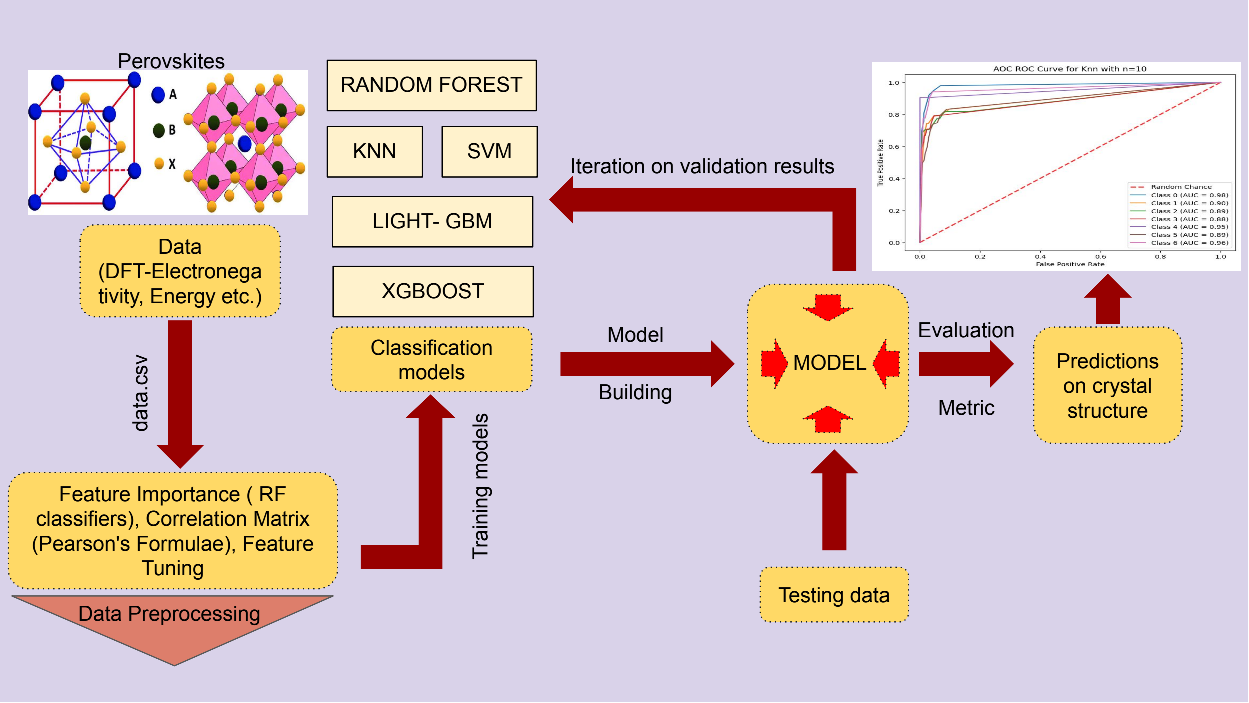
Graphical Abstract

Determination of lattice structure of perovskites using ML

Kaling Vikram Singh, S Sunil Raja, Subhankar Mishra



Highlights

Determination of lattice structure of perovskites using ML

Kaling Vikram Singh, S Sunil Raja, Subhankar Mishra

- Database collection based on DFT model.
- Machine learning models to study mixed perovskites.
- Prediction of the crystal structure of perovskites.

Determination of lattice structure of perovskites using ML

Kaling Vikram Singh[a], S Sunil Raja[a], Subhankar Mishra[b]

[a]*School of Physical Sciences, National Institute of Science Education and Research (NISER), Bhubaneswar, Odisha, 752050, India*
[b]*School of Computer Sciences, National Institute of Science Education and Research (NISER), Bhubaneswar, Odisha, 752050, India*

Abstract

Perovskites have been studied extensively in fields of solid state physics, inorganic chemistry and ceramic engineering due to their special properties. The most significant factor that affects these properties is the crystal structure. The determination of crystal structure through modern methods such as X-ray diffraction (X-RD) or Density functional theory (DFT) are time consuming, computationally expensive and costly. through this study, we suggest machine learning algorithms as an alternative method to correctly classify the crystal structure. Basic properties of the materials such as valency, atomic radii, bang gap, etc are used to predict the crystal structure. Various classification models were implemented, using which, the crystal structures Were predicted. Support vector machines (SVM), random forest (RF), Light gradient boosting machine (LGBM), XGBoost, Convolution neural networks (CNN), and K-nearest neighbors (KNN) are some of the models are used for the study. It was observed that weighted SVMs were the best model for determination of structure of mixed perovskites, with an average accuracy of 92.03% while random forest was the best model to predict the crystal structure of mixed perovskites with an average accuracy of 94.40%. Through this study, we suggest a time and cost effective alternative for crystal structure determination of perovskites.

Keywords: Perovskites, Machine learning, Crystal structure, Density functional theory (DFT), Classification model, Materials Project API

* Correspondence : kalingvikram.singh@niser.ac.in

Preprint submitted to Computational material sciences *March 5, 2024*

1. Introduction

Perovskites are class of compounds having crystal structures similar to calcium titanium, ABX_3, where A and B represent large and small cations respectively with X being an anion. The perovskites with X as oxygen are termed as ”oxygen perovskite” while X as halides or other elements, but not excluding oxygen as ”mixed perovskites”. Their structures have been extensively studied in inorganic chemistry, ceramic engineering and material physics due to their special properties[1],[2]. Perovskites are found with different structures (fig: 1), viz. Monoclinic, Orthorhombic, Tertahedral, Hexagonal and Rhombohedral based on combinations of A and B. Such varied structures are a result of: (i) displacement of the cations, (ii) distortion of the orbitals and (iii) tilting of the orbitals due to steric and electrostatic effects, which are instability driven factors and are dependent on the properties of A and B.

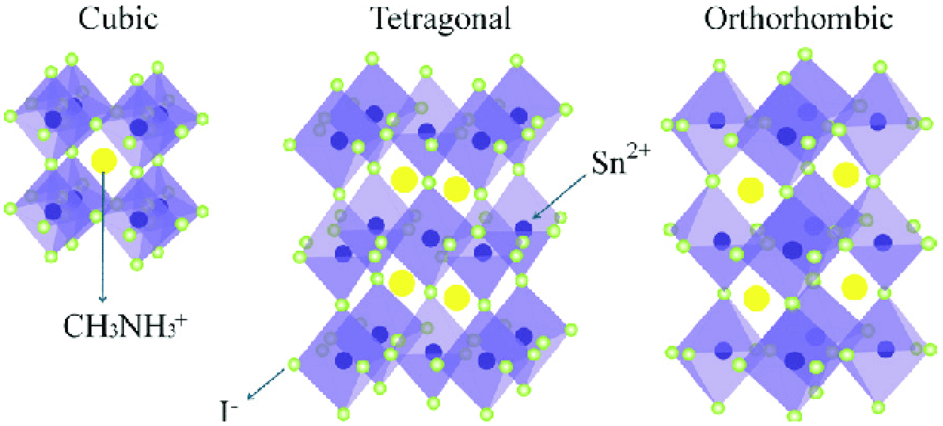


Figure 1: Different crystal structure of perovskites.[3]

Perovskite halides exhibit special characteristics, such as exceptional absorption coefficients, high carrier mobility, and prolonged charge carrier lifetimes, that make them promising for optoelectronic and photovoltaic applications. Perovskites have multiple industrial applications, including fabrication of solar panel, sensors, photovoltaic and memory devices. Perovskite halide solar cells have demonstrated noteworthy power conversion efficiencies, which are comparable to conventional silicon-based solar cells. Using perovskite halides in materials also exhibits certain obstacles, including stability problems and potential toxicity due to the use of lead in some formulations. Thus, in order to use perovskites in real life, the properties and structures of these complex compounds should be studied thoroughly.

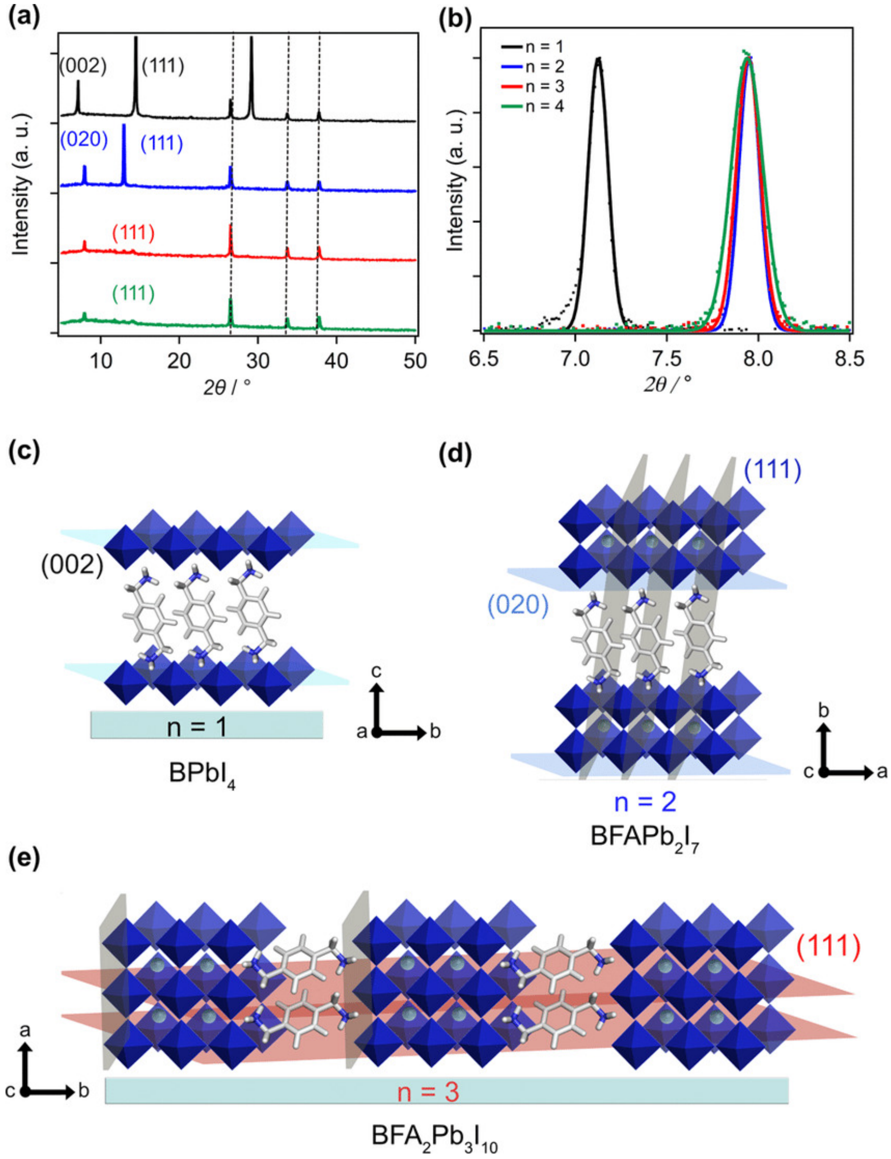


Figure 2: XRD : To find structural properties of 2-D perovskite.[4]

Materials in which the current flows through movement of oxide ions through the crystal lattice are called oxide ion conductors. Perovskites have good oxide ion conductors which results in their distorted structures. In cubic perovskites (with no distortion), the 3D framework leads to corner-sharing of BO_6 octahedra, and the A-cation is enclosed within 12 equidistant atoms forming a dodecahedra. The coordination number of O is 2 and is low since the A-O distance is almost 1.4 times the B-O bond distance. All these properties contribute to the crystal structure and hence, they can be used to determine the crystal structure.

Crystal structures are found using two power intensive and time consuming techniques: (a) Density functional theory (DFT) and (b) X- ray diffraction (X-RD) (fig: 2). Along with being costly, these techniques are highly sensitive to fluctuations. The cost, time and inaccuracies on crystal structure prediction can be reduced if classification based Machine learning (ML) models are used. Classification-based models such as KNN, RF-classifiers, CNN, Boosting algorithms, etc, can be utilized to obtain the required classifications.

1.1. Related works

Behera et al.[5], published the first work on crystal structure prediction using machine learning that included 675 instances to predict the crystal structure. The data used was biased toward orthorhombic structures due to data scarcity. Different models that Behera et al.[5] used included XgBoost, SVM, Light BGM, and Random Forest (RF) with an average accuracy of 74.8%, 76.6%,80.3%, and 62.8% respectively. The work accounts for the tolerance factor (τ), derived from a radius of A,B. They reported the best accuracy for RBF (which was explained using the density of states function which is also RBF) kernel in SVM and Light GBM. The sampled data were cut down to features with finite feature values. Use of τ lead to omitting of some features such as radius of B, "In-literature",etc.

Jarin et al.[6] reported models with crystal structures prediction ∼95% without oversampling and using genetic algorithm support vector regression. They also utilized various Neural networks to achieve the best possible accuracy. The tolerance factor (τ), along with some other features were removed that were of less importance. The reported accuracy was higher than achieved by Behera et al.[5] due to use of neural networks in boosting.

Ahmad et al.[7] reported 5- fold cross validation and random forest to give accuracy of 90.53% and 98.57% respectively. They utilized data used by Behera et al. [5] along with the use of factor τ. They reported better accuracy than previous study. The reason of the better accuracy is unclear, however, it may due to use of different features with some assigned weights and use of τ factor.

Ryan et al.[8] reported use of deep learning models to learn the topology of the materials to study the crystal structure. They utilised atomic fingerprints, a coordination topological model around the crystallographic sites as the input to use neural networks to predict the likelihood of crystal formation by different elements using the structural template provided during the training. They reported that in 30% of cases, there was a literature compound among the top 10 of the predicted element and structure. This route of crystal prediction is time consuming and required a lot of resources and computation power.

Cheng et al.[9] developed three optimization algorithm accelerate the search for crystal structure with the lowest formation enthalpy.The algorithms were tested on two databases with GN(MatB)-BO showing promising results for crystal structure prediction of 29 compounds. The computational cost using the algorithm reduced by three folds compared to conventional methods. They suggested the use of GN(MatB)-BO for data-driven crystal structure prediction.

Most of the previous study is done to predict crystal structure of oxygen perovskites or crystal structure in general, but no study is reported on mixed perovskites, perovskites containing anions different, but not excluding oxygen.

2. Baselines

Classification models such as Light GBM, XgBoost, SVMs, KNN, RF, and CNN were used to obtain the crystal structures from the input feature values. Boosting algorithms work on decision trees in which sequential tree growth using gradient boosting improves performance by correcting categorization errors made by earlier trees. SVMs are simple machines that work on the principle of support vectors along with the use of kernel functions. The corresponding kernel functions are used to get the best maximum accuracy while maintaining the AUC-ROC. Further, classifiers are used to get the importance matrix through which the corresponding importance of each feature is visualized. KNN are algorithms that detect the k nearest neighbors and classify them based on the K value. It works by joining k nearest neighbors to the node based on distance, therefore, classifying them. CNN are neural networks that work like the human brain. Information is transferred to the input layer, where the model learns parts of the classification. This is transferred to the next layers where errors are minimized and learning occurs repeatedly. Finally, at the output layer, the model makes a prediction. This is used to get the prediction on complex models.

Boosting methods were used on decision trees that were used to predict results on the test data. The boosting methods use sequential corrections to calculate the loss and get the best possible predictions using a decision tree. In the case of SVM, different kernels along with weights were used to study the models.

The main goal of the project was to build a classification model that classifies a perovskite into its various crystal structures. The implementation is as follows:

(i) Database collection.
(ii) Feature selection and data preprocessing.
(iii) Model selection.
(iv) Hyper-parameter optimization through metric.
(v) Testing for accuracy.

The basic implementation was similar for verification study and the study on mixed perovskites.

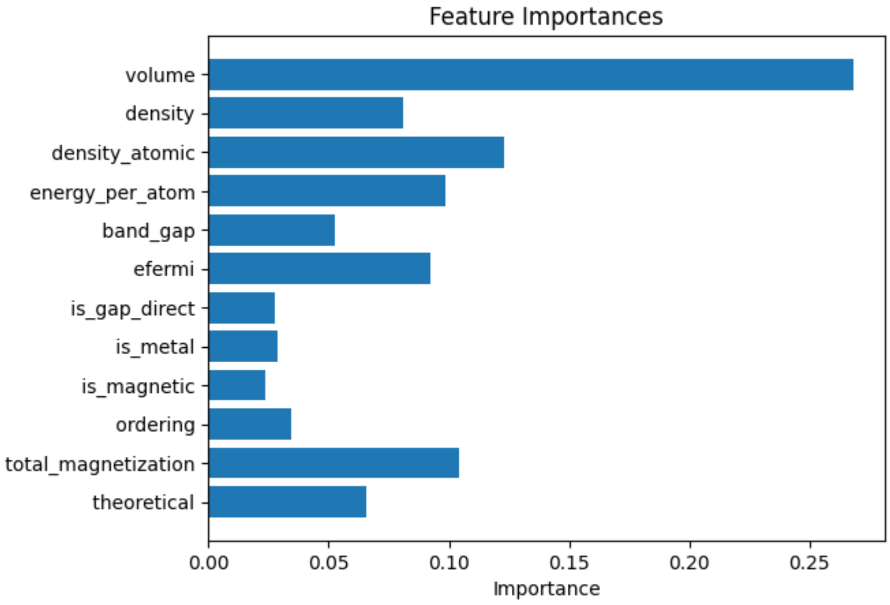


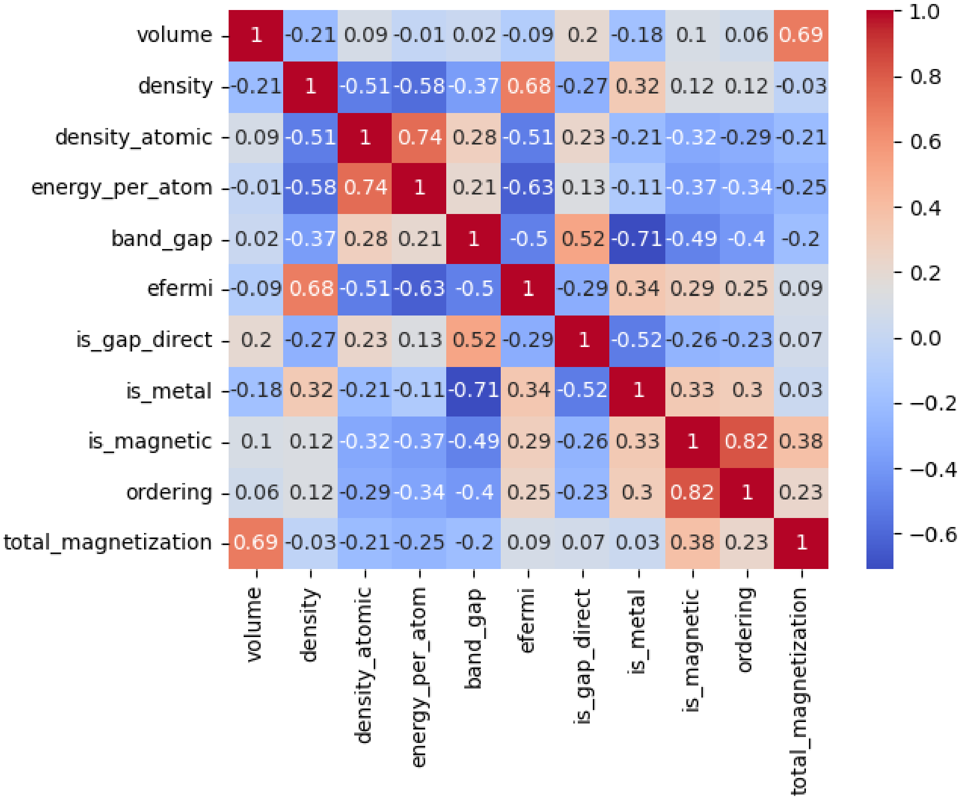
Figure 3: Feature importance using LGB classifiers

2.1. Model Environment

Python 3 was used for this work. Scikit-learn, Pandas, Imblearn, Numpy, etc libraries were used for data processing and implementation of different models such as XgBoost, SMOTE, SVM, Light GBM, CNN, Random Forests, Multi layer Perception, and Recurrent Neural Networks

2.2. Feature selection

Models were run based on data from behera et al.[5] to achieve an accuracy close to 80.3% for verification. The only difference was the use of all features, except radius of A, B, angle of A, angle of B and τ.



	volume	density	density_atomic	energy_per_atom	band_gap	efermi	is_gap_direct	is_metal	is_magnetic	ordering	total_magnetization
volume	1	-0.21	0.09	-0.01	0.02	-0.09	0.2	-0.18	0.1	0.06	0.69
density	-0.21	1	-0.51	-0.58	-0.37	0.68	-0.27	0.32	0.12	0.12	-0.03
density_atomic	0.09	-0.51	1	0.74	0.28	-0.51	0.23	-0.21	-0.32	-0.29	-0.21
energy_per_atom	-0.01	-0.58	0.74	1	0.21	-0.63	0.13	-0.11	-0.37	-0.34	-0.25
band_gap	0.02	-0.37	0.28	0.21	1	-0.5	0.52	-0.71	-0.49	-0.4	-0.2
efermi	-0.09	0.68	-0.51	-0.63	-0.5	1	-0.29	0.34	0.29	0.25	0.09
is_gap_direct	0.2	-0.27	0.23	0.13	0.52	-0.29	1	-0.52	-0.26	-0.23	0.07
is_metal	-0.18	0.32	-0.21	-0.11	-0.71	0.34	-0.52	1	0.33	0.3	0.03
is_magnetic	0.1	0.12	-0.32	-0.37	-0.49	0.29	-0.26	0.33	1	0.82	0.38
ordering	0.06	0.12	-0.29	-0.34	-0.4	0.25	-0.23	0.3	0.82	1	0.23
total_magnetization	0.69	-0.03	-0.21	-0.25	-0.2	0.09	0.07	0.03	0.38	0.23	1

Figure 4: Correlation matrix

For the new data generated using materials project API, equal importance was given to every feature irrespective of their importance matrix value. The

LGB classifier was used to get the relative feature importance (fig: 3). Column 1 representing volume was the most important feature. The correlation matrix was plotted using the data (fig: 4). In the case of weighted SVM, more importance was given to the instances which are found in nature and which have balanced atoms. Only, compound names were omitted from the calculations as they did not show any importance. The bond angles and radius were also omitted as they can be used to predict the structures without use of models.

It was decided to use Volume, density, and band gap as some of the features for models with mixed and oxide perovskites. Radius of A and B were not used as features as for same A,B the X anion may vary forming different structures like ABX_3 or ABO_3 leading to inaccuracies in the model. Similarly, It was found that using volume is a better feature than the volume per atom as it gives an idea of the atoms in the lattice.

3. Experiment

3.1. Data-Pre processing

For replication of experiment by Behera[5], the data for oxygen perovskites consisted of 5429 data-points which included data that could not be used for the model due to missing feature values and thus, such data-points were deleted. The data was obtained through DFT calculations and includes various chemical and physical properties of compounds. The data set was highly imbalanced which could lead to overfitting of a particular type of dataset. SMOTE was used to equalize the number of instances of each label in the training values to remove minority class problem. The features were selected based on the properties. Since the values were not highly scattered, normalization techniques were not utilized.

The new data set was obtained from Materials Project API,[10], using robocrys and MPRester libraries, and 14542 data points that consist of various features were generated. These were DFT-based calculations and predictions were made on data generated using DFT. The data was pre-processed by removing any missing feature values. No oversampling was done as the there as there were enough data-points for each label. Also, the number of structures increased to 6. The data had low variance, thus standardization techniques were not used.

4. Model: Oxygen perovskites - Verification

Table 1 depicts the accuracy along with the AUC-ROC curves achieved for oxygen perovksites, run with data used by Behera[5]. The accuracy achieved was higher than reported by the authors.

Table 1: Accuracy of different models for oxygen perovskites

Accuracy vs Model		
Model	Accuracy	AUC-ROC
SVM	91.33%	0.96
Weighted SVM	92.03%	0.96
Light GBM	88.26%	0.90
XGBoost	90.91%	0.90
CNN	67.75%	0.76
KNN	85.79%	0.86

4.1. Light GBM

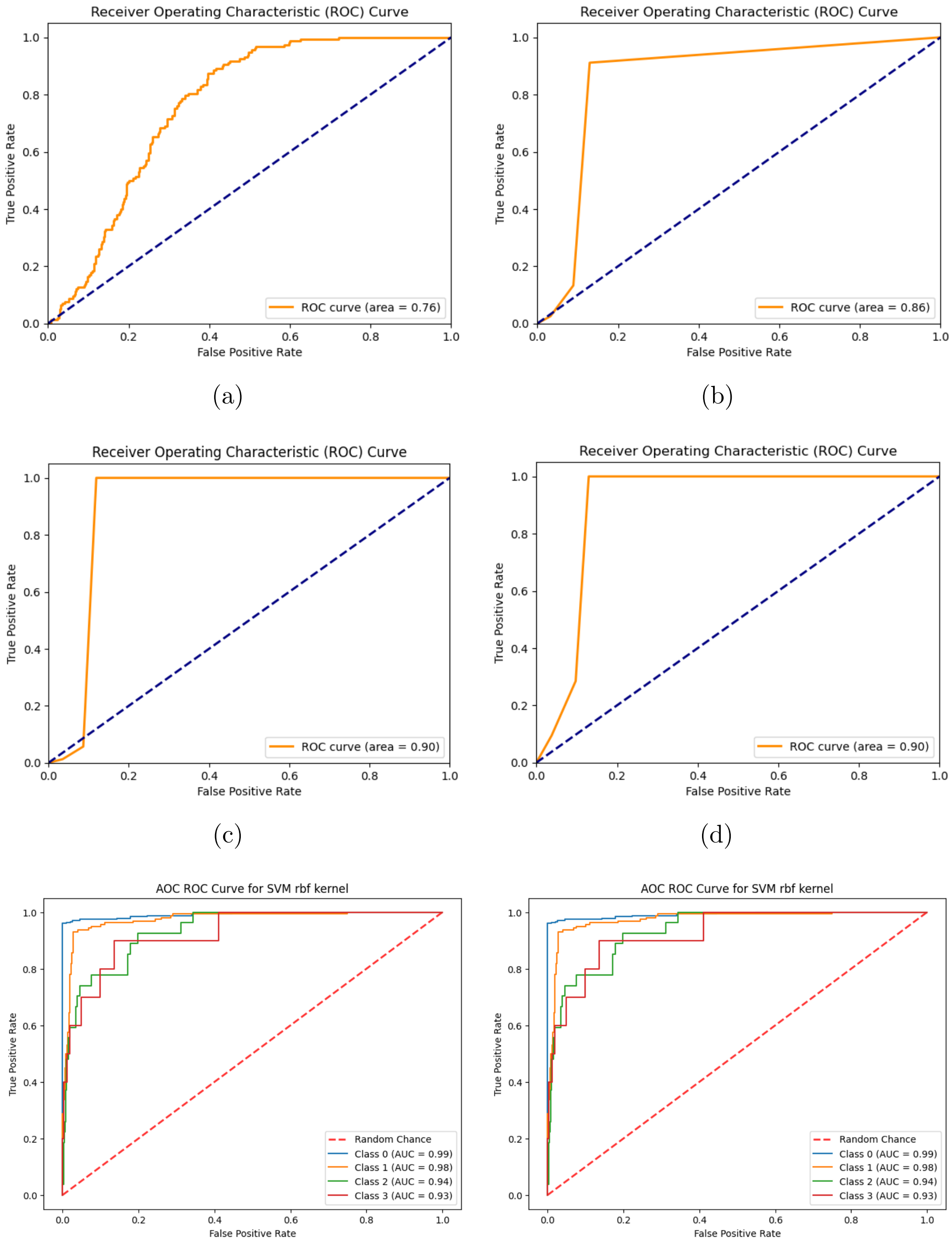
The height of the decision tree and the learning rate were the two hyper-parameters used in the model. The validation data set was prepared from the training data set which was used to get the hyper-parameter values. The optimum height was found to be 2 and the learning rate was 0.001. The losses were calculated on a logarithmic loss function. The hyperparameters were tuned to get the best accuracy of 88.26% with an AUC-ROC of 0.90 (fig 5(c)).

4.2. XgBoost

Height, learning rate, and number of epochs were the hyper-parameter used to study the model. The losses were calculated on logarithmic losses. 20 epochs with a tree height of 3 and a number of epoch 20 were found to have the highest AUC-ROC of 0.90. The accuracy achieved using this method was 90.91% (Fig 5(d)).

4.3. CNN

Implementation of CNN included tuning the loss function. Out of various loss functions, the mean square loss function was found to give the best AUC-ROC curve of 0.76. Other hyperparameters included the number of layers,



(a) (b)

(c) (d)

(e) (f)

Figure 5: AUC-ROC of different models run on mixedhalide perovskites: (a) AUC-ROC for CNN; (b) AUC-ROC for KNN; (c) AUC-ROC for L-GBM; (d) AUC-ROC for XgBoost; (e) AUC-ROC for SVM-RBF kernel;and, (f) AUC-ROC for Weighted-SVM-RBF kernel.

learning rate, batch sizes, and activation functions. The number of layers was fixed to be 3 and the input vector size was 13x1 dimensional. The CNN was configured to have 64 nodes in the first layer, 32 in the second, and 1 in the third with rectified linear activation unit (relu) in the first two layers and linear activation in the third layer. After tuning, the number of epochs was found to be 10, with a batch size of 8, and a learning rate of 0.001. The accuracy achieved was 69.68% (Fig 5(a)).

4.4. KNN

KNN was used to find the classification of compounds. K was the only hyper-parameter used in the model. After tuning, the value of k was 2 with an AUC-ROC of 0.86. Using these parameters, the accuracy achieved was 85.79% (Fig 5(b)).

4.5. SVM

C(Penalty parameter), gamma, and type of kernels(Linear, RBF, Sigmoid, and Polynomial) were used as the hyperparameters. After hyperparameter tuning, the RBF kernel (fig 3) was found to give the best AUC-ROC curve with C = 209.5 and an AUC-ROC of 0.96. The gamma parameter was found to be 0.01. The accuracy of the test data using SVM was found to be 91.33% (Fig 5(e)).

4.6. Weighted-SVM

C(Penalty Parameter), type of kernels, gamma, and the weights of the instances were used as the hyperparameters. It was decided that instances that occur in nature have higher weights compared to those which were produced using SMOTE. After tuning, the penalty parameter was found to be 208 while the kernel used was RBF with gamma 0.1. The weights used were: 5,1,0.5. The accuracy of the test data using SVM was found to be 92.03% (Fig 5(f)).

Contribution

The data was generated through materials project API. There was no such work done previously on such large-scale data with mixed perovskites. This work aims to classify various perovskites into their crystal structures based on their basic properties(computed through DFT).

5. Model: Mixed Perovskites

Table 2 depicts the accuracy and the AUC-ROC achieved for various classification models run on the data.

5.1. SVM-rbf kennel

The kernel and C (Penalty parameter) were used as the hyper-parameters in the model. After tuning, the kernel used was RBF with the penalty parameter C = 19. The average AUC-ROC (fig: 7(a)) obtained on the data was 0.86 with an accuracy percentage of 55.26%.

5.2. Random Forests

n estimators, criterion, max depth, and min sample split were used as the hyper-parameters in the model. After tuning, n estimators were found to be 250, with criterion entropy, max depth of 100, and the min sample split of 2. The average AUC-ROC(fig: 7(b)) obtained on the model was 0.98 with an accuracy percentage of 94.40%.

5.3. KNN

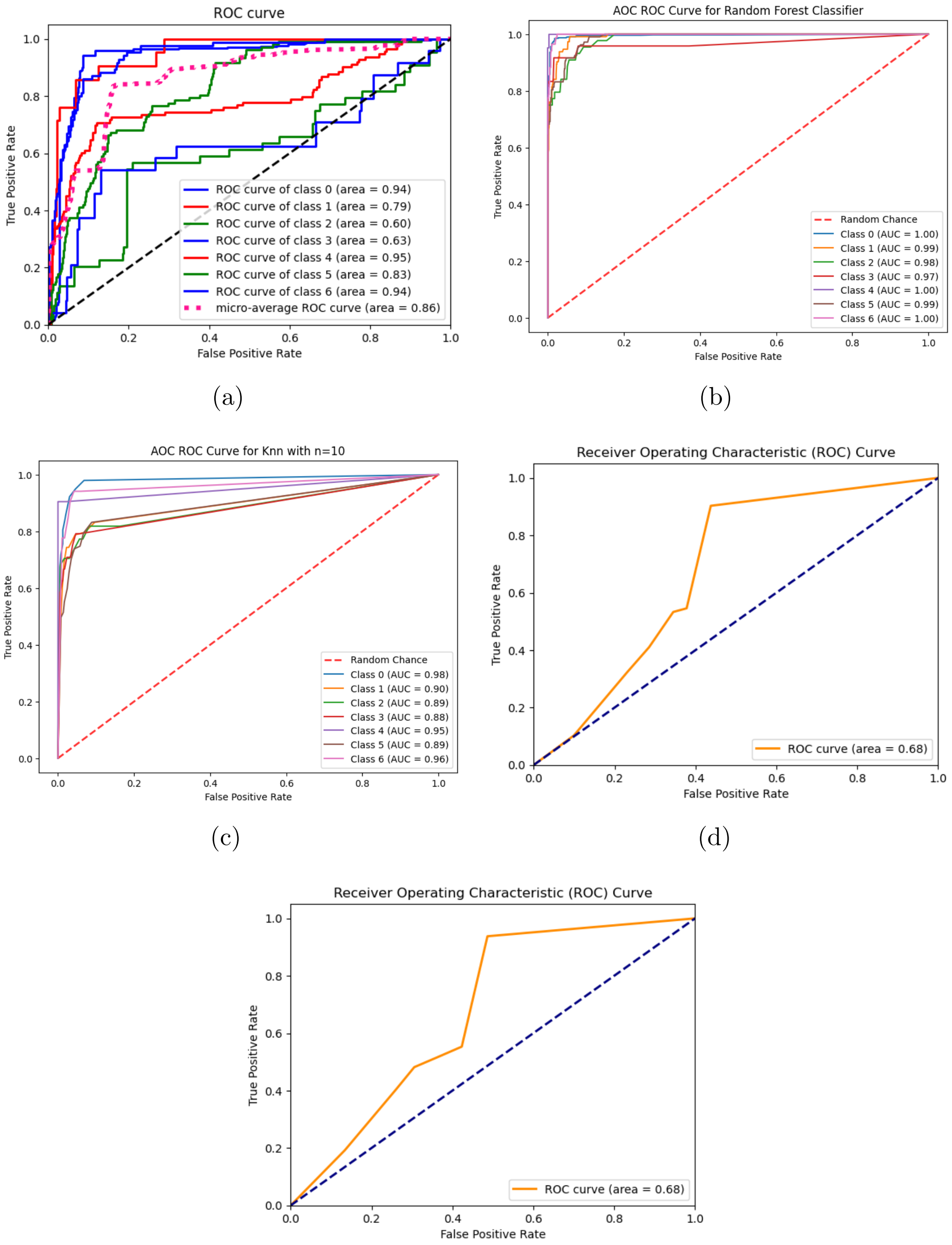
K was the only hyper-parameter used in the model. After tuning, k was found to be 10. The average AUC-ROC (fig: 7(c)) obtained was 0.92 with an accuracy percentage of 82.00%.

5.4. Light GBM

The hyperparameters used were the height of the decision tree and the learning rate. The validation data set was prepared from the training data set which was used to get the hyper-parameter values. The optimum height was found to be 3 with a learning rate of 0.01. The losses were calculated on a logarithmic loss function. The hyperparameters were used to get the best accuracy of 60.21% with an AUC-ROC of 0.68 (fig: 7(d)).

5.5. XgBoost

The parameters used were the height of the decision tree, the learning rate, and the number of epochs. After hyperparameter tuning, 1 epoch with a 0.7 learning rate and tree height 4 was found to be the best fit. The loss function used was logarithmic. The accuracy percent found through this model was 60.86% with an AUC-ROC of 0.68 (fig: 7(e)).



(a)

(b)

(c)

(d)

(e)

Figure 6: AUC-ROC of different models run on mixed perovskites: (a) AUC-ROC for SVM-RBF kernel; (b) AUC-ROC for Random forest; (c) AUC-ROC for KNN; (d) AUC-ROC for Light GBM; and, (e) AUC-ROC for XgBoost.

Table 2: Accuracy of different models for mixed perovskites

Accuracy vs Model		
Model	Accuracy	AUC-ROC
SVM	55.26%	0.86
Random Forests	94.40%	0.98
Light GBM	60.21%	0.68
XGBoost	60.86%	0.68
KNN	82.00%	0.92

6. Conclusion

The data (oxygenoxygen perovskites) analyzed and predicted had higher overall higher accuracy than those reported by behera et al.[5]. The best results were obtained using weighted SVM with an accuracy of 92.03% and AUC-ROC of 0.96. This can be used in real life to predict the structures of perovskites fairly accurately. In the case of mixed perovskites, the predictions through models, other than random forest and KNN were inaccurate. This can be attributed to a change in the data set from the last data set consisting of only oxygen perovskites and the absence of some of the features that were earlier present in the data.

The classification of crystals was best achieved by random forest with an accuracy of 94.40% and an AUC-ROC of 0.98. This model can be reasonably used in practical applications to predict crystal structures with high accuracy. The accuracy for the mixed perovskites had a higher accuracy and thus can be used to get accurate predictions for all perovskites.

Dataset

(i) Data for the prediction of oxide perovkithalides (verification dataset) was obtained from DFT calculations from Emery et al[11]. Approximately 5500 instances were recorded.
(ii) The dataset for mixed perovskite prediction was obtained from Materials Project API[10], using robocrys and MPRester libraries. Approximately 14,000 instances were obtained this way. The feature values were then substituted with numerical values in the data frame.

References

[1] M. A. Peña, J. L. G. Fierro, Chemical structures and performance of perovskite oxides, Chemical Reviews 101 (2001) 1981–2018. URL: https://doi.org/10.1021/cr980129f. doi:10.1021/cr980129f.

[2] H. M. Ghaithan, Z. A. Alahmed, S. M. H. Qaid, A. S. Aldwayyan, Density functional theory analysis of structural, electronic, and optical properties of mixed-halide orthorhombic inorganic perovskites, ACS Omega 6 (2021) 30752–30761. URL: https://doi.org/10.1021/acsomega.1c04806. doi:10.1021/acsomega.1c04806.

[3] M. Dawson, C. Ribeiro, M. Morelli, A review of three-dimensional tin halide perovskites as solar cell materials, Materials Research 25 (2022). doi:10.1590/1980-5373-mr-2021-0441.

[4] L. Yang, J. Milic, A. Ummadisingu, J.-Y. Seo, J.-H. Im, H.-S. Kim, Y. Liu, M. I. Dar, S. Zakeeruddin, P. Wang, A. Hagfeldt, M. Grätzel, Bifunctional organic spacers for formamidinium-based hybrid dion–jacobson two-dimensional perovskite solar cells, Nano Letters 19 (2018). doi:10.1021/acs.nanolett.8b03552.

[5] S. Behara, T. Poonawala, T. Thomas, Crystal structure classification in abo3 perovskites via machine learning, Computational Materials Science 188 (2021) 110191. URL: https://www.sciencedirect.com/science/article/pii/S0927025620306820. doi:https://doi.org/10.1016/j.commatsci.2020.110191.

[6] S. Jarin, Y. Yuan, M. Zhang, M. Hu, M. Rana, S. Wang, R. Knibbe, Predicting the crystal structure and lattice parameters of the perovskite materials via different machine learning models based on basic atom properties, Crystals 12 (2022). URL: https://www.mdpi.com/2073-4352/12/11/1570. doi:10.3390/cryst12111570.

[7] M. U. Ahmad, A. A. R. Akib, M. M. S. Raihan, A. B. Shams, Abo3 perovskites' formability prediction and crystal structure classification using machine learning, 2022. URL: https://arxiv.org/abs/2202.10125. doi:10.48550/ARXIV.2202.10125.

[8] K. Ryan, J. Lengyel, M. Shatruk, Crystal structure prediction via deep learning, Journal of the American Chemical Society 140 (2018) 10158–10168. URL: https://doi.org/10.1021/jacs.8b03913. doi:10.1021/jacs.8b03913.

[9] G. Cheng, X.-G. Gong, W.-J. Yin, Crystal structure prediction by combining graph network and optimization algorithm, Nature Communications 13 (2022). URL: https://doi.org/10.1038/s41467-022-29241-4. doi:10.1038/s41467-022-29241-4.

[10] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, Text-mined dataset of inorganic materials synthesis recipes, Scientific Data 6 (2019). URL: https://doi.org/10.1038/s41597-019-0224-1. doi:10.1038/s41597-019-0224-1.

[11] A. A. Emery, C. Wolverton, High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites, Scientific Data 4 (2017). URL: https://doi.org/10.1038/sdata.2017.153. doi:10.1038/sdata.2017.153.