JOURNAL OF AERONAUTICS AND SPACE TECHNOLOGIES
(ISSN : 1304-0448)
July 2021 Volume 14 Number 2
jast.hezarfen.msu.edu.tr

Research Survey

Ozone Level Prediction with Machine Learning Algorithms

Atınç YILMAZ[1*]

[1] *Beykent University, Computer Engineering Department, 34398 Sarıyer , Istanbul, Turkey*, atincyilmaz@beykent.edu.tr,
http://www.orcid.org/0000-0003-0038-7519
* *Corresponding Author*

Article Info

Received: *April 01, 2021*
Accepted: *June 10, 2021*
Online: *July 26, 2021*

Keywords: Machine Learning Algorithms, Ozone Level, Risk Prediction, Proposed Model, K-Fold Cross Validation

Abstract

The ozone level in the atmosphere affects the quality of life of all living things as well as it can be a hazard to human health and the environment. Ozone is a gas that absorbs most of the ultraviolet radiation reaching the Earth from the Sun. However, when the ozone level exceeds a certain threshold, risks would be exacerbated. Using machine learning algorithms can help to reduce risks, making inferences from earlier obtained data even for situations, which have not encountered before. In this study, a two-phased hybrid machine learning algorithm is proposed. It helps to predict the ozone level prospectively and reduce the risks. In the first stage, clustering is made with the method of genetic algorithms and the clustering result is transmitted as an introduction to the XGBoost classifier method. To check that the proposed model is applicable, support vector machine, random forest, multi-layered neural networks and XGBoost methods, which are among the frequently used machine learning methods, have been applied and the results were compared. After the 10-fold validation applied, the proposed model reached the most successful accuracy rate with 94%.

To Cite This Article: Atınç YILMAZ, "Ozone Level Prediction with Machine Learning Algorithms", Journal of Aeronautics and Space Technologies, Vol. 14, No. 2, pp. 177-183, July. 2021.

Makine Öğrenme Algoritmaları ile Ozon Seviyesi Tahmini

Makale Bilgisi

Geliş: *01 Nisan 2021*
Kabul: *10 Haziran 2021*
Yayın: *26 Temmuz 2021*

Anahtar Kelimeler: Makine Öğrenmesi, Ozon Seviyesi, Risk Tahmini, Önerilen Model, K-Fold Çapraz Doğrulama

Öz

Atmosferde bulunan ozon seviyesi tüm canlıların yaşam kalitesini etkilemek ile birlikte ölüme varan sonuçlara sebebiyet vermektedir. Ozon güneş ışınlarından gelen radyason değerinden konurulmasını sağlayan bir gazdır. Bu nedenle ozon seviyesi belirli bir eşiği aştığında risk durumu çoğalmaktadır. Makine öğrenmesi algoritmaları, uygulanacak problem hakkında yeterli verinin elde edilmesi halinde bu verilerden daha önce karşılaşmadığı durumlara karşın çıkarım yapma yeteneğine sahiptir. Bu çalışmada hibrit bir makine öğrenmesi algoritması önerilerek ozon seviyesinin tahmini ile risk oluşmadan buna engel olunması hedeflenmiştir. Önerilen hibrit model iki aşamalı olarak sonuç almaktadır. İlk aşamasında genetik algoritmalar yöntemi ile kümeleme yapılarak küme sonucu XGBoost sınıflandırıcı yöntemine giriş olarak iletilmektedir. Önerilen modelin uygulanabilir olduğunun gösterilmesi için sık kullanılan makine öğrenmesi yöntemlerinden destek vektör makineleri, random forest, çok katmanlı yapay sinir ağları ve XGBoost yöntemleri aynı probleme uygulanmıştır. 10-fold validation uygulanmasının ardından en başarılı doğruluk oranına %94 ile önerilen model ulaşmıştır.

1. INTRODUCTION

Ozone is a gas that helps to preserve Earth from the sun's ultraviolet radiation which can risked the lives of the creatures on earth [1, 2]. Tropospheric ozone in the lower atmosphere is a major air pollutant in urban areas. Air quality problem, which is included in atmospheric studies, is a popular subject studied within interdisciplinary. The reason is that the quality of the air is causally related to the life activities of all living things. When the ozone level exceeds a certain threshold, it creates a risk factor for human health [3, 4]. As a result of exposure, it causes a decrease in respiratory function, triggering chronic lung diseases, bronchitis, asthma, and even death. In addition, it seriously affects the activities in daily routine life [5-7]. It will be possible to minimize the risks that may occur by predicting the ozone level prospectively. For this estimation, various simulation systems and parametric models have been studied in the literature. However, as the problem is difficult to predict, the accuracy rates are not high enough, besides the models are costly even by using few parameters. Sufficient complete information on ozone formation has not yet been discovered.

Machine learning algorithms make it possible to learn rational behaviors through data [8]. With the application of algorithms, the system can make

rational decisions even for complex problems. Machine learning is associated with fields such as computer science, computational science, artificial intelligence, and data mining. It is aimed to take early precautions before risk occurs by estimating the ozone level with machine learning methods.

In this study, the most widely used machine learning algorithms such as XGBoost (eXtreme Gradient Boosting), SVM (Support Vector Machine), Random Forest and multilayer artificial neural networks model will be used for the estimation of ozone levels. In addition, the proposed hybrid model will be presented and utilized to the same problem. K-Fold cross validation will use to determine applied models' training sets.

The presented study consists of five chapters. In the second chapter, the related work is presented. In the third part, information about the data set, proposed hybrid model, and applied machine learning algorithms are shown. Implementation and conclusion are mentioned in the last sections.

2. RELATED WORK

In this section, related studies in the literature were summarized. Zhu et al. in 2020 [9]; used Deep Convolutional Neural Network to detect ozone depletion in Antarctica. They achieved %99.23 accuracy rate. Gagliardi et al. in 2020 [10]; performed machine learning modeling to investigate surface ozone behavior. By modeling the boosted regression trees algorithm in the study; compared with a multiple linear regression model in terms of performance. Mere et al. in 2020 [11]; they used machine learning techniques for ozone level prediction in Marrakesh. The aim of this study was to analyze the main factors associated with ozone formation. Liu et al. in 2020 [12]; they examined surface ozone levels between 2005 and 2017 in China with a machine learning approach. As a machine learning method eXtreme Gradient Boosting (XGBoost) algorithm was applied; 0.87 was obtained as the highest R-Square value. Wang et al. in 2020 [13]; predicted ground-level ozone by using hybrid deep learning method. As a result, it has been shown that monitoring stations in urban areas get better prediction results than remote stations. AlOmar et al. in 2020 [14]; used robust artificial intelligence to predict surface ozone. They obtained 0.93 RMSE Root Mean Square Error) value as the highest one using the proposed hybrid model. According to Jumin et al. in 2019; Ozone concentrations were estimated using a linear regression model and machine learning algorithms. Linear regression, artificial neural networks and boosted decision tree methods were used in the study.Among the applied models, the boosted decision tree model achieved the best performance with an R-Square value of 0.91. Watson et al. in 2019; performed machine learning-based modeling which predicts ozone exposure during bushfires. In the study, main machine learning methods such as gradient boosting, neural networks, support vector machines, random forest, elastic net, K-NN, K-Means were compared. Gradient boosting method reached the lowest RMSE value of 0.228. Ghoneim et al. in 2017 [17], performed predictive modeling of ozone concentration in smart cities using the deep learning method. In the deep learning model, grid search technique was used to learn the ozone level properties. Zhang et al. in 2007 [18]; has been applied to predict days with a high ozone threshold via data mining method.

3. METHODOLOGY

3.1. Dataset

The dataset used in this study is the "Ozone Level Detection Data Set" from the UCI (University of California-Irvine) open access data repository [19]. In the dataset, there are 2536 sample data with 73 features. While some of the input characteristics are temperature values measured at different times during the day; some of them are wind speeds at various times.

The most important features in the dataset highlighted by the Texas Environmental Quality Commission (TCEQ) are in the form of local ozone peak forecast, the ozone background level against the wind, leading emissions factor, maximum temperature in degrees F, the base temperature at which ozone production begins, the sum of solar radiation for the day, wind speed near sunrise and wind speed in the middle of the day.

3.2. Machine Learning Algorithms

Support vector machine (SVM), one of the widely used classification methods, is based on statistical learning [20]. The method is logically based on separating the data of two classes from each other in the most appropriate way. According to the nature of the problem, there are two types of support vector machines as linear and nonlinear. While class distinction can be made with a line for linear method; in a non-linear dataset, SVMs cannot draw a linear hyper-plane. For this reason, kernel tricks are used. The kernel method significantly increases machine learning efficiency in nonlinear data.

Random forest (RF) method is based on the principle of classifying the sample with the results produced by multiple classifiers instead of one classifier [21]. It makes this classification with multiple trees created using randomly generated samples. Multiple decision trees are created in the random forest architecture. The method combines these decision trees to obtain more

accurate and stable predict. One of the advantages of the random forest method is that it can be used to solve both classification and regression problems.

One of the decision tree-based method is XGBoost that is the optimized version of the Gradient Boosting method by bringing it to higher performance [22]. The important advantages of the XGBoost algorithm are that it can predict with high accurate, protect from overfitting, manage empty data and process with high performance.

Multi-layer neural networks (MLNN) is a learning-based method [23]. Thanks to its parallel calculation and learning capabilities, it is a method that can achieve high performance and accuracy even in nonlinear complex problems. Therefore, it has a very suitable algorithm for classification problems. However, it has disadvantages such as being hardware dependent and encountering excessive learning and memorization.

3.3. Proposed Hybrid Model

In the proposed model in this study, it was considered to cluster of parameters by examining the relationships between them before the classify. This approach allows to reduce the cost of operation due to the excessive input parameters, as well as increasing the accuracy rate by ignoring the parameters that have a negligible effect on the result.

Consequently, as the first step to make an effective clustering the Genetic Algorithm (GA) method will be used. The search capability of GA is used to determine the fixed number of cluster centers and to cluster the ungrouped data in the most appropriate way. GA's task is to search for suitable cluster centers so that the size of clustering is minimized. The clustering steps with the genetic algorithm are as follows:

1. Begin
2. K = 0
3. Begin the population for P(K).
4. Compute the fitness function for P(K).
5. Increase K by 1; K = K + 1
6. If it reaches the threshold of completion, go to step 11.
7. Select the value of P(K) from P(K-1).
8. Crossover the value of P(K)
9. Mutate P(K) value
10. Go to step 4.
11. The best ascent result is obtained.
12. End

Encoding of chromosomes is the first decision with modelling GA. Encoding depends on the type of problem. Permutation encoding was used in the proposed hybrid model. Permutation coding is useful for sorting. The fitness function reveals how close the solution to be applied is to the optimum solution of the problem. Fitness value is calculated repeatedly in GA. Therefore, it should be fast enough. Davies-Bouldin index (DBI) examines the differences within the cluster and the differences between clusters to obtain better clustering results. For this reason, DBI was used as a fitness function in the study. DBI focuses on the ratio of within-cluster scatter and between cluster separation. It exhibits good clustering when there is a small value. The small DBI value indicates more intra-cluster feature and bigger intra-cluster distance. First, the average distance (S_i) from the data in the class to the cluster center should be found in the CBI calculation. Si represents the degree of dispersion. The distribution of data points in the i-th class is as in Equation 1. A_i is the centroid of cluster of i, X_j is hte jth data point of cluster i, T_i is the amount of data in the equation 1.

$$S_i = \left(\frac{1}{T}\sum_{j=1}^{T_i}|x_j - A_i|\right)^{1/p} \quad (1)$$

Afterwards, M_{ij} representing the distance between cluster classes is calculated (Equation 2).

$$M_{i,j} = \|A_i - A_j\|_p \quad (2)$$

After calculating Equations 1 and 2, the Rij value is calculated to measure similarity (Equation 3).

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (3)$$

Fitness function calculation for the ith chromosome R_{ij} based on DBI is shown in Equation 3. With using Equation 3, nested n^2 loops are created based on the number of n cluster classes. Afterwards, the maximum R_{ij} is calculated for each cluster class. And in the last step, the average of the maximum similarity of all classes is taken to obtain the DBI index (Equation 4).

$$D_i \equiv max_{j\neq i} R_{i,j};\ DBI = \frac{1}{n}\sum_{i=1}^{N} D_i \quad (4)$$

Labels are exchanged between the clusters in the relabel procedure that are as follow:

Input:
Clustering X of N vectors into Cluster S_i and S_j
DBI threshold, ΔDBI
The centroid diameters difference thresholds, ΔM_1 and ΔM_2
Output:
Relabel if conditions are met.

Begin
if (DBI=0) && (Sj=isEmpty):
    Relabel(X)
else if (DBI>ΔDBI) && (M_1>M_2+ΔM_1):

Relabel(X)
else if (DBI<ΔDBI) && (M_1+ΔM_2<M_2):
Relabel(X)

Elitism Selection is used for selection method in proposed hybrid method. It uses partial reproduction strategies to acquire better parameters. It is a selection method in which a small fraction of the best individuals from the last generation are passed on to the next [24].

In the applied genetic algorithm phase, the fittest chromosomes were selected for crossover and standart single-point crossover was used. Gaussian distribution is done for mutation process. The fitness function is used as 1/DBI. The termination and threshold criterias are the maximum number of iterations and the number of consecutive iterations that do not update the population.

In the proposed model, the result obtained because of clustering processes feeds the classification model. Among the machine learning methods applied for classification, the XGBoost method, which achieves the highest accuracy, was preferred. The stages of the proposed model are shown in Figure 1.

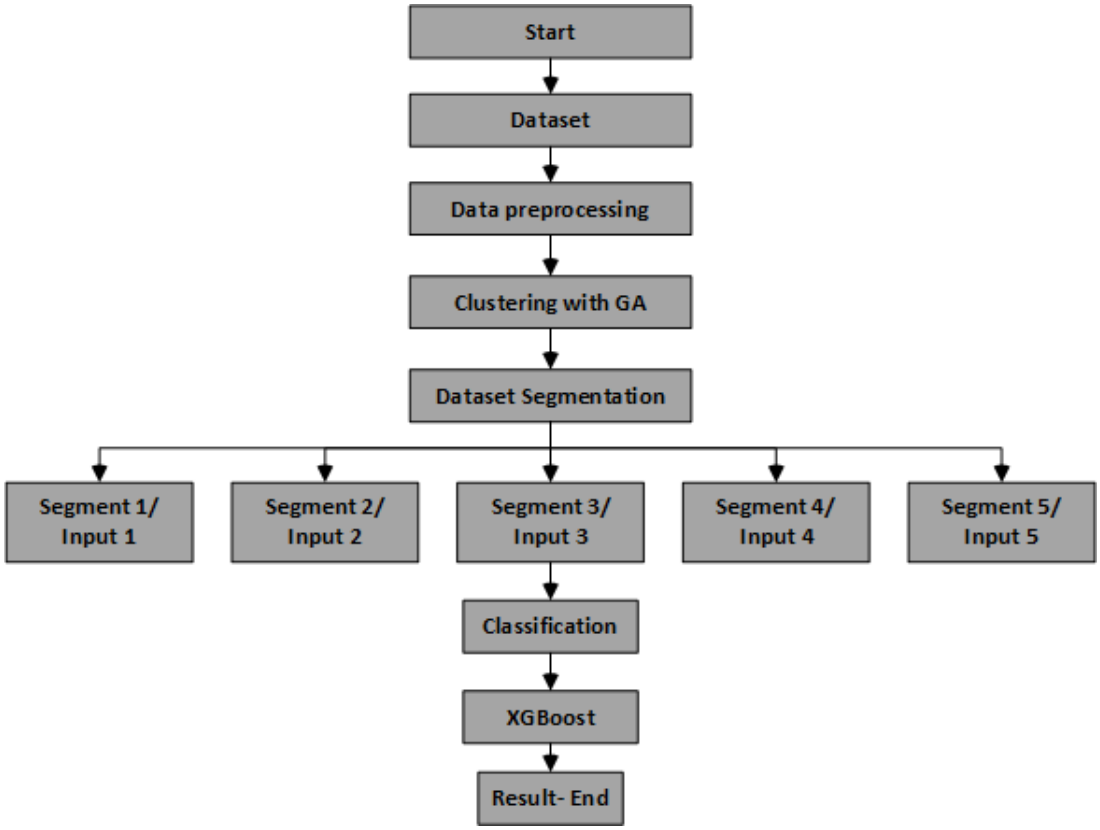


Figure 1. Proposed model's diagram.

XGBoost achieves high classification performance thanks to the loss function and regularization that used in the learning process. Loss function calculates the difference of each prediction value made by the model from the target class value. Regularization controls the complexity of the model, which eliminates the problem of overfitting in the model. In proposed hybrid model, the softmax multi-class function was used in XGBoost phase. The most suitable parameter values were obtained by grid search algorithm in proposed hybrid model. The grid search algorithm trains all combinations for the number of sheets and learning rate parameters and calculates the model accuracy with the cross-validation technique [25].

4. IMPLEMENTATION

All applications in the study were carried out with Python programming language over the Jupyter Notebook application. The dataset was examined before modeling machine learning methods. Since there are NaN values in the dataset, firstly, the data were preprocessed. It was determined that 300 of 2536 data were NaN data. For this reason, data with NaN values were determined and removed from the dataset.

As a result, 11.8% of data was extracted from the dataset, and a total of 2236 data was studied. In the dataset, 160 data is the data for the days that exceed the ozone threshold (Figure 2). It is seen that the target class distribution is not balanced in the current dataset. For preprocessing, min-max normalization was also applied to the dataset.

The K-fold validation method has been applied to the dataset for modeling. For the K value, the evaluations made on the 5, 10 and 15 values were also obtained for the highest performance K = 10. Using 10-fold validation provides the opportunity to create the optimum process.

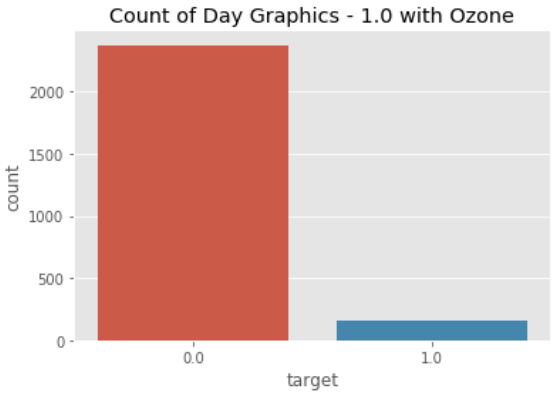


Figure 2. Count graphics.

After preprocessing, support vector machine (SVM), random forest (RF), XGBoost, multilayer neural network (MLNN) methods were applied to the problem. The applicability of the proposed model will be revealed by comparing the applied models in terms of performance.

After the clustering process step the proposed model shown in Figure 1 was completed, 73 parameters were reduced to 17 parameters. When the correlation of the 17 parameters suggested by the proposed model as the most relevant to the result, to the class label was examined with the Pearson's method, the most important 17 parameters were prepared for the clustering (Figure 3).

In the last phase of the proposed model, 17 input parameters are clustered in 5 different parts by GA, feeding the XGBoost classifier (Figure 4).

5. CONCLUSION

The study focused on estimating the ozone level, which may adversely affect the health of all living things and the routine life of the society. When the ozone level rises above a certain level are correctly predicted, fatal situations will be prevented. For this reason, machine learning algorithms have been applied based on samples of the ozone level at different times and temperatures. The machine learning algorithms applied in the study were selected as XGBoost, support vector machine, random forest and multilayer artificial neural networks methods which are high-performance and the most frequently used in the literature. In addition, a hybrid model using the genetic algorithm and XGBoost methods together has been proposed in the study.

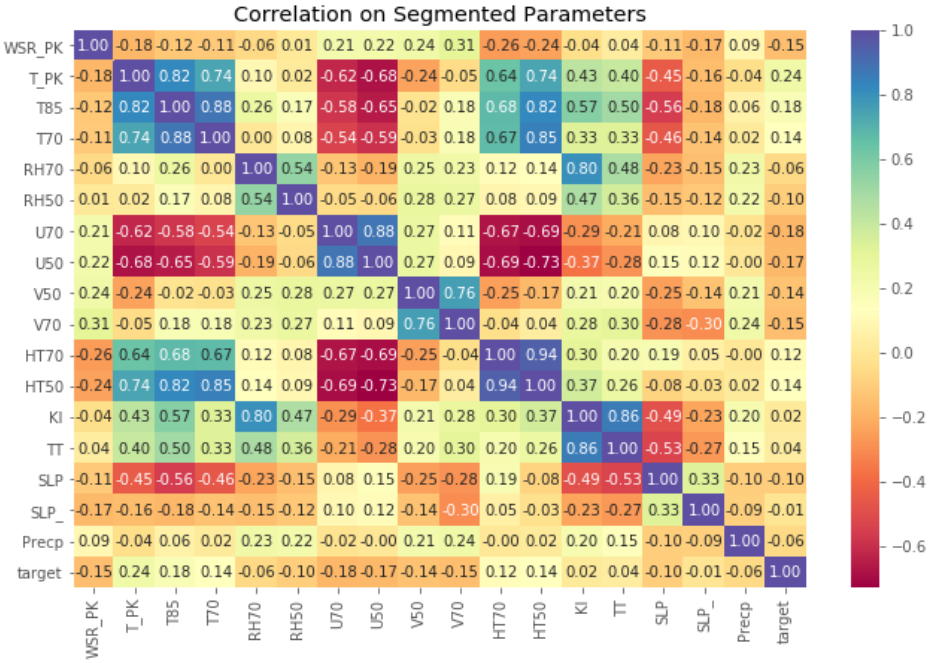


Figure 3. Correlation matrix.

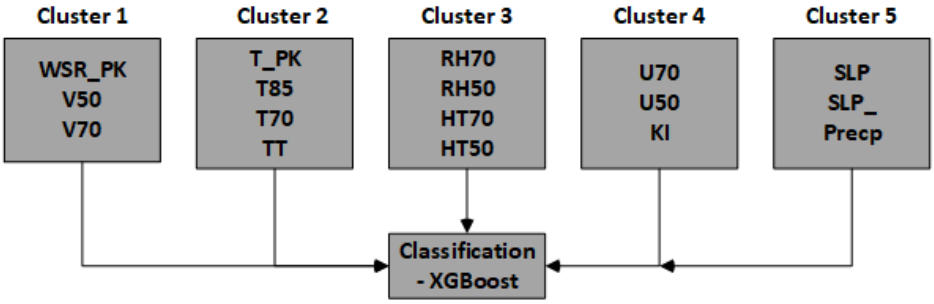


Figure 4. Proposed model's classification phase.

In this model, after clustering with the genetic algorithm, the results obtained from clustering were used as input for the XGBoost method to classify. Compared to other methods the proposed model provides a faster and more efficient result by working with fewer parameters thanks to its clustering capability. In addition, because the classification was made on the features that are more effective for the result, the accuracy rate was also higher.

The accuracy rates and error rates of all the methods applied are shown in Table 1. For 10-fold validation, the proposed model achieved the highest accuracy rate with 94%. Considering the accuracy rate and RMSE (Root Mean Square Error) values shown in Table 1, the applicability of the proposed model for ozone level determination has been revealed. As well, it was observed in the study that the proposed model does not have any disadvantages compared to other methods in terms of working time.

Table 1. Comparison of applied methods' performances.

Method	Accurate (Best %)	RMSE
SVM	0.902	4.51
RF	0.897	5.07
XGBoost	0.923	2.65
MLNN	0.903	4.32
Proposed Model	0.942	1.21

In future works, accuracy could be increased by using newly feature selection methods to select the most important input parameters. In addition, balancing the count of class samples would affect the success. Also, future works may focus on the detection of carbon emissions.

6. REFERENCES

[1] S. I. V. Sousa, M. C. M. Alvim-Ferraz and F. G. Martins, "Health effects of ozone focusing on childhood asthma: What is now known – a review from an epidemiological point of view" *Chemosphere*, vol. 90, pp. 2051-2058, June 2013.

[2] Y. C. Wang, F. C. Sung, Y. J. Chen, C. P. Cheng and Y. K. Lin, "Effects of extreme temperatures, fine particles and ozone on hourly ambulance dispatches" *Science of the Total Environment*, vol. 765, April 2021.

[3] Y. C. Wang, F. C. Sung, Y. J. Chen, C. P. Cheng and Y. K. Lin, "Ozone pollution in China: Background and transboundary contributions to ozone concentration & related health effects across the country" *Science of the Total Environment*, vol. 761, pp. 144131, March 2021.

[4] M. Pascal, V. Wagner, A. Alari, M. Corso and A. La Tertre, "Extreme heat and acute air pollution episodes: A need for joint public health warnings?" *Atmospheric Environment*, vol. 249, March 2021.

[5] J. Bosch, S. Elvira, C. Sausor, J. Bielby, I. F. Gonzalez and V. Bermejo, "Increased tropospheric ozone levels enhance pathogen infection levels of amphibians" *Science of the Total Environment*, vol. 759, March 2021.

[6] R. Long, A. Whitehill, A. Habel, S. Urbanski, H. Halliday, M. Colon, S. Kaushik and M. Landis, "Comparison of ozone measurement methods in biomass burning smoke: an evaluation under field and laboratory conditions" *Atmospheric Measurement Techniques*, vol. 14, pp. 1783-1800, March 2021.

[7] A. Gao, J. Wang, J. Luo, A. Li, K. Chen, P. Wang, Y. Wang, J. Li, J. Hu and H. Zhang, "Temporal variation of PM2.5-associated health effects in Shijiazhuang, Hebei" *Frontiers of Environmental Science & Engineering*, vol. 15, pp. 1783-1800, December 2020.

[8] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care" *Neural Computing & Applications*, vol. 32, pp. 18069-18083, December 2020.

[9] H. Zhu, S. Thedoridis, K. Lam and H. Lung, "A new method for ozone depletion detection over antarctica by deep convolutional neural network" in *Proc. of the* 2020 *IEEE 5th International Conference*

on Signal and Image Processing, ICSIP 2020, pp. 506-511, October 23-25, 2020.

[10] R. V. Gagliardi and C. Andenna, "A machine learning approach to investigate the surface ozone behavior" *Atmosphere*, vol. 11, pp. 1173, October 2020.

[11] J. Ordieres-Meré, J. Ouarzazi, B. El Johra, and B. Gong, "Predicting ground level ozone in Marrakesh by machine-learning techniques" *Journal of Environmental Informatics*, vol. 36, pp. 93-106, December 2020.

[12] R. Liu, Z. Ma, Y. Liu, Y. Shao, W. Zhao and J. Bi "Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach" *Environmental International*, vol. 142, September 2020.

[13] H. W. Wang, X. B. Li, D. Wang, J. Zhao, H. D. He, and Z. R. Peng, "Regional prediction of ground-level ozone using a hybrid sequence-to-sequence deep learning approach" *Journal of Cleaner Production*, vol. 253, April 2020.

[14] M. K. AlOmar, M. M. Hameed and M. A. AlSaadi, "Multi hours ahead prediction of surface ozone gas concentration: Robust artificial intelligence approach" *Atmospheric Pollution Research*, vol. 11(9), pp. 1572-1587, September 2020.

[15] E. Jumin, N. Zaini, A. N. Ahmed, S. Abdullah, M. İsmail, M.Sherif, A. Sefelnasr and A. El-Shafie "Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction" *Engineering Applications of Computational Fluid Mechanics*, vol. 14, May 2020.

[16] G. L. Watson, D. Telesca, C. E. Reid, G. G. Pfister and M. Jerrett, "Machine learning models accurately predict ozone exposure during wildfire events" *Environmental Pollution*, vol. 254, November 2019.

[17] O. A. Ghoneim, H. Doreswamy and B. Manjunatha, "Forecasting of ozone concentration in smart city using deep learning," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), ICACCI 2017, Udupi, India, September 13-16, 2017*.

[18] G. L. Watson, D. Telesca, C. E. Reid, G. G. Pfister and M. Jerrett, "Forecasting skewed biased stochastic ozone days : Analyses, solutions and beyond" *Knowledge and Information Systems*, vol. 14, pp. 299-326, July 2007.

[19] UCI Machine Learning Repository, Ozone Level Detection Dataset, https://archive.ics.uci.edu/ml/data setsozone+level+detection [Accessed: March. 2021].

[20] M. J. Wang and H. L. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis" *Applied Soft Computing*, vol. 88, March 2020.

[21] W. Chen, Y. Li, W. F. Xue, H. Shahabi, S. J. Li, H. Y. Hong, X. J. Wang, H. Y. Bian, S. Zhang, B. Pradhan and B. B. Ahmad, "Modeling flood susceptibility using data-driven approaches of naive Bayes tree, alternating decision tree, and random forest methods" *Science of the Total Environment*, vol. 701, January 2020.

[22] J. P. Lin, C. W. Qi, H. L. Wan, J. Y. Min, J. J. Chen, K. Zhang and L. Zhang, "Prediction of cross-tension strength of self-piercing riveted joints using finite element simulation and XGBoost algorithm" *Chinese Journal of Mechanical Engineering*, vol. 34, December 2021.

[23] M. J. Wang and H. L. Chen, "An improved grey wolf optimization-based learning of artificial neural network for medical data classification" *Journal of Information and Communication Technology*, vol. 20, pp. 213-248, April 2021.

[24] K. Jebari and M. Madiafi, "Selection methods for genetic algorithms" *International Journal of Emerging Sciences*, vol. 3, pp. 333-344, December 2013.

[25] M. Üstüner, S. Abdikan, G. Bilgin and F. B.Şanlı, "Crop Classification Using Light Gradient Boosting Machines" *Turkish Journal of Remote Sensing and GIS*, vol. 1, pp. 97-105, September 2020.

VITAE

Atınç YILMAZ received his PhD. degree in Computer and Information Engineering from Sakarya University, Turkey in 2015. He is currently working at Beykent University, Computer Engineering Department as faculty member-Asst.Prof.Dr.