Prediction of Crystal Structures for Lithium-Ion Silicate Cathodes Using Machine Learning Algorithms

Yogesh Yadav(M21PH209),Naman Sharma (M24IRM015),Ritul Kr Choudhri (M24IRM005)

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

Abstract

The physical and chemical characteristics of cathodes, used in batteries, are derived from the crystalline arrangement of the lithium-ion silicate cathodes, and this is pivotal to the overall battery performance. Therefore, the correct prediction of the crystal system is essential in estimate the properties of cathodes used in battery technologies. This study applies various machine learning classification algorithms for predicting the crystal systems, namely monoclinic, orthorhombic, and triclinic, that are related to Li–Si–(Mn, Fe, Co)–O-based silicate cathodes. This work has been generated with computational data for the Materials Project's DFT calculations. Feature evaluation showed that cathode properties are quite dependent on the crystal structure, and optimized classification strategies lead to better predictability. Of the methods considered, Random Forests provided the best predictions for crystal systems.

Keywords: Crystal Systems, DFT

Contents

1 Introduction

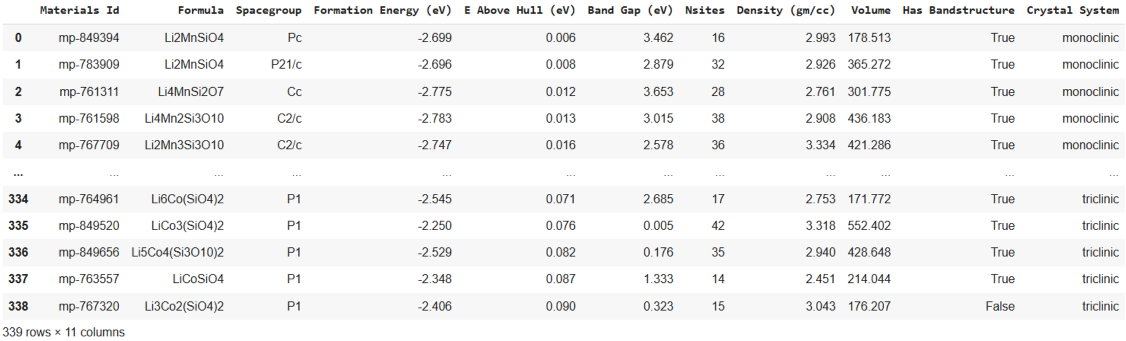
Recent advances in computational methods coupled with rising computational power have made it possible to calculate physical and chemical properties for a wide range of materials. The Materials Project releases publicly calculated information based on density functional theory. Density functional theory is a powerful tool that describes the electronic structures of a wide range of materials quite well. This database is so extensive that researchers can explore detailed relationships between material properties that were quite impossible to study before. While traditional statistical models usually fail to identify these relationships, the results achieved from the ML algorithm have been to map intricate patterns and correlations.

ML has successfully been applied to many material science problems, such as crystal structure prediction, an area of critical relevance to optimize materials for Li-ion batteries.

This paper discusses the prediction of the crystal structure of Li–Si–(Mn, Fe, Co)–O cathodes, specifically determining whether they belong to one of three types of dominant crystal systems, namely monoclinic, orthorhombic, and triclinic. We used various machine learning classification algorithms-including neural networks, support vector machines, k-nearest neighbors, and random forests-to build predictive models. In that regard, the results were best with the random forest algorithm with up to 70 percent. This paper aims to identify those features which are most influential to the prediction of crystal structure, and also helps in providing insight into the relationship between crystal structure and material performance. The results may make a major contribution to the advancement of more efficient and cost-effective lithium-ion batteries.

2 The Dataset

The dataset contains results from density functional theory (DFT) calculations for 339 cathode materials with Li–Si–(Mn, Fe, Co)–O compositions, sourced from the Materials Project.The dataset contains the following number of datapoints for each crystal system: monoclinic with 139 datapoints, orthorhombic with 128 datapoints, and triclinic with 72 datapoints.. The dataset includes various properties such as chemical formula, space group, formation energy, Bandgap, density, and crystal structure. It is characterized by a diverse range of complex structures and chemical compositions, with no apparent correlation between the features and crystal systems.



	Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Nsites	Density (gm/cc)	Volume	Has Bandstructure	Crystal System
0	mp-849394	Li2MnSiO4	Pc	-2.699	0.006	3.462	16	2.993	178.513	True	monoclinic
1	mp-783909	Li2MnSiO4	P21/c	-2.696	0.008	2.879	32	2.926	365.272	True	monoclinic
2	mp-761311	Li4MnSi2O7	Cc	-2.775	0.012	3.653	28	2.761	301.775	True	monoclinic
3	mp-761598	Li4Mn2Si3O10	C2/c	-2.783	0.013	3.015	38	2.908	436.183	True	monoclinic
4	mp-767709	Li2Mn3Si3O10	C2/c	-2.747	0.016	2.578	36	3.334	421.286	True	monoclinic
...	...	...	...	...	...	...	...	...	...	...	...
334	mp-764961	Li6Co(SiO4)2	P1	-2.545	0.071	2.685	17	2.753	171.772	True	triclinic
335	mp-849520	LiCo3(SiO4)2	P1	-2.250	0.076	0.005	42	3.318	552.402	True	triclinic
336	mp-849656	Li5Co4(Si3O10)2	P1	-2.529	0.082	0.176	35	2.940	428.648	True	triclinic
337	mp-763557	LiCoSiO4	P1	-2.348	0.087	1.333	14	2.451	214.044	True	triclinic
338	mp-767320	Li3Co2(SiO4)2	P1	-2.406	0.090	0.323	15	3.043	176.207	False	triclinic

339 rows × 11 columns

Figure 1: Data for some selected silicate cathodes.

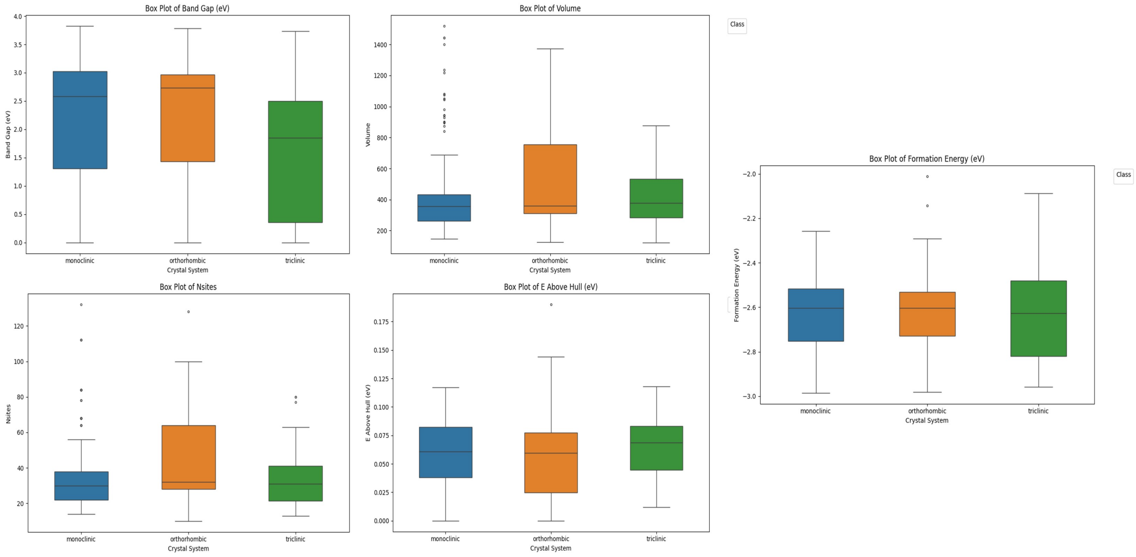


Figure 2: Box plots of different input features

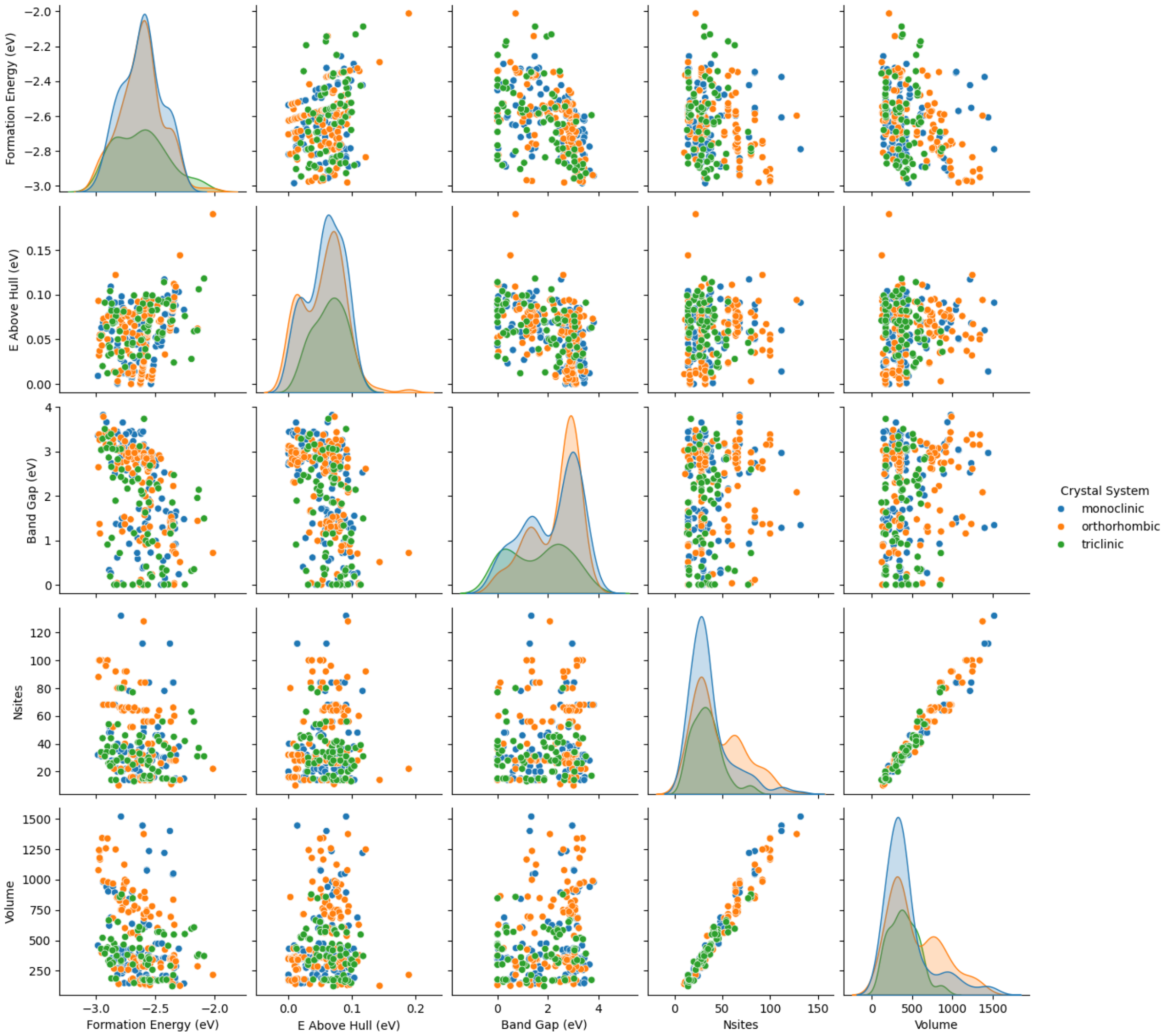


Figure 3: The pairs plot displays various properties of Li–(Mn, Fe, Co)–Si–O cathodes using data extracted from the Materials Project. The orange, green, and blue circles represent the orthorhombic, triclinic, and monoclinic crystal systems, respectively.

3 Methods of classification for machine learning

Classification in machine learning is a supervised method used to assign data to predefined classes. This study predicts crystal structures (CSs) of cathode materials—monoclinic, orthorhombic, and triclinic, using accuracy as the measure of correct predictions. The feature matrix (X) includes 339 samples and 5 features, while the response matrix (Y) indicates the corresponding crystal structure classes. The CS can be modeled based on variables such as volume (V), bandgap (Eg), number of sites (Ns), formation energy (Ef), and energy above the hull (EH).

3.1 Support Vector Machine

Support Vector Machine (SVM) is a classification algorithm that identifies the optimal hyperplane to distinguish between data points of different classes. The hyperplane is represented by the following equation:

$$w^T x + b = 0$$

where w is the weight vector and b is the bias term. SVM maximizes the margin, which is the distance between the hyperplane and the closest points (support vectors), given by:

$$\text{Margin} = \frac{2}{\|w\|}$$

The optimization problem to maximize the margin is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1$$

For non-linearly separable data, SVM uses a kernel function $K(x_i, x_j)$ to map the data into higher dimensions, enabling linear separation.

3.2 k-nearest neighbors

K-Nearest Neighbors (KNN) is a simple, non-parametric, and instance-based classification algorithm. In KNN, the class of a data point is determined based on the majority class of its K nearest neighbors in the feature space. The steps for classifying a point x are as follows:

1. Calculate the distance between the test point x and all training points. A common distance metric is the Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{ij})^2}$$

where x is the test point, x_i is a training point, and m is the number of features (dimensions).

2. Select the K nearest neighbors: After calculating the distances, the K closest training points to x are selected.

3. Majority Voting: The test point x is assigned to the class that appears most frequently among the K nearest neighbors. If there is a tie, different strategies (e.g., weighted voting or random choice) can be used to resolve it.

The decision rule for classification can be expressed as:

$$\hat{y} = \text{majority vote}\left(\{y_i : x_i \in N_k(x)\}\right)$$

where $\hat{y}$ is the predicted class, $N_k(x)$ denotes the set of K nearest neighbors of x, and y_i is the class label of each neighbor.

In the case of regression, the prediction $\hat{y}$ is computed as the mean (or weighted mean) of the target values of the nearest neighbors:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} y_i$$

where y_i is the target value of the i-th nearest neighbor.

3.3 Random forests

Random Forest is an ensemble learning strategy during which many decision trees are learned during the training phase using random subsets of data and features. Classification is based on majority voting over all the trees, whereas regression is performed by averaging the predictions. This approach naturally reduces overfitting and improves the accuracy through the integration of multiple diverse predictions from different trees. This method is rather effective in managing the immense datasets and provides worthwhile insight into the importance of features.

3.4 Artificial neural networks

The Artificial Neural Network has two hidden layers. The first hidden layer has 20 neurons and the next one has 10 neurons. Both of them are using ReLU activation function. For multi-class classification, the output layer will be using Softmax, which transforms the raw outputs into probabilistic format. This way, it assures the outputs to be in valid probability distribution across classes. The non-linearity comes from ReLU function to the model.

4 Result

In this study, we considered four different machine learning models: support vector machine (SVM), K-nearest neighbors (KNN), random forest (RF), and neural network (NN) for the classification of lithium-silicon-(manganese, iron, cobalt)-oxide (Li-Si-(Mn, Fe, Co)-O) cathode materials. SVM used the RBF kernel to perform its classification task with accuracy of 59%, which means it performs moderately. The KNN model performed slightly better; at its highest accuracy, it achieved 62%, which demonstrates its ability to fit some models of the given data. Meanwhile, RF, as an algorithm with 50 estimators, has outperformed other algorithms with a maximum accuracy of 70%, thus confirming its ability to handle complex and high-dimensional data. On the other hand, the NN using ReLU activation, Adam optimization algorithm, and categorical cross-entropy loss function was able to achieve merely 53% in training accuracy while letting the two hidden layers of neurons be set at 10 and 30 units each. Conclusion In a nutshell, the best performance in the experiment was given by the model Random Forest.

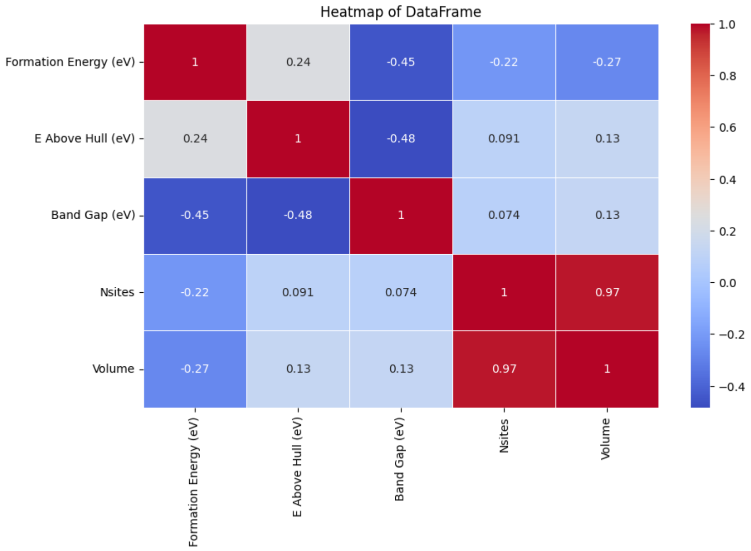


Figure 4: Correlation Matrix

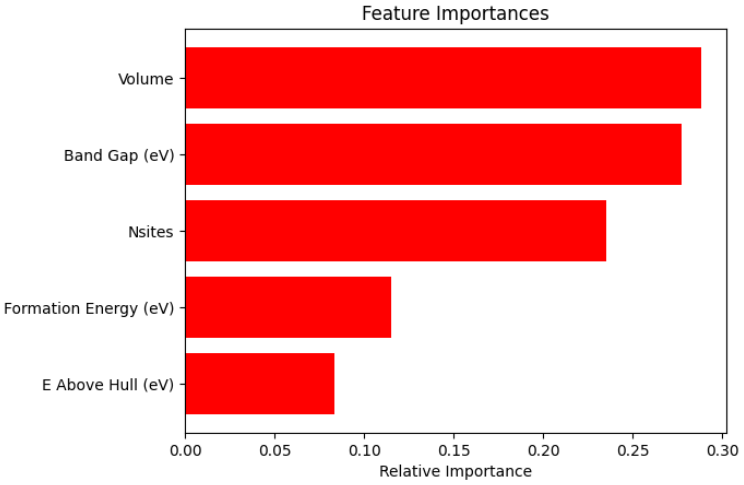


Figure 5: The importance of different features in a random forest model.

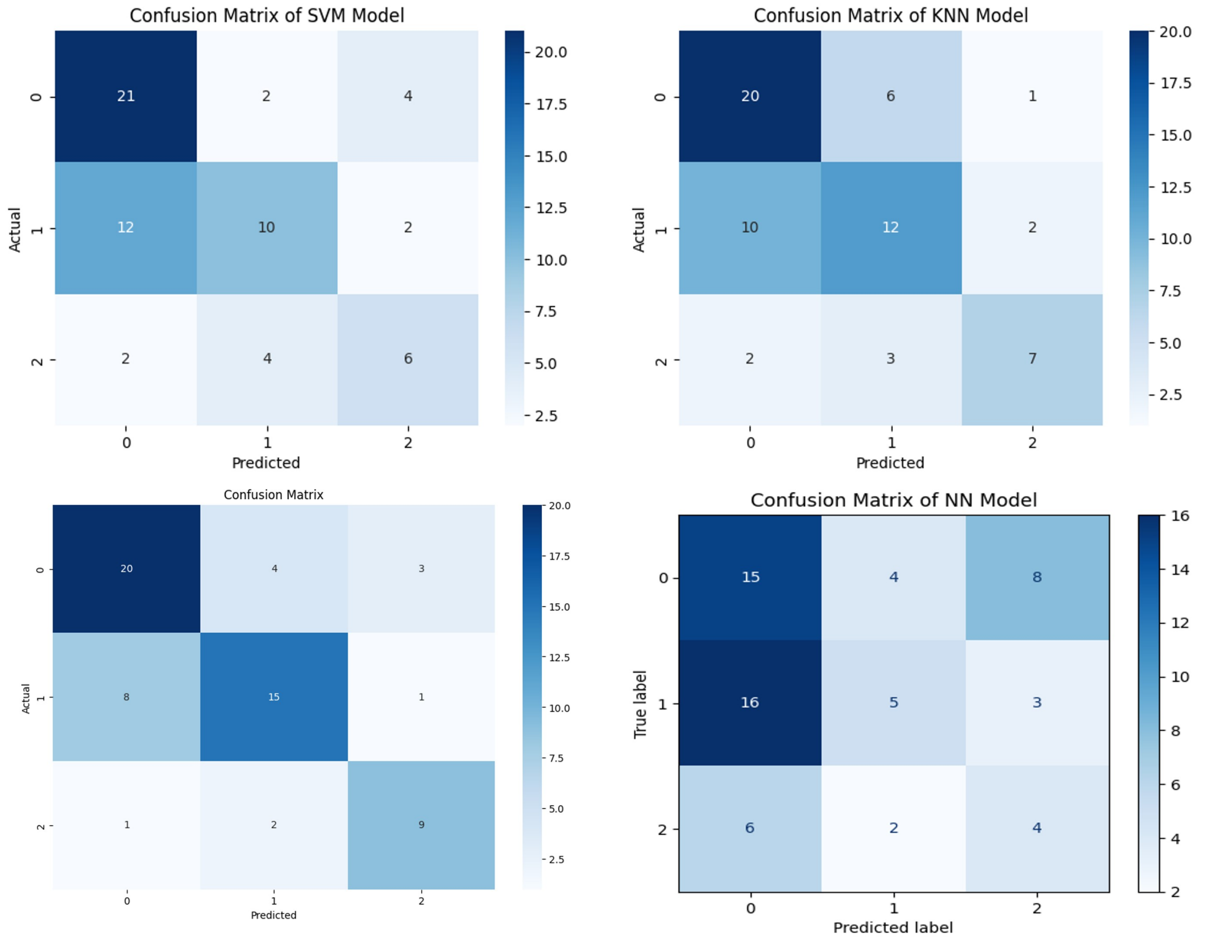


Figure 6: Confusion matrix of different models.

5 Conclusion

We have used different machine learning models like SVM, KNN, Random Forest, and Neural Networks to predict the crystal structure for different lithium-silicon-manganese, iron, and cobalt-oxygen cathode materials. In this project, Random Forest gave us an accuracy of 70%.

6 References

1. Shandiz, M. Attarian, and R. Gauvin. ”Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries.” Computational Materials Science 117 (2016): 270-278.