Himalayan Physics

Lattice parameters prediction of orthorhombic oxyhalides using machine learning

https://doi.org/10.3126/hp.v11i1.62178

Poojan Koirala, Madhav Prasad Ghimire*

Central Department of Physics, Tribhuvan University, Kirtipur – 44613, Kathmandu, Nepal

Abstract: Lattice parameters of orthorhombic oxyhalides with molecular formula AOX are predicted using KRR, LR, and GBR machine learning (ML) models. Seventeen data of orthorhombic oxyhalides are extracted from the Materials Project Database, and several features such as atomic radius, ionic radius, band gap, density, electro-negativity, and atomic mass are taken into account. After refining the data, they are used for ML training and testing processes. The actual values of the respective compounds' lattice parameters are compared with those predicted by different models. Then, the accuracy of their predictions is checked by calculating MAE, MSE, and R^2. The GBR model is more efficient in predicting lattice parameters 'b' and 'c', whereas KRR is found to be more more efficient in predicting 'a'. Further, using the random forest regression model, the features importance plot is also observed to understand which features play an important role in predicting the lattice parameters.

Keywords: Artificial intelligence • Machine learning model • Orthorhombic oxyhalides • Lattice parameters • Gradient boosting regression model

Recieved: 2023-10-17 **Revised: 2024-01-06** **Published: 2024-01-24**

I. Introduction

In this materialistic world, there has been a continuous exploration of various materials driven by the pursuit of discovering intriguing properties. Oxyhalides are a fascinating material with applications ranging from optoelectronic devices to photocatalysis. They are compounds consisting of one or more than one element bonding with both oxygen and halogen. They are represented by the chemical formula AO_mX_n, where A is a transition element, the main group element or lanthanides and actinides, O is the oxygen, and X is the halogen (Fig. 1) [1]. They have seven crystal structures, including cubic, triclinic, trigonal, tetragonal, orthorhombic, monoclinic, and hexagonal. Here, we are interested in dealing with the orthorhombic structures only. All these crystalline structures have a variety of structures typically characterized by several atomic positions and lattice constants.

* *Corresponding Author: madhav.ghimire@cdp.tu.edu.np*

Lattice constants are crucial in recognizing crystalline materials and determining their chemical, physical, and electronic properties. It is influenced by number of factors like ionic radii, electro-negativity, number of valence electrons, etc. [2]. So, getting accurate values of these lattice constants is of utmost importance. There are both theoretical and experimental ways of obtaining lattice parameters. Experimentally, it can be measured using techniques like X-ray, neutron, or electron diffraction, but these techniques are quite challenging, costly, and take a long time [3]. The lattice constants of the crystals can also be predicted computationally, which helps in finding and predicting their properties and crystal structures. So far, different lattice parameter prediction methods have been developed. Artificial intelligence (AI) development has set a new trend in this field. It mainly consists of building a model using different algorithms and using it to predict the lattice parameters of the desired materials [2]. The lattice prediction techniques can also be categorized based on the different machine learning models used.

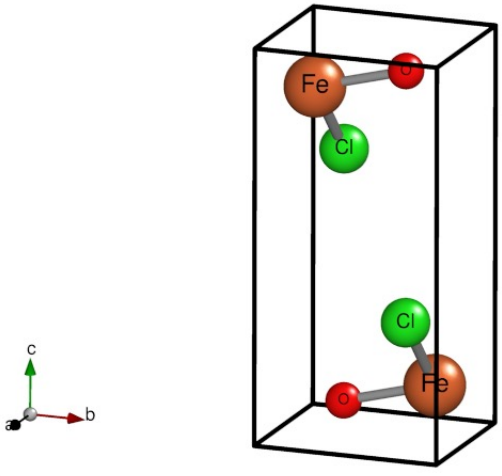

Figure 1. Crystal structure of AO_mX_n (A=Fe, X=Cl, m=n=1).

Machine Learning (ML), the branch of AI, is developed to create software that can learn from past data, gather expertise without human involvement, and thus generate predictions based on fresh data [4]. Recently, it has been developed into an effective and significant device for analyzing and discovering novel materials [5]. In addition, ML has now become the approach of choice in AI for several important tasks like creating useful software for robotics, image and speech recognition, banking services, computer vision, online fraud detection, product recommendations, social media features, and other applications [6]. With all these universal applications, ML is not just taking the place of earlier algorithms but is also found to be better at performing work, initially well handled by people [7].

Fig. 2 shows the basic workflow of machine learning. The whole dataset is split mainly into two parts: training and testing. The training dataset is used to train the model, whereas the testing dataset makes predictions based on the model prepared using the training dataset.

Now, to understand the types of data ML is effective at handling, there are three main types of learning: Supervised, unsupervised, and reinforcement learning [8]. In supervised learning, a known set of input and output data are fed into the system, and a particular ML algorithm is allowed to discover and, thus, infer patterns from the datasets. This makes it possible to predict upcoming events or data that wasn't part of the training dataset [9]. Support Vector Machine (SVM), linear regression, decision trees, neural networks, etc, are some of the widely used algorithms in supervised learning [8]. In unsupervised learning, the system tries to detect hidden patterns in the given input data set and applies them to the newly introduced data [10]. Moreover, in reinforcement learning, an artificial agent exists that acts in an environment to enhance rewards. Here, the environment allows the machine to learn continuously through trial and error, thus allowing the system to learn from the past to gather experiences to run new data [8].

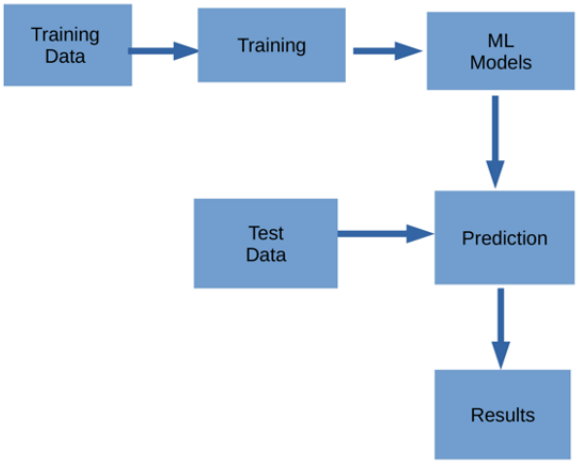

Figure 2. Machine learning workflow [11].

Ongoing through the literature on lattice constant predictions using machine learning models, we found plenty of studies on perovskite's materials properties prediction using different models. However, no literature was related to the lattice parameters prediction of orthorhombic oxyhalides. So here, we intend to use the ML model to predict the lattice parameters of orthorhombic oxyhalides of the general form AOX.

II. Methodology

In this section, the procedural steps for predicting the lattice parameters of orthorhombic oxyhalides using different programming models are outlined as below:

Data collection and feature engineering

At first data are collected mainly from the Materials Project Website. As we dealt with oxyhalide compounds, data of only one crystal structure, i.e., Orthorhombic of type AOX, are considered. Also, from the AOX family, only one space group of Pmmn is taken for the study. A total of only seventeen data are taken for model development. We tabulated all these compounds with their lattice constants, atomic radii, ionic radii, density, electro-negativity, band gap, and formation energy. We then saved the collected data as a csv file for loading into the program. As orthorhombic structures have different values of lattice parameters a, b, and c, but the lattice angles are all equal, i.e., $\alpha = \beta = \gamma = 90^{\circ}$, we will predict a, b, and c only.

Following data collection, one needs to extract the appropriate attributes for target prediction, known as feature engineering. In feature engineering, features are extracted from raw data and supplied as input to the algorithms. It is a key factor in the entire ML model [12]. The right features can incorporate crucial information, and its quality directly affects the model's predictability [5]. In the case of oxyhalides, AO_mX_n, while predicting lattice parameters, atomic radii, ionic radii, density, electro-negativity, band gap, formation energy, and atomic mass of the elements A and X are considered as descriptors or features. At the same time, lattice parameters (a, b, and c) are set as target variables.

ML model and model validation

Linear regression

The regression model is a type of supervised learning used to model continuous variables and thus make predictions. Linear regression (LR) is the most basic and most straightforward type of regression [13]. It focuses on the spatial distribution of a dependent variable using the independent variables [14]. The simple regression model can be represented in terms of the equation as,

$$y = a_o + a_1 x + e \tag{1}$$

where y = dependent variable, a_o = constant which is the intercept of the regression line on the vertical axis, a_1 = regression coefficient or the slope of the regression line, x = independent variable and e = random error [15]. Here, the training testing ratio was set to 70:30. This implies that 70 % of available data is used for training the model, and 30 % is used for testing the result.

Kernel ridge regression

Kernel Ridge Regression (KRR) is simply the extension of linear regression. It is a kernel-based regression technique that addresses the over-fitting issues by using regularization and the kernel method in non-linear input variables [16]. It is one of the most extensively used models in predicting material properties. It is a similarity-based regression algorithm whose results can be presented as a weighted sum

over kernel functions, which can be explained in terms of Euclidean distance across the data points [17]. KRR works more efficiently than other models in cases where there are huge numbers of independent predictors and there is confusion about which predictor affects the target in a significant way [18].

In this project, the hyper-parameters of this model are alpha (α) and kernel. Here, α is set to 0.05, and the kernel is set to rbf, which is the radial basis function. α is the regularization strength parameter, which maintains the balance between training data and its over-fitting. The kernel is the choice of kernel function being selected for the model.

Gradient boosting regression

Gradient Boosting Regression (GBR) method is an ensemble method capable of handling both linear and non-linear relationships of the data. The central theme of this model is to combine different simple models iteratively with better prediction accuracy than the other simple models [19]. So here, multiple models are developed sequentially with improved prediction accuracy. Each new base model focuses on fixing the errors created by the earlier base models [20]. Thus, through successive iterations, GBR mainly aims to minimize errors to make prediction effective. Therefore, GBR model is straightforward, making it possible to test various model architectures. In addition to real-world applications, this method has demonstrated significant effectiveness in various machine learning and data mining tasks [21]. GBR emphasizes minimizing the loss function, i.e., this model tries to reduce the deviation of the predicted value from the true value. So, it is an optimization strategy that reduces the loss function as much as possible by building on a base model at each stage [20].

Here in this project, the hyper-parameters of this model are $n\_estimators$ and max_depth , which are set to 100 and 3, respectively. $n\_estimators$ gives the idea of the number of weak learners used in the model. And max_depth specifies the maximum depth of each individual tree used in the model.

Random forest regression

Random Forest(RF) is a widespread and excellent algorithm technique. It is based on model aggregation ideas and is suitable for classification and regression cases [22]. RF is an accumulation of several classification or regression trees. So, the name forest comes from this fact [23]. Here RF is used simply for getting the features importance plot as it can easily calculate the importance of each feature present in the dataset and thus can convincingly evaluate the contribution of several features in predicting the target [24].

In this project, we will use the KRR, GBR, and LR models to predict lattice parameters. The training-test ratio was set to 70:30 for every single model. Then, the Random Forest Regressor (RFR) model is used to observe the feature's importance in predicting lattice parameters of the orthorhombic system of the Oxyhalides family. The features importance score provides information about different features contributions in predicting the results.

Model accuracy

After the prediction has been made, one needs to observe the model's accuracy and validity to find out how well the model has made the prediction. For doing so, several parameters like mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2) are calculated. The formulas for the calculation of these parameters are given below:

$$MSE = \frac{1}{m}\sum (x_i - x_l)^2 \tag{2}$$

$$MAE = \frac{1}{m}\sum |(x_i - x_l)| \tag{3}$$

here m being the number of observations, x_i the actual value and x_l the predicted value.

R^2 is then given by,

$$R^2 = \left| \frac{1}{M} \frac{\sum (Y_j - \overline{Y})(X_j - \overline{X})}{\sigma_x \sigma_y} \right|^2 \tag{4}$$

here M is the number of observations, σ_x and σ_y are the standard deviation of X and Y respectively. X_j and $\overline{X}$ are the observed values and the mean of the observed values, respectively. Whereas Y_j and $\overline{Y}$ are the calculated values and the mean of the calculated values respectively [25].

III. Results and Discussion

Here, the lattice parameters a, b and c are determined using KRR, LR and GBR. About seventeen data are taken of the orthorhombic oxyhalides AOX form, considering the space group Pmmn only. The data are then used for training and testing of the model. The actual and the predicted values of lattice parameters using different models are compared, and the validity of the models is observed by calculating MAE, MSE and R^2.

Table 1. Result of lattice parameters prediction

S.N.	Compounds	Actual			KRR			GBR			LR		
		a	b	c	a	b	c	a	b	c	a	b	c
1	InOCl	3.55	4.10	8.41	3.524	4.067	8.606	3.268	4.119	8.695	3.409	3.998	8.318
2	InOBr	3.65	4.08	8.91	3.522	3.966	8.810	3.450	3.965	8.676	3.452	4.041	8.609
3	CrOF	3.00	3.89	5.68	2.974	3.857	6.542	3.050	4.039	6.510	3.046	4.047	6.508
4	TmClO	3.71	4.15	9.20	3.658	4.112	8.826	3.749	4.155	9.025	3.762	4.129	9.034
5	ScBrO	3.56	3.96	8.94	3.561	3.983	8.878	3.749	3.965	9.012	3.499	4.075	8.685
6	HoBrO	3.84	4.16	9.18	3.642	3.996	8.987	3.749	4.089	9.129	3.709	4.166	9.104

Lattice parameters prediction

A total of only seventeen data of orthorhombic oxyhalides of the AOX form of the space group Pmmn were available on the materials project website, and all those available data are used. Different models like KRR, GBR and LR have been used to predict the lattice parameters. The training and testing ratio was set to 70:30 in every model. In the orthorhombic system case, we have different values of a, b and c. The Table 1 shows the compounds with their actual lattice parameters values and those predicted by considered models.

Model validity

The three different models and their respective values of the parameters MAE, MSE and R^2 in predicting a, b and c separately are presented in the Table 2. From the Table 2, it is seen that in the prediction of 'a', KRR is the one with least value of MAE and MSE whereas highest value of R^2.

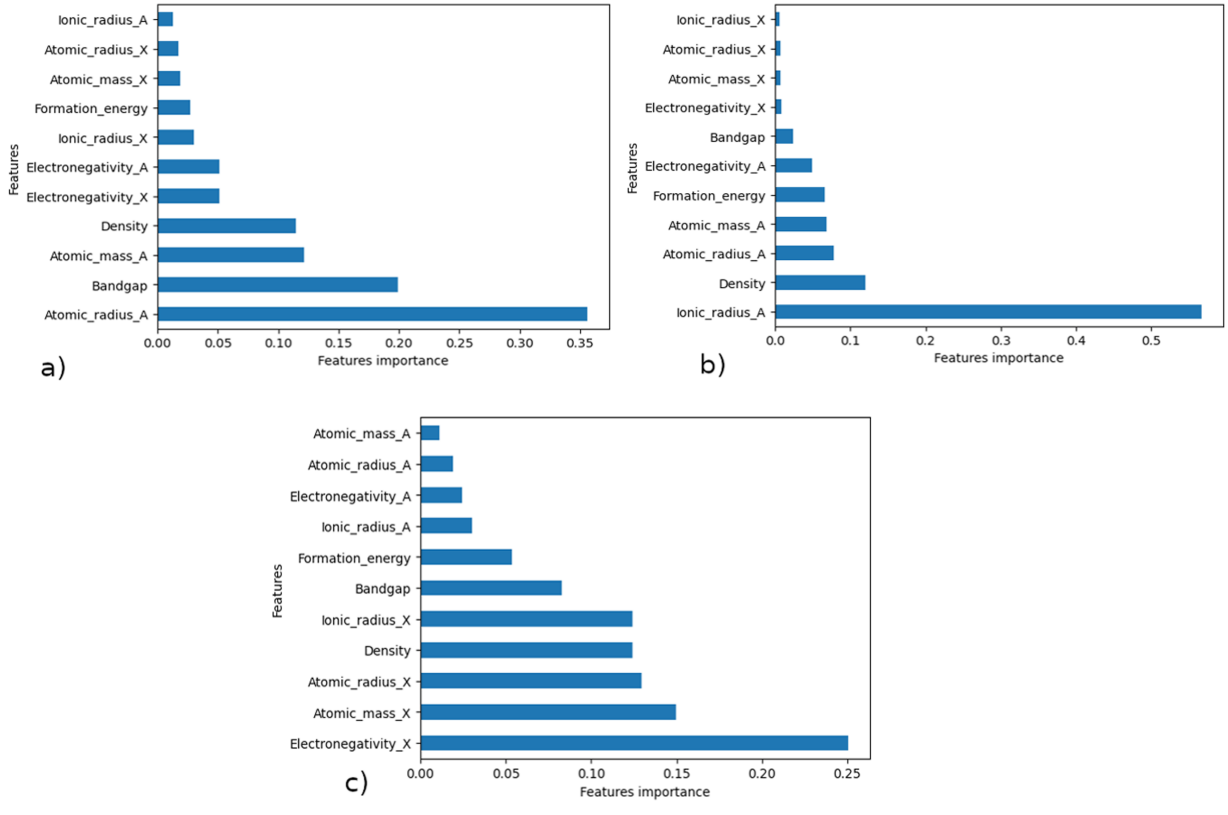

Figure 3. Features vs Features importance plot for determining lattice parameter (a) 'a' (b) 'b' and (c) 'c.'

Thus, KRR shows better accuracy in predicting 'a' than other model. While predicting 'b', the

Table 2 shows that the GBR has the lowest value of MAE and MSE, whereas the highest value of R^2 shows better prediction than another corresponding model. Lastly, on predicting c, again, the GBR model has the lowest value of MAE and MSE with the highest value of R^2. Thus, GBR shows better results in predicting c as well.

Table 2. MAE, MSE and R^2 for three models

Parameters	a ($\AA$)			b ($\AA$)			c ($\AA$)		
Model	KRR	GBR	LR	KRR	GBR	LR	KRR	GBR	LR
MAE	0.072	0.142	0.105	0.068	0.061	0.073	0.298	0.274	0.286
MSE	0.009	0.028	0.014	0.007	0.007	0.008	0.162	0.143	0.147
R2	0.858	0.603	0.799	0.255	0.303	0.147	0.894	0.906	0.904

Features importance plot

Out of all the input features taken, the features that play an important role in predicting the lattice parameters are figured out from the bar diagram. In total, three bar diagram are obtained to observe the several features and their varying importance while individually predicting the lattice parameters a, b and c. The plot is observed using the random forest regressor model. From Fig. 3(a), it is seen that the features that play the most important role in predicting 'a' is atomic radius of element A. Similarly band gap, atomic mass of element A follow after. Here, ionic radius of element A is the one less contributing in predicting 'a'. Similarly, from Fig. 3(b), it is seen that the ionic radius of element A is the main factor contributing to predicting 'b'. Density, atomic radius of A follow after. The ionic radius of element X contributes least to predicting 'b'.

Again, from Fig. 3(c), it is seen that electro-negativity of element X is the main factor contributing in predicting 'c' followed by the atomic mass and radius of element X. Atomic mass of element A is less contributing to predicting 'c'.

IV. Conclusions

The lattice parameters of orthorhombic oxyhalides of the AOX form of the space group Pmmn are predicted using three different models namely KRR, GBR and LR. The train and test data split ratio is set to 70:30 for every model. The validity of the models is observed by calculating MAE, MSE and R^2. On comparing the results, GBR is found to be more efficient in predicting lattice parameters 'b' and 'c', whereas KRR is more efficient in predicting 'a'. These models showed lesser values of MAE and MSE and, at the same time, higher values of coefficient of determination than other models.

Further, the features importance graph is also observed using Random forest regressor model. The features are set in the Y axis and their corresponding importance on the X axis on the features vs features

importance bar-plot. It is found that the atomic radius of element A is the main factor contributing to predicting lattice parameter 'a'. Similarly, the ionic radius of element A and the electro-negativity of element X are the main factors contributing to predicting 'b' and 'c', respectively. This work can be further expanded to the much-needed exploration of oxyhalides, which has yet to be investigated. Such exploration may yield materials with enhanced photocatalytic activity and more efficient optoelectronic properties.

V. Acknowledgements

MPG was supported by the University Grants Commission, Nepal under UGC Grant No. CRG-78/79 S&T-03. This work was partly supported by a grant from UNESCO-TWAS and the Swedish International Development Cooperation Agency (SIDA) with award number 21-377 RG/PHYS/ASG. The views expressed herein do not necessarily represent those of UNESCO-TWAS, SIDA or its Board of Governors. We acknowledge NVIDIA for providing the license package of Deep learning, a part of machine learning.

References

[1] Wells AF, O'Brien T. Structural Inorganic Chemistry. The Journal of Physical Chemistry. 1946;50(5):443-3.

[2] Nait Amar M, Ghriga MA, Ben Seghier MEA, Ouaer H. Prediction of lattice constant of a2xy6 cubic crystals using gene expression programming. The Journal of Physical Chemistry B. 2020;124(28):6037-45.

[3] Ubic R, Subodh G. The prediction of lattice constants in orthorhombic perovskites. Journal of alloys and compounds. 2009;488(1):374-9.

[4] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019;9(4):e1312.

[5] Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. Nanoscale Advances. 2020;2(8):3115-30.

[6] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255-60.

[7] Erik B, Andrew M. The Business of Artificial Intelligence: What It Can—and Cannot—Do for Your Organization. Harvard Business Review Digital Articles. 2017;7:3-11.

[8] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, et al. Machine learning and the

physical sciences. Reviews of Modern Physics. 2019;91(4):045002.

[9] Boehnlein A, Diefenthaler M, Sato N, Schram M, Ziegler V, Fanelli C, et al. Colloquium: Machine learning in nuclear physics. Reviews of Modern Physics. 2022;94(3):031003.

[10] Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR)[Internet]. 2020;9(1):381-6.

[11] Osman AF. Radiation oncology in the era of big data and machine learning for precision medicine. Artificial Intelligence-Applications in Medicine and Biology IntechOpen. 2019:41-70.

[12] Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-58.

[13] Ray S. A quick review of machine learning algorithms. In: 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE; 2019. p. 35-9.

[14] Forkuor G, Hounkpatin OK, Welp G, Thiel M. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PloS one. 2017;12(1):e0170478.

[15] Rong S, Bao-Wen Z. The research of regression model in machine learning field. In: MATEC Web of Conferences. vol. 176. EDP Sciences; 2018. p. 01033.

[16] Tao H, Salih SQ, Saggi MK, Dodangeh E, Voyant C, Al-Ansari N, et al. A newly developed integrative bio-inspired artificial intelligence model for wind speed prediction. IEEE Access. 2020;8:83347-58.

[17] Cherukara MJ, Mannodi-Kanakkithodi A. Deep learning the properties of inorganic perovskites. Modelling and Simulation in Materials Science and Engineering. 2022;30(3):034005.

[18] Ali M, Prasad R, Xiang Y, Yaseen ZM. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. Journal of Hydrology. 2020;584:124647.

[19] Phankokkruad M, Wacharawichanant S. Prediction of mechanical properties of polymer materials using extreme gradient boosting on high molecular weight polymers. In: Complex, Intelligent, and Software Intensive Systems: Proceedings of the 12th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2018). Springer; 2019. p. 375-85.

[20] Zhang Y, Haghani A. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies. 2015;58:308-24.

[21] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 2013;7:21.

[22] Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern recognition letters. 2010;31(14):2225-36.

[23] Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, et al. Pathway analysis using random forests classification and regression. Bioinformatics. 2006;22(16):2028-36.

[24] Jiang F, Kutia M, Sarkissian AJ, Lin H, Long J, Sun H, et al. Estimating the growing stem volume of

coniferous plantations based on random forest using an optimized variable selection method. Sensors. 2020;20(24):7248.

[25] Harishkumar K, Yogesh K, Gad I, et al. Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. Procedia Computer Science. 2020;171:2057-66.