Makale Türü
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(2) (2022) 106-118
DOI: 10.38016/jista.1054331

Ozon Konsantrasyonlarını Modellemek için Makine Öğrenmesi ve Derin Öğrenme Yöntemlerinin Karşılaştırılması

Şevket Ay[1]*, Ekin Ekinci[1],

[1]Sakarya Uygulamalı Bilimler Üniversitesi, Bilgisayar Mühendisliği Bölümü, Sakarya, Türkiye

sevketay09@gmail.com, ekinekinci@subu.edu.tr

Öz

Hava kirliliği günümüz için önemli bir problem olmakla birlikte sanayileşme, orman yangınları, egzoz gazları, kalitesiz yakıt kullanımı gibi sebepler gelecek nesilleri de tehdit edecek ciddi bir problem ile bizleri yüzleştirmektedir. Bu sebepler içerisinde ise yoğun sanayileşme hava kirliliğinde rol oynayan en önemli faktörlerden birisidir. Bölgesel sanayi gelişimi şehirlerde hava kalitesini etkilemektedir. Sanayinin gelişmesi ile birlikte bazı kirleticilerin miktarı azalmakta iken, ozon seviyelerinde artış yaşanmaktadır. Önümüzdeki yıllarda hava kirliliğini neden olacağı problemleri daha fazla hissetmemek, hava kalitesini yönetmek ve risklere karşı önlem almak için hava kirliliğinin tahmin edilmesi kaçınılmaz hale gelmektedir. Bu çalışmada sanayinin gelişmiş olduğu Kocaeli ve Sakarya illeri ile sanayinin çok fazla gelişmediği Çanakkale illeri için 2018-2021 arası saatlik ozon seviyelerini tahmin etmek amacıyla zaman serilerine dayalı makine öğrenmesi ve derin öğrenme yöntemleri uygulanmıştır. Uygulanan modeller Ortalama Mutlak Hata (MAE), Bağıl Mutlak Hata (RAE) ve R-kare (R^2) metrikleri kullanılarak karşılaştırılmış ve en etkin yöntemin belirlenmesi amaçlanmıştır.

Anahtar kelimeler: Makine Öğrenmesi, Derin Öğrenme, Aşırı Gradyan Arttırma (Xgboost), Yapay Sinir Ağları (YSA), Uzun Kısa Süreli Bellek (LSTM), Zaman Serileri

Comparison of Machine Learning and Deep Learning Methods for Modeling Ozone Concentrations

Abstract

Although air pollution is an important problem for today, reasons such as industrialization, forest fires, exhaust gases, poor quality fuel use confront us with a serious problem that will threaten future generations. Among these reasons, intensive industrialization is one of the most critical factors that play a role in air pollution. Regional industrial development affects air quality in cities. While the amount of some pollutants decreases with the development of the industry, there is an increase in ozone levels. In the coming years, it becomes inevitable to predict air pollution in order not to feel the problems that air pollution will cause more, to manage air quality, and to take precautions against risks. In this study, machine learning and deep learning methods based on time series were applied to predict hourly ozone levels between 2018 and 2021 for the provinces of Kocaeli and Sakarya, where the industry is developed, and Çanakkale, where the industry is not developed much. The applied models were compared using Mean Absolute Error (MAE), Relative Absolute Error (RAE), and R-square (R^2) metrics, and it was aimed to determine the most effective method.

Keywords: Machine learning, Deep learning, Extreme Gradient Boosting (Xgboost), Artificial Neural Network (ANN), Long-Short Term Memory (LSTM), Time Series

1. Giriş (Introduction)

Soluduğumuz hava kalitesinin sağlığımıza doğrudan etkisi vardır. Normal olarak havanın %78.084'ü Azot (N_2), % 20.946'sı Oksijen (O_2), %0.934'ü Argon (Ar), %0.035'i Karbondioksitten (CO_2) oluşturmaktadır. Günümüzün önemli tehditlerinden birisi olan hava kirliliği geçmişten günümüze çevresel değişiklikler, endüstriyel kirlilik, fosil yakıtların kontrolsüz tüketimi, kente göç vb. nedenlerden ötürü ortaya çıkmaktadır.

* Sorumlu yazar.
E-posta adresi: ekinekinci@subu.edu.tr

Alındı : 6 Ocak 2022
Revizyon : 27 Şubat 2022
Kabul : 20 Mart 2022

Hava kirliliği tehlikeli boyutlara ulaşırken hava kirliliği ile mücadele etmek elzem olmaktadır. Bu amaçla sürekli ölçümler yapılmaktadır. Kriter olarak ölçülmesi gereken kirleticiler ise, Karbon monoksit (CO), Kükürt dioksit (SO_2), Ozon (O_3), Partikül madde (PM), Azot oksitler (NOX) olarak belirtilmektedir.

Günümüzde yapay zekada yaşananan gelişmeler sadece bilgisayar bilimcileri değil diğer bilim dallarında çalışan araştırmacıların da ilgisini çekmeye başlamıştır. Hava kirliliği tahmininde yapay zekanın kullanılması literatürde önemli bir yer tutmaya başlamıştır.

Literatürde, O_3 konsantrasyonların modellenmesi için makine öğrenmesi ve derin öğrenme tabanlı farklı çalışmalar bulunmaktadır. Makine öğrenmesi doğrusal olmayan ve yüksek boyutlu veri kümeleri üzerinden kararlı ve performansı yüksek bilgi çıkarımı yapmaktadır (Bilgin, 2021; Yıldırım vd., 2021). Makine öğrenmesi yöntemlerinden çok katmanlı algılayıcı (ÇKA) (Paoli vd., 2011; Chattopadhyay vd., 2019; Yang vd., 2021; Bekesiene vd., 2021; Makarova vd., 2021), destek vektör makineleri (DVM) (Chelani, 2010; Tanaskuli, 2019; Mehdipour ve Memarianfard; 2019), lineer regresyon (Alipio, 2020; Allu vd., 2020; Matasović vd., 2021), Xgboost (Ding vd., 2020; Liu vd., 2020), rastgele orman (Liu vd., 2020; Ma vd., 2021) yöntemleri ile yapılmış çalışmalar mevcuttur.

Büyük veri analizi ve Grafik İşleme Biriminin (GPU) kullanılmasından bu yana, derin öğrenme büyük ilgi görmekte ve makine öğrenmesinin uygulandığı her alana uygulanmaktadır (Çağıl ve Yıldırım, 2020; Darendeli ve Yılmaz, 2021). Derin öğrenme yöntemleri ile yapılan çalışmalarda derin sinir ağları (DSA) (Wang vd., 2020; Felix vd., 2021), oto kodlayıcı (Nghiem vd., 2021), özyinelemeli sinir ağları (Adnane vd., 2021), Uzun Kısa Süreli Bellek (LSTM) (Alghieth vd., 2021; Ekinci vd. 2021; Zhang vd., 2021), konvolüsyonel sinir ağları (CNN) (Eslami vd., 2020; Sayeed vd., 2021) kullanılmıştır.

Bu çalışmanın amacı saatlik O_3 konsantrasyonlarını modellemede makine öğrenmesi ve derin öğrenme yaklaşımlarını etkinliğini değerlendirmektir. Bu amaçla kirliliğe sebep olan parametrelerden (PM10, SO_2, NO, NO_2 ve O_3) oluşan zaman serisi veri kümesi kullanılarak Xgboost, YSA ve Uzun Kısa Süreli Bellek (LSTM) yöntemleri karşılaştırılmıştır. Yapılan deneyler sonucunda ozon seviyesini tahmin etmede LSTM yönteminin diğer iki makine öğrenmesi yöntemine kıyasla daha başarılı olduğu gözlemlenmiştir.

Makalenin geri kalan kısmı ikinci bölümde veri kümesi ve uygulanan yöntemlerin anlatıldığı materyal ve yöntemler kısmıdır. Üçüncü bölümde veri önişleme, kullanılan hata metrikleri, modellerin tasarımı ve elde edilen deneysel sonuçlar ayrıntılı şekilde verilmiştir. Son bölümde ise sonuçlar ve öneriler yer almaktadır.

2. Materyal ve Yöntemler (Materials and Methods)

2.1. Veri kümesi (Dataset)

Marmara bölgesinde özellikle Kocaeli ve Sakarya illerinin sanayilerinin gelişmesi ile birlikte bu illerde hava kirliliği oldukça yüksektir. Bu çalışmada, Kocaeli ve Sakarya ile birlikte sanayinin yoğun olmadığı Çanakkale illeri için T.C. Çevre ve Şehircilik Bakanlığı Hava Kalitesi İzleme Ağı'ndan[1] sürekli ölçümler yapılarak elde edilen verilerden oluşan bir veri kümesi oluşturulmuştur.

İstasyonda ölçülen meteorolojik parametreler 10 μm'nin altındaki parçacıkları ifade eden PM10, azot oksitlerden NO, NO_2, NOX, SO_2 ve O_3 şeklindedir. Bu parametreler içerisinden O_3 konsantrasyonunu tahmin etmek için PM10, SO_2, NO, NO_2 ve O_3 parametrelerine dayalı bir tahmin yapılması hedeflenmiştir. Kocaeli, Sakarya ve Çanakkale illeri için 2018 Kasım ile 2021 Kasım arası saatlik ölçülen zaman serisi değerleri kullanılmıştır. 4 saatlik bir pencere boyutu ile (yani 4 zaman noktası) 5. saat için O_3 konsantrasyonlarının tahmini gerçekleştirilmiştir.

2.2. Yöntemler (Methods)

2.2.1. Xgboost (Xgboost)

Xgboost (Chen vd., 2016) karar ağacı temelli topluluk öğrenimi algoritmasıdır. Algoritmanın çalışma mantığı, değişkenlere farklı ağırlıklar vererek elde edilen ağaç topluluğundan çıkarımlar yapmaktır. İlk etapta tüm değişkenler eşit ağırlığa sahiptir. Ağaç topluluğu büyümeye başladıkça, problem bilgisine bağlı olarak ağırlıklar düzenlenmektedir. Yanlış sınıflandırılan gözlemlerin ağırlığı yükseltilirken, doğru sınıflandırılan gözlemlerin ağırlığı düşürülmektedir. Bu sayede ağaçlar zor durumlar karşısında kendini düzenleyebilme yeteneği kazanmaktadır. Fazla uyumu azaltan ve genel performansı artıran çeşitli düzenlemeler içermektedir (Ekinci vd., 2020). Bu özelliğinden dolayı "düzenli artırma" tekniği olarak da isimlendirilmektedir. Xgboost algoritması çeşitli düzenlemeler ile doğruluğu arttıran, paralel işleme ile hızlı sonuçlar verebilen, eksik değerlerin kullanımı için standart bir yapıya sahip olan, yükseltme işleminin yineleme aşamalarının her birinde çapraz doğrulama yapan bir algoritmadır.

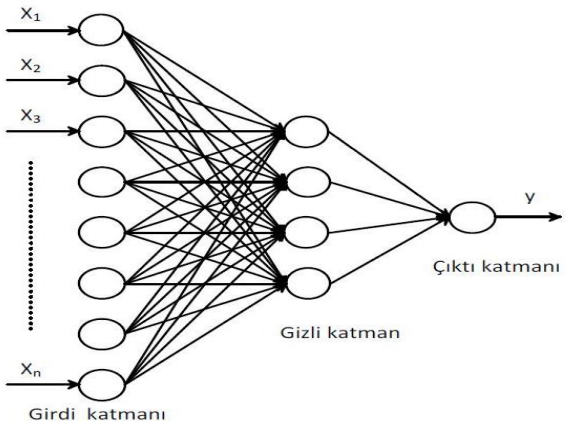
2.2.2. Yapay Sinir Ağları (Artificial Neural Networks)

Beyin insan vücudunun yapı taşı olup girdileri sinyal şeklinde alan, işleyen ve çıkış sinyallerini gönderen biyolojik sinir ağıdır. Beynin temel birimi nörondur. Beyin 200 milyar nörondan oluşmaktadır. Nöronlar dentrit, soma, akson ve sinapsis olmak üzere dört temel kısımdan oluşmaktadır. Nöron, dentritlerden sinyal

[1] http://sim.csb.gov.tr/Services/AirQuality/

toplamaktadır, soma hücreleri ise bu sinyallerin tümünü toplamaktadır ve toplam eşik değerine ulaştığında sinyal aksondan diğer nöronlara geçmektedir. Sinapslar ağırlıkları temsil etmektedir. İşte en sık tercih edilen yöntem olan YSA bu biyolojik sinir ağını taklit etmektedir ve bilgileri ağırlıklarda saklamaktadır (Garip vd., 2016; Şen, 2018). YSA'nın yapısı Şekil 1'de verilmektedir.

YSA doğrusal olmama, genelleme yapabilme, çok sayıda değişken ve parametre kullanabilme özelliklerine sahiptir. Her katman beynimizin nöronlarını taklit eden düğümlerden oluşmaktadır. Giriş katmanı, sinir ağının işleyebileceği bilgileri girdi olarak alan katmandır. Her düğüm bir özelliği yani bilgi parçasını temsil etmektedir.



Şekil 1. YSA Mimarisi (ANN Architecture)

Giriş katmanındaki her bir düğüm bir sonraki katmanda bulunan düğüme bağlanmaktadır. Ara katmanlar bir diğer deyişle gizli katmanlar giriş katmanından gelen bilgileri işlemektedir ve çıkış katmanına göndermektedir. Çıkış katmanı ise ağın son ara katmanındaki bilgileri bir araya getirmektedir ve bu şekilde gerekli tüm bilgileri çıkarmakta ve dış dünyaya göndermektedir.

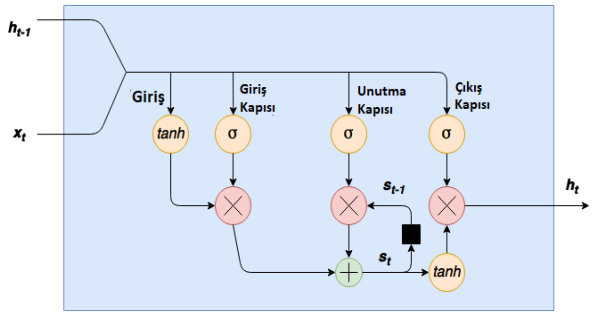
2.2.3. Uzun Kısa Süreli Bellek (LSTM)

Tekrarlı Yapay Sinir Ağları (RNN) aldıkları girdiyle ilgili önemli bilgileri hatırlayabilen ve bir sonraki adımın ne olacağını tahmin etmede çok hassas olan bir sinir ağıdır. Zaman serisi, konuşma, metin, finansal veri, ses, video, hava durumu ve daha fazlası gibi sıralı veriler için tercih edilen algoritma olmasının nedeni bu özelliğidir. RNN sadece çıktıları beslemez aynı zamanda kendi kendine geri besleme sağlamaktadır. Çünkü RNN'in dahili belleği vardır. LSTM ise RNN'in özelleşmiş bir yapısıdır (Sepp vd., 1997). Standart ileri beslemeli sinir ağlarının aksine, LSTM'in geri bildirim bağlantıları vardır. Yalnızca tek veri noktalarını değil aynı zamanda tüm veri dizilerini de işleyebilmektedir. RNN'den en önemli farkı uzun bir hafızası olmasıdır. RNN yakın geçmişi hafızasında saklayabilirken LSTM'de ise uzun bir hafıza vardır. Örneğin, LSTM ayrılmamış, bağlı el yazısı tanıma (Liwicki vd., 2009), konuşma tanıma (Sak vd., 2014) ve ağ trafiğinde veya izinsiz giriş tespit sistemlerinde (IDS) anormal durumları algılama gibi görevlere uygulanabilmektedir.

Ortak bir LSTM birimi bir hücreden, bir giriş geçidinden, bir çıkış geçidinden ve bir unutma kapısından oluşmaktadır. Hücre, keyfi zaman aralıkları boyunca değerleri hatırlamaktadır ve bu üç kapı, hücrenin içine ve dışına bilgi akışını düzenlemektedir.

Giriş kapısı, yeni bir değerin hücreye ne kadar aktığını kontrol etmektedir; unutma kapısı bir değerin hücrede ne kadar kalacağını kontrol etmektedir ve çıkış kapısı, hücredeki değerin LSTM ünitesinin çıkış aktivasyonunu hesaplamak için kullanılma derecesini kontrol etmektedir. LSTM kapılarının aktivasyon fonksiyonu genellikle lojistik sigmoid fonksiyondur. LSTM kapılarının bağlantıları vardır. Eğitim sırasında öğrenilmesi gereken bu bağlantıların ağırlıkları kapıların nasıl çalıştığını belirler.

Sigmoid katmanı, her bir bileşenden ne kadarının geçmesi gerektiğini tanımlayan sıfır ile bir arasında rakamlar vermektedir. Sıfır değeri geçiş izni yok demek iken, bir değeri geçişe izin var demektir. Aşağıdaki denklemlerde ifade edilen değişkenler vektörleri temsil etmektedir. LSTM mimarisi Şekil 2'de verilmiştir.



Şekil 2. LSTM Mimarisi (LSTM Architecture)

Şekil 2'de σ sigmoid katmanını, b_g girdi bias değerini, U_g girdi için ağırlık değerini, V_g önceki hücre çıktısı için ağırlık değerini, tanh ise aktivasyon fonksiyonunu temsil etmektedir.

$$g = \tanh\left(x_t U_g + h_{t-1} V_g + b_g\right) \quad (1)$$

$$i = \sigma(x_t U_i + h_{t-1} V_i + b_i) \quad (2)$$

$$f = \sigma(x_t U_f + h_{t-1} V_f + b_f) \quad (3)$$

$$s_t = f \circ s_{t-1} + g \circ i \quad (4)$$

$b_i$girdi kapısı için bias değerini, U_i girdi kapısı için ağırlık değerini, V_i önceki hücrenin çıktısının ağırlığını, $g \circ i$ girdi bölümünün çıktısını ifade etmektedir. $b_f$unutma kapısı için bias değerini U_f unutma kapısı için ağırlık değerini, V_f önceki hücrenin çıktısının ağırlığını temsil etmektedir. b_oçıktı kapısı için bias değerini, U_o çıktı kapısı için ağırlık değerini, V_o önceki hücrenin çıktısını ve h_t çıkışı ifade etmektedir.

$$h_t = \sigma(x_t U_o + h_{t-1} V_o + b_o) \circ \tanh(s_t) \quad (5)$$

3. Deneysel Çalışma (Experimental Study)

3.1. Veri Önişleme (Data Pre-processing)

Girdi olarak kullanılan değişkenler arasındaki değer farkının önemli ölçüde fazla olması modelde yanlılığa sebep olabilmektedir ve model verileri genelleştirememektedir. Bu sebeple eksik öğrenme durumu oluşabilmektedir. Çözüm olarak verilerin normalize edilmesi tavsiye edilmektedir. Normalleştirme yöntemlerinden Min-Maks normalizasyonu kullanılmıştır. Min-Maks normalizasyonu değişkenlerin değerini belirlenen bir aralığa dönüşmektedir. Bu çalışma kapsamında veri [0,1] aralığına çekilmiştir. Böylece herhangi bir bilgi kaybı yaşamadan modelin her özniteliğe eşit şekilde yaklaşması sağlanmıştır. Min-Maks normalizasyon formülü aşağıdaki gibidir (8):

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (6)$$

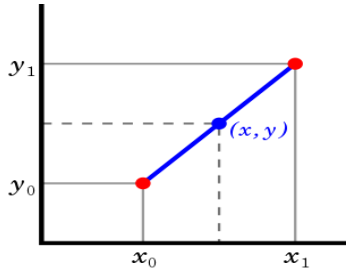
x_i değişkenin orijinal değeri, x_{min} değişkenin aldığı minimum değeri, x_{max} değişkenin aldığı maksimum değeri, y_i ise değişkenin normalize edildikten sonraki değerini temsil etmektedir.

Veriler üç il için eğitim ve test kümelerine bölünmüştür. Eldeki veri kümesinin %90'ı eğitim için %10'u ise test için kullanılmıştır. Tablo 1'de her il için eğitim ve test kümesinde bulunan örnek sayısı verilmiştir.

Tablo 1. Veri Dağılımı (Data Distribution)

Şehir 1	Eğitim Kümesi	Test Kümesi
Kocaeli	21500	4079
Sakarya	21500	4079
Çanakkale	21500	4079

Zaman serisi verilerinde eksik değerleri ortalama, ortanca vb. yöntemler ile doldurmak tehlikeli olabilmektedir. Veri seti incelendiğinde Kocaeli için %5, Sakarya için %4 ve Çanakkale için %5 oranında eksik değer bulunmaktadır. Veri adedi göz önüne alındığında bu oranlar düşük sayılabilir ve doldurulduğu zaman verilerin dağılımı bozulmayabilir. Enterpolasyon yöntemi komşular yardımıyla eksik değerleri doldurmaya yaramaktadır. Zaman serisi verilerinde eksik değerleri doldurmak için enterpolasyon yöntemi tercih edilmektedir. Çeşitli enterpolasyon metotları bulunmaktadır. Bunlar doğrusal, zamansal, polynomial gibi yöntemlerdir. Bu çalışmada doğrusal enterpolasyon kullanılmıştır. Doğrusal enterpolasyon, en yakın tanımlanmış iki veri noktası arasında eksik değerleri doğrusal olarak aralıklı değerlerle değiştirir. Şekil 3'te doğrusal enterpolasyon grafiği gösterilmiştir.



Şekil 3. Doğrusal Enterpolasyon (Linear Interpolation)

$$\frac{y-y_0}{x-x_0} = \frac{y_1-y_0}{x_1-x_0} \quad (7)$$

Formülde (x_0, y_0) ve (x_1, y_1) koordinat düzleminde bilinen iki noktayı temsil etmektedir. (x_0, x_1) aralığındaki x değeri için, y enterpolasyon sonucunda oluşan değerdir.

3.2 Hata Metrikleri (Error Metrics)

Çalışmada modellerin performansını değerlendirmek için MAE, R^2 ve RAE kullanılmıştır. MAE, tahmin edilen değer ile asıl değer arasındaki farkın mutlak değerinin ortalamasıdır. R^2 bir değişkenin varyansının ikinci değişkenin varyansını ne ölçüde açıkladığını gösterir. R^2 [0,1] arasında değer almaktadır. 1'e yakın bir R^2 değeri, model performansının iyi olduğunu gösterir. RAE ise tahmin edilen değer ile beklenen değerin mutlak farkının, her bir beklenen değer ile beklenen değerlerin mutlak farkına bölünmesi ile elde edilir. Sıfıra ne kadar yakınsa modelin başarılı tahminler yaptığı söylenebilir. Hata metriklerinin formülleri Eşitlik 9, 10 ve 11 ile verilmiştir.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)2} \quad (10)$$

$$RAE = \frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{\sum_{i=1}^{n}|y_i-\bar{y}_i|} \quad (11)$$

Formüllerde $\hat{y}_i$ tahmin edilen değeri, y_i beklenen değeri, $\bar{y}_i$ değişkenlerin ortalama değeri, n değişken sayısını ifade etmektedir.

3.3. Modellerin Gerçekleştirimi (Realization of Models)

Bu çalışmada her üç il için Xgboost, YSA ve LSTM modelleri kullanılarak performansları karşılaştırılmıştır. Modelleri geliştirirken Sckit-learn[2] ve Keras[3] kütüphaneleri kullanılmıştır.

XGBOOST modeli için ilk etapta GridSearch yöntemi ile ağacın derinliği, öğrenme oranı, ağaçları

[2] https://scikit-learn.org/stable/

[3] https://keras.io/

oluştururken alt örnekleme oranı ve ağaç sayısı parametreleri için en uygun değerler bulunmuştur. Parametre optimizasyonu sonucu ağaç derinliği değeri 4, öğrenme oranı 0.1, alt örnekleme oranı 1 ve ağaç sayısı 500 olarak belirlenmiştir.

YSA ve LSTM için ise katman sayısı, katmandaki nöronlar, gizli katman sayısı, çıktı katmanının boyutu, seyreltme değeri, optimize türü, dönem sayısı ve öğrenme tur sayısı gibi ayarlanması gereken parametreler ayarlanmıştır. YSA için iki adet yoğunluk katmanı kullanılmıştır. İlk yoğun katman için 128 nöron, ikinci yoğun katman için ise 64 nöron kullanılmıştır. Yoğunluk katmanlarında aktivasyon fonksiyonu relu tercih edilmiştir. Relu verilerdeki karmaşık ilişkilerin öğrenilmesine izin veren lineer olmayan bir fonksiyondur. Aşırı öğrenmeyi engellemek amacıyla 0.2 oranında seyreltme kullanılmıştır. Seyreltme işlemi rastgele bir şekilde bilgi azaltım yapmaktadır ve büyük oranlara sahip seyreltme işlemi önemli bilgilerin atılmasına sebebiyet verebilmektedir. LSTM için ise bir adet LSTM katmanı kullanılmıştır. LSTM katmanında 64 nöron tercih edilmiştir ve aktivasyon fonksiyonu tanh kullanılmıştır. YSA ve LSTM 0.001 öğrenme oranı, 16 parti boyutu, 500 öğrenme tur sayısı ve RMSProp kayıp fonksiyonu optimizasyonu kullanılmıştır. Kayıp fonksiyonu optimizasyonu derin öğrenme modelleri için oldukça büyük öneme sahiptir. RMSProp, gradyan tabanlı bir optimizasyondur. Kaybolan gradyan problemini önlemek için geliştirilmiştir. Dikey yönde salınımları kısıtlamaktadır ve yatay yönde hızlı yakınsama sağlamaktadır.

3.4. Deneysel Sonuçlar (Experimental Results)

Çalışma ile amacımız modellerin iller özelinde performansını karşılaştırmaktır. Sonuçlar Tablo 2, 3 ve 4 ile gösterilmiştir.

Sonuçlar incelendiğinde her üç modelinde birbirine yakın performans gösterdiği söylenebilir. Modeller Kocaeli ve Çanakkale illeri için yüksek performans göstermişlerdir. Kocaeli için R^2 skoru Xgboost için 0.93, YSA ve LSTM için ise 0.94'tür. Çanakkale için R^2 skoru Xgboost, YSA ve LSTM için 0.94'tür. Sakarya için R^2 skoru Xgboost için 0.88, YSA için 0.87 ve LSTM için ise 0.83'tür. R^2 değerlerine ait grafikler Şekil 5'te verilmiştir. Modeller Sakarya ili için daha düşük performans göstermiştir. Bunun sebebi Şekil 4 (c)'de görüldüğü üzere SO_2 değişkeni ile NO_2 ve NOX değişkeninin oldukça düşük bağımlılıkta olması olarak söylenebilir. Şekil 4 incelendiğinde Kocaeli ve Çanakkale illerindeki değişkenlerin birbiri ile daha iyi korelasyon içinde olduğu söylenebilir. İller bazında beklenen ve gerçekleşen değerlere ait grafikler ise Şekil 6, 7 ve 8 ile verilmiştir.

Tablo 2. XGBOOST modeli için performans değerlendirmesi (Performance evaluation for the XGBOOST model)

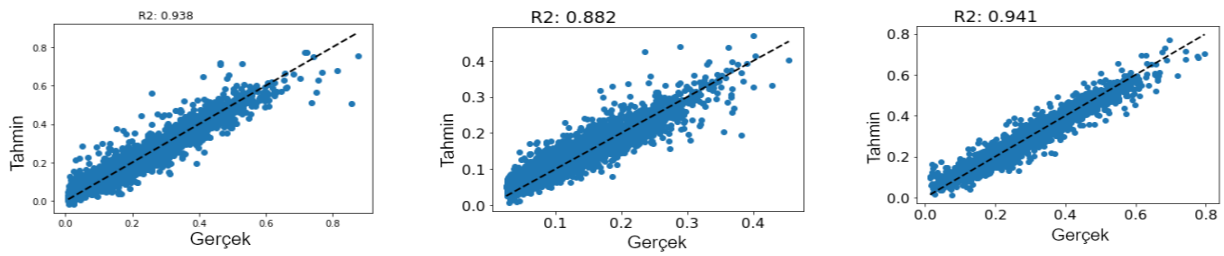
Metrik	Kocaeli	Sakarya	Çanakkale
R^2	0.93	0.88	0.94
MAE	0.026	0.019	0.020
RAE	0.19	0.30	0.21

Tablo 3. YSA Modeli için performans değerlendirmesi (Performance evaluation for the ANN model)

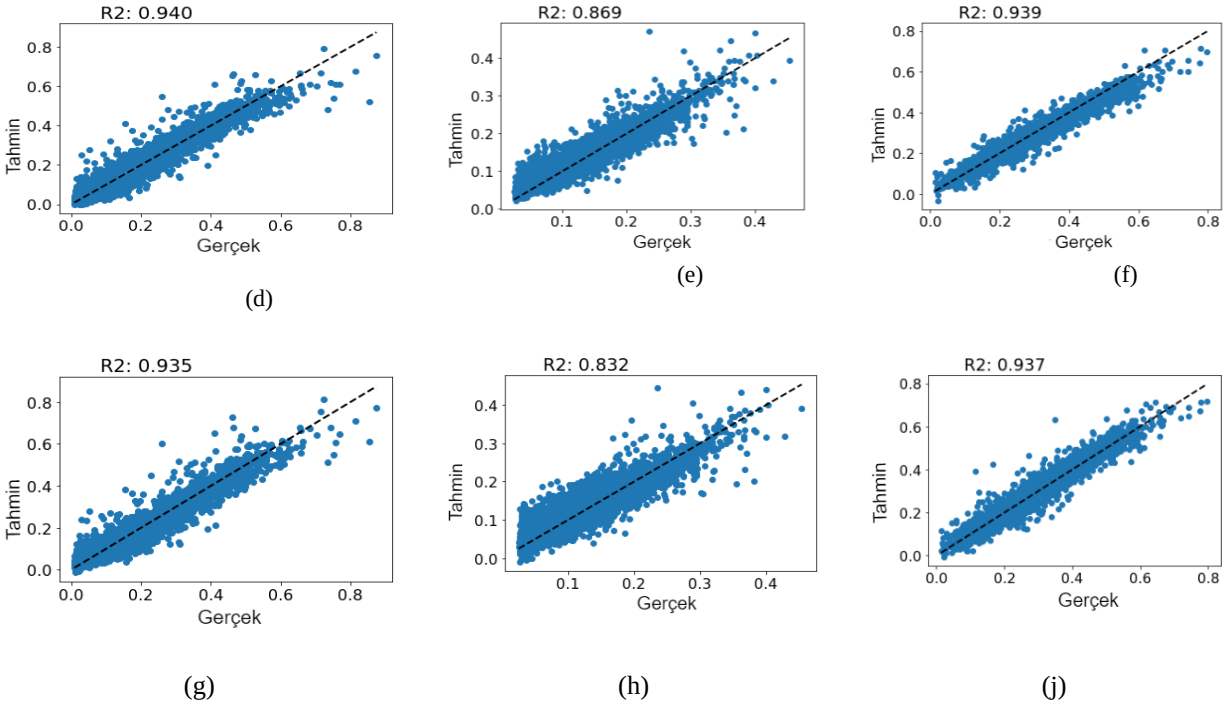
Metrik	Kocaeli	Sakarya	Çanakkale
R^2	0.94	0.87	0.94
MAE	0.026	0.021	0.021
RAE	0.19	0.32	0.22

Tablo 4. LSTM Modeli için performans değerlendirmesi (Performance evaluation for the LSTM model)

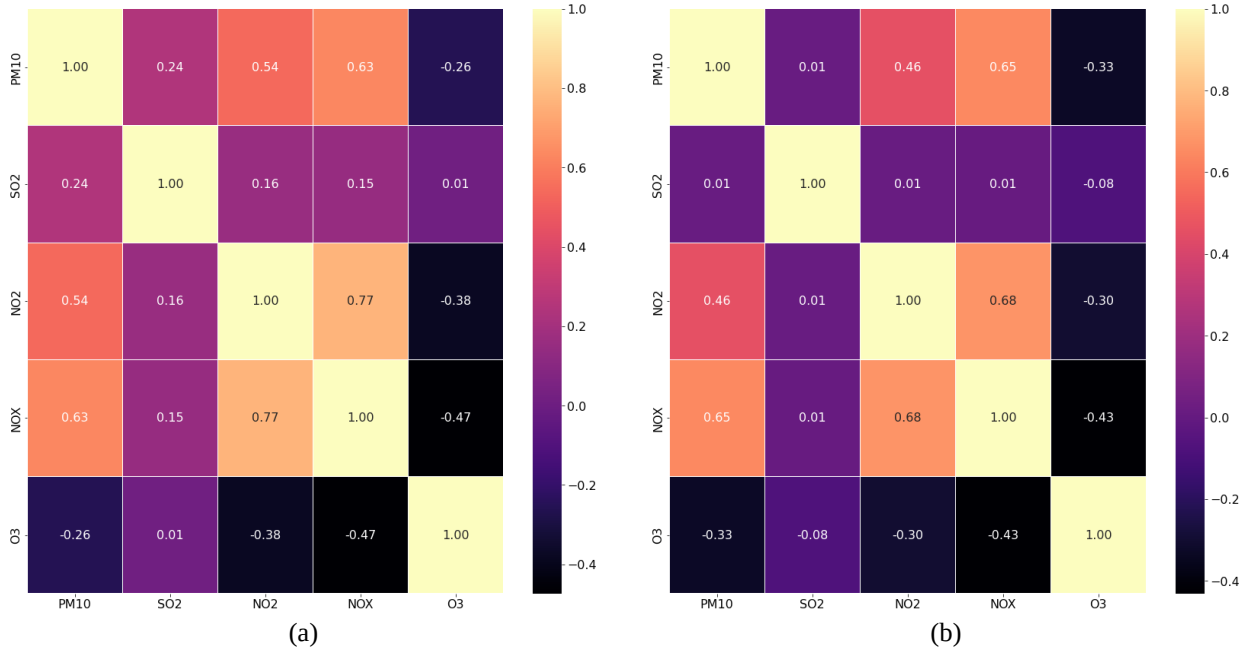
Metrik	Kocaeli	Sakarya	Çanakkale
R^2	0.94	0.83	0.94
MAE	0.027	0.022	0.020
RAE	0.20	0.34	

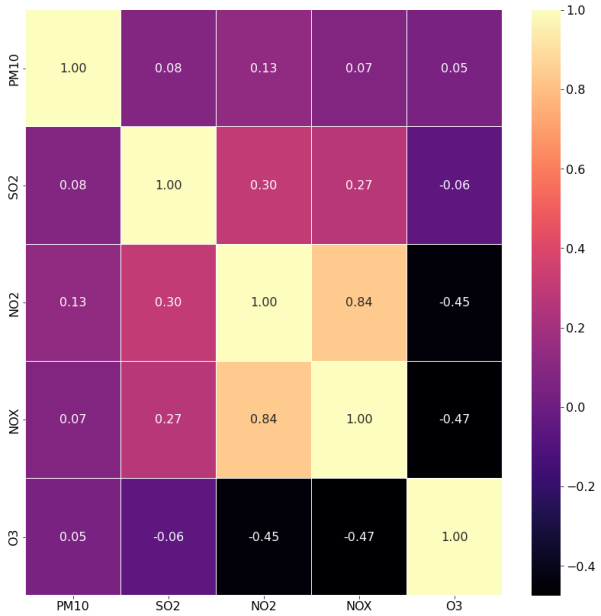


(a) (b) (c)



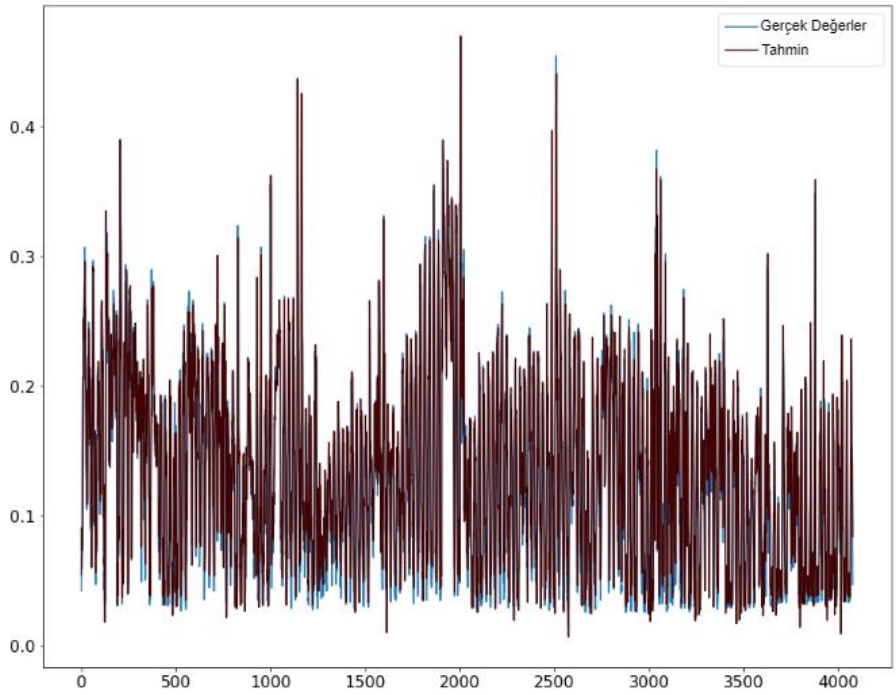
Şekil 5. Tahmin Edilen O_3 Değerleri ile Ölçülen O3 Değerleri XGBOOST; (a) Kocaeli (b) Sakarya (c) Çanakkale, ANN; (d) Kocaeli (b) Sakarya (c) Çanakkale, LSTM; (a) Kocaeli (b) Sakarya (c) Çanakkale. (Predicted O3 Values vs. Measured O3 Values XGBOOST; (a) Kocaeli (b) Sakarya (c) Çanakkale, ANN; (d) Kocaeli (b) Sakarya (c) Çanakkale, LSTM; (a) Kocaeli (b) Sakarya (c) Çanakkale)



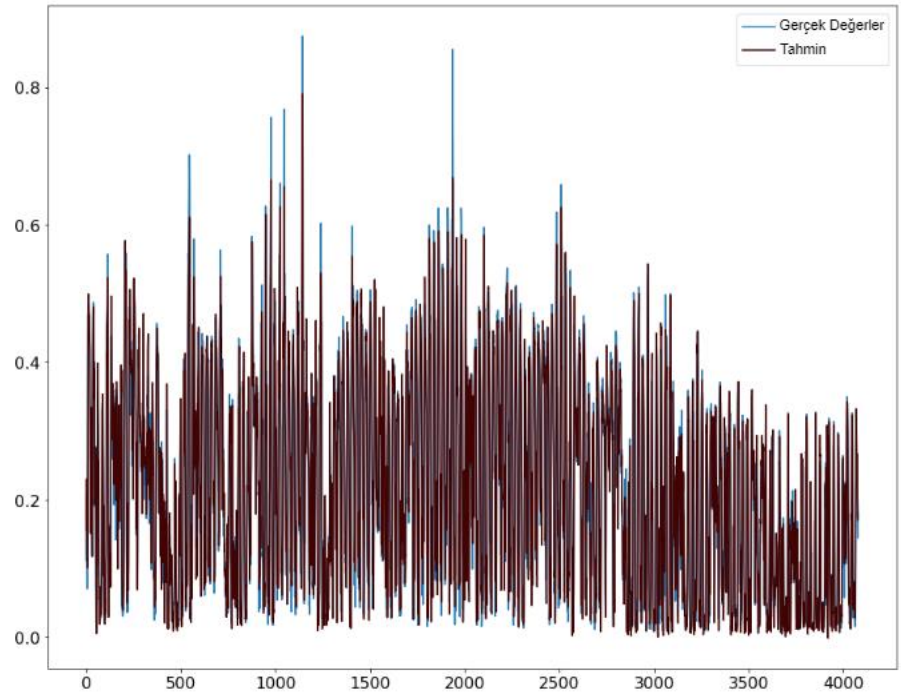


(c)

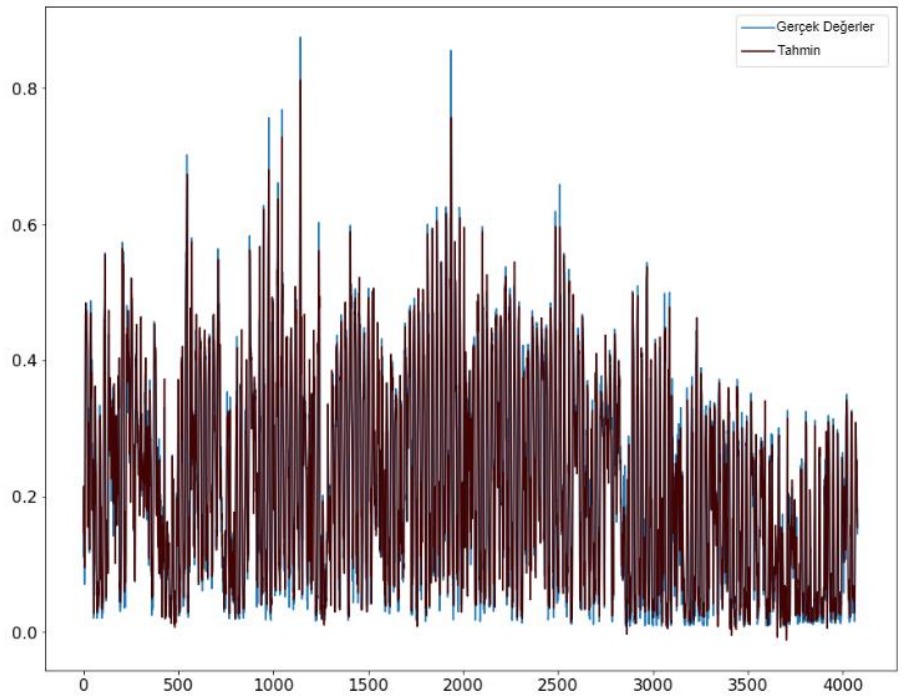
Şekil 4. (a) Kocaeli ili için değişkenlerin korelasyonu (b) Sakarya ili için değişkenlerin korelasyonu (c) Çanakkale ili için değişkenlerin korelasyonu ((a) Correlation of variables for Kocaeli province (b) Correlation of variables for Sakarya province (c) Correlation of variables for Çanakkale province)



(a)

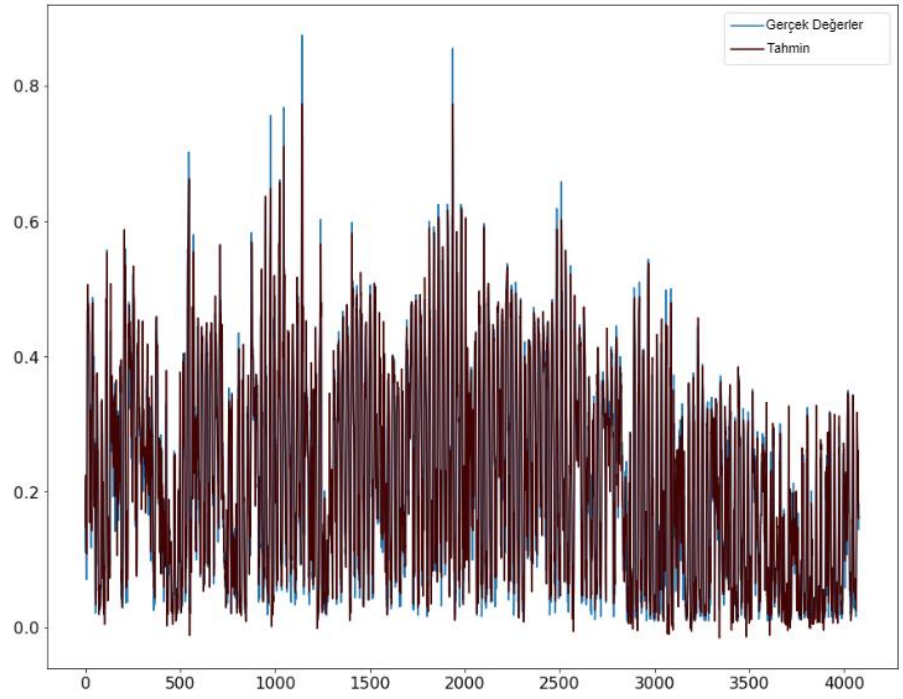


(b)

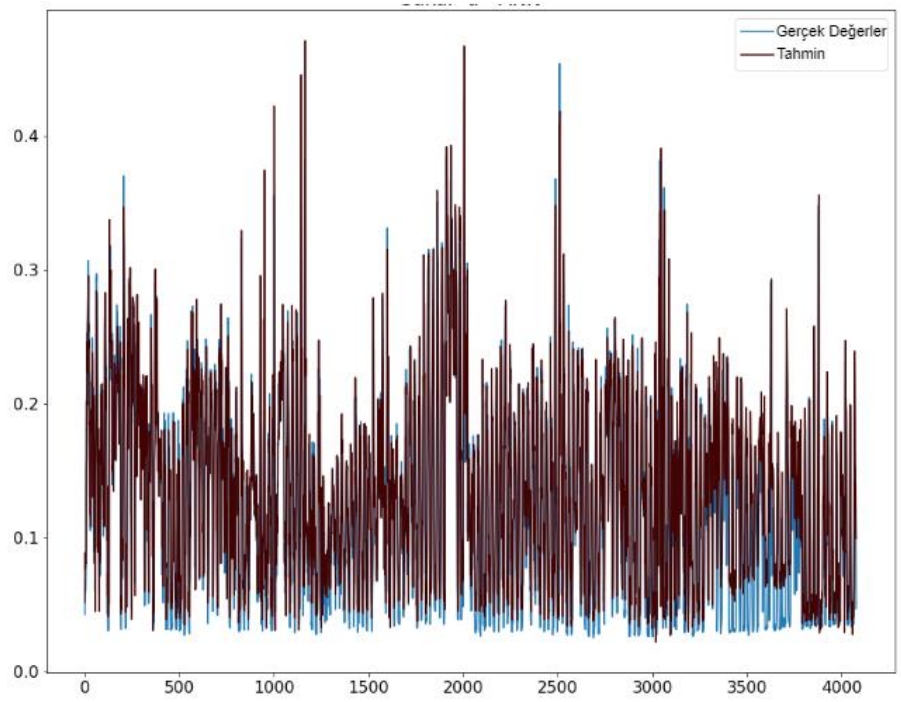


(c)

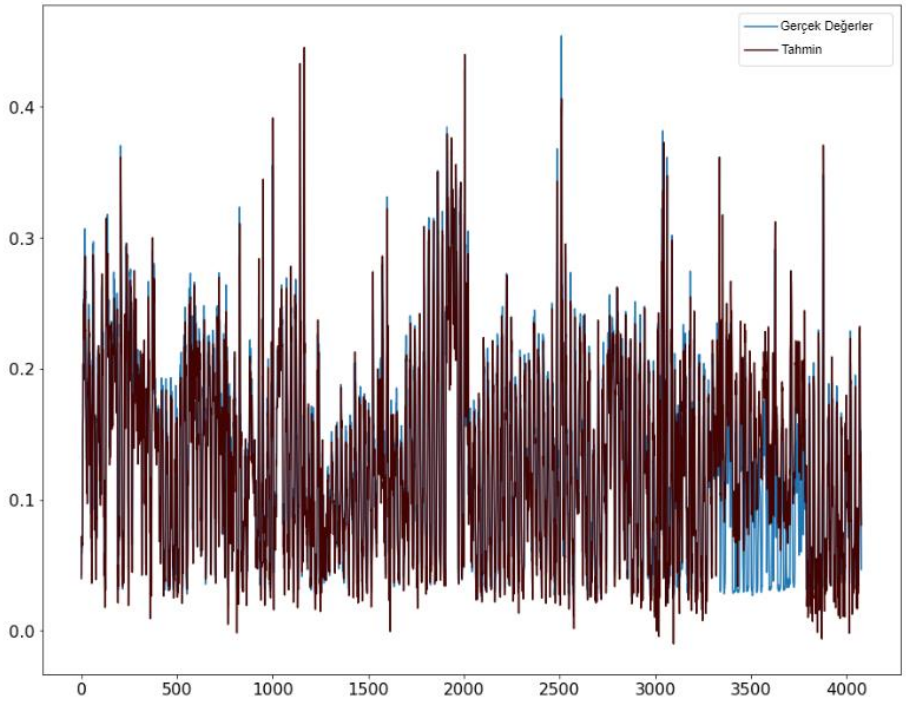
Şekil 6. Kocaeli ili için Ölçülen Değerler ve Tahmin Edilen Değerler (a) XGBOOST (b) ANN (c) LSTM (Measured Values and Estimated Values for Kocaeli Province (a) XGBOOST (b) ANN (c) LSTM)


Gerçek Değerler
Tahmin
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
3500
4000

(a)

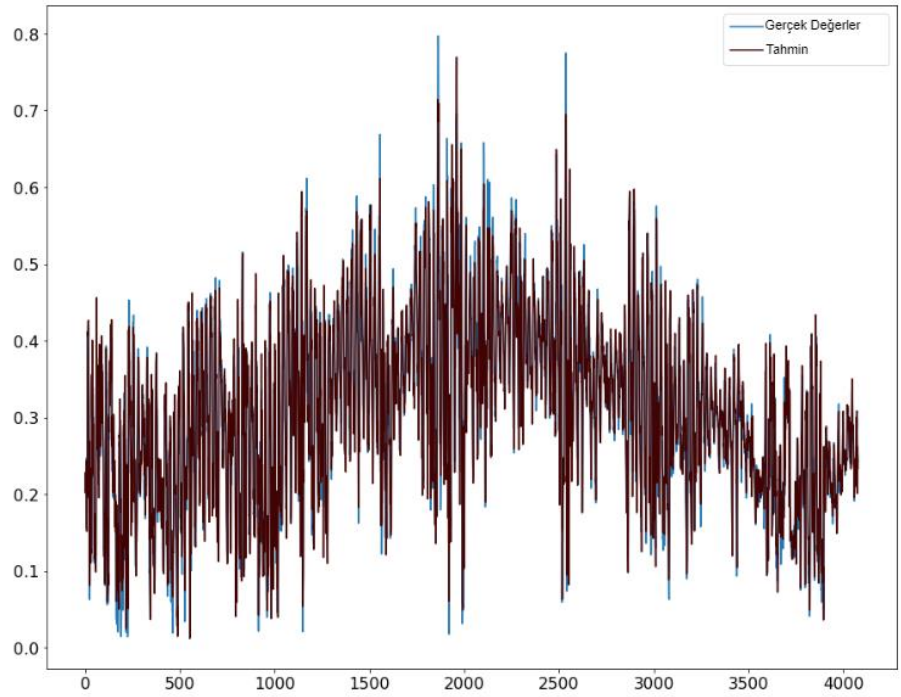

Gerçek Değerler
Tahmin
0.4
0.3
0.2
0.1
0.0
0
500
1000
1500
2000
2500
3000
3500
4000

(b)

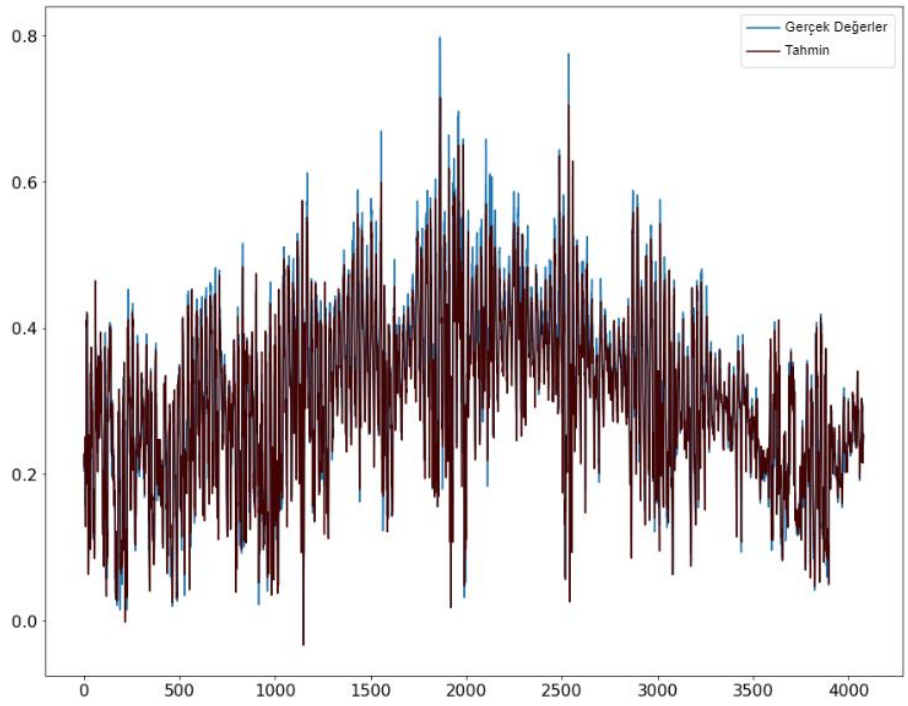


(c)

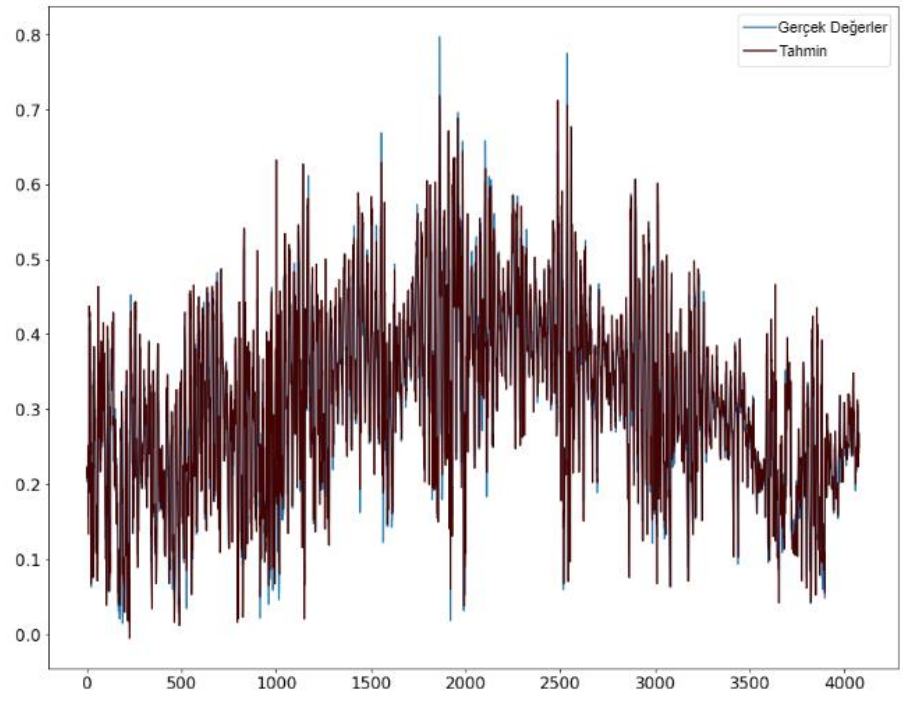
Şekil 7. Sakarya ili için Ölçülen Değerler ve Tahmin Edilen Değerler (a) XGBOOST (b) ANN (c) LSTM (Measured Values and Estimated Values for Sakarya Province (a) XGBOOST (b) ANN (c) LSTM)



(a)



(b)



(c)

Şekil 8. Çanakkale ili için Ölçülen Değerler ve Tahmin Edilen Değerler (a) XGBOOST (b) ANN (c) LSTM (Measured Values and Estimated Values for Çanakkale Province (a) XGBOOST (b) ANN (c) LSTM)

4. Sonuçlar (Results)

Bu çalışmanın amacı Türkiye'nin önde gelen sanayi şehirlerinden Sakarya, Kocaeli ile nispeten sanayisi daha az gelişmiş olan Çanakkale ilinin O_3 konsantrasyonlarını modellemektir. Bu amaçla makine öğrenmesi yöntemlerinden XGBOOST ve YSA ve derin öğrenme yöntemlerinden LSTM uygulanmıştır. Modelleme için girdi olarak kirliliğe sebep olan parametreler PM10, SO_2, NO, NO_2 ve O_3 kullanılmıştır. Test edilen modeller arasında herhangi bir ayrım yapmak zordur fakat karmaşık doğrusal olmayan sistemlerin modellemesinde ve zaman serisi problemlerindeki başarımından dolayı LSTM

kullanılması tavsiye edilir. Bu çalışmanın sonuçları, sanayilerinin gelişmişlik seviyelerine göre şehirlerin 0_3 seviyelerini tahmin etmek için bilgilendirici olabilir. Bundan sonraki çalışmalarda BiLSTM, CNN-LSTM ve Stacked LSTM gibi gelişmiş modeller denenebilir.

Kaynakçalar (References)

Adnane, A., Leghrib, R., Chaoufi, J., & Chirmata, A., 2020. The Use of a Recurrent Neural Network for Forecasting Ozone Concentrations in the City of Agadir (Morocco). Journal of Atomic, Molecular, Condensed Matter and Nano Physics, 7(3), 197-206.

Alghieth, M., Alawaji, R., Saleh, S. H., Alh, S., 2021. Air Pollution Forecasting Using Deep Learning. International Journal of Online & Biomedical Engineering, 17(14).

Alipio, M. M., 2020. Do latitude and ozone concentration predict Covid-2019 cases in 34 countries?. medRxiv.

Allu, S. K., Srinivasan, S., Maddala, R. K., Reddy, A., Anupoju, G. R., 2020. Seasonal ground level ozone prediction using multiple linear regression (MLR) model. Modeling Earth Systems and Environment, 6, 1981-1989.

Bekesiene, S., Meidute-Kavaliauskiene, I., Vasiliauskiene, V., 2021. Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks. Mathematics, 9(4), 356.

Bilgin, G., 2021. Investigation of The Risk of Diabetes in Early Period using Machine Learning. Journal of Intelligent Systems: Theory and Applications, 4(1), 55-64.

Chattopadhyay, G., Midya, S. K., Chattopadhyay, S., 2019. MLP based predictive model for surface ozone concentration over an urban area in the Gangetic West Bengal during pre-monsoon season. Journal of Atmospheric and Solar-Terrestrial Physics, 184, 57-62.

Chelani, A. B., 2010. Prediction of daily maximum ground ozone concentration using support vector machine. Environmental monitoring and assessment, 162(1), 169-176.

Çağıl, G., Yıldırım, B., 2020. Detection of an Assembly Part with Deep Learning and Image Processing. Journal of Intelligent Systems: Theory and Applications, 3(2), 31-37.

Darendeli, B. N., Yılmaz, A., 2021. Convolutional Neural Network Approach to Predict Tumor Samples Using Gene Expression Data. Journal of Intelligent Systems: Theory and Applications, 4(2), 136-141.

Ding, J., Liu, M., Ma, Z., Liu, R., Bi, J., 2020. Spatial and temporal trends in the mortality burden of ozone pollution in China: 2005-2017. ISEE Conference Abstracts, 24-27 August 2020.

Ekinci, E., İlhan Omurca, S., Özbay, B., 2021. Comparative assessment of modeling deep learning networks for modeling ground-level ozone concentrations of pandemic lock-down period. Ecological Modelling, 457, 1-11.

Ekinci, E., İlhan Omurca, S., Sevim, S., 2020. Improve Offensive Language Detection with Ensemble Classifiers. International Journal of Intelligent Systems and Applications in Engineering, 8(2), 109-115.

Eslami, E., Choi, Y., Lops, Y., Sayeed, A., 2020. A real-time hourly ozone prediction system using deep convolutional neural network. Neural Computing and Applications, 32(13), 8783-8797.

Garip Batık, Z., Büyükbıçakçı, E., 2016. Klasik Enterpolasyon Yöntemleri ve Yapay Sinir Ağı Yaklaşımları ile Matematiksel Denklemlerin Karşılaştırılmalı Çözümü İçin Arayüz Tasarımı, 4th International Symposium on Innovative Technologies in Engineering and Science, 3-5 November 2016, Antalya, Turkey, pp. 1379-1383.

Kleinert, F., Leufen, L. H., Lupascu, A., Butler, T., Schultz, M. G., 2021. Representing chemical history for ozone time-series predictions-a method development study for deep learning models. EGU General Assembly Conference Abstracts, 19-30 April, pp. EGU21-12146.

Liu, H., Liu, J., Liu, Y., Ouyang, B., Xiang, S., Yi, K., Tao, S., 2020. Analysis of wintertime O3 variability using a random forest model and high-frequency observations in Zhangjiakou—an area with background pollution level of the North China Plain. Environmental Pollution, 262, 114191.

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., Bi, J., 2020. Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. Environment international, 142, 105823.

Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (2009). "A Novel Connectionist System for Improved Unconstrained Handwriting Recognition". (IEEE Transactions on Pattern Analysis and Machine Intelligence. 31 (5): 855

Ma, R., Ban, J., Wang, Q., Zhang, Y., Yang, Y., He, M. Z., Li, S., Shi, W., Li, T., 2021. Random forest model based fine scale spatiotemporal O3 trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017. Environmental Pollution, 276, 116635.

Ma, Z., Liu, R., Bi, J., 2019. Spatiotemporal distributions of ground ozone levels in China from 2005 to 2016: a machine learning approach. AGU Fall Meeting Abstracts, 9-13 December 2019, San Francisco, USA, pp. A41J-2709.

Makarova, A., Evstaf'eva, E., Lapchenco, V., Varbanov, P. S., 2021. Modelling tropospheric ozone variations using artificial neural networks: A case study on the Black Sea coast (Russian Federation). Cleaner Engineering and Technology, 5, 100293.

Matasović, B., Pehnec, G., Bešlić, I., Davila, S., Babić, D., 2021. Assessment of ozone concentration data from the northern Zagreb area, Croatia, for the period from 2003 to 2016. Environmental Science and Pollution Research, 1-11.

Mehdipour, V., Memarianfard, M., 2019. Ground-level O3 sensitivity analysis using support vector machine with radial basis function. International Journal of Environmental Science and Technology, 16(6), 2745-2754.

Nghiem, T. D., Mac, D. H., Nguyen, A. D., Lê, N. C., 2021. An integrated approach for analyzing air quality monitoring data: a case study in Hanoi, Vietnam. Air Quality, Atmosphere & Health, 14(1), 7-18.

Paoli, C., Notton, G., Nivet, M. L., Padovani, M., Savelli, J. L. 2011. A neural network model forecasting for prediction of hourly ozone concentration in Corsica. 2011 10th International Conference on Environment and Electrical Engineering, 1-7 May 2011, Rome, Italy, pp. 1-4.

Sak, Hasim; Senior, Andrew; Beaufays, Francoise (2014). "Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling"

Sayeed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Salman, A. K., Lee, J. B., Park, H. J., Choi, M. H. (2021). A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. Scientific reports, 11(1), 1-8.

Sepp H., Jürgen S., 1997. Long short-term memory.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Şen, Z., 2018. Significance of Artificial Intelligence in Science and Technology. Journal of Intelligent Systems: Theory and Applications, 1(1), 1-4.

T. Chen, C. Guestrin, M. Assoc Comp, XGBoost: a scalable tree boosting system, 2016.

Tanaskuli, M., Ahmed, A. N., Zaini, N., Abdullah, S., Borhana, A. A., Mardhiah, N. A., 2020. Ozone prediction based on support vector machine. Indonesian Journal of Electrical Engineering and Computer Science, 17(3), 1461-1466.

Wang, H. W., Li, X. B., Wang, D., Zhao, J., & Peng, Z. R., 2020. Regional prediction of ground-level ozone using a hybrid sequence-to-sequence deep learning approach. Journal of Cleaner Production, 253, 119841.

Yang, X., Zhang, M., Zhang, B., 2021. A Generic Model to Estimate Ozone Concentration from Landsat 8 Satellite Data Based on Machine Learning Technique. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 7938-7947.

Yıldırım, A. E., Kadıoğlu, Ö. F., Kavak, H., Salman, K., Uçar, M. K., Uçar, Z., Bozkurt, M. R., 2021. Gender-Based Artificial Intelligence Based Detection of Basal Metabolic Rate by Electrocardiography Signal. Journal of Intelligent Systems: Theory and Applications, 4(2), 168-176.