Prediction of lattice parameters of tetragonal oxyhalides AOX

Kashmira Malla* and Madhav Prasad Ghimire*

*Central Department of Physics, Tribhuvan University, Kirtipur, Kathmandu, Nepal.

Abstract: Machine learning enables computers to emulate human intelligence for complex data analysis and pattern recognition. This work utilizes machine learning to predict lattice parameters in tetragonal oxyhalide compounds with molecular formula AOX. Four supervised learning methods - random forest regression, gradient boosting regression, support vector regression, and kernel ridge regression - are employed to forecast lattice parameters from features including atomic radii, ionic radii, atomic masses, electronegativities, band gaps, formation energies, and densities. Model accuracy is evaluated using mean absolute error and R^2 as regression scoring measures. An analysis of gradient boosting regression determines the predictive capacity of distinct features toward lattice parameters. Comparisons identify kernel ridge regression as optimal for predicting lattice constant *a,* with the highest R^2 of 0.840; whereas gradient boosting shows superior in modeling lattice parameter *c* with a maximum R^2 reaching 0.948. This research demonstrates the successful application of machine learning methodologies for predicting material properties, enabling the estimation of lattice parameters in tetragonal oxyhalides.

Keywords: Lattice parameters; Machine learning; Oxyhalides; Supervised learning; Tetragonal system.

Introduction

The increasing use of oxyhalides as photocatalysts for environmental remediation drives efforts to discover new variants exhibiting optimal optoelectronic performance[1-3]. Oxyhalides with the molecular formula AOX, where A comprises main group elements, transition metals, post-transition metals, metalloids, lanthanides, and actinides, O represents oxygen, and X represents halogens (F, Cl, Br, and I), crystallize in all seven crystal systems - cubic, orthorhombic, tetragonal, hexagonal, monoclinic, trigonal, and triclinic. The specific crystal structure expressed depends on the six lattice parameters: the length dimensions of the unit cell sides *a*, *b*, *c*, and the inter-axial angles *α*, *β*, and *γ*. Tetragonal oxyhalides predominantly crystallize in the space groups P4/mmm or P4/nmm, which have $a=b\neq c$ lattice parameters and right-angle $\alpha=\beta=\gamma=90°$ interaxial angles. The schematic diagrams of the unit cell, tetragonal unit cell, and structure of one of the tetragonal oxyhalides, BiOCl (generated from full-potential local-orbital (FPLO) computational software) are shown in Figure 1. Prediction of the lattice parameters is one of the major challenges for both experimental and computational study. Advancements from traditional trial-and-error methods to density functional theory (DFT) calculations still face many difficulties in discovering new materials with desired properties[4].

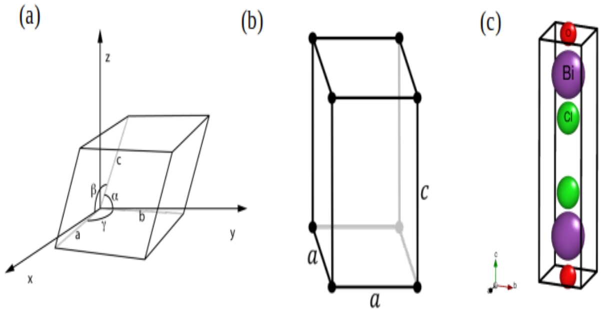

Figure 1: The schematic diagrams of (a) Unit cell[5] and lattice parameters: *a, b, c,* and *α*, *β*, *γ*, (b) Tetragonal unit cell: $a=b\neq c$ and $\alpha=\beta=\gamma=90°$ [6], and (c) Structure of BiOCl (generated from FPLO).

Author for correspondence: Madhav Prasad Ghimire, Central Department of Physics, Tribhuvan University, Kirtipur, Kathmandu, Nepal.
Email: madhav.ghimire@cdp.tu.edu.np; https://orcid.org/0000-0003-2783-4008

Received: 20 Mar, 2024; Received in revised form: 30 Apr, 2024; Accepted: 05 May 2024.
Doi: https://Doi.org/10.3126/sw.v17i17.66435

The advent of artificial intelligence (AI) and machine learning (ML) has helped overcome these research challenges by enabling high-throughput screening and prediction of suitable materials candidates[7]. ML is the mapping from input to output done by providing a large number of examples. There are four main types of ML algorithms: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, the machine has input (sample) as well as output (label) data and hence it predicts output with the help of experiences and examples. In this type of learning, target outputs are present and predicted outputs are compared with actual outputs[8]. In unsupervised learning, the machine is provided with some inputs but the output is unknown. This helps in drawing inferences and may not always provide the correct result as supervised learning. Semi-supervised lies intermediate between supervised and unsupervised learning which works based on few labels and unlabeled data. Reinforcement learning is feedback-based machine learning that performs based on the interaction between algorithm and environment without specific patterns of data-sets[9]. There are several methods to solve the problem using machine learning Regression analysis, Naive Bayes classifiers, Support vector machine (SVM), Decision tree and random forest (RF), Artificial neural network (ANN), Deep learning (DL), linear regression (LR), etc. The choice of method depends on the nature of the available data and the desired analytical outcomes·[10].

Chonge et al. adopted LR and ANN methods to predict the lattice parameters of ABO_3 perovskite. They used ionic radii for the LR model and ionic radii, cation electronegativities, and oxidation states for the ANN model. Further lattice constants a, *b*, and *c* were used to test the ANN model[11] . Majid et al. predicted lattice constants of double perovskite of type $A_2BB'O_6$ by using support vector regression (SVR), ANN, Multiple LR, and SPuDS program[12]. Ganose et al. predicted equilibrium lattice parameters of BiOX with an error of 1% using structural optimization[13]. Ahmad et al. applied ANN and vector regression models to predict Half Heusler compound lattice parameters of general formula XYZ (X and Y atoms have a distinct cationic character and Z has an anionic character) with ionic radii as descriptors, achieving 1.35 % average error[14]. Ma et al. predicted 2D octahedra oxyhalides using SVR, random forest regression (RFR), bagging, and gradient boosting regression (GBR), finding GBR best with the least error and highest coefficient of determination[15]. Williams et al. predicted cubic inorganic perovskite lattice parameters using DL and Hirshfeld surface fingerprints containing geometric and bonding information along with ionic radii, electronegativities, and oxidation states[16]. Zhang & Xu predicted lattice parameters for orthorhombic distorted-perovskite (ABO_3) with high accuracy using Gaussian process regression with ionic radii, electronegativities, and oxidation states as descriptors[17]. Li et al. predicted the generic lattice constant across all the crystal structures reporting the worst performance for monoclinic systems and the best for cubic, concluding higher symmetry enables better prediction. They predicted the cubic parameter *a* using RF and 18,000 samples with $R^2 = 0.973$[5]. Alede et al. studied relating pyrochlore ($A_2B_2O_7$) properties to lattice constants, predicting constants with ANN and SVR using ionic radii and electronegativity, finding SVR more accurate[18] .

Previous works on parameter prediction suggest that the choice of ML algorithms and descriptor data impacts model performance. Many studies have compared prediction accuracy across different crystal systems, highlighting the role of symmetry. However, direct comparisons of different models on the same crystal system are rare. Additionally, most existing research has focused on either generic systems or specifically perovskites, while studies on other systems like oxyhalides have been limited. To address these gaps and given the growing applications of tetragonal oxyhalides, we have proposed this comparative study. We developed and tested four ML models: RFR, GBR, SVR, and kernel ridge regression (KRR) for predicting lattice parameters of tetragonal oxyhalides. By evaluating multiple models on the same crystal system, our work provides insight into the relative performance of different regression algorithms for parameter prediction in oxyhalides.

Materials and Methods

Data Processing and Feature Engineering

Data processing includes data collection, data cleaning, and normalization[9]. Data are collected from the materials project database[19] (https://materialsproject.org/) and installed gelemental 1.2.0 package of ubuntu. A total of 52 AOX with space groups P4/mmm and P4/nmm (where, A is main group element, transition metals, post-transition metals, metalloids, lanthanides and actinides, X is halogens (such as F, Cl, Br, and I)) were selected. These data are then normalized by using min-max normalization given by equation (1)[20].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad \ldots(1)$$

In the feature engineering section, the chosen descriptors (features) were atomic radii, ionic radii, atomic masses, electronegativities, band gaps (BG), formation energy (FE), and the densities of AOX. The target parameters, meaning the parameters that needed to be predicted and compared to actual values, were the lattice parameters *a* and *c*. The values in the collected data for the actual lattice parameter *a* range from 3.57 Å to 4.31 Å, while the values for c range from 5.17 Å to 9.91 Å. The data collected are listed in Table 1.

Machine Learning Models: RFR, GBR, SVR, and KRR

Lattice parameters prediction was performed using four ML regression methods: two ensemble methods - RFR and GBR - as well as two kernel-based techniques - SVR and KRR. RF is the extension form of bagging in which base models are aggregated to reduce the variance of base models without increasing the bias whereas, the correlation between individual members limits the reduction in variance. The variance reduction can be achieved through RFR by assuming the base models as regression trees in which additional randomness is injected while constructing each tree to reduce correlation among the base models[21]. After building K numbers of regression trees T(x) and averaging the result, the RFR predictor is given in equation (2)[22].

$$f_{rf}^{K} = \frac{1}{K}\sum_{k=1}^{K} T(x) \qquad \ldots(2)$$

Boosting is an ensemble of base models in which each base model is trained sequentially such that weak base models are converted to one strong model. Different types of boosting include AdaBoost, gradient boosting, and XGBoost. Gradient boosting can be used for classification and regression problems with a gradient-descent-based formulation to build a statistical framework. There are three main components of GBR: loss function, weak learner, and additive model. The loss function is optimized to reduce error, weak learners are used to make predictions. In the additive model, an error is reduced by adding decision trees[23]. From the additive model, the function approximator is given in equation (3).

$$f^{n}(x) = \sum_{i=1}^{n} \alpha^{(i)} f^{(i)}(x) \qquad \ldots(3)$$

where, $\alpha^{(i)}$ are real-valued coefficients and $f^{(i)}$ are basis functions.

In sequential boosting method i^{th} iteration is defined in equation (4).

$$f^{i}(x) = f^{(i-1)}(x) + \alpha^{(i)} f^{(i)}(x) \qquad \ldots(4)$$

SVR and KRR belong to kernel-based non parametric supervised learning ML model[22] where SVR find the hyperplane of *N-1* dimension for *N* dimensional group of points[10].

Suppose $\{(x_{1,}y_1), \ldots, (x_n, y_n)\} \subset \chi \times R$ be the training data where *x* represent feature and *y* represent target variable. Thus in linear epsilon-SVR, we have to find function *f (x)* such that error is less than *ε*.

The linear function *f* is defined in equation (5)[24].

$$f(x) = \langle w, x\rangle + b; where, w\epsilon\chi, b\epsilon R \quad .(5)$$

KRR is a special case or the simplified version of SVR[25]. Through the KRR model nonlinear problems can be solved by using the kernel function to replace the dot product in the SVR model[26]. The algorithms of four models were implemented using Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Tensorflow, and Keras libraries in Python 3.9

Table 1. Collected data of tetragonal oxyhalides AOX for parameters prediction.

		Atomic radii (Å)		Ionic radii (Å)		Atomic masses (g/mol)		EN (Pauli)					Lattice parameters (Å)	
S.N.	Compounds	A	X	A	X	A	X	A	X	BG (eV)	FE (eV/atom)	Density (g/cc)	a	c
1	ScOCl	1.6	1	0.723	1.81	44.955912	35.453	1.36	3.16	3.61	-3.279	3.16	3.57	7.96
2	YOCl	1.8	1	0.893	1.81	88.90585	35.453	1.22	3.16	5.09	-3.475	4.48	3.93	6.73
3	LaOCl	1.95	1	1.016	1.81	138.90547	35.453	1.10	3.16	4.07	-3.605	5.32	4.13	6.96
4	CeOCl	1.85	1	1.034	1.81	140.116	35.453	1.12	3.16	0.00	-3.410	5.56	4.06	6.93
5	PrOCl	1.85	1	1.013	1.81	140.90765	35.453	1.13	3.16	4.70	-3.458	5.50	4.11	6.88
6	NdOCl	1.85	1	0.995	1.81	144.242	35.453	1.14	3.16	4.77	-3.471	5.72	4.07	6.84
7	SmOCl	1.85	1	0.964	1.81	150.36	35.453	1.17	3.16	4.88	-3.490	6.12	4.01	6.79
8	EuOCl	1.85	1	0.95	1.81	151.964	35.453	1.20	3.16	0.00	-2.886	6.11	4.03	6.81
9	GdOCl	1.8	1	0.938	1.81	157.25	35.453	1.20	3.16	3.04	-3.478	6.52	3.97	6.75
10	DyOCl	1.75	1	0.908	1.81	162.5	35.453	1.22	3.16	5.17	-3.515	6.91	3.91	6.72
11	HoOCl	1.75	1	0.894	1.81	164.93032	35.453	1.23	3.16	5.18	-3.513	7.07	3.89	6.72
12	ErOCl	1.75	1	0.881	1.81	167.259	35.453	1.24	3.16	5.19	-3.510	7.22	3.88	6.67
13	TmOCl	1.75	1	0.87	1.81	168.93421	35.453	1.25	3.16	5.09	-3.514	7.38	3.85	6.70
14	AcOCl	1.95	1	1.18	1.81	227	35.453	1.10	3.16	4.45	-3.586	7.06	4.27	7.28
15	PuOCl	1.75	1	1.08	1.81	224	35.453	1.28	3.16	0.00	-3.084	8.96	3.99	6.87
16	BiOCl	1.6	1	0.96	1.81	208.9804	35.453	2.02	3.16	2.57	-1.696	7.21	3.91	7.83
17	ScOF	1.6	0.5	0.723	1.33	44.955912	18.9984032	1.36	3.98	4.33	-4.043	3.90	3.63	5.17
18	YOF	1.8	0.5	0.893	1.33	88.90585	18.9984032	1.22	3.98	4.95	-4.220	4.96	3.88	5.52
19	LaOF	1.95	0.5	1.016	1.33	138.90547	18.9984032	1.10	3.98	54.57	-4.257	5.89	4.09	5.86
20	CeOF	1.85	0.5	1.034	1.33	140.116	18.9984032	1.12	3.98	0.00	-4.083	6.20	4.03	5.77
21	GdOF	1.8	0.5	0.938	1.33	157.25	18.9984032	1.20	3.98	2.93	-4.184	7.40	3.93	5.60
22	PuOF	1.75	0.5	1.08	1.33	224	18.9984032	1.28	3.98	0.00	-3.751	10.27	3.97	5.73
23	BiOF	1.6	0.5	0.96	1.33	208.9804	18.9984032	2.02	3.98	2.89	-2.216	8.66	4.01	5.81
24	ScOBr	1.6	1.15	0.723	1.96	44.955912	79.904	1.36	2.96	3.35	-3.124	4.13	3.64	8.57
25	YOBr	1.8	1.15	0.893	1.96	88.90585	79.904	1.22	2.96	4.48	-3.315	4.66	3.87	8.79
26	LaOBr	1.95	1.15	1.016	1.96	138.90547	79.904	1.10	2.96	3.73	-3.412	5.90	4.15	7.67
27	CeOBr	1.85	1.15	1.034	1.96	140.116	79.904	1.12	2.96	0.00	-3.229	5.93	4.02	8.18
28	PrOBr	1.85	1.15	1.013	1.96	140.90765	79.904	1.13	2.96	4.46	-3.260	6.10	4.10	7.67
29	NdOBr	1.85	1.15	0.995	1.96	144.242	79.904	1.14	2.96	4.48	-3.284	5.81	4.02	8.50
30	SmOBr	1.85	1.15	0.964	1.96	150.36	79.904	1.17	2.96	4.52	-3.311	6.01	3.95	8.70
31	EuOBr	1.85	1.15	0.95	1.96	151.964	79.904	1.20	2.96	0.00	-2.718	6.29	3.96	8.36
32	GdOBr	1.8	1.15	0.938	1.96	157.25	79.904	1.20	2.96	0.08	-3.312	6.28	3.91	8.78
33	HoOBr	1.75	1.15	0.894	1.96	164.93032	79.904	1.23	2.96	4.50	-3.357	6.64	3.83	8.88
34	ErOBr	1.75	1.15	0.881	1.96	167.259	79.904	1.24	2.96	4.38	-3.355	6.94	3.82	8.62
35	TmOBr	1.75	1.15	0.87	1.96	168.93421	79.904	1.25	2.96	4.35	-3.363	7.07	3.80	8.61
36	YbOBr	1.75	1.15	0.858	1.96	173.04	79.904	1.10	2.96	0.00	-2.377	7.21	3.68	9.13
37	LuOBr	1.75	1.15	0.85	1.96	174.967	79.904	1.27	2.96	4.40	-3.357	7.26	3.75	8.81
38	AcOBr	1.95	1.15	1.18	1.96	227	79.904	1.10	2.96	4.24	-3.396	7.50	4.31	7.70
39	PuOBr	1.75	1.15	1.08	1.96	244	79.904	1.28	2.96	0.00	-2.915	8.49	3.94	8.56
40	BiOBr	1.6	1.15	0.96	1.96	208.9804	79.904	2.02	2.96	2.23	-1.555	7.29	3.95	8.88
41	YOI	1.8	1.4	0.893	2.2	88.90585	126.90447	1.22	2.66	3.41	-3.063	5.14	3.95	9.59
42	LaOI	1.95	1.4	1.016	2.2	138.90547	126.90447	1.10	2.66	3.29	-3.170	5.53	4.17	9.75
43	PrOI	1.85	1.4	1.013	2.2	140.90765	126.90447	1.13	2.66	3.66	-3.025	5.60	4.12	9.89
44	NdOI	1.85	1.4	0.995	2.2	144.242	126.90447	1.14	2.66	3.57	-3.038	5.93	4.09	9.63
45	SmOI	1.85	1.4	0.964	2.2	150.36	126.90447	1.17	2.66	3.32	-3.059	6.33	4.04	9.43
46	EuOI	1.85	1.4	0.95	2.2	151.964	126.90447	1.20	2.66	0.00	-2.500	6.28	4.08	9.38
47	HoOI	1.75	1.4	0.894	2.2	164.93032	126.90447	1.23	2.66	3.45	-3.098	6.89	3.93	9.60
48	TmOI	1.75	1.4	0.87	2.2	168.93421	126.90447	1.25	2.66	3.39	-3.105	6.91	3.89	9.91
49	LuOI	1.75	1.4	0.85	2.2	174.967	126.90447	1.27	2.66	3.27	-3.094	7.16	3.86	9.89
50	NpOI	1.75	1.4	1.1	2.2	237	126.90447	1.36	2.66	0.00	-2.451	8.48	4.00	9.30
51	PuOI	1.75	1.4	1.08	2.2	244	126.90447	1.28	2.66	0.00	-2.021	8.44	4.01	9.47
52	BiOI	1.6	1.4	0.96	2.2	208.9804	126.90447	2.02	2.66	1.47	-1.335	7.38	4.03	9.76

Model Validation and Evaluation

While predicting parameters from ML, it is equally important to validate the models to test the flexibility of that model with the data selected. In order to validate each model, 5-fold cross-validation along with grid search best model hyper-parameters technique has adopted such that this process is repeated 5 times to get average estimated parameters and performance score. The value for maximum depth is 4, number of estimators is 500 and random state 5 has been taken in all four models. Two scoring measures: mean absolute error (MAE) and coefficient of determination (R^2) were used to compare the performance accuracy of different models. The MAE given by equation (6) calculates the average magnitude of errors. Lower MAE values indicate better model performance. The R^2 score, given by equation (7), represents the proportion of variance in the dependent variable that is predictable from the independent variables. R^2 ranges from 0 to 1, with higher values indicating more variance explained by the model[5].

$$MAE = \frac{1}{n}\sum_{i=1}|y_i - \hat{y}_i| \qquad \ldots(6)$$

$$R^2 = \left(\frac{n\sum(y_i\hat{y_i}) - \sum y_i \sum \hat{y}_i}{\sqrt{[n\sum y_i{}^2 - (\sum y_i)^2]\,[n\sum \hat{y}_i{}^2 - (\sum \hat{y}_i)^2]}}\right)^2 \qquad \ldots(7)$$

where, y_i and $\hat{y}_i$ are estimated and predicted values respectively.

Results

Parameters Prediction

Lattice parameters a and c were predicted using four ML methods – RFR, GBR, SVR, and KRR, validated through a 5-fold cross-validation approach. Of the 52 total data points, 10 were retained as the test set for model evaluation. The actual and predicted lattice constants are tabulated in Table 2 and visualized in Figure 2. Examining the scatter plots, prediction of parameter a in Figure 2(a) shows the SVR outputs exhibiting maximal deviation from the true values, while KRR demonstrates superior congruence, with RFR and GBR having intermediate clustering closer to the actual values. However, prediction performance is reversed for the c parameter in Figure 2(b). In this case, KRR produces greatest variability versus the reliable GBR equivalency mapping to actual values. This implies that no individual method maintains consistent accuracy over both outputs, affirming need for comparison. Quantitative scoring of predictor errors reveals a full ranking of effectiveness, further discussed in the following section.

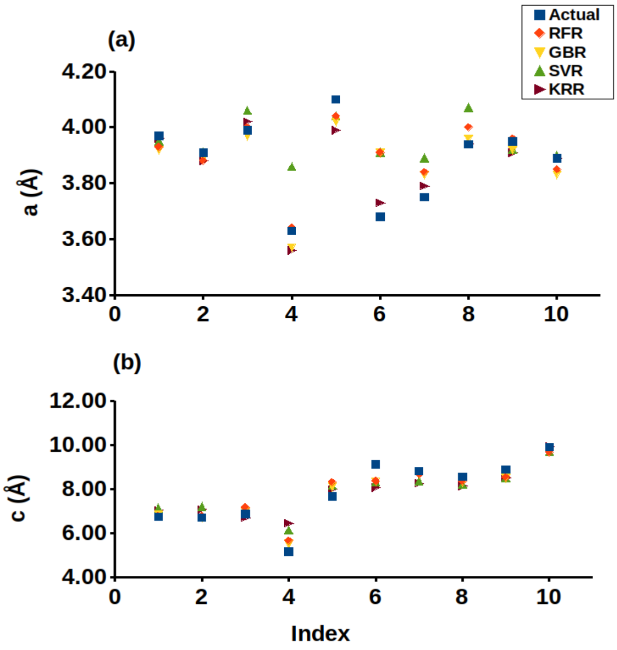

Figure 2: Graph of actual and predicted lattice parameters (a) *a* and (b) *c* using four models.

Table 2. Actual and predicted lattice parameters from four models.

				Predicted							
		Actual		RFR		GBR		SVR		KRR	
S.N	Compounds	a (Å)	c (Å)	a (Å)	c (Å)	a (Å)	c (Å)	a (Å)	c (Å)	a (Å)	c (Å)
1	GdOCl	3.97	6.75	3.93	6.79	3.92	6.85	3.95	7.13	3.96	7.04
2	DyOCl	3.91	6.72	3.88	6.76	3.89	6.72	3.91	7.21	3.88	7.08
3	PuOCl	3.99	6.87	4.00	7.17	3.97	6.98	4.06	7.04	4.02	6.71
4	ScOF	3.63	5.17	3.64	5.67	3.57	5.54	3.86	6.14	3.56	6.45
5	PrOBr	4.10	7.67	4.04	8.31	4.02	8.10	4.04	8.17	3.99	7.99
6	YbOBr	3.68	9.13	3.91	8.36	3.91	8.35	3.91	8.32	3.73	8.07
7	LuOBr	3.75	8.81	3.84	8.68	8.83	8.69	3.89	8.36	3.79	8.27
8	PuOBr	3.94	8.56	4.00	8.36	3.96	8.30	4.07	8.20	3.94	8.15
9	BiOBr	3.95	8.88	3.96	8.53	3.92	8.51	3.92	8.48	3.91	8.15
10	TmOI	3.89	9.91	3.85	9.66	3.83	9.74	3.90	9.71	3.89	9.92

Accuracy and Validation

To evaluate model prediction accuracy on the validation data, MAE and R^2 were computed as measures of discrepancy and equivalence between the actual data and ML-predicted lattice parameters. Table 3 reports the average values across the 5 folds for each method's MAE and R^2 score, while the bar plots in Figure 3 visually profile comparative model performance.

Table 3. MAE and R^2 in the prediction of lattice parameters from four models.

Parameters	*a* (Å)				*c* (Å)			
Model	RFR	GBR	SVR	KRR	RFR	GBR	SVR	KRR
MAE	0.051	0.047	0.060	0.048	0.231	0.225	0.379	0.481
R^2	0.753	0.798	0.655	0.840	0.945	0.948	0.881	0.867

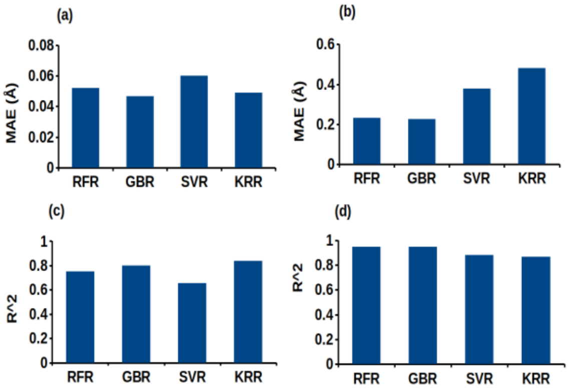

Figure 3: Bar graph of (a) MAE for *a*, (b) MAE for *c*, (c) R^2 for *a*, and (d) R^2 for *c* for four ML models.

In the evaluation of various ML models for predicting lattice parameters, Figure 3(a) illustrates that the GBR model yields the minimum MAE when predicting parameter *a* while the SVR model exhibits the maximum MAE. Conversely, for the prediction of parameter *c* (Figure 3(b)), the GBR model demonstrates the minimum MAE, while the KRR model exhibits the maximum. Turning to the R^2 values, Figure 3(c) indicates that the KRR model achieves the highest R^2 value in predicting parameter *a*, while the SVR model attains the least. Meanwhile, in the prediction of parameter *c* (Figure 3(d)), the GBR model outperforms others with the highest R^2 value, and the KRR model records the lowest.

The observed MAE and R^2 values generally suggest an inverse relationship - lower MAE tends to correspond to higher R^2. However, this correlation is not universally applicable. Given that MAE only represents the absolute difference between actual and predicted values of lattice parameters, emphasis is placed on R^2 values for accuracy assessment. Notably, the KRR model emerges as the most accurate predictor for parameter *a* with an R^2 value of 0.840, while the GBR model excels in predicting parameter *c* with an R^2 value of 0.948.

To delve into the impact of the 11 features (atomic radius of A, atomic radius of X, ionic radius of A, ionic radius of X, atomic mass of A, atomic mass of X, electronegativity of A, electronegativity of X, band gap, formation energy, and density) on lattice parameters prediction, Figure 4 depicts the feature importance using the GBR model. This analysis aids in understanding the relative significance of each feature in the predictive performance of the model.

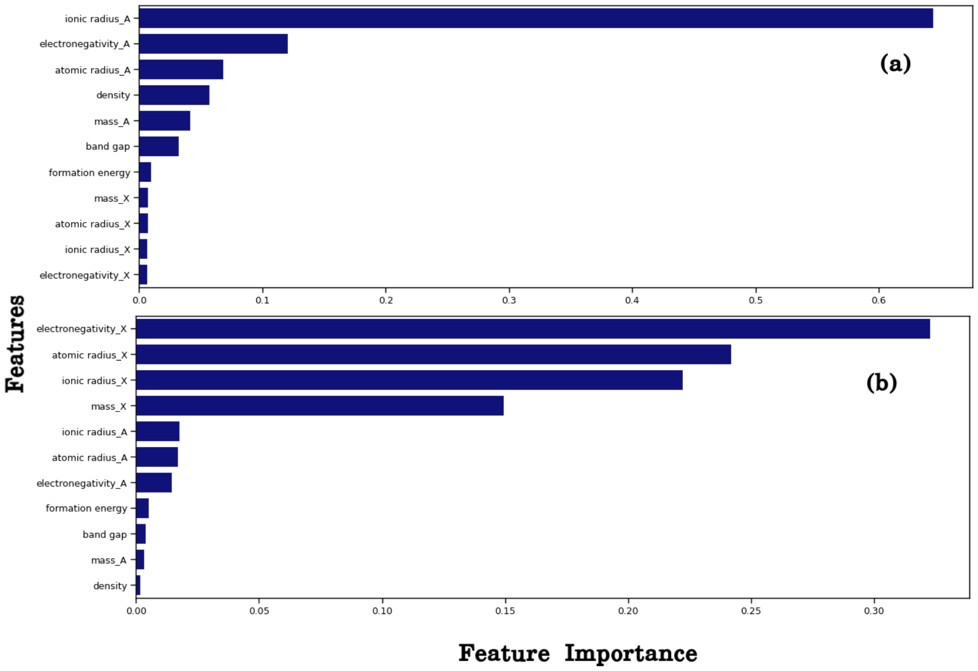

Figure 4: Features importance for the estimation of lattice parameters (a) *a* and (b) *c* by GBR model.

In Figure 4(a), the feature importance plot derived from the GBR method for predicting lattice parameter *a* reveals distinctive contributions from various features. Approximately 65% of the predictive power for parameter *a* is attributed to the ionic radius of A, indicating its dominant role. The electronegativity of A follows closely, contributing around 12% to the predictive accuracy. Additional contributors include atomic radius and density, each accounting for approximately 8-9% of the overall predictive influence. The mass of A and band-gap features play a moderate role, each contributing about 5% to the predictive capability. On the other hand, the contribution of formation energy and features associated with element X (including mass, atomic radius, ionic radius, and electronegativity of X) is relatively minor, each contributing less than 2% to the overall predictive importance.

In Figure 4(b), the feature importance plot for predicting lattice parameter *c* provides insights into the key contributors. Notably, about 34% of the predictive influence for parameter *c* is attributed to the electronegativity of X, establishing it as the primary contributor. Following closely, the atomic radius of X contributes significantly, accounting for approximately 24% of the overall predictive power. The ionic radius of A emerges as another substantial factor, contributing about 22% to the accurate prediction of lattice parameter *c*. The mass of X plays a notable role as well, contributing around 15% to the predictive capability. Conversely, the ionic radius, atomic radius, and electronegativity of A each make more modest contributions, approximately 2% each. Formation energy, band gap, mass of A, and density contribute less than 1% individually.

This analysis underscores the distinct contributions of various features in predicting lattice parameters *a* and *c*. Specifically, it highlights that the ionic radius of A and the electronegativity of X play pivotal roles in predicting *a* and *c* respectively. Furthermore, the delineation of feature contributions suggests that features associated with element A have a more substantial impact on the prediction of *a*, while features from element X contribute more significantly to the prediction of *c*.

Discussions

The comparative analysis reveals that the KRR outperforms SVR, RFR, and GBR models in predicting lattice parameter *a*. Conversely, for the prediction of lattice parameter *c*, the GBR model exhibits superior performance compared to SVR, KRR, and RFR.

The observed higher MAE in determining parameter *c* as opposed to *a* across all four models can be attributed to the inherent differences in the parameter range for *a* and *c*. Given that MAE provides only the absolute error between actual and predicted values, its limitations in adequately describing model accuracy are acknowledged. Consequently, R^2 is incorporated into the assessment. The R^2 analysis aligns with the comparison of actual and predicted data, supporting that the R^2 is a more comprehensive measure for evaluating the prediction of lattice parameters. Notably, R^2 proves to be comparatively better in predicting parameter *c* than *a* for all four models. This discrepancy can be attributed to the increased number of contributing features for the prediction of *c* as illustrated in Figure 4.

It's noteworthy that the exclusion of oxygen from AOX in determining lattice parameters is a deliberate choice in this study. The absence of oxygen in the consideration is justified based on the results and the focus on the contributing features from elements A and X.

Conclusions

This work demonstrates ML techniques for predicting tetragonal oxyhalide lattice parameters. Four models were developed - RFR, GBR, SVR, and KRR to forecast the *a* and *c* lattice parameters of AOX compounds. Results show kernel-based methods exhibit superior accuracy in modeling lattice constant *a*, while ensemble models perform best predicting parameter *c*. The models themselves prove robust and stable predictors. However, the dataset size was limited to 52 compounds, restricting model training and evaluation. While findings clearly establish the potential of employing different machine learning approaches to optimize property estimates for materials discovery, more data is needed to further improve generalization performance and confirm predictive capabilities. Extending this framework across additional crystal systems could also further validate predictive capacities given expanded data availability. To comprehensively capture the diverse range of potential lattice configurations, it is essential to conduct a more extensive sampling of the feature space. This broader exploration will ensure a more representative representation of the various lattice arrangements and contribute to a more robust understanding of the system under study. Overall model accuracy measures were reasonable but not near perfect, indicating room for advancement with more and better quality training data. Additionally, model selection was tailored to individual outputs, while a single multi-output model could provide a more elegant approach. Ultimately, this research enables targeted oxyhalide exploration and design by accelerating the identification of variants with enhanced functionality as optoelectronic and photocatalytic materials. The lattice parameter estimations serve as useful inputs to subsequent DFT computations, with model selection tailored to maximize predictive accuracy.

Acknowledgments

MPG was supported by the University Grants Commission, Nepal under UGC Grant No. CRG-78/79 S&T-03. This work was also supported in part by a grant from UNESCO-TWAS and the Swedish International Development Cooperation Agency (SIDA) with award number 21-377 RG/PHYS/ASG. We acknowledge NVIDIA for providing the license package of deep learning, a part of machine learning. We thank Dr. Rajendra Adhikari, Kathmandu University for the technical assistance.

References

1. Di, J., Xia, J., Li, H., Guo, S. and Dai, S. 2017. Bismuth oxyhalide layered materials for energy and environmental applications. *Nano Energy.* **41**: 172–192.
   Doi: https://doi.org/10.1016/j.nanoen.2017.09.00
2. Wang, Z., Chen, M., Huang, D., Zeng, G., Xu, P., Zhou, C., Lai, C., Wang, H., Cheng, M. and Wang, W. 2019. Multiply structural optimized strategies for bismuth oxyhalide photocatalysis and their environmental application. *Chem. Eng. J.* **374**: 1025–104.
   Doi: https://doi.org/10.1016/j.nanoen.2017.09.008
3. Chen, X. and Ok, K. M. 2022. Metal oxyhalides: an emerging family of nonlinear optical materials. *Chem. Sci.* **13**(14): 3942–3956.
   Doi: 10.1039/D1SC07121A
4. Juan, Y., Dai, Y., Yang, Y. and Zhang, J. 2021. Accelerating materials discovery using machine learning. *J. Mater. Sci. Technol.* **79**: 178–190.
   Doi: https://doi.org/10.1016/j.jmst.2020.12.010
5. Li, Y., Yang, W., Dong, R. and Hu, J. 2021. Mlatticeabc: generic lattice constant prediction of crystal materials using machine learning. ACS *Omega.* **6**(17): 11585–11594.
   Doi: https://doi.org/10.1021/acsomega.1c00781
6. Mayer, D. and Stannered. 2007. Tetragonal crystal structure [Digital image]. Retrieved: 01-Jan-2023.
   URL:https://commons.wikimedia.org/wiki/File:Tetragonal.svg
7. Liu, Y., Zhao, T., Ju, W. and Shi, S. 2017. Materials discovery and design using machine learning. *J Mater.* **3**(3): 159–177.
   Doi: https://doi.org/10.1016/j.jmat.2017.08.002
8. Natekin, A. and Knoll, A. 2013. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**:21.
   Doi: 10.3389/fnbot.2013.00021
9. Ayodele, T. O. 2010. Types of machine learning algorithms, new advances in machine learning. *Yagang Zhang edition, InTech.*
10. Cai, J., Chu, X., Xu, K., Li, H. and Wei, J. 2020. Machine learning-driven new material discovery. *Nanoscale. Adv.* **2**(8): 3115–3130.
    Doi: 10.1039/D0NA00388C
11. Chonghe, L., Yihao, T., Yingzhi, Z., Chunmei, W. and Ping, W. 2003. Prediction of lattice constant in perovskites of $GdFeO_3$ structure. *J. Phys. Chem. Solids.* **64**(11): 2147–2156.
    Doi: https://doi.org/110.1016/S0022-3697(03)00209-9
12. Majid, A., Farooq Ahmad, M. and Choi, T. S. 2009. Lattice constant prediction of $A_2BB'O_6$ type double perovskites. *In Computational Science and Its Applications-ICCSA 2009.* Seoul Korea. Pp: 82-92.
13. Ganose, A. M., Cuff, M., Butler, K. T., Walsh, A. and Scanlon, D. O. 2016. Interplay of orbital and relativistic effects in bismuth oxyhalides: BiOF, BiOCl, BiOBr, and BiOI. *Chem. Mater.* **28**(7): 1980–1984.
    Doi: https://doi.org/10.1021/acs.chemmater.6b00349
14. Ahmad, R., Gul, A. and Mehmood, N. 2019. Artificial neural networks and vector regression models for prediction of lattice constants of half-heusler compounds. *Mater. Res. Express.* **6**(4): 046517.
    Doi: 10.1088/2053-1591/aafa9f
15. Ma, X.-Y., Lewis, J. P., Yan, Q.-B. and Su, G. 2019. Accelerated discovery of two-dimensional optoelectronic octahedral oxyhalides via high-throughput ab initio calculations and machine learning. *J. Phys. Chem.* Lett. **10**(21): 6734–6740.
    Doi: https://doi.org/10.1021/acs.jpclett.9b02420

16. Williams, L., Mukherjee, A. and Rajan, K. 2020. Deep learning based prediction of perovskite lattice parameters from hirshfeld surface fingerprints. *J. Phys. Chem.* Lett. **11**(17): 7462–7468. Doi: https://doi.org/10.1021/acs.jpclett.0c02201

17. Zhang, Y. and Xu, X. 2021. Predicting lattice parameters for orthorhombic distorted-perovskite oxides via machine learning. *Solid State Sci*. **113**: 106541. Doi: https://doi.org/10.1016/j.solidstatesciences.2021.106541

18. Alade, I. O., Oyedeji, M. O., Rahman, M. A. A. and Saleh, T. A. 2022. Prediction of the lattice constants of pyrochlore compounds using machine learning. *Soft Comput.* **26**(17): 8307–8315. Doi: https://doi.org/10.1007/s00500-022-07218-1

19. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. et al. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater*. **1**(1): 011002. Doi: https://doi.org/10.1063/1.4812323

20. Han, J., Kamber, M. and Pei, J. 2011. Data Mining: Concepts and Techniques. *Elsevier*.

21. Lindholm, A., Wahlström, N., Lindsten, F. and Schön, T. B. 2022. *Machine Learning: A First Course for Engineers and Scientists.* Cambridge University Press.

22. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev*. **71**: 804–818. Doi: 10.1016/j.oregeorev.2015.01.001

23. Alade, I. O., Zhang, Y. and Xu, X. 2021. Modeling and prediction of lattice parameters of binary spinel compounds (am 2x4) using support vector regression with bayesian optimization. New J. Chem. **45**(34): 15255–15266. Doi: https://d0i.org/10.1039/D1NJ01523K

24. Smola, A. J. and Schölkopf, B. 2004. A tutorial on support vector regression. *Stat Comput.* **14**(3): 199–222. Doi: https://doi.org/10.1023/B:STCO.0000035301.49549.88

25. Vovk, V. 2013. *Empirical Inference.* V edition, Springer, Berlin, Heidelberg.

26. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. and Data, M. 2005. Practical machine learning tools and techniques. Third edition *Elsevier*. Amsterdam, Netherlands.

[redacted]