yüksek Başarımlı ve Grid Hesaplama Merkezi'ne (TRUBA) teşekkürü borç bilirler. (The authors are grateful to Bursa Technical University High-Performance Computing Laboratory, TUBITAK ULAKBIM High Performance and Grid Computing Center (TRUBA).)

ETİK STANDARTLARIN BEYANI
(DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Batuhan NURAY: Problemin belirlenmesi ve araştırılması, makalenin yazımı, analizlerin yapılması ve sonuçların değerlendirilmesi. (Identifying and researching the problem, writing the article, conducting analyses, and evaluating the results.)

Volkan ALTUNTAŞ: Problemin belirlenmesi, makalenin düzeltilmesi, yayına hazırlanması ve kontrolü. (Identifying the problem, revising the article, preparing it for publication, and checking it.)

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur. / There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] P. Acera Mateos, Y. Zhou, K. Zarnack, E. Eyras, ve Y. Zhou contributed equally, "Concepts and methods for transcriptome-wide prediction of chemical messenger RNA modifications with machine learning", ***Briefings in Bioinformatics***, c., 1-14. (2023).

[2] Y. Zhang *vd.*, "StackRAM: a cross-species method for identifying RNA N 6-methyladenosine sites based on stacked ensemble", (2022).

[3] L. He, H. Li, A. Wu, Y. Peng, G. Shu, ve G. Yin, "Functions of N6-methyladenosine and its role in cancer".

[4] W. Chen, H. Tran, Z. Liang, H. Lin, ve L. Zhang, "Identification and analysis of the N 6-methyladenosine in the Saccharomyces cerevisiae transcriptome OPEN", ***Nature Publishing Group***, (2015).

[5] A. K. Sangaiah *vd.*, "M6AMRFS: Robust Prediction of N6-Methyladenosine Sites With Sequence-Based Features in Multiple Species", (2018).

[6] A. Khan, H. U. Rehman, U. Habib, ve U. Ijaz, "Detecting N6-methyladenosine sites from RNA transcriptomes using random forest", ***Journal of Computational Science***, c. 47, 101238, (2020).

[7] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, ve X. Gao, "Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites", ***Neurocomputing***, c. 324, 3-9, (2019).

[8] A. Maity ve B. Das, "N6-methyladenosine modification in mRNA: Machinery, function and

implications for health and diseases", ***FEBS Journal***, c. 283, 1607-1630, (2016).
[9] J. Luo, T. Xu, ve K. Sun, "N6-Methyladenosine RNA Modification in Inflammation: Roles, Mechanisms, and Applications", ***Frontiers in Cell and Developmental Biology***, c. 9, 670711, (2021).
[10] M. U. Rehman, K. J. Hong, H. Tayara, ve K. T. Chong, "m6A-NeuralTool: Convolution Neural Tool for RNA N6-Methyladenosine Site Identification in Different Species".
[11] M. F. Sabooh, N. Iqbal, M. Khan, M. Khan, ve H. F. Maqbool, "Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC", ***Journal of Theoretical Biology***, c. 452, 1-9, (2018).
[12] Z. Chen *vd.*, "iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets", ***Nucleic Acids Research***, c. 50, W434-W447, (2022).
[13] Z. Chen *vd.*, "iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization", ***Nucleic Acids Research***, c. 49, (2021).
[14] A. El Allali, Z. Elhamraoui, ve R. Daoud, "Machine learning applications in RNA modification sites prediction", ***Computational and Structural Biotechnology Journal***, c. 19, 5510-5524, (2021).
[15] H. Wang, S. Wang, Y. Zhang, S. Bi, ve X. Zhu, "A brief review of machine learning methods for RNA methylation sites prediction", ***Methods***, c. 203, 399-421, (2022).
[16] T. H. Nguyen-Vo, Q. H. Nguyen, T. T. T. Do, T. N. Nguyen, S. Rahardja, ve B. P. Nguyen, "IPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features", ***BMC Genomics***, c. 20, 1-11, (2019).
[17] M. O. Arowolo, M. O. Adebiyi, C. Aremu, ve A. A. Adebiyi, "A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector", ***Journal of Big Data***, c. 8, 1-17, (2021).
[18] C. Lan, H. Peng, G. Hutvagner, ve J. Li, "Construction of competing endogenous RNA networks from paired RNA-seq data sets by pointwise mutual information", ***BMC Genomics***, c. 20, 1-10, (2019).
[19] Y. Bengio, O. Delalleau, N. Le Roux, J. F. Paiement, P. Vincent, ve M. Ouimet, "Learning Eigenfunctions Links Spectral Embedding and Kernel PCA", ***Neural Computation***, c. 16, 2197-2219, (2004).
[20] Y. Liang, S. Zhang, H. Qiao, ve Y. Yao, "iPromoter-ET: Identifying promoters and their strength by extremely randomized trees-based feature selection", ***Analytical Biochemistry***, c. 630, 114335, (2021).
[21] X. Zhu, T. Ching, X. Pan, S. M. Weissman, ve L. Garmire, "Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization", ***PeerJ***, c. 2017, e2888, (2017).
[22] N. Yu, M. J. Wu, J. X. Liu, C. H. Zheng, ve Y. Xu, "Correntropy-Based Hypergraph Regularized NMF for Clustering and Feature Selection on Multi-Cancer Integrated Data", ***IEEE Transactions on Cybernetics***, c. 51, 3952-3963, (2021).
[23] Y. Liang, S. Zhang, H. Qiao, ve Y. Yao, "iPromoter-ET: Identifying promoters and their strength by extremely randomized trees-based feature selection", ***Analytical Biochemistry***, c. 630, 114335, (2021).
[24] X. Zhou, J. Zhu, K. Y. Liu, B. L. Sabatini, ve S. T. C. Wong, "Mutual information-based feature selection in studying perturbation of dendritic structure caused by TSC2 inactivation", ***Neuroinformatics 2006 4:1***, c. 4, 81-94, (2006).
[25] R. Qi, A. Ma, Q. Ma, ve Q. Zou, "Clustering and classification methods for single-cell RNA-sequencing data", ***Briefings in Bioinformatics***, c. 21, 1196-1208, (2020).
[26] S. Karasu ve Z. Saraç, "Güç Kalitesi Bozulmalarının Hilbert-Huang Dönüşümü, Genetik Algoritma Ve Yapay Zeka/Makine Öğrenmesi Yöntemleri İle Sınıflandırılması", ***Journal of Polytechnic***, c. 23, 1219-1229, (2020).
[27] M. O. Arowolo, M. Adebiyi, A. Adebiyi, ve O. Okesola, "PCA Model for RNA-Seq Malaria Vector Data Classification Using KNN and Decision Tree Algorithm", ***2020 International Conference in Mathematics, Computer Engineering and Computer Science, ICMCECS 2020***, (2020).
[28] H. Liu, H. Q. Tian, Y. F. Li, ve L. Zhang, "Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions", ***Energy Conversion and Management***, c. 92, 67-81, (2015).
[29] H. H. Patel ve P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", ***International Journal of Computer Sciences and Engineering***, c. 6, 74-78, (2018).
[30] A. Kulkarni ve B. Lowe, "Random Forest Algorithm for Land Cover Classification", ***Computer Science Faculty Publications and Presentations***, (2016). Erişim: 04 Eylül 2024. [Çevrimiçi]. Erişim adresi: https://scholarworks.uttyler.edu/compsci_fac/1
[31] Z. U. Rehman, M. T. Mirza, A. Khan, ve H. Xhaard, "Predicting G-Protein-Coupled Receptors Families Using Different Physiochemical Properties and Pseudo Amino Acid Composition", ***Methods in Enzymology***, c. 522, 61-79, (2013).
[32] İ. Keskin, M. Yadgar AHMED, A. Makalesi, ve R. Article Mohammed Yadgar AHMED, "A simulation on soil structure interaction with ABAQUS; effect on the behavior of a concrete building of soil layers and earthquake properties", ***Journal of Polytechnic***, c. 27, 749-757, (2024).
[33] D. Chicco ve G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation".