Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 32, S. 262-267, Aralık 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 32, pp. 262-267, December 2021
Copyright © 2021 EJOSAT
Research Article

Derin Öğrenme Teknikleri Kullanarak İkili ve Çok Etiketli Sınıflandırma İle Enzimatik Fonksiyon Tahmini

Münevver Baran[1*], Mustafa Öztürk[2], Fatma Latifoğlu[3]

[1*] Erciyes Üniversitesi, Eczacılık Fakültesi, Temel Bilimler AD., Kayseri, Türkiye, (ORCID: 0000-0003-0369-1022), munevverbaran@erciyes.edu.tr
[2] Erciyes Üniversitesi, Sağlık Bilimleri Enstitüsü, İlaç Araştırma ve Geliştirme ve Uygulama AD., Kayseri, Türkiye, (ORCID: 0000-0001-9911-2499), eczmustafaozturk@gmail.com
[3] Erciyes Üniversitesi, Mühendislik Fakültesi, Biyomedikal Müh. Bölümü, Kayseri, Türkiye (ORCID: 0000-0003-2018-9616), flatifoglu@erciyes.edu.tr

(International Conference on Design, Research and Development (RDCONF) 2021 – 15-18 December 2021)

(**DOI:** 10.31590/ejosat.1041643)

ATIF/REFERENCE: Baran, M., Öztürk, M. & Latifoğlu, F. (2021). Derin Öğrenme Teknikleri Kullanarak İkili ve Çok Etiketli Sınıflandırma İle Enzimatik Fonksiyon. *Avrupa Bilim ve Teknoloji Dergisi*, (32), 262-267.

Öz

Biyolojik katilazör olarak görev yapan enzimler katalizlediği tepkime türüne ve mekanizmasına göre sınıflandırılırken her sınıf altında substrat seçiciliği durumlarına göre de alt sınıflar oluşturulmuştur. Aynı zamanda enzimlerin sınıflandırılmasında yapısal, kimyasal ve bağlantısallık özellikleri önemli olmaktadır. Enzim fonksiyonunu tahmini yeni enzimlerin tasarlamalarına yardımcı olmak ve enzimle ilişkili hastalıkları teşhisinde önemli olmaktadır. Enzimlerin önemli bir çoğunluğu belirli reaksiyonları gerçekleştiriken, sınırlı sayıda enzim farklı reaksiyonlar gerçekleştirebilmektedir. Bu nedenle birden fazla enzimatik fonksiyonla doğrudan ilişkilendirilebilmektedir. Gerçekleştirilen bu çalışmada enzimatik fonksiyonun ikili ve çok etiketli sınıflandırma ile tahmini amaçlanmıştır. Enzimlerin sınıflandırılmasında daha başarılı sonuçların kimyasal özelliklerin kullanılmasında ortaya çıktığı görülmüştür. Ancak tüm özelliklerin kullanılması durumunda sınıflandırma performansının daha da arttığı görülmüştür. Enzimatik fonksiyon tahmnine yönelik kullanılan modellerin başarısı incelendiğinde Derin Öğrenme modellerinin hem ikili hemde çok etiketli sınıflandırma performansının daha yüksek olduğu görülmüştür. Sonuç olarak önerilen modellerinin enzimatik fonksiyonların sınıflandırılmasında önemli bir araç olduğu ortaya konmuştur.

Anahtar Kelimeler: Enzimatik fonksiyon, Enzim Komisyonu numaraları, Derin Sinir Ağları, Makine Öğrenmesi.

Enzymatic Function Estimation with Binary and Multilabel Classification Using Deep Learning Techniques

Abstract

Enzymes that act as biological catalysts are classified according to the reaction type and mechanism they catalyze, while subclasses are formed under each class according to their substrate selectivity. At the same time, structural, chemical and connectivity features are important in the classification of enzymes. Predicting enzyme function is important in helping to design new enzymes and in diagnosing enzyme-related diseases. While a significant majority of enzymes carry out certain reactions, a limited number of enzymes can perform different reactions. Therefore, it can be directly associated with more than one enzymatic function. In this study, it was aimed to predict the enzymatic function by binary and multi-label classification. It has been observed that more successful results the use of chemical properties in have emerged in the classification of enzymes. However, it was observed that the classification performance increased even more when all features were used. When the success of the models used for enzymatic function estimation was examined, it was seen that the Deep Learning models had higher both binary and multi-label classification performance. As a result, it has been demonstrated that the proposed models are an important tool in the classification of enzymatic functions.

Keywords: Enzymatic function, Enzyme Commission numbers, Deep Neural Networks, Machine Learning.

[*] Sorumlu Yazar: munevverbaran@erciyes.edu.tr

1. Giriş

Enzimlerin karakterizasyon aktiviteleri, biyolojik ve biyomedikal uygulamaların ilerlemesinde önemli bir rol oynamaktadır. Biyolojik katalizörler olarak görev yapan enzimler, kimyasal reaksiyonları hızlandırılmasında etkin olup hücresel yaşamın şekillenmesine ve kontrol edilmesine önemli katkılarda bulunmaktadır (Garcia-Viloca vd., 2004). Örneğin kinazlar, sinyal iletimi ve hücre regülasyonu için vazgeçilmezdir ve ATPazlar, hücresel süreci bloke eden toksinleri, atıkları ve çözeltileri ihraç etmek için taşımayı aktive etmekle ilgilidir (Zhou vd., 2017; Feltcher ve Braunstein, 2012). Uluslararası Biyokimya Birliği Adlandırma Komitesi, enzimleri katalize ettikleri reaksiyonlara göre sınıflandırmaktadır. Her Enzim Komisyonu (EC) numarası dört haneli sayısal bir temsilden oluşmakta olup, ilk seviyede, sistem ana enzimatik sınıfları (yani EC1: oksidoredüktazlar, EC2: transferazlar, EC3: hidrolazlar, EC4: lyazlar, EC5: izomerazlar ve EC6: ligazlar) göstermektedir. Buna göre herhangi bir EC numarasındaki ilk hane, açıklamalı enzimin altı ana sınıftan hangisine ait olduğunu, ikinci hane alt sınıf sınıfını, üçüncü hane alt alt sınıf sınıfını ve dördüncü hane enzimin substratını temsil etmektedir (Feltcher ve Braunstein, 2012). Her bir sonraki basamak, önceki basamaklarla birleşerek bir enzimin işlevini daha spesifik olarak tanımlamaktadır.

Karakterize edilmemiş proteinlerin enzimatik fonksiyonlarının otomatik tahmini, laboratuvara dayalı işlevsel tanımlama prosedürlerinin hem yüksek maliyetleri hem de zaman alması nedeniyle biyoinformatik alanında önemli bir konudur. EC terminolojisinin hiyerarşik yapısı, tahmin çalışmaları için uygundur. Enzimleri sınıflandırmak için çeşitli yöntemler ve araçlar günümüzde kullanılmaktadır (Dalkiran vd., 2018; Che vd., 2016; Li vd., 2018; Lu vd., 2007). Bu yöntemler arasında tek işlevli enzim işlevini tahminine yönelik çalışmalar olmasına rağmen, tüm enzimlerin nispeten büyük bir bölümünü oluşturan çok işlevli enzim işlevinin tahmini üzerinde oldukça sınırlı sayıda çalışma bulunmaktadır (De Ferrari vd., 2012; Zou vd., 2013; Che vd., 2016; Zou ve Xiao, 2016; Amidi vd., 2017).

Ayrıca literatür çalışmalar incelendiğinde, yapısal ve bağlantısallık özelliklerine bağlı olarak derin öğrenme ve makine öğrenme algoritmaları kullanılarak sınıflandırma çalışmaları yapılmıştır (Amidi vd., 2017; Roy vd., 2012; Quester ve Schomburg, 2011; Shen ve Chou, 2007; Li vd., 2018; Zou vd., 2019).

Sınıflandırma yöntemlerini kullanarak enzimlerin fonksiyonlarını tahmin etmek için girdi örnekleri (yani proteinler), fiziksel, kimyasal ve biyolojik özelliklerini yansıtan nicel vektörler olarak temsil edilmelidir. Literatürde çeşitli tipte protein özellik gösterimleri önerilmiştir ve enzimatik fonksiyonların tahmini için kullanılan başlıca öngörüler homoloji, fizikokimyasal özellikler, amino asit dizisine dayalı özellikler ve yapısal özellikler olarak kategorize edilebilir (Dalkiran vd., 2018).

Enzimlerin sınıflandırılmasında temel olarak yapısal, kimyasal ve bağlantısallık özellikleri göz önünde bulundurulmaktadır.

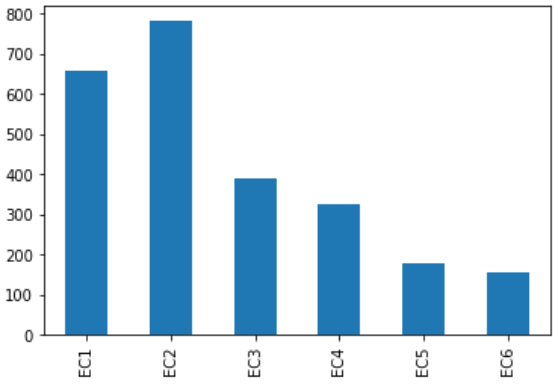
Gerçekleştirilen bu çalışmada enzimlerin yapısal, kimyasal ve bağlantısallık özellikleri ayrı ayrı ve birlikte kullanılarak tek işlevli ve çok işlevli tahmini gerçekleştirilmesi amaçlanmıştır. Bu amaç doğrultusunda derin sinir ağları ve makine öğrenimi yaklaşımı kullanılarak literatürde ilk defa çoklu analiz tekniklerinin kullanıldığı ve sınıflandırma işleminin gerçekleştirildiği bir çalışma ortaya konmuştur.

2. Materyal ve Metot

2.1. Veriseti ve Materyal

Bu çalışmada 2021 yılında halka açık olarak yayınlanan "Multi-label Classification of enzyme substrates" veriseti kullanılmıştır (https://www.kaggle.com/gopalns/ec-mixed-class). Bu veri seti içerisinde EC1-EC6 enzim tipi substrat içeren 1039 moleküle ait Kimyasal (196 adet), Yapısal (512 adet) ve Bağlantısallık (512 adet) özellikleri bulunmaktadır.

Hedef enzim sınıfı substratı olabilme özellikleri 6 sınıf halinde (EC1, EC2, EC3, EC4, EC5, EC6) Şekil1 de görüldüğü gibi çoklu etiket olarak verilmiştir.



Şekil 1 Sınıflara Ait Etiket Oranları

Sınıflandırma çalışması Google Colaboratory ortamında yapılmış olup Python 3.7.11, Tensorflow 2.6.0, Keras 2.6.0, NumPy 1.19.5, Pandas 1.1.5, scikit-learn 0.22.2.post1, Matplotlib 3.2.2 araçları kullanılmıştır (https://colab.research.google.com; https://github.com/fchollet/keras; Van Rossum ve Drake, 1995; Harris vd., 2020; McKinney, 2010; Hunter, 2007; Chen ve Guestrin, 2016).

2.2. Kullanılan Sınıflandırma Modelleri

Makine Öğrenmesi ve Derin Öğrenme'de kullanılan sınıflandırma yöntemleri 3 grupta toplanabilir.

- İkili (Binary): Hedef olarak sadece iki ihtimal mevcuttur. Negatif/Pozitif veya 0/1 gibi.
- Çoklu Sınıf (Multiclass): Hedef birden fazla sınıftan sadece bir tanesine aittir. Bir hastalığın derecesi olarak 0, 1, 2, 3 gibi.
- Çoklu Etiket (Multilabel): Hedef birden fazla sınıfa ait olabilir. Bir kişinin hem diyabet hem de tansiyon hastası olması gibi.

Bu çalışmada Enzimlerin ikili ve çoklu etiketli sınıflandırılmasına yönellik modeller geliştirilmiştir. Bu amaçla, Makine öğrenmesi yöntemlerinden XGBoost, AdaBoost, Rastgele Orman (Random Forest-RO), Karar Destek Makinesi (Support Vector Machine-KDM) ve Derin Öğrenme Yöntemlerinden Evrişimsel Sinir Ağları (Convolutional Neural Networks- ESA) ve Derin Sinir Ağları (Deep Neural Networks, DSA) mimarileri kullanılmıştır.

XGBoost, gradyan artırma yaklaşımı kullanan karar ağacı tabanlı bir topluluk Makine Öğrenimi algoritması olup küçük ve orta

ölçekli verilerde iyi performans sergilemektedir (Freund ve Schapire, 1997).

Adaboost algoritması, bir istatistiksel sınıflandırma meta-algoritması olup çeşitli zayıf algoritmaların bir araya getirilerek ya da karşılaştıtılarak oluşturulan bir modeldir (Breiman, 2001).

RO algoritması, doğruluğu ve bağımsızlığı en yüksek olan karar ağaçlarının birleşiminden oluşan bir modeldir. Veri setinin özelliklerine ve karar noktalarına göre dallanma gösterir (Jiang ve Zhou, 2004).

RO algoritması için max_depth parametresi 5 olarak alınmıştır.

KDM algoritması veri içerisindeki noktaları kendisine en yakın noktalar ile karşılaştırılarak kümeler oluşturan gözetimsiz bir algoritmadır (Angelova vd., 2015).

KDM algoritmasının çekirdek fonksiyonu varsayılan olan rbf (Radyal Tabanlı Fonksiyon) olarak seçilmiştir.

Derin öğrenme bir veya daha fazla gizli katman içeren yapay sinir ağları ve benzeri makine öğrenme algoritmalarını içermektedir. Derin sinir ağları ise yapay sinir ağlarının çok katmanlı halidir.

Evrişimsel sinir ağları son zamanlarda nesne tanıma, sınıflandırma ve modelleme gibi pek çok alanda uygulanabilen bir derin öğrenme yöntemidir (Baran vd., 2021).

Makine öğrenmesi modellerine ek olarak ekibimiz tarafından geliştirilen "Model 2" ve "Model B" olarak adlandırdığımız 2 adet DSA modeli ve "Model 0" olarak adlandırdığımız 1 adet 1 boyutlu ESA modeli de çalışmaya dahil edilmiştir.

Model 2, Model B ve Model 0'ın giriş ve çıkış katmanlarıdaki veri boyutları verisetine uygun olacak şekilde çalışma zamanı içinde değiştirilmiştir. Örnek olarak Kimyasal bilgileri içeren verisetinin multilabel sınıflaması için giriş ve çıkış katmanları boyutu sırayla (1,196) ve (6,) iken tüm bilgileri içeren verisetinin ikili sınıflaması için (1,1220) ve (2,) olarak değiştirilmiştir.

Model 2 mimarisi giriş, flatten, dense, droupout, dense, droupout, dense, dense katmanlarından oluşmaktadır.

Model B mimarisi giriş, flatten, dense, dense, dense katmanlarından oluşmaktadır.

Model 0 mimarisi dört adet paralel olarak gerçekleştirilen evrişimsel sinir ğları modellerinin birleşmesinden oluşmuştur Matthews, 1995).

Çapraz doğrulama için fold sayısı 5 olarak seçilmiştir. Her fold testi sırasında data içinden rastgele %20'lik bir kısım doğrulama (validation) için ayrılmıştır. DSA ve ESA modelleri için EarlyStopping kullanılmış ve en iyi model seçilerek ağırlıkları kaydedilmiştir.

2.3. Değerlendirme Kriteri

Bu çalışmada ikili ve çoklu sınıflandırma gerçekleştirilmek üzere makine öğrenmesi modelleri kurularak performansları analiz edilmiştir. Bu amaç doğrultusunda performans kriterleri olarak Matthews korelasyon katsayısı (Mkk) kullanılmıştır.

Bu çalışmada performans kriteri olarak, gerçek negatifleri de hesaba katması ve dengesiz dağılımda bile kendi içinde rastgele dağıtılmış veri için 0'a yakın skor vermesinden dolayı Mkk skoru tercih edilmiştir.

$$Mkk = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Burada TP gerçek pozitif (true positive), TN gerçek negatif (true negative), FP yanlış pozitif (false positive), FN yanlış negatifdir (false negative).

Makine öğrenmesi modellerinde multilabel sınıflama mümkün olmadığı için her hedef için sadece ikili sınıflama yapılmış ve ayrı ayrı Mkk skorları değerlendirme kriteri olarak alınmıştır.

Derin öğrenme modellerinde ise hem makine öğrenmesindeki gibi ikili sınıflama hem de multilabel sınıflama çalışılmıştır. Multilabel sınıflama sonucunda her sınıfa ait olan Mkk skoru hesaplanarak değerlendirme kriteri olarak alınmıştır.

Tablo 1 İkili Sınfılama Test Sonuçları

		Skor						Standart Sapma					
Veriseti	**Model**	**MCC-EC1**	**MCC-EC2**	**MCC-EC3**	**MCC-EC4**	**MCC-EC5**	**MCC-EC6**	**MCC-EC1**	**MCC-EC2**	**MCC-EC3**	**MCC-EC4**	**MCC-EC5**	**MCC-EC6**
Tümü Birleştirilmiş	**AdaBoost**	0,51	0,21	0,47	0,35	0,34	0,34	0,077	0,090	0,077	0,050	0,141	0,069
	DecisionTree	0,42	0,25	0,46	0,35	0,28	0,36	0,088	0,075	0,040	0,064	0,079	0,064
	Model_0	0,53	0,41	0,57	0,46	0,39	0,38	0,049	0,059	0,054	0,064	0,115	0,104
	Model_2	0,49	0,30	0,56	0,48	0,40	0,41	0,085	0,121	0,041	0,025	0,065	0,065
	Model_B	0,55	0,38	0,58	0,50	0,43	0,50	0,040	0,068	0,018	0,028	0,099	0,080
	RandomForest	0,45	0,14	0,50	0,32	0,27	0,11	0,053	0,075	0,046	0,099	0,046	0,127
	SVM	0,52	0,23	0,53	0,38	0,32	0,18	0,051	0,058	0,048	0,036	0,069	0,123
	XGBoost	0,53	0,25	0,54	0,43	0,35	0,31	0,040	0,022	0,050	0,065	0,054	0,054
Kimyasal	**AdaBoost**	0,41	0,18	0,42	0,33	0,30	0,30	0,044	0,046	0,054	0,139	0,086	0,048
	DecisionTree	0,40	0,28	0,40	0,34	0,31	0,24	0,029	0,074	0,043	0,093	0,076	0,050
	Model_0	0,50	0,11	0,52	0,46	0,36	0,36	0,056	0,093	0,043	0,070	0,058	0,126
	Model_2	0,52	0,27	0,54	0,46	0,37	0,36	0,023	0,172	0,052	0,055	0,078	0,051
	Model_B	0,47	0,28	0,52	0,37	0,36	0,39	0,064	0,066	0,073	0,120	0,083	0,094

	RandomForest	0,43	0,16	0,46	0,36	0,21	0,23	0,086	0,043	0,067	0,027	0,045	0,040
	SVM	0,39	0,18	0,43	0,32	0,10	0,00	0,030	0,076	0,049	0,014	0,105	0,034
	XGBoost	0,50	0,24	0,52	0,46	0,40	0,33	0,058	0,085	0,053	0,018	0,076	0,072
Bağlantısal	**AdaBoost**	0,20	0,11	0,25	0,12	0,13	0,24	0,057	0,032	0,074	0,013	0,123	0,071
	DecisionTree	0,22	0,13	0,13	0,11	0,17	0,17	0,089	0,053	0,077	0,051	0,087	0,049
	Model_0	0,25	0,22	0,31	0,15	0,29	0,35	0,070	0,076	0,058	0,116	0,081	0,102
	Model_2	0,31	0,18	0,35	0,26	0,25	0,27	0,057	0,066	0,081	0,059	0,033	0,060
	Model_B	0,35	0,19	0,31	0,23	0,31	0,26	0,023	0,060	0,038	0,057	0,112	0,082
	RandomForest	0,20	0,02	0,12	0,00	0,00	-0,01	0,079	0,071	0,034	0,000	0,000	0,012
	SVM	0,26	0,13	0,26	0,10	0,12	0,05	0,039	0,062	0,026	0,066	0,118	0,076
	XGBoost	0,21	0,11	0,32	0,10	0,20	0,12	0,087	0,046	0,099	0,097	0,104	0,054
Yapısal	**AdaBoost**	0,19	0,03	0,22	0,17	0,11	0,17	0,076	0,071	0,044	0,132	0,069	0,043
	DecisionTree	0,14	0,14	0,15	0,13	0,19	0,07	0,056	0,063	0,083	0,052	0,060	0,141
	Model_0	0,23	0,17	0,22	0,19	0,22	0,19	0,075	0,091	0,104	0,120	0,195	0,112
	Model_2	0,22	0,21	0,23	0,27	0,28	0,33	0,054	0,047	0,077	0,096	0,112	0,109
	Model_B	0,30	0,17	0,22	0,18	0,24	0,26	0,057	0,054	0,089	0,091	0,057	0,126
	RandomForest	0,27	-0,05	0,06	0,00	0,00	0,00	0,064	0,059	0,041	0,000	0,000	0,000
	SVM	0,22	0,08	0,24	0,03	0,09	0,05	0,096	0,054	0,066	0,061	0,067	0,087
	XGBoost	0,23	0,11	0,17	0,11	0,05	0,09	0,065	0,096	0,089	0,085	0,076	0,112

Tablo 2 Multilabel Sınıflama Test Sonuçları

		Skor						Standart Sapma					
Veriseti	**Model**	**MCC-EC1**	**MCC-EC2**	**MCC-EC3**	**MCC-EC4**	**MCC-EC5**	**MCC-EC6**	**MCC-EC1**	**MCC-EC2**	**MCC-EC3**	**MCC-EC4**	**MCC-EC5**	**MCC-EC6**
Tümü Birleştirilmiş	**Model_0**	0,36	0,09	0,44	0,27	0,13	0,18	0,101	0,078	0,088	0,055	0,096	0,169
	Model_2	0,44	0,15	0,38	0,29	0,11	0,12	0,084	0,070	0,086	0,175	0,151	0,124
	Model_B	0,40	0,09	0,37	0,31	0,16	0,21	0,225	0,082	0,210	0,181	0,136	0,132
Kimyasal	**Model_0**	0,36	0,08	0,36	0,19	0,12	-0,02	0,073	0,087	0,050	0,106	0,132	0,038
	Model_2	0,35	-0,02	0,34	0,32	0,03	0,05	0,075	0,046	0,068	0,101	0,040	0,154
	Model_B	0,31	-0,01	0,39	0,24	0,03	0,07	0,195	0,085	0,188	0,188	0,044	0,088
Bağlantısal	**Model_0**	0,13	0,01	0,08	0,03	0,03	0,01	0,100	0,060	0,108	0,091	0,064	0,026
	Model_2	0,24	0,12	0,13	0,05	0,04	0,06	0,059	0,069	0,137	0,072	0,079	0,137
	Model_B	0,21	0,14	0,20	0,05	0,00	0,00	0,029	0,029	0,104	0,103	0,000	0,000
Yapısal	**Model_0**	0,13	0,03	0,03	0,06	0,02	0,04	0,171	0,088	0,074	0,125	0,034	0,091
	Model_2	0,21	0,13	0,10	0,00	0,00	0,00	0,076	0,075	0,104	0,000	0,000	0,000
	Model_B	0,15	0,09	0,15	-0,02	0,03	0,00	0,070	0,069	0,062	0,026	0,060	0,010

3. Araştırma Sonuçları

Bu çalışmada EC1-EC6 enzim sınıflaması ikili (Tablo 1) ve çok etiketli (Tablo 2) olarak sadece kimyasal, yapısal özellikler, bağlantısallık özellikleri ve tüm özellikler birleştirilmiş veriseti kullanılarak gerçekleştirilmiştir.

Bu özelliklere göre sınıflandırma performansı incelendiğinde kimyasal özelliklerin sınıflandırma performansının yapısal ve bağlantısallık özelliklerine göre daha yüksek olduğu görülmektedir. Ancak tüm özelliklerin birlikte modellere giriş olarak uygulanması durumunda sınıflandırma performansının en yüksek olduğu görülmektedir. Bu nedenle enzim sınıflandırıklmasında tüm özelliklerin birlikte sınıflandırıcılara giriş olarak uygulanmasının daha doğru olacağı görülmektedir.

Tüm modellerin ikili sınıflandırma performansı incelendiğinde tüm özellikler kullanılarak elde edilen Mkk skoru

Model B den elde edilen sınıflandırma sonucuna göre EC1 için 0.55, EC2 için 0.38, EC3 için 0.58, EC4 için 0.50, EC5 için 0.43 ve EC6 için 0.50 olarak elde edilmiştir. Diğer modellerin ortalama Mkk skorunun Model B ye göre daha düşük olduğu görülmektedir.

Çok etiketli sınıflandırma performansı değerlendirildiğinde Model 2'nin performansının EC1 için 0.44, EC2 için 0.15, EC3 için 0.38, EC4 için 0.29, EC5 için 0.11 ve EC6 için 0.12 Mkk skoru ile Model 0 Model B ye göre daha yüksek olduğu görülmektedir.

4. Tartışma

Enzim fonksiyonlarını tahmininden proteinin yapısı ve işlevi arasında önemli bir bağlantı olması sebebiyle, enzimlerin sınıflandırılması son zamanların en popüler konularından biri haline gelmiştir.

Bir enzim ya çoklu katalitik bölgelerin varlığından ya da tek bir bölgenin düzenlenmesinden dolayı farklı reaksiyonlar gerçekleştirebilir. Bu nedenle çok etiketli sınıflandırma da ön plana çıkmaktadır.

Enzimlerin sınıflandırılmasında kimyasal özelliklerin yapısal ve bağlantısallık özelliklerine göre daha önemli olduğu ve sınıflandırma performansını tüm özellikler kullanılarak daha da artırıldığı görülmüştür.

Ayrıca Derin öğrenme yöntemlerinin AdaBoost, XGBoost, RO ve KDM modellerine göre daha iyi olduğu görülmüştür. Önerdiğimiz modeller kullanılarak elde edilen çok etiketli sınıflandırma performasının ikili sınıflandırmaya göre daha düşük olduğu görülmektedir. Ancak ikili sınıflandırma modelleri ile birleştirme (ensemble) yaklaşımı kullanılarak çok etiketli sınıflandırma probleminin performansınıda artırabileceği görülmektedir.

İleriki çalışmalarda Auto encoder yöntemi ile istenen enzim grubuna ait olabilecek substratın belirlenmesi ile uygun ilaç molekülünde olması gereken özellikler ortaya konulabieceği düşünülmektedir.

5. Teşekkür

Verisetini yayınlanlayan araştırmacılara teşekkür ederiz.

Kaynakça

Amidi, S., Amidi, A., Vlachakis, D., Paragios, N., & Zacharaki, E. I. (2017). Automatic single-and multi-label enzymatic function prediction by machine learning. PeerJ, 5, e3095.

Angelova, A., Krizhevsky, A., & Vanhoucke, V. (2015, May). Pedestrian detection with a large-field-of-view deep network. In 2015 IEEE international conference on robotics and automation (ICRA) (pp. 704-711). IEEE.

Baran M, Öztürk M, Latifoğlu F. (2021). Gaita mikrobiyotasının hastalıklarla ilişkisinde öğrenmemodellerinin karşılaştırılması. MAS International European Conference on Mathematics-Engineering-Natural&Medical Sciences-XV. September 2021 ADANA, 7-8.

Breiman, L. (2001). Random forest. Mach. Learn, 45: 5–32.

Che, Y., Ju, Y., Xuan, P., Long, R., & Xing, F. (2016). Identification of multi-functional enzyme with multi-label classifier. PloS one, 11(4), e0153503.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Dalkiran, A., Rifaioglu, A. S., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2018). ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC bioinformatics, 19(1), 1-13

De Ferrari, L., Aitken, S., van Hemert, J., & Goryanin, I. (2012). EnzML: multi-label prediction of enzyme classes using InterPro signatures. BMC bioinformatics, 13(1), 1-12.

Feltcher, M. E., & Braunstein, M. (2012). Emerging themes in SecA2-mediated protein export. Nature Reviews Microbiology, 10(11), 779-789.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Garcia-Viloca, M., Gao, J., Karplus, M., & Truhlar, D. G. (2004). How enzymes work: analysis by modern rate theory and computer simulations. Science, 303(5655), 186-195.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

https://www.kaggle.com/gopalns/ec-mixed-class.

https://colab.research.google.com.

https://github.com/fchollet/keras.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Jiang, Y., & Zhou, Z. H. (2004, August). Editing training data for kNN classifiers with neural network ensemble. In International symposium on neural networks (pp. 356-361). Springer, Berlin, Heidelberg.

Li, Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L., & Gao, X. (2018). DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics, 34(5), 760-769.

Lu, L., Qian, Z., Cai, Y. D., & Li, Y. (2007). ECS: an automatic enzyme classifier based on functional domain composition. Computational biology and chemistry, 31(3), 226-232.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451.

McKinney, W. (2010, June). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, 445, pp. 51-56.

Quester, S., & Schomburg, D. (2011). EnzymeDetector: an integrated enzyme function prediction tool and database. BMC bioinformatics, 12(1), 1-13.

Roy, A., Yang, J., & Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic acids research, 40(W1), W471-W477.

Shen, H. B., & Chou, K. C. (2007). EzyPred: a top–down approach for predicting enzyme functional classes and subclasses. Biochemical and biophysical research communications, 364(1), 53-59.

Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Zhou, X. X., Fan, L. Z., Li, P., Shen, K., & Lin, M. Z. (2017). Optical control of cell signaling by single-chain photoswitchable kinases. Science, 355(6327), 836-842.

Zou, Q., Chen, W., Huang, Y., Liu, X., & Jiang, Y. (2013). Identifying multi-functional enzyme by hierarchical multi-label classifier. Journal of Computational and Theoretical Nanoscience, 10(4), 1038-1043.

Zou, H. L., & Xiao, X. (2016). Classifying multifunctional enzymes by incorporating three different models into Chou's general pseudo amino acid composition. The Journal of membrane biology, 249(4), 551-557.

Zou, Z., Tian, S., Gao, X., & Li, Y.(2019). mldeepre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. Frontiers in genetics, 9, 714.

.