Determination of lattice structure of perovskites using ML

Kaling Vikram Singh
kalingvikram.singh@niser.ac.in
NISER

S Sunil Raja
sunilraja.s@niser.ac.in
NISER

Abstract

The determination of crystalline structure is an important aspect of studying materials and determining their uses in everyday life. The various techniques for the detection of these properties are time-consuming and costly. Thus, one alternative to correctly estimating the crystal structure is using machine learning. Perovskites are a special class of elements with $AB0_3$ type configuration. Basic properties such as valency, atomic radii, bang gap, etc are used to predict the crystal structure. Various classification models are implemented, using which, the crystal structures are predicted. Support vector machines (SVM), random forest (RF), Light gradient boosting machine (LGBM), XGBoost, Convolution neural networks (CNN), and K-nearest neighbors (KNN) are used for the study. It was observed that, for oxide perovskites, weighted SVMs were the best model, with an accuracy of 92.03% while for mixed perovskites, the best model to predict structure was a random forest with an accuracy of 94.4%.

1 Introduction

Perovskites are a class of elements that have a similar crystal structure as the compound calcium titanium, ABX_3. These are used in various industries and their main applications include creating solar panels that could be coated on various surfaces. These materials are lightweight and cheap and are used in the photovoltaic industry. These have high variations in A, and B and can be found in various structures such as cubic, monoclinic, orthorhombic, tetrahedral, hexagonal, or rhombohedral. The most significant reasons for the change in shapes include (i) displacement of the cation, (ii) distortion of the octahedra and (iii) tilting of the octahedra. The displacement and distortions are instability-driven factors.

Perovskites have high anion-cation interactions and hence, their properties affect the overall crystal structure. Good oxide ion conductivity is an important property of a cubic perovskite structure's reduced distortion. In a cubic perovskite, the 3D framework leads to corner-sharing of BO_6 octahedra, and the A-cation is enclosed within 12 equidistant atoms. 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.

Perovskite halides exhibit distinct characteristics, such as exceptional absorption coefficients, high carrier mobilities, and prolonged charge carrier lifetimes, that make them promising for optoelectronic and photovoltaic applications. In fact, 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.

There is extensive work in progress that includes Density Functional Theory (DFT) calculations and analytical techniques such as X-ray diffraction (XRD) to get the crystalline structure of materials. These are power-intensive and costly processes. These processes require more than 15-16 hours to

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

get accurate data on crystal structures. The cost and time on crystal predictions can be reduced if 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

Santosh and Taher et al.[1], provided the basis and included 675 data points and the data was biased towards orthorhombic structures. The methods used include XgBoost, SVM, Light BGM, and Random Forest (RF) with an accuracy of 74.8%, 76.6%,80.3%, and 62.8% respectively. The paper accounts for the tolerance factor (τ), derived from a radius of a,b, which has been neglected in the current paper, and the model is built with the available data set. They reported the best accuracy for RBF (which is 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.

Jarin et al.[2] reported models with crystal structures prediction $\sim$95% without oversampling 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.

2 Baselines

Simple and reliable models such as Light BGM, XgBoost, SVMs, KNN, RF, and CNN to obtain the crystal structures. Boosting algorithms work on decision trees in which sequential tree growth using gradient boosting improves performance by correcting categorization errors made by earlier trees. Further, SVMs are simple machines that work on the principle of support vectors along with kernel functions. The corresponding kernel functions are used to get the best maximum accuracies. 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 and based on distance, 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.

The main aim is to build a classification model that classifies a perovskite. The implementation is as follows:

(i) Database collection

(ii) Feature selection and data-processing

(iii) Model selection

(iv) Hyper-parameter optimization

(v) Testing for accuracy.

2.1 Model Environment

Python 3 was used for this project. 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, Multilayer Perception, and Recurrent Neural Networks.

2.2 Feature selection

Importance was given to every feature even if their importance was low. The LGB classifier was used to get the relative feature importance (fig 1). Column 7 representing the ionic radii of A was the most important feature. The correlation matrix was plotted using the data (fig 2). 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 lattice edge length were omitted as they are precursors of crystalline structure.

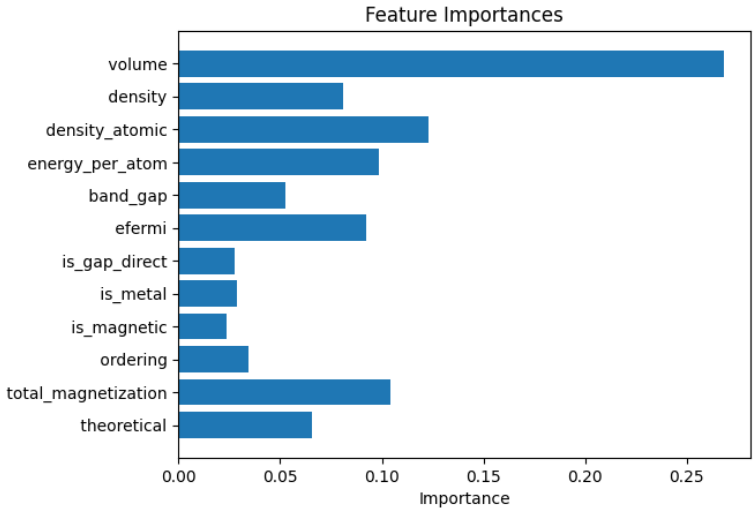

Figure 1: Feature importance for Random forests

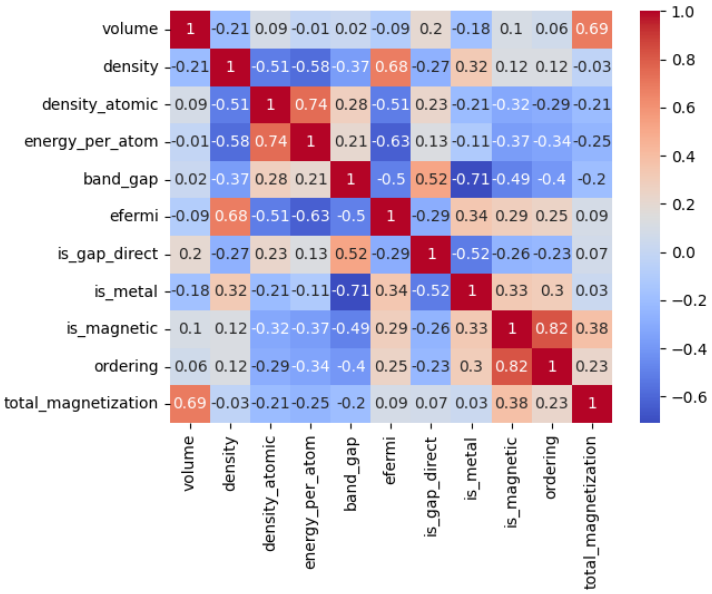

Figure 2: Correlation matrix

It was decided to use Volume, density, and band gap as some of the features for models with halides and oxide perovskites. In the midway code, we used radii A, B, and valency A, B as features in the model, This could only be done as only oxides were present in the list. If radii A and B were used as features in the model for halides and oxides, we will obtain some inaccuracies as ABO3 and ABF3 (here O and F represent Oxygen and fluorine) have the same A and B ions but different anions which may lead to a different structure.
Similarly, It was found that using volume is a better feature than the volume per atom (used in midway code) as it gives an idea of the atoms in the lattice.

3 Experiment

GitHub repository: `https://github.com/dimo192/ML`

3.1 Data-Pre processing

The data for mixed perovskites consists of 14542 datasets which include data that cannot be used for our model and thus, such rows were deleted. The data was obtained through DFT calculations and includes various chemical and physical properties of compounds. The dataset is highly imbalanced which can 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. The features were selected beforehand. Since the values were not highly scattered, normalization techniques were not utilized.

We have used Boosting methods on decision trees that are 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, we have utilized different kernels to study the best fit. Further, we used weighted SVM to calculate the predictions.

4 Models (Oxygen perovskites)

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

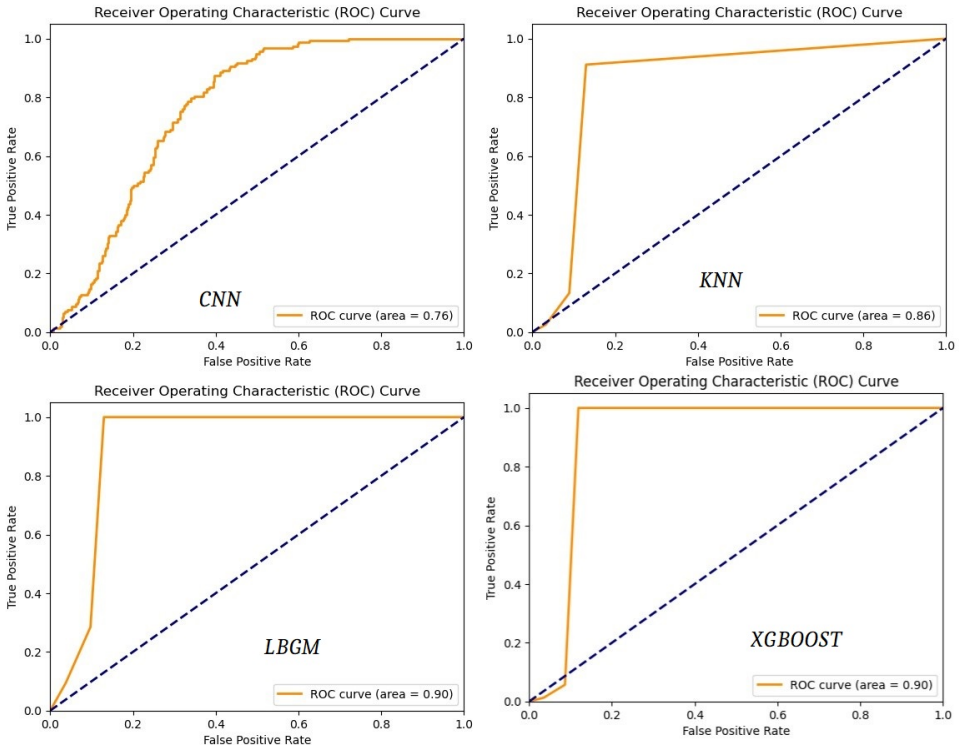

Figure 3: AUC-ROC curve for models.

4.1 Light GBM

The parameters used were the height of the decision tree and the learning rate. The validation dataset was prepared from the training dataset 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. Finally, the hyperparameters were used to get the best accuracy of 88.26% with an AUC-ROC of 0.90 (fig 3).

4.2 XgBoost

The parameters used were the height of the decision tree, the learning rate, and the number of epochs. After hyperparameter tuning, 20 epochs with a 1.25 learning rate and tree height 3 were found to be the best fit. The loss was again logarithmic. The accuracy found through this method was 90.91% with an AUC-ROC of 0.90.

4.3 CNN

Implementation of CNN included deciding on the loss function. Out of various loss functions, the mean square loss function was found to give the best AUC-ROC curve. Further, the number of layers was fixed to be 3 and the vector was 15x1 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. The number of epochs was found to be 10, batch size 8, and learning rate 0.001. The accuracy achieved was 69.68 with an AUC-ROC of 0.76

4.4 KNN

KNN was used to find the classification of compounds. It was found that for k = 2, the AUC-ROC metric was maximum at 0.86, and the accuracy obtained was 85.79%.

4.5 SVM

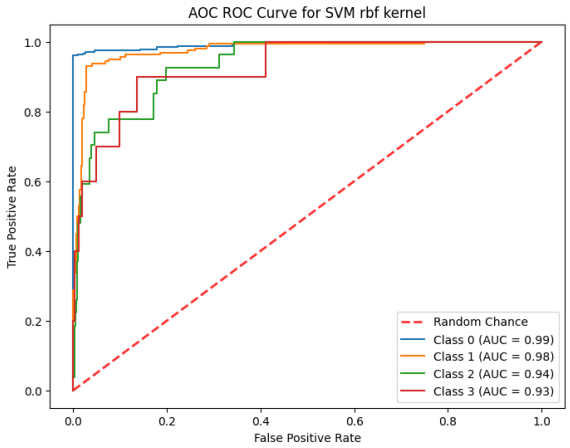

Figure 4: AUC-ROC curve for SVM (oxygen perovskite)

The parameters used were C(Penalty parameter) and kernels(Linear, RBF, Sigmoid, and Polynomial). After Hyperparameter tuning, RBF (fig 3) was chosen to be the kernel, and the value of C was 209.5. Upon further tuning the value of gamma for RBF was chosen to be 0.01.

4.6 Weighted SVM

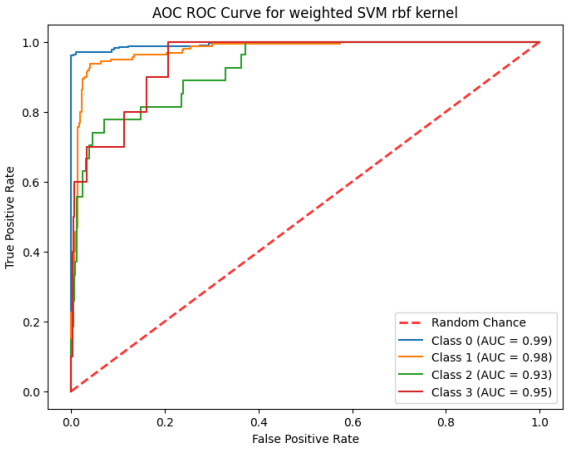

Figure 5: AUC-ROC curve for weighted SVM (oxygen perovskite)

The parameters used were C(Penalty Parameter), kernels, and the weights of the instances. It was decided that instances that occur in nature have higher weights compared to those which were produced using SMOTE. The final parameter which was obtained is 208 for the penalty parameter, kernel as RBF(with gamma as 0.1), and the value of weights as 5,1,0.5.

Contribution

The data was generated through API. There was no such work done previously on such large-scale data with so many different types of perovskites. This work tries to classify various perovskites based on their basic properties.

5 Models (Mixed Perovskites)

5.1 SVM

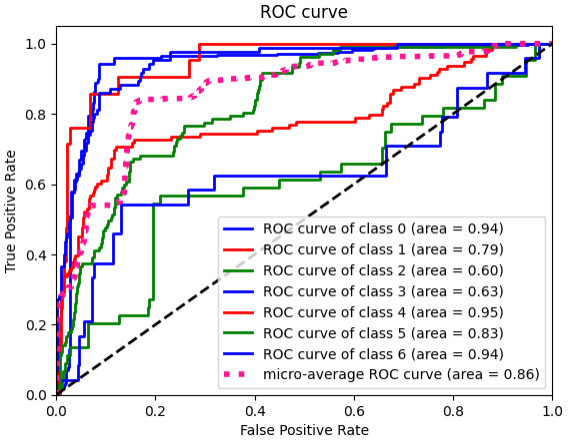

Figure 6: AUC-ROC curve for SVM (mixed perovskite)

The kernel and Penalty parameter were used as the hyper-parameters in the model and the output obtained was that the kernel is RBF(Radial Basis Function) and C as 19. The average AUC-ROC obtained was 0.86 and the accuracy percent was 55.26%.

5.2 Random Forests

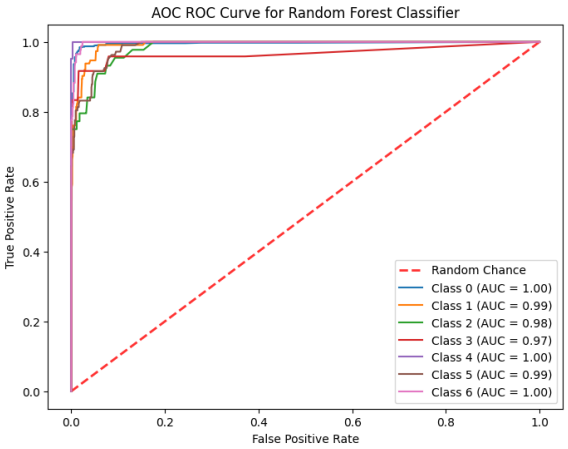

Figure 7: AUC-ROC curve for random forest (mixed perovskite)

n estimators, criterion, max depth, and min sample split were used as the hyper-parameters in the model and the output obtained was that the kernel is 250, entropy, 100, and 2 respectively. The average AUC-ROC obtained was 0.98 and the accuracy percent was 94.40%.

5.3 KNN

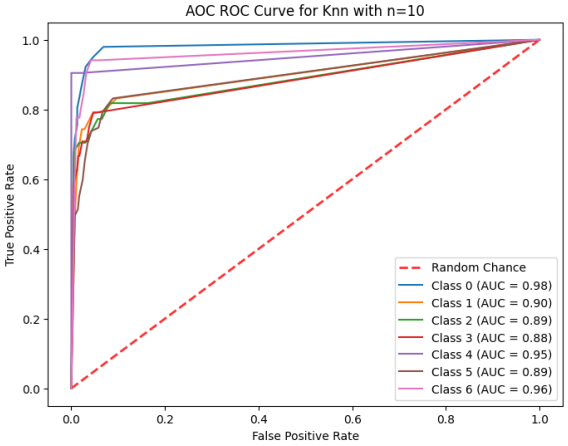

Figure 8: AUC-ROC curve for KNN (mixed perovskite)

The value of the only hyper-parameter (N neighbors) was obtained to be 10. The average AUC-ROC obtained was 0.92 and the accuracy percent was 82%.

5.4 Light GBM

The parameters used were the height of the decision tree and the learning rate. The validation dataset was prepared from the training dataset which was used to get the hyper-parameter values. The optimum height was found to be 3 and the learning rate was 0.01. The losses were calculated on a logarithmic loss function. Finally, the hyperparameters were used to get the best accuracy of 60.21% with an AUC-ROC of 0.68.

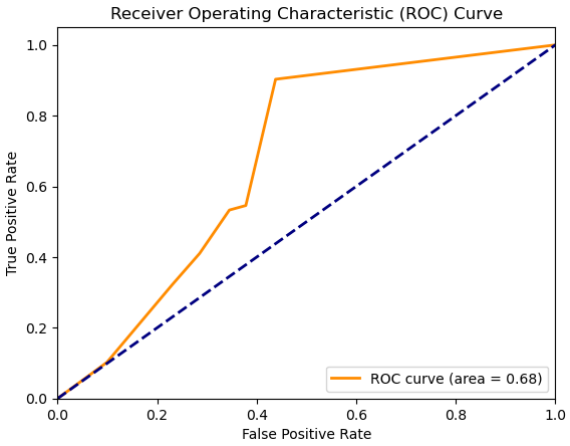

Figure 9: AUC-ROC curve for LGBM (mixed perovskite).

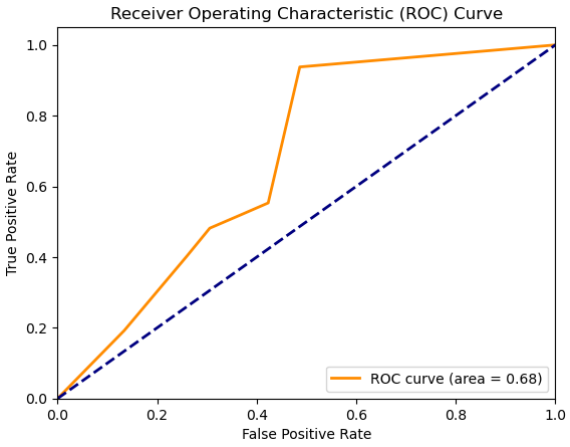

Figure 10: AUC-ROC curve for XGBOOST (mixed perovskite)

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 was again logarithmic. The accuracy found through this method was 60.86% with an AUC-ROC of 0.68.

Table 2: Accuracy of different models for mixed perovskites

Accuracy vs Model		
Model	Accuracy	AUC-ROC
SVM	55.26%	0.86
Random Forests	94.4%	0.98
Light GBM	60.21%	0.68
XGBoost	60.86%	0.68
KNN	82%	0.92

6 Conclusion

The data (oxygen perovskites) analyzed and predicted had higher overall higher accuracies than those reported by Santosh et al. 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 not accurate. This can be attributed to a change in the data set from the earlier dataset 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 fairly used in everyday life to predict crystal structures with high accuracy.

Dataset

(i) Data for the prediction of oxide halides was obtained from DFT calculations from Emery et al[3]. Approximately 5500 instances were recorded.
(ii) The dataset was obtained from Materials Project API[4], 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] Santosh Behara a, et al. “Crystal Structure Classification in ABO3 Perovskites via Machine Learning.” Computational Materials Science, Elsevier, 1 Dec. 2020.

[2] Jarin, S.; Yuan, Y.; Zhang, M.; Hu, M.; Rana, M.; Wang, S.; Knibbe, R. Predicting the Crystal Structure and Lattice Parameters of the Perovskite Materials via Different Machine Learning Models Based on Basic Atom Properties. Crystals 2022, 12, 1570. https://doi.org/10.3390/cryst12111570

[3] Emery, Antoine & Wolverton, Chris. (2017). High-Throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO 3 perovskites. Scientific Data. 4. 170153. 10.1038/sdata.2017.153.

[4] A. Jain*, S.P. Ong*, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson (*=equal contributions) The Materials Project: A materials genome approach to accelerating materials innovation. doi:10.1063/1.4812323