JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING RESEARCH
Bilim, Teknoloji ve Mühendislik Araştırmaları Dergisi
ISSN (Online) 2717-8404
Available *online* at https://dergipark.org.tr/en/pub/jster

RESEARCH ARTICLE

The Impact of Lempel-Ziv Distance Metric in Fuzzy K-Nearest Neighbor Algorithm: A Case Study on Classification of Growth Factors

[1] Berk Tolga Çifci, [1] Ramazan Kabadayı, [1*] Çağın Kandemir Çavaş

[1]Dokuz Eylül Üniversitesi, Fen Fakültesi, Bilgisayar Bilimleri Bölümü, İzmir,Türkiye
berktolga.cifci@ogr.deu.edu.tr, Orcid.0000-0001-9779-270X
ramazan.kabadayi@ogr.deu.edu.tr,Orcid. 0000-0002-5114-291X
cagin.kandemir@deu.edu.tr, Orcid.0000-0003-2241-3546

Citation:
Çiftçi, B.T., Kabadayı, R., Çavaş, Ç.K. (2024). *The Impact of Lempel-Ziv Distance Metric in Fuzzy K-Nearest Neighbor Algorithm: A Case Study on Classification of Growth Factors*, Journal of Science, Technology and Engineering Research, 5(2):148-162. DOI: 10.53525/jster.1573661

HIGHLIGHTS

- *The functions of growth factors in cellular processes and their analysis using bioinformatics methods*
- *The use of the Lempel-Ziv distance metric in the FKNN algorithm and the evaluation of classification success*
- *Presentation of the results of the classification method using the Lempel-Ziv distance and its advantages over the traditional Euclidean distance*

Article Info

Received : 25.10.2024
Accepted : 25.11.2024

DOI:
10.53525/jster.1573661

***Corresponding Author:**

Çağın Kandemir Çavaş
cagin.kandemir@deu.edu.tr
Phone: +90 232 3019512

ABSTRACT

Cellular events occur as a result of the activities of proteins. Different amino acid sequences create distinct protein structures, which in turn affect the activities in cellular events. Therefore, classifying protein sequences structurally or functionally is valuable for gaining insights into their roles in cellular events. Growth factors are proteins involved in processes such as cell proliferation, differentiation, repair, and maintenance. The bioinformatics study of growth factors can provide faster results at a lower cost. Neurotrophins are a family of growth factors that influence the growth, proliferation, differentiation, and functions of nerve cells. In this study, the fuzzy classification of NGF and BDNF, which share a common ancestor and have highly similar protein structures, as well as NT-3, is performed. The performance of the k-Nearest Neighbors (KNN) algorithm, widely used in bioinformatics, significantly depends on the distance measure used. For the Fuzzy KNN (FKNN) algorithm, the distance measurements are important for calculating the degree of fuzziness. In this study, Lempel-Ziv Complexity is proposed as a new distance measure to improve the classification success of the FKNN algorithm. When the Lempel-Ziv distance metric is used in the Fuzzy k-Nearest Neighbors Algorithm, and with data obtained from the Uniprot database, the classification performance reaches 83% when the number of neighbors (K) is 12. The highest classification performance obtained using Euclidean Distance was 75%. At the point where maximum accuracy was achieved, the algorithm's processing time using Euclidean Distance was 0.0054 ms, while it was 0.0038 ms when using the Lempel-Ziv distance. The results show that using Lempel-Ziv distance in the Fuzzy K-Nearest Neighbors Algorithm not only provides faster processing time compared to Euclidean distance but also significantly improves classification performance. In this context, it is believed that our approach could contribute significantly to bioinformatics research.

Keywords: *Fuzzy K-nearest Neighbor Algorithm, Growth Factor, Lempel-Ziv Complexity, Neurotrophin, Protein Sequence*

JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING RESEARCH
Bilim, Teknoloji ve Mühendislik Araştırmaları Dergisi
ISSN (Online) 2717-8404
Available *online* at https://dergipark.org.tr/en/pub/jster

ARAŞTIRMA MAKALESİ

Bulanık K-En Yakın Komşuluk Algoritmasında Lempel-Ziv Mesafe Ölçütünün Etkisi: Büyüme Faktörlerinin Sınıflandırılması Örneği

Berk Tolga Çifci, [1] Ramazan Kabadayı, [1*] Çağın Kandemir Çavaş

[1]Dokuz Eylül Üniversitesi, Fen Fakültesi, Bilgisayar Bilimleri Bölümü, İzmir,Türkiye
berktolga.cifci@ogr.deu.edu.tr, Orcid.0000-0001-9779-270X
ramazan.kabadayi@ogr.deu.edu.tr,Orcid. 0000-0002-5114-291X
cagin.kandemir@deu.edu.tr, Orcid.0000-0003-2241-3546

Alıntı / *Citation* **:**
Çiftçi, B.T., Kabadayı, R., Çavaş, Ç.K. (2024). *The Impact of Lempel-Ziv Distance Metric in Fuzzy K-Nearest Neighbor Algorithm: A Case Study on Classification of Growth Factors*, Journal of Science, Technology and Engineering Research, 5(2):148-162. DOI: 10.53525/jster.1573661

ÖNE ÇIKANLAR / *HIGHLIGHTS*

- *Büyüme faktörlerinin hücresel süreçlerdeki işlevleri ve biyoenformatik yöntemlerle incelenmesi*
- *FKNN algoritmasının Lempel-Ziv mesafe metriğinin kullanılması ve sınıflandırma başarılarının değerlendirilmesi*
- *Lempel-Ziv uzaklığının kullanıldığı sınıflandırma yönteminin sonuçlarının ve geleneksel Öklid uzaklığına göre avantajlarının ortaya konulması.*

Makale Bilgileri / *Article Info*

Geliş Tarihi : 25.10.2024
Kabul Tarihi : 25.11.2024

DOI: 10.53525/jster.1573661

*** Sorumlu Yazar:**

Çağın Kandemir Çavaş
cagin.kandemir@deu.edu.tr
Phone: +90 232 3019512

ÖZET / *ABSTRACT*

Hücresel olaylar, proteinlerin aktiviteleri sonucunda gerçekleşir. Amino asitlerin farklı dizilimleri farklı protein yapılarının oluşmasına neden olur. Yapılarına göre hücresel olaylardaki aktiviteleri de değişiklik gösterir. Bu nedenle protein dizilerinin yapısal veya işlevsel olarak sınıflandırılması hücresel olaylardaki rolleri hakkında bilgi edinmek için oldukça değerlidir. Büyüme faktörleri; hücreler üzerinde çoğalma, farklılaşma, onarım ve bakım gibi birçok süreçte yer alan proteinlerdir. Büyüme faktörlerinin biyoenformatik alanında incelenmesi, düşük maliyetle daha hızlı sonuçlara ulaşılmasını sağlayabilir. Nörotrofinler; sinir hücrelerinin büyümesi, çoğalması, farklılaşması ve fonksiyonları üzerinde etkili olan büyüme faktörü ailelerinden biridir. Çalışmamızda, ortak bir atadan gelen ve çok benzer yüksek dereceli protein yapısına sahip olan NGF ve BDNF'nin, ayrıca NT-3'ün bulanık sınıflandırılması yapılmaktadır. Biyoenformatik alanında yaygın olarak kullanılan k-En Yakın Komşuluk (KNN) algoritmasının performansı önemli ölçüde kullanılan mesafeye bağlıdır. Bulanık KNN (FKNN) algoritması için de mesafe ölçümleri, bulanıklık derecesini hesaplamak için önemlidir. Bu çalışmada, FKNN algoritmasının sınıflandırma başarısını artırmak amacıyla yeni bir mesafe ölçütü olarak Lempel-Ziv Karmaşıklığı önerilmiştir. Bulanık K-En Yakın Komşuluk Algoritması'nda Lempel-Ziv mesafe metriği kullanıldığında, Uniprot veri tabanından alınan verilerle birlikte FKNN algoritmasında K komşu sayısının 12 olması durumunda, sınıflandırma performansı %83 olarak elde edilmiştir. Öklid Uzaklığı kullanıldığında elde edilen en yüksek sınıflandırma performansı ise %75'tir. Maksimum doğruluk oranını elde ettiğimiz noktada Öklid uzaklığını kullandığımızda algoritmamızın çalışma süresi 0.0054 ms iken Lempel-Ziv uzaklığı kullandığımızda 0.0038 ms'dir. Elde edilen sonuçlara göre, Bulanık K-En Yakın Komşuluk Algoritması'nda Lempel-Ziv mesafesinin kullanılması, Öklid mesafesine kıyasla daha hızlı bir işlem süresi sağlamakla kalmamış, aynı zamanda sınıflandırma performansını da önemli ölçüde iyileştirmiştir. Bu bağlamda, önerdiğimiz yaklaşımın biyoenformatik çalışmalarına önemli katkılar sunabileceği düşünülmektedir.

Anahtar Kelimeler: *Bulanık k-en yakın komşu algoritması, büyüme faktörü, Lempel-Ziv karmaşıklığı, nörotrofinler, protein sekansı.*

I. GİRİŞ [INTRODUCTION]

Organizmaların incelenmesi üzerine yapılan araştırma ve projeler sonucunda zengin biyolojik veriler elde edildi ve artan bir şekilde elde edilmeye devam ediliyor. Deneysel yöntemler, oldukça fazla yoğun iş gücü, uzun zaman ve maliyet gerektirdiği için elde edilen yeni verilerin incelenmesi, yorumlanması maliyetli ve zaman alıcıdır. Biyoenformatik, deneysel yöntemlere göre daha az zaman ve maliyetle verileri incelemeye ve sonuçlar elde etmeye olanak sağlar. Biyoenformatik, bilgisayar bilimi yöntemleriyle biyolojik verileri depolamak, analiz etmek, modellemek için yöntemler geliştiren disiplinler arası bir alandır. Biyoenformatik; veri madenciliği teknikleri, görselleştirme, makine öğrenimi teknikleri gibi yöntemleri kullanarak verileri verimli bir şekilde analiz etmeyi amaçlar. Bu alanda verilerin sınıflandırılması da önemli bir konu olarak yer alır.

Hücre içerisinde gerçekleşen çoğunlukla birçok olay, proteinlerin aktiviteleri sonucu gerçekleşir. Ayrıca vücuttaki organ ve dokuların yapı ve işlevlerinde etkilidirler. Proteinlerin dört farklı yapısı bulunmaktadır. Bunlardan birincil yapısı, amino asitlerin doğrusal dizisinden oluşur [1]. Proteinler, birçok bileşenden oluşsa da farklılaşmasını sağlayan yan zincir olarak adlandırılan amino asitlerdir. Amino asitlerin bir araya gelmesiyle beraber aynı zamanda farklı bağlarla bir araya gelmeleri çeşitli üç boyutlu yapılara katlanmalarına neden olur. Katlanan yapılar da proteinin doğrusal amino asit dizisine bağlıdır ve bu yapılar da proteinin işlevini belirler [2]. Dolayısıyla yapısal veya işlevsel olarak yapılacak sınıflandırma o proteinin hangi hücresel olaylarda nasıl görev aldığını anlamak açısından büyük öneme sahiptir. Sınıflandırmanın diğer bir önemli yanı da homolojiyi bulmaktır. Homoloji, dizilerin ortak bir atadan geldiklerini dolayısıyla ortak işlevsel yönlerinin olduğunu gösterir.

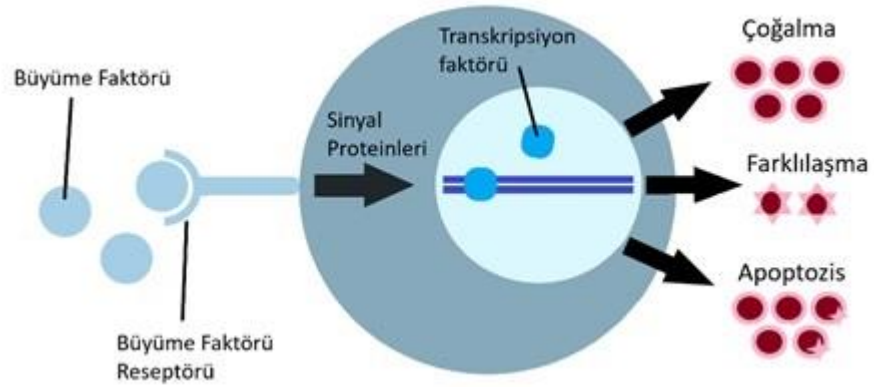
Biyoenformatik çalışmalarında k-en yakın komşu algoritması (KNN) yaygın olarak kullanılmaktadır. KNN, etiketli olmayan verileri, mesafe ölçümü kullanarak en yakın k adet etiketli verilerin sınıflarına göre sınıflandıran makine öğrenmesi algoritmalarından biridir. K adet en yakın komşudan çoğunluk hangi sınıf ise yeni değer o sınıfa atanır. Ancak KNN algoritması, her komşuya aynı önemi verir [3]. Klasik KNN algoritmasında bir veri, bir sınıfa atandığında verilerin arasındaki uzaklık durumuna göre bu atama gerçekleşir. Bu durumda o sınıfa üyeliğin gücünün hiçbir etkisi yoktur. Etkili olan yalnızca mesafedir. Bu noktada KNN algoritmasına bulanıklığın eklenmesi gerekmektedir. Bulanık KNN algoritması, en başta sınıfı bilinmeyen veri ile diğer verileri arasındaki mesafeyi hesaplar ve en yakın k adet komşuyu seçer ancak atama işlemi mesafeye göre değil, sınıfa ait üyelik derecesine göre yapılır. Algoritma içerisinde elde edilen üyelik değerleri, sonuçtaki sınıflandırmaya eşlik edecek bir güvence düzeyi sağlar. Bu sayede algoritma içerisinde keyfi atamalar yapılmaz [4].

KNN algoritması yukarıda da bahsedilen verilere eşit önem verme durumu nedeniyle bazen istenilen sonuçlara ulaşmada zorluklara neden olmaktadır. Örneğin, etiketli tüm verilere eşit önem verilmesi, mesafe temelli bir sınıflandırma yapılmasına neden olur. Protein dizisi o proteine özgüdür ve proteinin yapısını belirler. Mesafeye göre yapılan bir sınıflandırma, protein dizisinin yapısını göz ardı eder. Bulanık yöntem, protein dizisinin yapısı nedeniyle sahip olduğu sınıfın da işleme dâhil edilmesini sağlayarak sınıflandırmanın mesafe ile birlikte proteinin yapısının da göz önünde bulundurularak yapılmasını sağlar. KNN algoritmasında bulanık yaklaşımın kullanıldığı çalışmalar incelendiğinde; proteinin yapısal sınıf tahmini [5], protein çözücü erişebilirliğinin tahmini [6], hücre altı konumlarının tahmin edilmesi [7], proteinlerin ikincil yapısının tahmin edilmesi [8], DNA mikro dizi sınıflandırması [9], gen fonksiyon sınıflandırılması [10], gibi konularda bu yaklaşımın kullanıldığı görülmektedir. Farklı çalışmalarda, KNN'nin biyolojik ve tıbbi verilerin sınıflandırılmasında yaygın olarak kullanılmasına rağmen, henüz uygulanmamış başka araştırma alanlarının olduğunu ve bu alanlarda bulanık KNN ile yapılan deneysel sonuçlarda yüksek tahmin doğrulukları elde edildiği de görülmektedir [11]. İncelenen birçok çalışmada bulanık k-en yakın komşu algoritmasının çoğunlukla proteinlerde tahmin amacıyla kullanıldığı görülmektedir. Kendi çalışmamızda bulanık KNN algoritması, protein dizilerinin sınıflandırılması amacıyla kullanılmaktadır.

KNN, eğitim verilerine olan mesafenin hesaplanmasına dayanan bir sınıflandırma yöntemidir. Dolayısıyla KNN algoritmasının performansı önemli ölçüde kullanılan mesafeye bağlıdır [12]. Öklid uzaklığı KNN'de kullanılan en yaygın uzaklık metriğidir. Ayrıca farklı çalışmalarda, KNN'de, Manhattan, Minkowski, Kosinüs, Hamming, Jaccard,

Mahalanobis, City Block, Chebyshev uzaklık modellerinin [13]; bulanık KNN'de, Manhattan, Kosinüs, Hamming, City Block uzaklık modellerinin [14] kullanıldığı görülmektedir. Mesafe ölçümleri, bulanıklık derecesini hesaplamak için önemlidir [15]. Lempel-Ziv, kayıpsız bir veri sıkıştırma algoritmasıdır. Ek olarak Lempel-Ziv karmaşıklığı bir dizideki farklı yapıların sayısını vermektedir. Lempel-Ziv, DNA dizilerinin karmaşıklığını incelemek için kullanılmıştır [16]. Çalışmamızda, bulanık KNN algoritmasında test verisi ile eğitim verileri arasındaki mesafeyi ölçmek ve bilgi kaybını önlemek için protein sekanslarının Lempel-Ziv karmaşıklık değerlerinin bir tür mesafe ölçümü olarak kullanılması önerilmektedir.

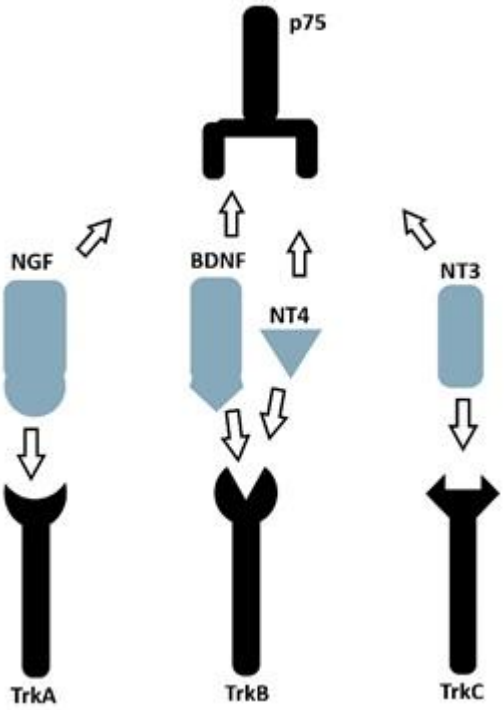
Şekil 1'de gösterildiği gibi, büyüme faktörleri, hücreler üzerinde çoğalma, farklılaşma, onarım ve bakım gibi birçok süreçte yer alan, hücre yüzeyindeki reseptörlere bağlanan proteinlerdir. Büyüme faktörleri, yapısal ve evrimsel olarak kendi içinde ilişkili olan protein ailelerinden oluşur. Her bir büyüme faktörü ailesi, farklı fonksiyonel özelliklere sahiptir ve farklı hücresel faktörleri yalnızca süreçlerde rol alırlar. Büyüme faktörleri, yalnızca onarım gibi süreçlerde değil ayrıca programlı hücre ölümü gibi süreçlerde de yer alabilir. Aktiviteleri, hücre tipine ve organizmanın gelişim aşamasına göre değişir [17]. Büyüme faktörlerinin in vivo çalışmaları kısa yarı ömre, zayıf bir dayanıklığa yol açar [18]. Etkili sonuç alabilmek için yüksek dozlar gereklidir ve bu da yüksek maliyete neden olmaktadır [19]. Büyüme faktörlerinin biyoenformatik alanında incelenmesi düşük maliyetle, daha hızlı sonuçlara ulaşılmasını sağlayabilir.



Şekil. 1. Büyüme faktörü (Growth factor)

Nörotrofinler (NGF), sinir hücrelerinin büyümesi, çoğalması, farklılaşması, fonksiyonları üzerinden etkisi olan ve sinaptik plastisitenin sağlanmasında etkili olan büyüme faktörü ailelerinden biridir. NGF'ler, biyolojik ve yapısal olarak birbirlerine benzemektedirler [20]. İlk olarak tükürük bezleri üzerinde yapılan deneyde molekül saflaştırılması sonucu NGF adı verilen bir protein keşfedilmiştir [21]. Daha sonraki deneyler NGF'nin, nöronların hayatta kalması ve farklılaşmasını destekleyen bir protein olduğunu göstermiştir [22]. NGF, sinir hücrelerinin ulaştığında hayatta kaldığı, ulaşamadığında öldüğü sınırlı miktarda üretilen bir proteindir [23]. NGF, sempatik ve duyusal nöronların hayatta kalması ve bakımı için önemliyken; beyinde türetilmiş sinir hücresi büyüme faktörü (BDNF), merkezi ve çevresel sinir sistemindeki nöronların gelişimi için önemlidir. Şekil 2'de gösterildiği üzere, NGF'ler, tirozin kirazlar (Trk) reseptörlerine ve ayrıca p75 reseptörüne de bağlanabilmektedirler. Örneğin, BDNF, düşük afinitede p75 reseptörüne bağlanarak hücre ölümüne neden olurken yüksek afinitede TrkB reseptörüne bağlanarak hücrenin yaşamasını sağlar, Uzun Süreli Güçlendirme'ye (Long Term Potentiation) neden olur. NGF'ler farklı reseptör bağlantılarında farklı işlevlere sahiptirler. Bu nedenle yapılarının, bağlantılarının incelenmesi önemlidir. Literatürdeki çalışmalar NGF ve BDNF'nin çok benzer bir yüksek dereceli protein yapısına sahip olduklarını göstermektedir [24]. "Her iki protein arasında maksimum hizalama için amino asit dizilerinde yalnızca birkaç boşluk eklenmesi gerekir" [25]. "Amino asit düzeyinde NGF ve BDNF'nin olgun formları arasında yaklaşık %50 amino asit özdeşliği mevcuttur" [24]. Çalışmalar her ne kadar bu proteinlere dair bilgiler sunsa da hücresel ve moleküler işlevlerinin hala iyi derecede anlaşılmadığını da göstermektedir [24] [26]. Ayrıca çalışmalar, bu iki nörotrofik faktörün farklı görevlerini ortaya koymasıyla beraber paralel ve karşılıklı rollere sahip olduklarını da göstermektedir [24] [27]. NGF ve BDNF ile ilgili çalışmaların moleküler biyoloji teknikleri kapsamında gerçekleştirildiği görülmektedir [24]. Biyoenformatik alanı temelinde literatürde NGF ve BDNF'nin sınıflandırılmasıyla ilgili herhangi bir çalışma bulunmamaktadır. Çalışmada, Lempel-Ziv karmaşıklığının

bulanık kNN algoritmasında uzaklık metriği olarak kullanılarak, ortak bir atadan gelen ve yapısal olarak birbirine benzeyen NGF ve BDNF'nin bulanık sınıflandırılması yapılmaktadır. Böylelikle bu çalışma, makine öğrenmesi tekniklerinin nörotrofinlerin sınıflandırılmasında ilk kez uygulanması açısından bir yenilik sunmaktadır.

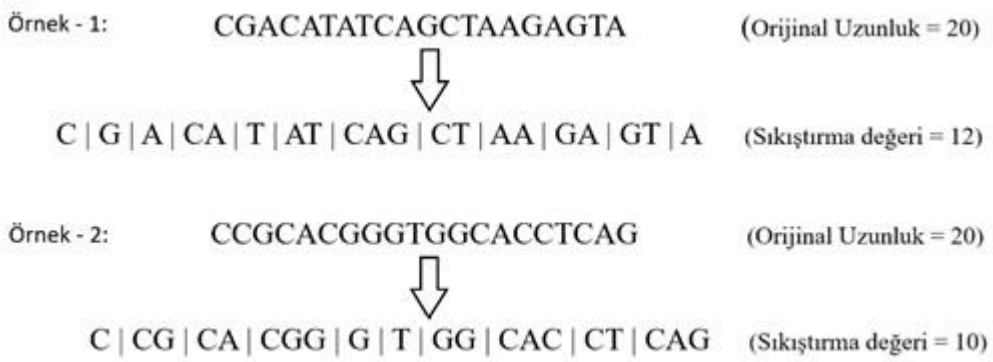


Şekil 2. Nörotrofinler (NGF)

II. MALZEME VE YÖNTEMLER [MATERIALS AND METHODS]

A. Lempel-Ziv karmaşıklığı (Lempel-ziv complexity)

Veri sıkıştırma, kayıplı ve kayıpsız olarak iki türdür. Kayıpsız sıkıştırma, istatiksel fazlalığı ortadan kaldırarak bitleri azaltır. Kayıpsız sıkıştırma, sıkıştırılmış verilerin yeniden yapılandırılarak orijinal verilerin elde edilmesini sağlayan bir tekniktir. Bu işlem sırasında hiçbir veri kaybolmaz. Kayıplı sıkıştırma, daha az önemli veya gereksiz bilgileri ortadan kaldırarak bitleri azaltır. Farklı sıkıştırma yöntemleri bulunmaktadır. Tek geçiş sıkıştırma veya uyarlamalı sıkıştırma, verileri yalnızca tek geçişte sıkıştırır. İki geçişli sıkıştırma veya yarı uyarlanabilir sıkıştırma, verileri bir geçişte analiz eder ve daha sonra başka bir geçişte sıkıştırır [28]. İki geçişli sıkıştırma algoritmalarından biri olan Huffmann Kodlaması, veri içerisindeki en çok kullanılan karakterler üzerine odaklanan bir algoritmadır. Ne kadar çok tekrar eden karakter varsa o kadar iyi sıkıştırma gerçekleşir. Ancak Huffmann Kodlaması'nın başarısı, kaynak dizinin dağılımına bağlıdır [29]. Her zaman kaynak dizinin dağılımı sık tekrar eden karakterlerden oluşmayabilir. Bu durumda, veri içerisindeki en çok tekrar eden karakterlerden ziyade tekrar eden yapılara odaklanan sözlük tabanlı ve kayıpsız veri sıkıştırma algoritmalarından biri olan Lempel-Ziv (LZ78) kullanılabilir. Huffman gibi teknikler, bir kerede tek bir sembolü kodlarken sözlük tabanlı sıkıştırıcı olan Lempel-Ziv, bir verideki belirli bir sembol grubunu kodlar [30]. Bu teknik, Huffman Kodlaması gibi statik tekniklerin aksine daha verimlidir [31].



Şekil 3. Lempel-Ziv Örnek Gösterimi

Lempel-Ziv algoritması, tek geçişte verileri sıkıştırır ve sıkıştırılacak veri hakkında önceden herhangi bir bilgiye ihtiyaç duymaz [32]. Temel olarak, bir dizi soldan sağa doğru incelendiğinde karşılaşılan farklı alt dizilerin sayısı olarak tanımlanabilir. Lempel-Ziv algoritması şu şekilde işler: İlgili verinin içerisindeki farklı karakterlerin olduğu bir sözlük mevcuttur. Algoritma veriyi karakter karakter okur ve sözlükte olup olmadığını kontrol eder. Eğer sözlükte yoksa sözlüğe ekler. Eğer sözlükte varsa, sözlükte olmayan karakterle karşılaşana kadar karakter karakter okumaya devam eder ve bu elde edilenleri bir desen yapıları olarak gruplandırmaya başlar. Bu sayede farklı yapı veya desenler aranır. Son olarak Lempel-Ziv sıkıştırma değeri, bulunan farklı yapıların veya desenlerin sayısına göre belirlenir. Lempel-Ziv, bir veri sıkıştırma algoritması olduğu kadar aynı zamanda veride bulunan kapıların ne kadar çeşitli olduğunu ölçmesi bakımından bir karmaşıklık ölçüsüdür.

```
Algoritma 1 – Lempel-Ziv
P = ilk girdi karakteri
WHILE okunacak karakter olduğu sürece
  C = bir sonraki girdi karakter
  IF P+C sözlükte varsa
    P = P + C
  ELSE
    P'nin indeksini yazdır
    P + C'yi sözlüğe ekle
    P = C
END WHILE
P'nin indeksini yazdır
```

Sınıflandırma için kullanılan algoritmalar, sayısal girdiler almak zorundadır. Elde edilen DNA, protein veya enzim gibi veriler sayısal veriler olmadığı için bu verilerin algoritmalarda kullanılması için sayısala dönüştürülmesi gerekir. Kimi zaman bu dönüştürme işlemleri verilerde kayıplara neden olabilmektedir. Kayıpsız veri sıkıştırma algoritmaları, biyolojik çalışmalardan elde edilen sekansların algoritmalarda verimli bir şekilde kullanılmasına olanak sağlar. Bu açıdan Lempel-Ziv algoritması, sayısal olmayan verileri kayıpsız şekilde sıkıştırarak sayısal bir şekilde algoritmada kullanılmasını sağlar. Ayrıca, özellikle çalışmamızdaki bulanık sınıflandırma işleminde protein dizisinin yapısı nedeniyle sahip olduğu sınıfın da işleme dahil edilmesini istemediğimiz için Lempel-Ziv sıkıştırma değeri veya Lempel-Ziv karmaşıklığı, protein verilerinin yapısını da yansıtmış olacaktır.

B. Lempel-Ziv karmaşıklığına dayalı mesafe ölçümü (Lempel-Ziv complexity based distance measure)

S = ACGAGCTATT ve R = TCCTCTTT şeklinde iki dizi verildiğinde her iki dizinin LZ karmaşıklığı c(S) = 7 ve c(R) = 5 olarak bulunur.

c(S): A | C | G | AG | CT | AT | T
c(R): T | C | CT | CTT | T

S ve R dizileri iki farklı sırayla birleştirilir:

S.R = ACGAGCTATTTCCTCTTT
R.S = TCCTCTTTACGAGCTATT

Daha sonra birleştirme sonucu elde edilen iki dizinin Lempel-Ziv karmaşıklıkları c(S.R) = 11 ve c(R.S) = 10 olarak bulunur.

c(S.R) = A | C | G | AG | CT | AT | T | TC | CTC | TT | T
c(R.S) = T | C | CT | CTT | TA | CG | A | G | CTA | TT

Lempel-Ziv karmaşıklıklarının farkına dayalı olarak iki dizi arasındaki benzerlik derecesini hesaplamak için aşağıdaki formül (1) kullanılır.

$$d(S,R) = \frac{\max[c(S.R) - c(S), c(R.S) - c(R)\ ]}{\max[c(S), c(R)]} \quad (1)$$

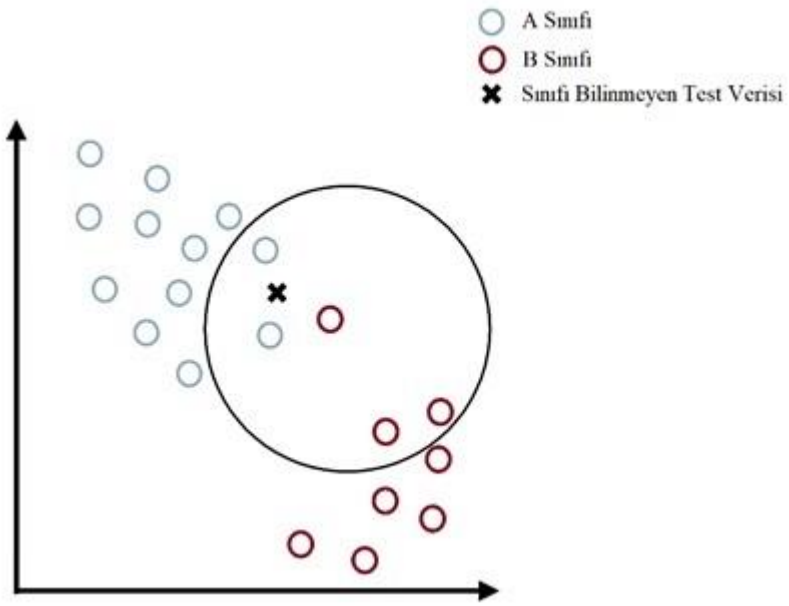
Bu formül (1), dizi uzunluğunun mesafe ölçümlerinin hesaplanması üzerindeki etkisini ortadan kaldırabilen normalleştirilmiş bir mesafe ölçümüdür [33].

Formüle dayanarak S ve R dizilerinin arasındaki Lempel-Ziv mesafesi şu şekilde hesaplanır:

$$d(S,R) = \frac{\max[11 - 7, 10 - 5\ ]}{\max[7,5]} = \frac{5}{7} = 0.71428 \quad (2)$$

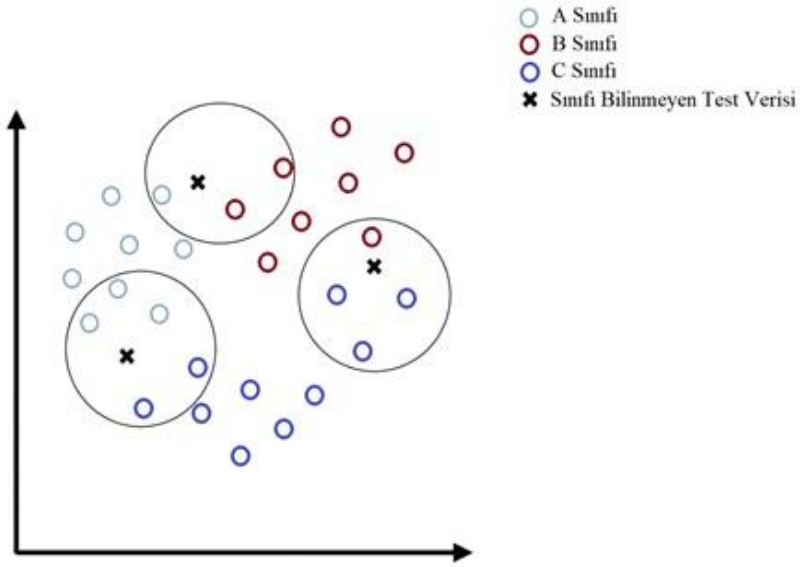
C. Bulanık k-en yakın komşu algoritması (Fuzzy k-nearest neighbor algorithm)

Makine öğrenimi teknikleri denetimli öğrenme ve denetimsiz öğrenme şeklinde iki alt kategoriye ayrılır. Denetimsiz öğrenmede tüm veriler etiketli değildir ve algoritmalar, veriler hakkında daha çok bilgi sahibi olmak için verilerdeki temel yapıyı modeller. Denetimli öğrenmede, etiketli veri kümesi kullanılarak algoritmalar, girdi verilerinden çıktıyı tahmin etmeyi öğrenir. K-en yakın komşu algoritması (KNN), etiketli olmayan veriyi, mesafe ölçümleri kullanarak kendisine en yakın k adet etiketli verilerin sınıflarına göre sınıflandıran denetimli makine öğrenmesi algoritmalarından biridir. K, en yakın komşu sayısını belirleyen önemli bir parametredir. Sabit bir değeri yoktur ve en etkili sonuç için algoritma farklı k değerleri için çalıştırılmalıdır. KNN'de, k adet en yakın komşudan çoğunluk hangi sınıf ise yeni değer o sınıfa atanır. Bu duruma çoğunluk oylaması veya oylama KNN kuralı [34] adı verilir. Komşu veriler, mesafeye göre belirlenir ve genellikle Öklid mesafesi kullanılır. Bayes sınıflandırıcı, sinir ağları gibi birçok sınıflandırma yöntemi, eğitim ve test veri kümlerinin aynı dağılıma sahip olduğu durumlarda çok iyi performans gösterirler. Ancak benzer ama farklı dağılımlarda performansları iyi değildir. KNN, verilerin dağılımlarının aynı olmasını gerektirmeyen bir yöntemdir [35]. Klasik KNN, seçilen tüm komşulara aynı önemi verir. Bu durum, algoritmanın en yakın k komşunun sınıflarından çoğunluk oylamasıyla sınıflandırmaya karar verirken her komşuya oy için aynı ağırlığı verdiği anlamına gelir. Sonuçta bu özellik, KNN'nin k değerine oldukça fazla duyarlı olduğunu gösterir [10].



Şekil 4. Bulanık k-en yakın komşu algoritması: belirsizlik, örnek 1 (Fuzzy k-nearest neighbor algorithm: uncertanity, example 1)

Klasik KNN'de, bir veri bir sınıfa atandığında, o sınıfa üyeliğin gücünün hiçbir etkisi yoktur. Sınıf üyelikleri olmaması nedeniyle belirsizliklerle baş edemez. KNN'nin performansını artırmak için bulanık mantığa [36] dayalı bulanık k-en yakın komşu algoritması (FKNN) geliştirilmiştir. FKNN, sınıf üyeliklerini dâhil eder ve rastgele atama yapılmamasını sağlayarak sınıflandırmada bir güvence düzeyi sağlar [4]. Algoritmanın temelinde, veriyi belirli bir sınıfa atamak yerine veriye sınıf üyeliği atamak vardır. Şekil 4, k değeri 5 olması karşılığında bir test verisinin en yakın 5 komşusu alındığını ve bu 5 komşudan 3 tanesi B sınıfına ait olduğu için test verisinin de B sınıfına ait olduğunu gösterir. Şekil 4'te test verisi mesafe olarak A sınıfındaki verilere daha yakın olsa da çoğunluk oylaması nedeniyle B sınıfına atanır. Bu durum da bir belirsizlik oluşturur [10]. Şekil 5, üç test örneğinin ve komşularının belirsiz alan oluşturduğu bir örnektedir. Klasik KNN, belirsizlikle baş edemeyeceği için FKNN'nin kullanılması gerekmektedir. Klasik KNN'de olduğu gibi FKNN'de de k değeri önemlidir. K değerinin sabit bir değer olması, belirsizlikler içeren eğitim örneklerinde FKNN'nin performansını olumsuz bir şekilde etkiler [35].



Şekil 5. Bulanık k-en yakın komşu algoritması: belirsizlik, örnek 2 (Fuzzy k-nearest neighbor algorithm: uncertanity, example 2)

Bulanık k-en yakın komşu algoritmasında belirsizlikle baş edebilmek için her test verisi, en yakın eğitim verileri kullanılarak yeniden temsil edilir. Örneğin şekil 5'teki test noktalarından biri k değeri 3 olduğunda, diğeri k değeri 4 olduğunda ve bir diğeri k değeri 5 olduğundaki komşulukları ile yeniden temsil edilirler. Daha sonra algoritma içerisinde bu komşulukların, sınıflara ait üyelik dereceleri hesaplanır.

Algoritma 2 – Bulanık k en yakın komşu algoritması
Algoritma öncesinde; tüm sekansların Lempel-Ziv karmaşıklıkları hesaplanır. Bu karmaşıklık değerleri, formül (1)'de kullanılarak Lempel-Ziv uzaklık değerleri hesaplanır. Lempel-Ziv uzaklıklarından simetrik matris oluşturulur ve algoritma bu matris üzerinden işlemleri gerçekleştirir.

Adım 1: Her bir test verisi ile diğer tüm eğitim verileri arasındaki k adet en yakın komşuyu klasik KNN kullanarak bul.
Seçilen k adet komşunun kaç tanesinin hangi sınıfa ait olduğunu bul.
Sınıf adedi frekans değerlerini k değerine böl.
Elde edilen değeri k değerine tekrar böl. Son değer $\left(\frac{V_i}{kInit}\right)$'tir.
Adım 2: Adım 1'de elde edilen $\left(\frac{V_i}{kInit}\right)$ değerini formül (3)'te kullanarak ilgili sınıfın kendi ve diğer sınıflara ait üyelik derecelerini hesapla.
Adım 3: Elde edilen k adet komşuların birbirleri arasındaki Lempel-Ziv uzaklıklarını matriste ara.
Adım 4: Adım 2'de elde edilen üyelik dereceleri ve Adım 3'te elde edilen Lempel-Ziv uzaklıklarını formül (4) içerisinde kullanarak test verisinin sınıflara ait üyelik derecelerini hesapla.
Adım 5: Hangi sınıf üyeliği daha yüksekse test verisini o sınıfa ata.

$$u_{ji} = \begin{cases} 0.51 + \left(\frac{V_i}{kInit}\right) * 0.49 & eğer\ i = j \\ \left(\frac{V_i}{kInit}\right) * 0.49 & aksi\ halde \end{cases} \quad (3)$$

i, en yakın komşunun gerçek sınıflarıdır. j, eğitim veri setinden alınan test verilerinin gerçek sınıf değerleridir. V_i, ilgili i sınıfının en yakın k komşu içerisindeki sayısıdır. kInit, toplam komşu değeri olan k'dır. $\left(\frac{V_i}{kInit}\right)$, k adet komşudaki sınıf frekans bilgisinin k değerine oranının tekrar k değerine oranı ile elde edilir. Her sınıfa üyelik bu formüle göre atanır.

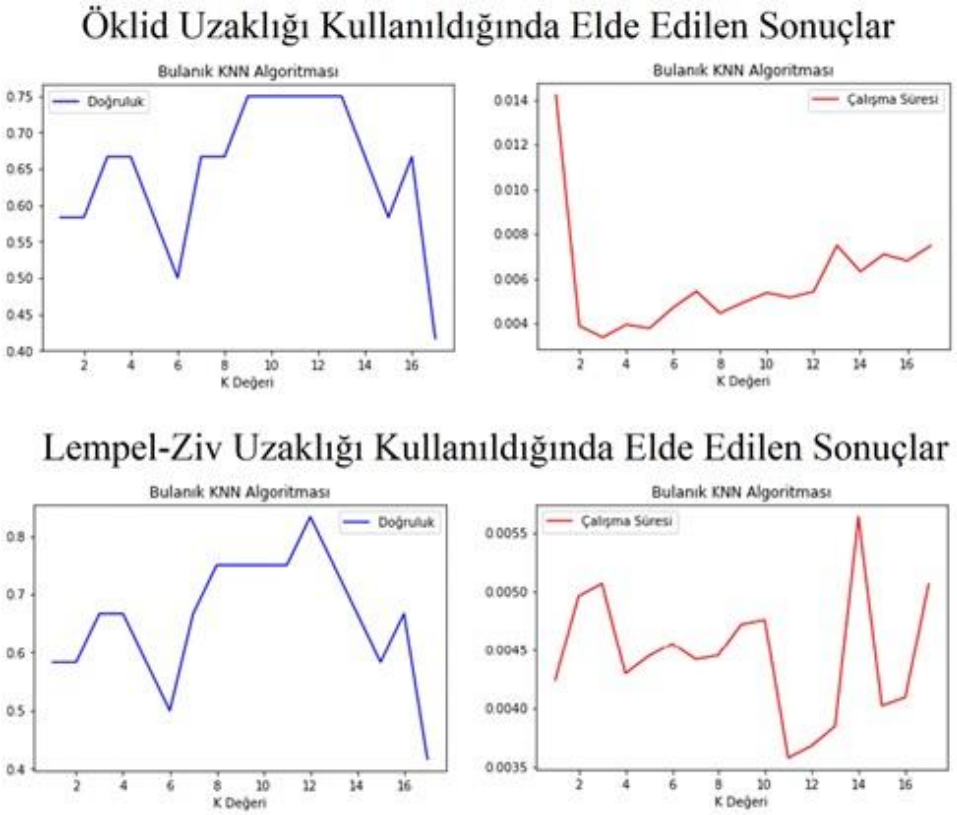
$$\mu^j(x) = \frac{\sum_{i=1}^{k} uji \left(\frac{1}{||\, x-y_i \,||^{2/(m-1)}}\right)}{\sum_{i=1}^{k} \left(\frac{1}{||\, x-y_i \,||^{2/(m-1)}}\right)} \quad (4)$$

$||\ x - y_i\ ||$, ilgili test verisinin j. komşudaki Lempel-Ziv uzaklık değeridir. m, ağırlıklandırma parametresidir. Bu parametre, bulanık üyelik değerinin hesaplanmasında her bir en yakın komşunun ağırlığını belirler. m, 2 olarak seçildiğinde her komşunun etkisi, uzaklığın tersi ile ağırlıklandırılır. Ters uzaklık, bir verinin incelenen veriden yakınsa daha fazla, uzaksa daha az üyeliğini ağırlıklandırmaya yarar. m, bire yaklaştıkça, yakın komşular uzaktakilerden çok daha fazla ağırlıklandırılır; m arttıkça, komşular daha eşit ağırlıkta olur ve göreli mesafeleri daha az etkiye sahip olur [4].

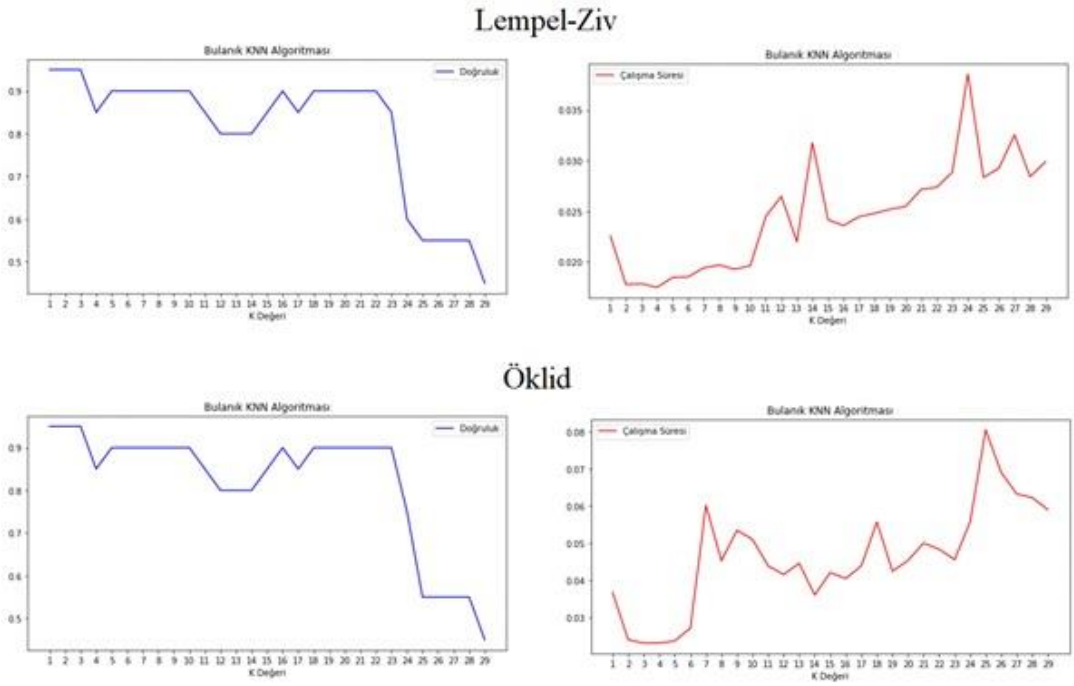
III. SONUÇ VE DEĞERLENDİRME [CONCLUSION]

Geliştirilen algoritmaların gerçek dünya verilerindeki performansını yansıtması açısından doğru değerlendirme yöntemlerinin kullanılması kritik öneme sahiptir. İstatistiksel tahmin yöntemlerinde, bağımsız bir veri seti, jackknife testi birçok araştırmacı tarafından modelin başarısını test etme amacıyla kullanılmaktadır. Jackknife testi, bu değerlendirmelerde yaygın olarak tercih edilen bir tekniktir çünkü her bir veri örneğini sırasıyla test verisi olarak kullanarak, modelin her bir veri noktasındaki başarısını objektif bir şekilde ölçmeye olanak tanır. Jackknife testi, matematiksel olarak kanıtlanmış olan daha anlamlı sonuçlar vermektedir [37-38]. Bu yaklaşım, özellikle biyoinformatik alanında, her protein veya gen dizisinin doğru bir şekilde sınıflandırılmasını sağlamak ve modelin genel doğruluğunu güvenilir bir şekilde belirlemek için kullanılır. Bu sayede, algoritmanın overfitting (aşırı uyum) veya underfitting (yetersiz uyum) gibi problemlere karşı daha dayanıklı olması sağlanabilir. Çalışmamızda, Jackknife testi, her bir sekans için algoritma performansını değerlendirmek için kullanılmaktadır; burada her bir sekans, test verisi olarak davranırken

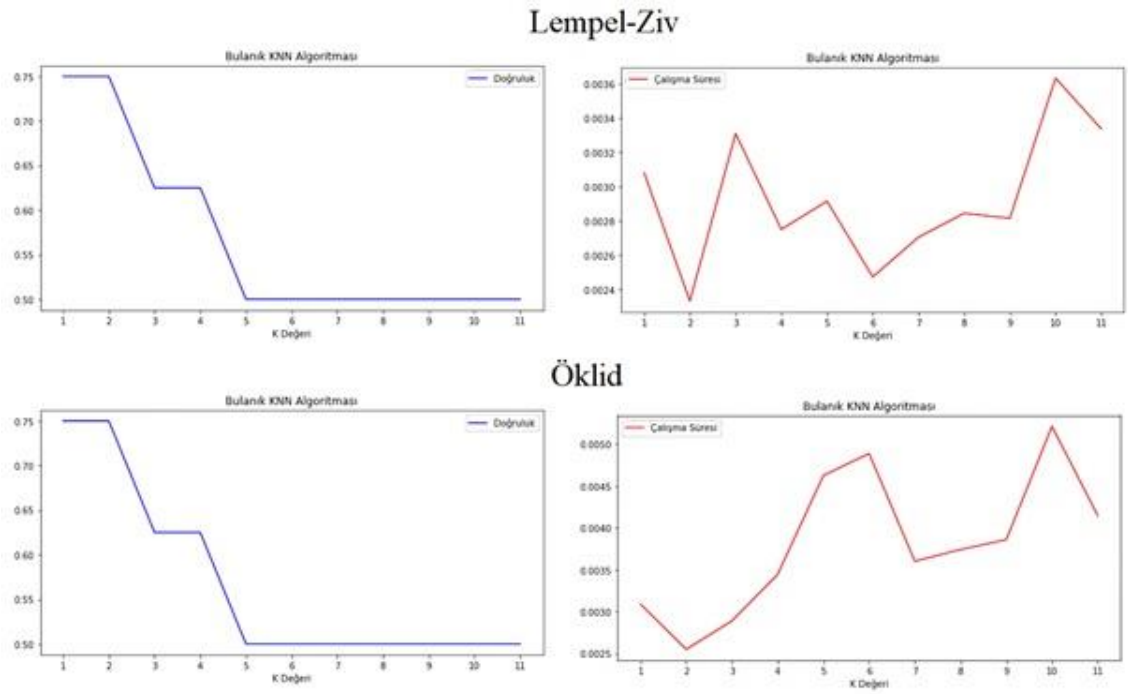
geriye kalanlar eğitim verisi olarak kullanılır. Veri tabanlarından alınan nörotrofin sekansları listesinde baştan sona kadar, her protein sırayla test verisi olarak görev yapmaktadır. Çalışmada, Uniprot veri tabanından Homo Sapiens için NGF ile ilişkili 15 tane ve BDNF ile ilişkili 15 tane sekans alındı. Lempel-Ziv uzaklığı ve Öklid uzaklığı kullanıldığında bulanık k-en yakın komşu algoritmasının sonucu Şekil 6'da gösterilmektedir.



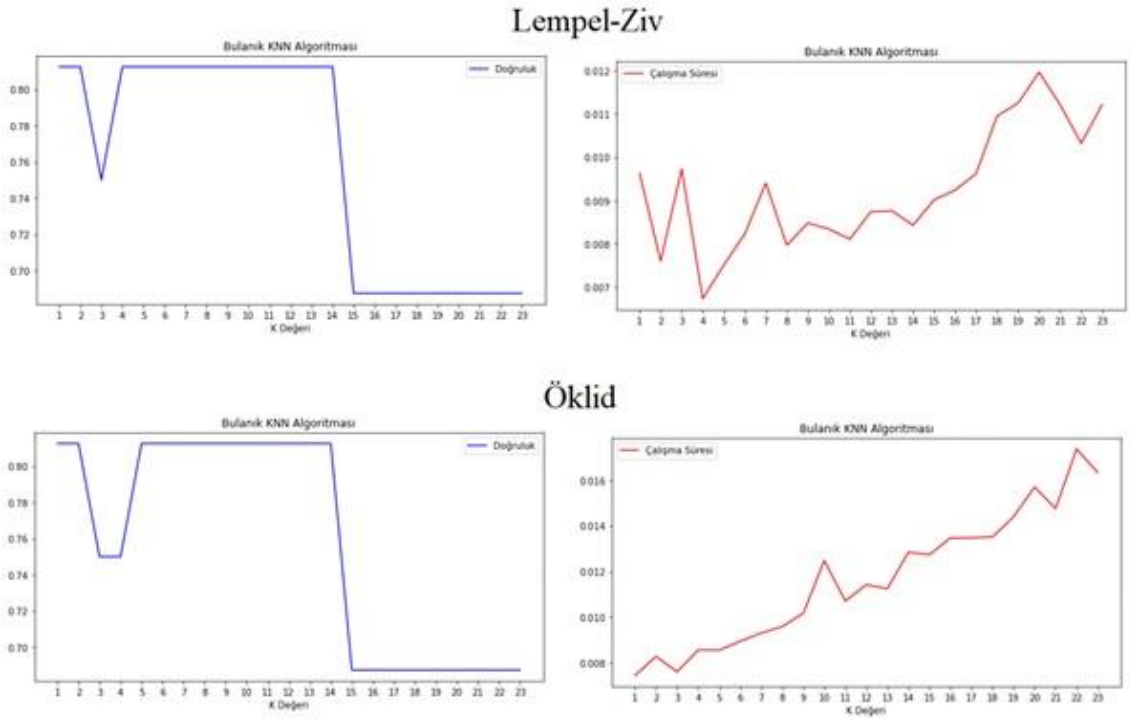
Şekil 6. Lempel-Ziv ile Öklid uzaklığı arasındaki fark 1 (Difference between Lempel-Ziv and Euclidean distance 1)



Şekil 7. Lempel-Ziv ile Öklid uzaklığı arasındaki fark 2 (Difference between Lempel-Ziv and Euclidean distance 2)



Şekil 8. Lempel-Ziv ile Öklid uzaklığı arasındaki fark 3 (Difference between Lempel-Ziv and Euclidean distance 3)



Şekil 9. Lempel-Ziv ile Öklid uzaklığı arasındaki fark 3 (Difference between Lempel-Ziv and Euclidean distance 3)

National Library of Medicine'den Homo Sapiens için 30 adet NGF, 30 adet BDNF ve 7 adet NT-3 sekansları FASTA formatında alındı. Çalışmamızda izoform olanları elenerek o sınıfa ait farklı veriler kullanıldı ancak veri sayısını yüksek tutmak açısından birkaç adet izoform da kullanıldı. 23 adet NGF, 21 adet BDNF ve 6 adet NT-3 şeklinde 3 farklı sınıf olarak toplamda 50 sekans algoritmamızda kullanıldı. Sonuçları Şekil 7'dedir. 28 adet NGF ve 12 adet BDNF şeklinde

2 farklı sınıf olarak toplamda 40 sekans algoritmamızda kullanıldı. Sonuçları Şekil 8'dedir. 13 adet NGF ve 7 adet NT-3 şeklinde 2 farklı sınıf olarak toplamda 20 veri algoritmamızda kullanıldı. Sonuçları Şekil 9'dadır.

Çalışmamızda ilk olarak yapısal olarak birbirine benzeyen iki nörotrofin olan NGF ve BDNF sekanslarının Lempel-Ziv karmaşıklıkları hesaplanmıştır. Lempel-Ziv karmaşıklığı, protein verilerinin orijinal yapısını korumaktadır. Bu karmaşıklıklara dayalı Lempel-Ziv uzaklık değerleri hesaplanmıştır. Çalışmamızda protein sekanslarının, bulunduğu konumdaki halini değiştirmeden ve herhangi bir bilgi kaybı olmadan kullanabilmek için Lempel-Ziv karmaşıklığına dayalı uzaklık değerleri kullanıldı. Uniprot veri tabanından alınan verilerle birlikte Bulanık K-En Yakın Komşu algoritmasında Lempel-Ziv uzaklığı kullanıldığında K komşu sayısının 12 olması karşılığında, sınıflandırma performansı %83 olarak elde edilmiştir. Öklid Uzaklığı kullanıldığında elde edilen en yüksek sınıflandırma performansı ise %75'tir. Maksimum doğruluk oranını elde ettiğimiz noktada Öklid uzaklığını kullandığımızda algoritmamızın çalışma süresi 0.0054 ms iken Lempel-Ziv uzaklığı kullandığımızda 0.0038 ms'dir.

National Library of Medicine'den alınan verilerde ise şu sonuçlar elde edilmiştir: 50 adet ve 3 farklı sınıf olarak (NGF, BDNF ve NT-3) verilerin, FKNN algoritmamızda kullanılması sonucunda Lempel-Ziv uzaklığı kullanıldığında çalışma süresinin 0.020 ms ile 0.035 ms arasında iken Öklid uzaklığı kullanıldığında çalışma süresinin 0.04 ms ile 0.08 ms arasında olduğu görülmektedir. 40 adet ve 2 farklı sınıf olarak (NGF ve BDNF) verilerin, FKNN algoritmamızda kullanılması sonucunda Lempel-Ziv uzaklığı kullanıldığında çalışma süresinin 0.007 ms ile 0.012 ms arasında iken Öklid uzaklığı kullanıldığında çalışma süresinin 0.008 ms ile 0.016 ms arasında olduğu görülmektedir. 20 adet ve 2 farklı sınıf olarak (NGF ve NT-3) verilerin, FKNN algoritmamızda kullanılması sonucunda Lempel-Ziv uzaklığı kullanıldığında çalışma süresinin 0.0024 ms ile 0.0036 ms arasında ilen Öklid uzaklığı kullanıldığında çalışma süresinin 0.0024 ms ile 0.0050 ms arasında olduğu görülmektedir.

Bulanık K-En Yakın Komşu (FKNN) algoritması, makine öğrenmesi ve veri madenciliği alanlarında sıklıkla kullanılan güçlü bir sınıflandırma yöntemidir. Protein sınıflandırmasında FKNN, özellikle amino asit dizilerinin ikincil yapılarını tahmin etmek amacıyla etkili bir şekilde uygulanmaktadır. Bu yöntem, verilerdeki belirsizlikleri ve düzensizlikleri daha iyi modelleyebilme kapasitesi sayesinde, geleneksel K-En Yakın Komşu (KNN) algoritmalarına kıyasla üstün performans sergileyebilir. FKNN algoritması, proteinlerin ikincil yapılarını sınıflandırırken, verinin bulanık doğasını dikkate alarak daha doğru ve güvenilir sonuçlar elde edilmesine olanak tanır. Literatürdeki birçok çalışma, FKNN'nin protein sınıflandırma problemlerinde başarıyla kullanıldığını ve sınıflandırma doğruluğunu artırdığını göstermektedir. Proteinlerin ikincil yapısının tahmin edilmesi için profiller ve Bulanık K-En Yakın Komşu algoritmasının kullanıldığı yeni bir yöntem, Bondugula ve arkadaşları tarafından önerilmiş ve pozisyon-ağırlıklı mutlak mesafe ölçüsü kullanılarak sınıflandırma doğruluğu %75.75 olarak elde edilmiştir [39]. Mirceva ve arkadaşları [40] tarafından yapılan çalışmada ise Bulanık K-En Yakın Komşuluk algoritmasında standart olarak Öklid mesafe ölçüsü kullanıldığında, protein sınıflandırma performansı en yüksek %79.97 olarak elde edilmiştir. Bu çalışmalar ışığında, çalışmamızda önerilen yaklaşımda ise, FKNN algoritmasının sınıflandırma performansını daha da yükseltecek yeni bir mesafe metriği olarak Lempel-Ziv Karmaşıklığı kullanılmıştır. Çalışmamızda önerilen algoritmadan elde edilen sonuçlara göre, Bulanık K-En Yakın Komşuluk Algoritması'nda mesafe metriği olarak Lempel-Ziv mesafesi kullanıldığında, Öklid mesafesine kıyasla daha hızlı bir çalışma süresi sağlandığı ve aynı zamanda sınıflama performansı olarak daha yüksek sonuçlar elde edildiği gözlemlenmiştir. Bu bağlamda, literatürde yapılan bu çalışmalara referans olabilecek şekilde, önerilen yaklaşımımızın yeni bir bakış açısı sunduğu ve daha yüksek sınıflama performansı sağladığı ortaya konmuştur.

Biyoenformatik, biyolojik verilerin analiz edilmesi ve yorumlanmasında kullanılan önemli bir alandır. Özellikle, büyük veri setlerinin işlenmesi gerektiğinde, veri işleme sürelerinin kısaltılması ve doğru sınıflandırma yapılabilmesi önemli bir zorluk oluşturur. Bu nedenle, verilerin büyük olmasından kaynaklanan zorlukları aşmak için geliştirilen çeşitli algoritmalar ve metrikler, bu alanda büyük önem taşımaktadır. Lempel-Ziv uzaklığı gibi yeni mesafe ölçütlerinin kullanılması, biyoenformatik çalışmalarda algoritma verimliliğini artırırken, aynı zamanda işlem sürelerini de optimize etmeye olanak tanımaktadır. Bu tür yöntemler, özellikle büyük ve karmaşık veri setlerinde, daha hızlı ve daha doğru analizler yapılmasını sağlayabilir. Önerilen yaklaşımımızın biyoenformatik çalışmalarına katkı sağlayacak potansiyeli bulunmaktadır. Çalışmamızın üç temel katkısı, önerilen yöntemin biyoenformatikteki mevcut tekniklerle nasıl uyumlu bir şekilde çalıştığını ve alanı nasıl ileriye taşıyabileceğini ortaya koymaktadır. Biyoenformatikte verilerin çok büyük olduğunu düşünecek olursak büyük verilerle çalıştığımızda süre daha önemli bir hale gelir. Sonuç olarak Lempel-Ziv

uzaklığının, biyoenformatik çalışmalarında ve uzaklık tabanlı sınıflandırma algoritmalarında kullanılması önerilmektedir.

Çalışmamızın üç ana katkısı bulunmaktadır. İlk olarak, biyoenformatik alanında yaygın olarak kullanılan makine öğrenimi tekniklerinden biri olan KNN algoritmasının bulanık versiyonu için Lempel-Ziv Karmaşıklığı'nın bir uzaklık metriği olarak kullanılması önerilmektedir. Bu yöntem, kayıpsız sıkıştırma ve farklı kalıpların çeşitliliğini ölçme açısından algoritmalarda önemli iyileştirmeler sağlamaktadır. İkinci olarak, algoritmamızda kullanılan veriler, yüksek yapısal benzerliklere sahip olmasının yanı sıra, farklı hücresel işlevlerde etkili ve paralel rolleri olan proteinlere ait olduğu için, bu verilerin sınıflandırılması, proteinlerin daha kolay analiz edilmesini ve anlamlı sonuçların elde edilmesini sağlayacaktır. Son olarak, bu çalışma, makine öğrenmesi yöntemlerinin nörotrofinlerin sınıflandırılmasında ilk kez uygulanmasıyla alana yenilikçi bir yaklaşım sunmaktadır.

KATKI ORANI BEYANI [STATEMENT OF CONTRIBUTION RATE]

Yazarların çalışmadaki katkı oranları eşittir.

ÇIKAR ÇATIŞMASI [CONFLICTS OF INTEREST]

Yazarlar arasında ve ilgili kurumları arasında herhangi çıkar çatışması olmadığını bildirmişlerdir.

ETİK KURALLARA UYGUNLUK [COMPLIANCE WITH ETHICAL RULES]

Yazarlar bu makalenin etik kurul onayı veya herhangi bir özel izin gerektirmediğini beyan ederler.

KAYNAKLAR [REFERENCES]

[1] "Protein structure", nature.com, 2014. [Online]. Available: https://www.nature.com/scitable/topicpage/protein-structure-14122136/. [Accessed: 6 June 2022].

[2] K. Ahern, I. Rahagopal, T. Tan, "2.3: Structure & fuction- proteins I", bio.libretext.org, Mar. 7, 2022. [Online]. Available: https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_For_All_(Ahern_Rajagopal_and_Tan)/02%3A_Structure_and_Function/203%3A_Structure__Function-_Proteins_I [Accessed: June. 6, 2022]

[3] J. Maillo, J. Luengo, S. Garcia, F. Herrera, I. Triguero, "Exact fuzzy k-nearest neighbor classification for big datasets", 2017 IEEE International Conference on Fuzzy Systems (FUZ-IEEE), July 09-12, 2017, Naples, Italy [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/document/8015686/authors#authors, [Accessed: 12 June 2022]

[4] James M. Keller, Michael R. Gray, James A. Givens, "A fuzzy k-nearest neighbor algorithm", IEEE Transactions on Systems, Man, and Cybernetics, vol: SMC-15, issue:4, pp. 580-585, July-Aug 1985, Doi: 10.1109/TSMC.1985.6313426. [Accessed: 15 June 2022]

[5] X. Zheng, C. Li, J. Wang, "An information-theoretic approach to the prediction of protein structural class", Journal of Computational Chemistry, vol. 31, issue 6, pp. 1201-1206, September 2009, Doi: 10.1002/jcc.21406. [Accessed: 28 June 2022]

[6] JY. Chang, JJ. Shyu, YX. Shi (2008). "Fuzzy k-nearest neighbor classifier to predict protein solvent accessibility" Ishikawa, M., Doya, K. Miyamoto, H., Yamakawa, T., Neural Information Processing. ICONIP 2007, vol 4985, pp. 837-845, Springer, Berlin, Heidelberg. [Online]. Doi: https://doi.org/10.1007/978-3-540-69162-4_87. [Accessed: 28 June 2022]

[7] Y. Huang, Y. Li, "Prediction of protein subcellular locations using fuzzy k-NN method", Bioinformatics, vol. 20, no. 1, pages. 21-8, 2004 Jan. [Online]. Doi: 10.1093/bioinformatics/btg366. [Accessed: 25 June 2022]

[8] R. Bondugula, O. Duzlevski, D. XU, "Profiles and fuzzy k-nearest neighbor algorithm for protein secondary structure prediction", Proceedings of the 3rd Asia-Pacific Bioinformatics Conference, pp. 85-94, January 2005, Singapore, [Online]. Doi: 10.1142/9781860947322_0009. [Accessed: 1 July 2022]

[9] M. Kumar, SK. Rath, "Microarray data classification using fuzzy k-nearest neighbor", International Conference on Contemporary Computing and Informatics (IC3I), Mysore, India, IEEE, pp. 1032-1038, November 2014, Doi: 10.1109/IC3I.2014.7019618 [Accessed: 1 July 2022]

[10] D. Li, JS, Deogun, K. Wang, "Gene function classification using fuzzy k-nearest neighbor approach", 2007 IEEE International Conference on Granular Computing (GRC 2007), Fremont, CA, USA, IEEE Xplore, pp. 644-644, November 2007, Doi: 10.1109/GrC.2007.99. [Accessed: 28 June 2022]

[11] J. Sim, SY. Kim, J. Lee, "Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method", Bioinformatics, vol. 21, issue 12, pages 2844-2849, April 2002 [Online] Doi: http://doi.org/10.1093/bioinformatics/bti423. [Accessed: 25 June 2022]

[12] HA. Abu Alfeilat, ABA. Hassanat, O. Lasassmed, AS. Tarawneh, MB. Alhasanat, HS. Eyal Salman, VBS. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review", Big Data, 7(4): 221-248, Dec 2019, Doi: 10.1089/big.2018.0175. Epub 2019 Aug 14. [Accessed: 8 July 2022]

[13] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, N. Kerdprasop, "An empirical study of distance metrics for k-nearest neighbor algorithm", Proceedings of the 3rd International Conference on Industrial Application Engineering 2015, Japan, pp. 280-285, Doi: 10.12792/iciae2015.051. [Accessed: 5 July 2022]

[14] P. Melin, E. Ramirez, G. Prado- Arechiga, "A new variant of fuzzy k-nearest neighbor using interval type-2 fuzzy logic", 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2018, pp. 1-7, Doi: 10.1109/FUZZ-IEEE.2018.8491472. [Accessed: 5 July 2022]

[15] PK. Jena, S. Chattopadhyay, "Comparative study of fuzzy k-nearest neighbor and fuzzy c-means algorithms", International Journal of Computer Applications, vol. 57, no. 7, pp. 22-32, November 2012, Doi: 10.1007/978-3-642-30157-5_45. [8 July 2022]

[16] F. Rosas, P. Mediano, "When and how to use Lempel-Ziv complexity", Information Dynamics, 26 June 2019, [Online]. Available: https://information-dynamics.github.io/complexity/information/2019/06/26/lempel-ziv.html. [Accessed: 10 July 2022]

[17] AW. Norman, HL. Henry, Hormones: Growth factors, Third Edition, Academic Press, 2015, pp. 363-379, Doi: 10.1016/B978-0-08-091906-5.000-3. [Accessed: 15 July 2022]

[18] AC. Mitchell, PS. Briquez, JA. Hubbell, JR. Cochran, "Engineering growth factors for regenerative medicine applications", Acta Biomater, Jan 2016, 1-12, Doi: 10.1016/j.actbio.2015.11.007. [Accessed: 16 July 2022]

[19] X. Ren, M. Zhao, B. Lash, MM. Martino, Z. Julier, "Growth factor engineering strategies for regenerative medicine applications", Frontiers in Bioengineering and Biotechnology journal, vol. 7, January 2020, Doi: 10.3389/fbioe.2019.00469. [Accessed: 16 July 2022]

[20] İB. Çitçi, DA. Jafari, B. Kosova, Sağlık Bilimleri Alanında Akademik Çalışmalar-II: Nörotrofin ailesi, Gece Kitaplığı, vol. 2, pp. 333-349, June 2020. [Accessed: 31 July 2022]

[21] S.Cohen, R. Levi-Montalcini, V. Hamburger, "A nerve growth-stimulating factor isolated from sarcom as 37 and 180", Proc Natl Acad Sci USA, 40 (10): 1014-1018, Oct 1954; Doi: 10.1073/pnas.40.10.1014. [Accessed: 10 August 2022]

[22] L. Aloe, "Rita Levi-Montalcini: the discovery of nerve growth factor and modern neurobiology", Trends in Cell Biology, vol. 14(7), pp. 395-9, Jul 2004, Doi: 10.1016/j.tcb.2004.05.011. [Accessed: 10 August 2022]

[23] M. Costandi, Nöroplastisite: Büyüme faktörleri ve hücre intiharı, Pan Yayıncılık, January 2019, pp. 49-51. [Accessed: 10 August 2022]

[24] U. Suter, C. Angst, CL. Tien, CC. Drinkwater, RM. Lindsay, EM. Shooter, "NGF/BDNF chimeric proteins: analysis of neurotrophin specificity by homolog-scanning mutagenesis", The Journal of Neuroscience: the official journal of the Society for Neuroscience, vol. 12, 1, pp. 306-318, Jan 1992, Doi: 10.1523/JNEUROSCI.12-01-00306.1992. [Accessed: 16 July 2022]

[25] J. Leibrock, F. Lottspeich, A. Hohn, M. Hofer, B. Hengerer, P. Masiakowski, H. Thoenen, YA. Barde, "Molecular cloning and expression of brain-derived neurotrophic factor", Nature, vol. 341, 6238, pp 149-152, Sep 1989, Doi: 10.1038/341149a0. [Accessed: 25 July 2022]

[26] J. Langhnoja, L. Buch, P. Pillai, "Potential role of NGF, BDNF, and their receptors in oligodendrocytes differentiation from neural stem cell: an in vitro study", Cell Biology International, vol. 45, issue 2, pp. 432-446, February 2021, Doi: 10.1002/cbin.11500. [Accessed: 25 July 2022]

[27] PC. Maisonpierre, L. Belluscio, B. Friedman, RF. Alderson, SJ. Wiegand, ME. Furth, RM. Lindsay, GD. Yancopoulos, "NT-3, BDNF, and NGF in the developing rat nervous system: Parallel as well as reciprocal patterns of expression", Neuron, vol. 5, issue 4, pp. 501-509, October 1990, Doi: 10.1016/0896-6273(90)90089-X. [Accessed: 25 July 2022]

[28] C. Zeeh, "The Lempel Ziv algorithm", uwaterloo.ca, January 16, 2003. [Online]. Available: https://ece.uwaterloo.ca/~ece611/LempelZiv.pdf. [Accessed: 10 July 2022]

[29] T. Weissman, "Chapter 1. Lempel-Ziv compression", web.stanford.edu, [Online]. Available: https://web.stanford.edu/class/ee376a/files/EE376C_lecture_LZ.pdf. [Accessed: 10 July 2022]

[30] E. Roberts, "Dictionary-based compressors", cs.stanford.edu, [Online]. Available: https://cs.stanford.edu/people/eroberts/courses/soco/projects/data-compression/lossless/lz78/index.htm. [Accessed: 10 July 2022]

[31] ST. Brink, "Lempel-Ziv compression", webdemo.inue.uni-stuttgart.de, [Online]. Available: https://webdemo.inue.uni-stuttgart.de/webdemos/03_theses/Lempel-Ziv-Compression/index.php?id=1. [Accessed: 10 July 2022]

[32] G. Sharma, "Analysis of Huffman Coding and Lempel-Ziv-Welch (LZW) coding as data compression techniques", International Journal of Scientific Research in Computer Science and Engineering, vol. 8, issue 1, pp. 37-44, Feb 2020. [Accessed: 12 July 2022]

[33] YT. Tan, BA. Rosdi, "FPGA-based hardware accelerator for the prediction of protein secondary class via fuzzy K-nearest neighbors with Lempel-Ziv complexity-based distance measure", Neurocomputing, vol. 148, pp. 409-419, January 2015, Doi: 10.1016/j.neucom.2014.06.001. [Accessed: 5 July 2022]

[34] HB. Shen, J. Yang, KC. Chou, "Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition", Journal of Theoretical Biology, 240, 9-13, June 2006, Doi: 10.1016/j.jtbi.2005.08.016. [Accessed: 20 July 2022]

[35] Z. Bian, CM. Vong, PK. Wong, S. Wang, "Fuzzy KNN method with adaptive nearest neighbors", IEEE Transactions on Cybernetics, vol. 52, no. 6, pp. 5380-5393, June 2022, Doi:10.1109/TCYB.2020.3031610. [Accessed: 20 July 2022]

[36] LA. Zadeh, "Fuzzy sets", Information and Control, vol. 8, issue 3, pp. 338-353, 1965, Doi: 10.1016/S0019-9958(65)90241-X. [Accessed 20 July 2022]

[37] KC. Chou, CT. Zhang, "Review: prediction of protein structural classes", Critical Reviews in Biochemistry and Molecular Biolog, 30, 275–349, 1995. [Accessed 25 November 2024]

[38] S. Sinharay, "Jackknife Methods", International Encyclopedia of Education (Third Edition), 229-231, 2010. [Accessed 25 November 2024]

[39] R. Bondugula, O. Duzlevski, D. Xu, "Profiles and fuzzy k-nearest neighbor algorithm for protein secondary structure prediction", In Proceedings of the 3rd Asia-Pacific Bioinformatics Conference pp. 85-94, 2005.

[40] G. Mirceva, A. Naumoski, A. Kulakov, "Classification of protein structures by using fuzzy KNN classifier and protein voxel-based descriptor", Mathematical Modeling, 2(3), 116-118, 2018.