DOI: 10.54254/2755-2721/52/20241564

Machine learning classification methods for battery cathode material crystal system

Zhengli Wu

Department of Material, University of Manchester, Manchester, United Kingdom

Zhengli.wu1@envision-aesc.com

Abstract. The optimization of lithium-ion battery material design is becoming increasingly pivotal in the context of the transition to new energy sources. This research delves into the classification of the crystal structures of 339 sets of lithium silicate cathode materials, depending on the chemical and physical parameters of the materials to classify the crystal system type. Five classification algorithms were employed in this study, and the classification accuracy was enhanced through algorithm optimization. Notably, the accuracy of the random forest model with PCA (Principal component analysis) =5 reached 74% accuracy with 80% of the data used for training. Simultaneously, this study optimized the model accuracy by fine-tuning parameters such as depth, k-fold, and learning rate. Through this research, the classification accuracy of cathode materials in batteries was further elevated, bearing significant implications for the development of new, suitable battery materials. This research aids in expediting the screening process for researchers and facilitating industrial-scale experimentation, thereby accelerating the application of next-generation batteries.

Keywords: Machine Learning, Battery Cathode Material Crystal System, PCA.

1. Introduction

At present, due to the crucial phase of energy transition, lithium-ion batteries are recognized as powerful tools for this transformation and are being increasingly applied on a large scale in industries such as electric vehicles, electric trains, and energy storage. More and more research is focusing on the technological advancements of lithium-ion batteries, researchers are aiming to significantly enhance the energy density of lithium-ion batteries, which has led to continuous study in cathode materials, the crystal structure of battery materials a focal point of research.

Machine learning (ML) applications in materials science have become increasingly prevalent. This is primarily driven by the establishment of numerous materials databases, For example, Material Project, AFLOW and NOMAD, which are populated with data from extensive laboratory experiments and measurements [1]. Currently, some algorithms have been successfully employed for predicting battery performance. Machine learning can leverage vast feature data for classification, considering various data dimensions to achieve high classification accuracy. When these algorithms are applied to the classification of materials, they can rapidly assess material performance and determine whether the composition of a material is suitable as a candidate for high-performance batteries.

In this study, a database containing 339 candidate orthosilicate cathode materials was established, complemented with data such as chemical energy and atomic density. Various classification algorithms were employed to categorize these materials and infer their crystal systems. Throughout this process, machine learning algorithm parameters were continuously adjusted to explore whether data processing and parameter tuning could further enhance algorithm accuracy. The results indicated that different machine learning algorithms applied to the data exhibited varying levels of accuracy, which were also correlated with the volume of data in the training dataset. Moreover, these algorithms significantly expedited the determination of crystal systems for battery cathode materials.

2. Literature Review

2.1. Machine Learning Algorithms in Material Science

Machine learning predictions in materials design and computation rely heavily on platforms with access to extensive databases. Nowadays, within the research community, several open platforms host a wealth of experimental numerical characterizations of material physical and chemical properties, which also facilitates the application of machine learning algorithms.

Before Machine Learning technology applied widely, many computing calculation of material prosperities used by Density Functional Theory (DFT). Obviously, DFT provides the big data for ML to predict the properties of material at a quantum level. ML, on the other hand, leverages data-driven approaches to predict material properties and screen large datasets efficiently. These two approaches are often used together, with ML helping to narrow down the possibilities generated by DFT and experimental work, accelerating the process of material design and discovery [2].

By way of illustration, during the study of battery conductivity, the researchers incorporated machine learning techniques to predict the conductivity of different compositions, combinations at 373k, which allows for efficient examination of experimental results for different composition ratios. Driven by algorithms, more experimental repetitive work can then be avoided [3].

More specifically by algorithms, some research used SVM (Support vector machine) model to classify the thermal stability of LLZO(Li7La3Zr2O12) compound. At the same time, logistic regression was applied for interface reaction energy. So machine learning is becoming a key method to deal with the data and to make prediction [4].

In recent years, innovation in technology has been primarily focused battery cathode material design. In this project, more concern are focused on battery cathode material. Normally, the common cathode material is choosen from tramsition metal oxides and polyanionic compounds, also with chalcogenides [5]. Recently, one type of polyanion compounds with the orthosilicate structure ($Li2XSiO4$, X = Mn, Fe, Co) are a leading contender as appropriate cathodes for Li-ion batteries [6].

On a more micro level, turning to cyrstal system, predicting the crystal system categorization of lithium metal compounds becomes crucial. As known, crystallographic structures significantly influence the characteristics and functionality of batteries cathode material. For instance, crystal structures are able to affect the safety, longevity, charging rate, and energy density of battery through the cathode material. As a reuslt, accurate crystal system classification helps in better understanding the performance of battery materials in advance.

2.2. Classification Algorithms

With respect to classification algorithms, generally speaking, they are categorized into multi-label tasks and multi-classification tasks. In the course of this project, which belongs to the multi-classification task, all the Cathode materials are classified into three types of crystal systems, which based on the arrangement of atoms.

Classification algorithms have a wide range of applications covering health, finance and environmental climate. Specifically, Casanova et al. researcher studied fundus photography data with classification algorithms, random forests and logistic regression were applied in this diabetic retinopathy classification study and better results were obtained [7].

In fact, high latitude data often need to be processed, which includes noise reduction and generalization to reduce overfitting, etc., so the feature selection of the data is very important [8].

Secondly, for classification algorithms, attention needs to be paid to the balance of the data to ensure that the final classification data reaches the optimal solution by evaluating each classifier data in a hierarchical manner [9]. Therefore, the feature selection of data in the classification algorithm is a very important step.

The model building after data processing, different classification algorithms have different advantages, for example, for Random Forest, has a certain robustness to noise, and is not easy to overfitting. For K-Nearest Neighbor, the learning and classification is done directly through examples, the mechanism is very simple, and the results are very intuitive.

In this project, different algorithms were tried to compare the accuracy, with the aim of obtaining better classification algorithms to create predictions for the classification of crystalline systems. This will help to better develop new battery cathode materials and help to further the development of new energy technologies.

3. Dataset

The data for this paper comes from the Kaggle website, tracing back to the original data which provided by the Material Project platform for 339 types of lithium-ion battery cathode materials in the lithium silicate series. The overall data size is 339 rows * 11 columns. Some Details are shown in Table 1.

Table 1. Overall data view

	Formation	E Above Hull (eV)	Band Gap (eV)	Nsites	Density(gm/cc)	Volums
Count	339.00	339.00	339.00	339.00	339.00	339.00
Mean	-2.62	0.06	2.08	38.84	2.98	467.77
Std	0.17	0.03	1.09	23.13	0.35	292.67
Min	-2.98	0.00	0.00	10.00	2.20	122.58
Max	-2.01	0.19	3.82	132.00	4.20	1518.85

The components of the data involve a number of physical quantities in materials science, so a detailed interpretation is necessary. At first, formation energy: In the context of materials science and chemistry, formation energy quantifies the energy required (endothermic) or released (exothermic) when atoms or ions come together to form a particular chemical compound or crystal structure; Band gap: The energy difference between the lowest conduction band's bottom and the valence band's top is known as the band gap;E above Hull:Energy if decomposition of material into most stable ones [10]; Space Group: The symmetries of crystal patterns are described by space groups, while the symmetry of the macroscopic crystal is represented by the space group's point group; Nsites: Number of atoms in the unit cell of the crystal; Bandstructure: Or the electronic band structures of a solid, explain the electrical characteristics of a material by describing the energy that electrons possess.

Before formal data reading and model building, the model needs to be cleansed. Cleaning the data can help to improve the quality of the data by eliminating errors, missing values, and inconsistencies in the data. Low-quality data may lead to inaccurate analysis results and may introduce misleading conclusions. After analyzing the data, there are no duplicate rows or empty data [11]. As a result, further analysis could be applied by these data.

The following Figure 1 is basic information about the dataset, including the numerical distribution of each feature. Among them, only 19.2% of the materials do not have bandstructure, and most materials have bandstructure. The data in Figure 1, densities show a positive distribution, and the most concentrated density distribution is around 3 gm/cc.

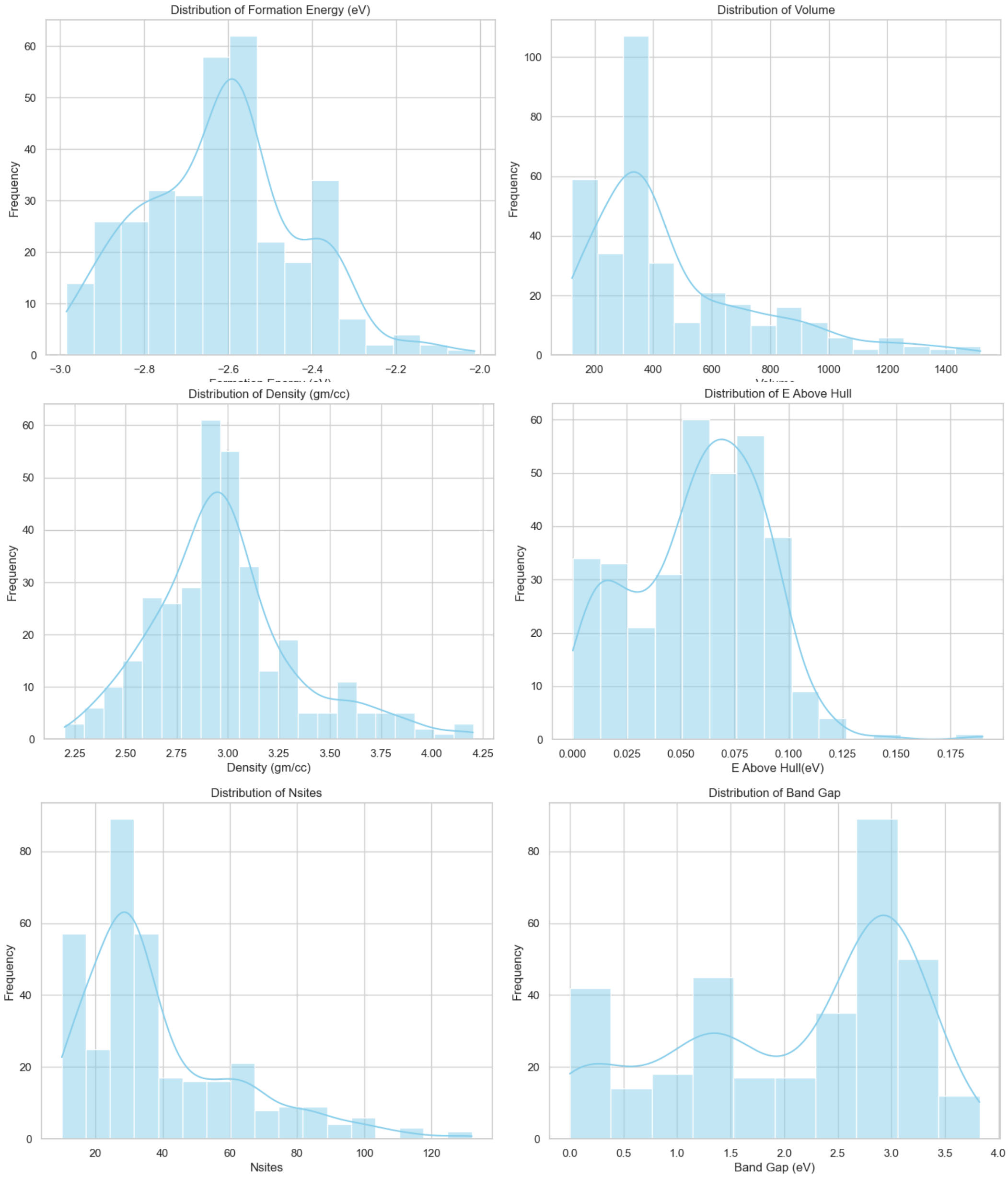

Figure 1. The distribution of data

Figure 2 shows that the intensity of the space group feature is significantly higher than other feature data, which is supported the conclusion from the UCL (University College London). Interpreting the relationship of the data in terms of the heat map, it can be observed that Nites and Volume are correlated close to a value of 1, because in physics, a larger volume also indicates contains a larger number of atoms; in terms of the interrelationships of the other data, there is no obvious strong correlation.

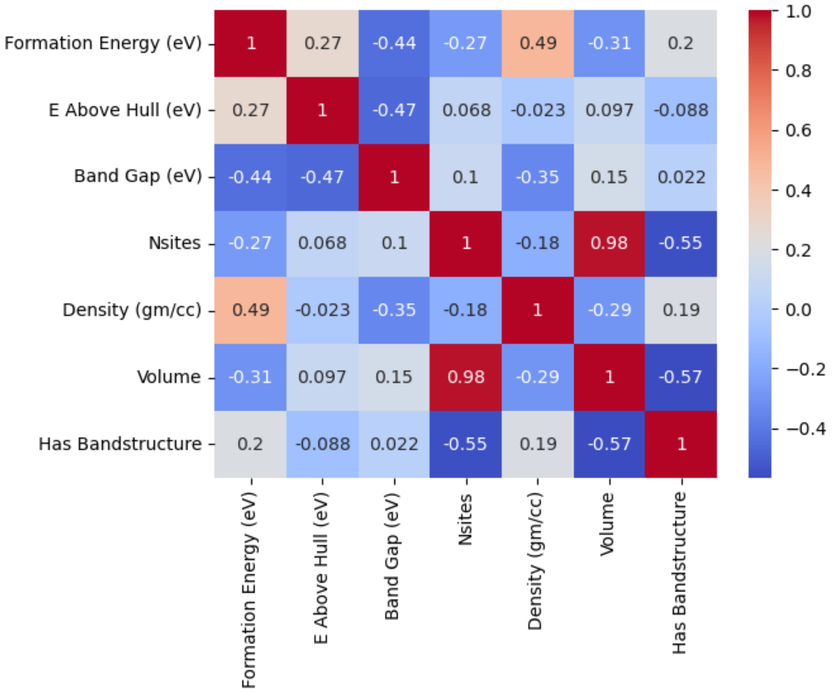

Figure 2. Heatmap of features

4. Algorithms and Results

In this project, 5 different classification algorithms are applied for the processing of the crystal system. In the following exposition, the mechanism of each model is described, as well as the process of processing the data. In this process, the working mechanism of the algorithms is better understood. The final accuracy will be presented graphically, and the accuracy scenarios for different numbers of training datasets will be explored.

4.1. Normal Random Forest

Random forest, this classifier has an endless number of decision trees that may be added to it. These decision trees are then integrated in a complimentary or weighted way to create a new classifier that is referred to as the random decision forest [12].

In the process of classification algorithm, a noteworthy observation is that when including the space group as a feature, the predictive accuracy soars to an impressive 99%. Upon an in-depth examination of the relevant literature, the significant explanation is: space group plays a pivotal role in the unequivocal definition of the crystal system [13]. This finding underscores a substantial correlation between space group and crystal system. According the explanation by Univerisity of London, the ordering of the Lau calss for space group is designed to describe the crystal system. For instance, consider $Li2MnSiO4$, which belongs to space group $P21/C$. This classification is due to the lattice center's designation as P, the crystal class as 2/m, and the Laue class as 2/m. Therefore, it could be considered that the spacegroup is directly connected to crystal system (See Figure 3).

In light of this discovery, other data feature, excluding space group, have been selected for crystal system classification prediction in the next step. This strategic adjustment allows for a more comprehensive understanding of the influential factors, such as band gap and Nsites, in the determination of crystal categories.

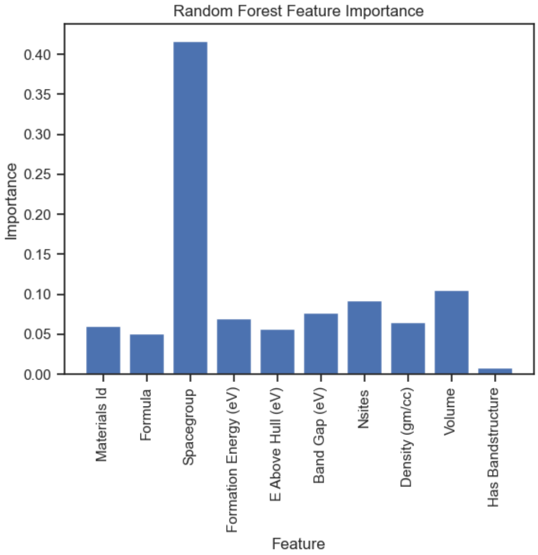

Figure 3. The feature importance of random forest

Back to the random forest alogrithim, in the initial parameters, the number of estimators was chosen to be 100, and the maximum depth was 10, which gave an accuracy of 60%, better than the decision tree (See Figure 4). After that, for the optimal depth to be explored, you can see the Figure 4 that when the depth reaches 13, and all other parameters remain unchanged, there is a slight increase in the accuracy, to 63%. Controling the depth is to avoid the random forest to a overfitting status, which can be obsevred the higher depths instead lead to a decrease in accuracy.

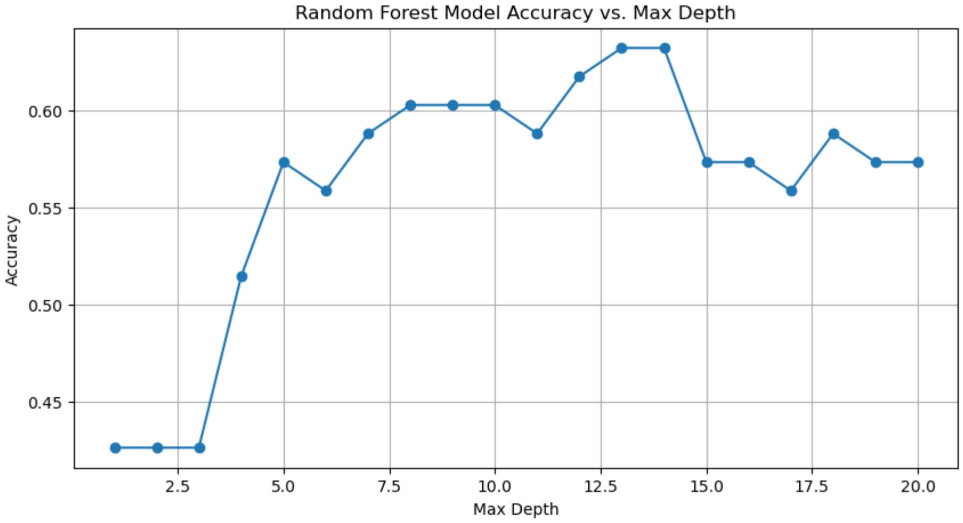

Figure 4. Best Depth for random forest

4.2. Decision Tree

Move to Decision Tree, it is a top-down model that will start splitting based on the data features and keep making decision judgments for different groups in order to achieve classification. Nowadays, the popular one is Cart Decision Tree model, which makes judgment based on Gini coefficient, the smaller Gini coefficient means more ideal classification judgment.

As shown in Figure 5, if the decision tree with spacegroup data is attached, the bifurcated branches are very few, and the classification of the crystal system can be completed quickly based on the data features and achieve a high accuracy rate. When the spacegroup is removed, the decision tree starts to have more bifurcations and the accuracy decreases. This is because the correlation between the feature data is no longer strong, and more decision decisions need to be added to the decision tree.

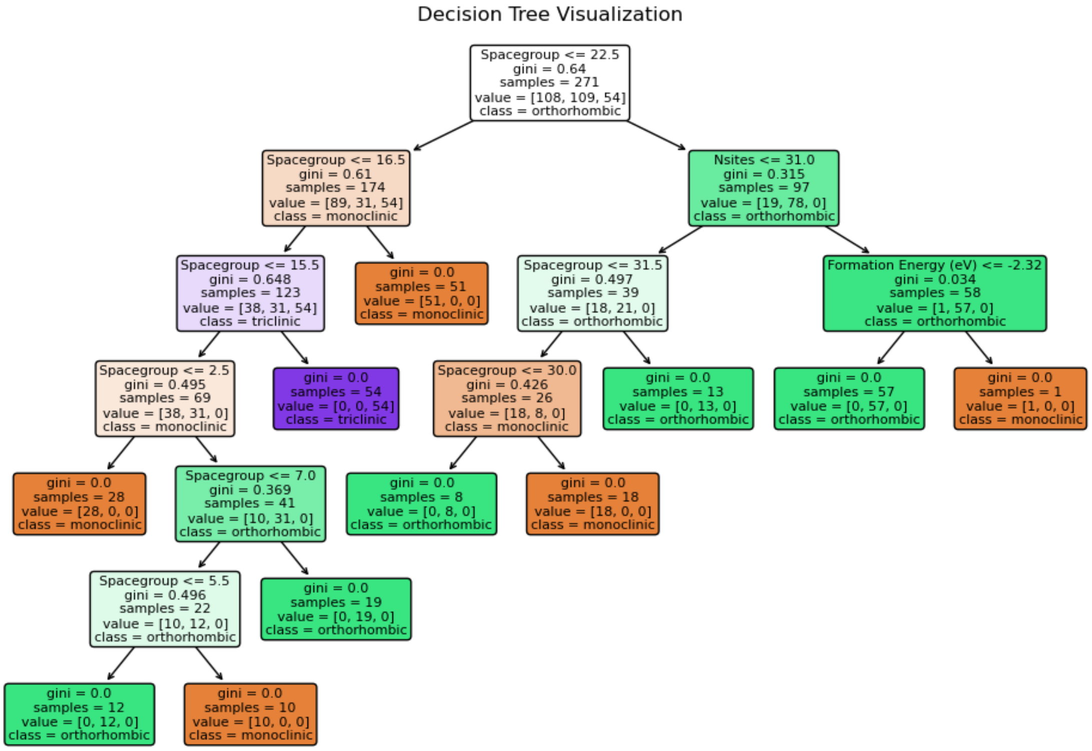

Figure 5. Decision tree schematic diagram with space group data

During the coding process of decision tree, the target value is assigned to crystal system, and others data act as selected feature. The dataset has been split to training and testing, and the fitting process would be carried in testing data to evaluate the model accuracy. From the results of the decision tree, a classification accuracy of 51% was obtained using 80% training data and random state=42, which is not a good accuracy.

4.3. XGboost

XGboost (Extreme gradient boosting) is an implementation of the Gradient boosting decision tree machine learning algorithm that incorporates regular terms to control the complexity of the model. Authored by Dr. Tianqi Chen from the University of Washington, this algorithm performs a second-order Taylor's formula expansion of the cost function, which allows for the use of both first- and second-order derivatives.

During the XGboost algorithm, Since the crystal system needs to be classified into three categories, multi:softmax is imported into the algorithm. multi:softmax's loss function calculates the loss mainly by comparing the predicted value of each sample with its actual category. XGBoost tries to minimize this loss in order to find the best classification model.

An accuracy of 51% was achieved using this code, while also exploring the deep learning rate and the number of estimators to finalize the boost to 52% (learning rate=0.01, max depth=6, n_estimators=100), but the effect of the whole one accuracy boost was not very significant.

4.4. Random Forest with PCA Process

PCA (Principal component analysis) is a dimensionality reduction technique designed to reduce the dimensions of a dataset. A smaller number of PCA components can reduce computational complexity and mitigate the risk of overfitting. In this process, the proportion of redundant information is reduced, and the proportion of important original features is further increased. The detailed operation is: For the labeled data, remove the label and then perform PCA data dimensionality reduction, then use the downscaled data for model training.

In this research, for PCA was introduced to build a random forest model with an accuracy of 74% and gave the best Number of Principal Component equal to 5. As Figure 6 shows, the number of PCA will affect the accuracy, and initially with the number increasing, the accuracy increases as well. The optimal accuracy of the curve stops at n=5 and then the accuracy starts to decrease again (See Figure 6).

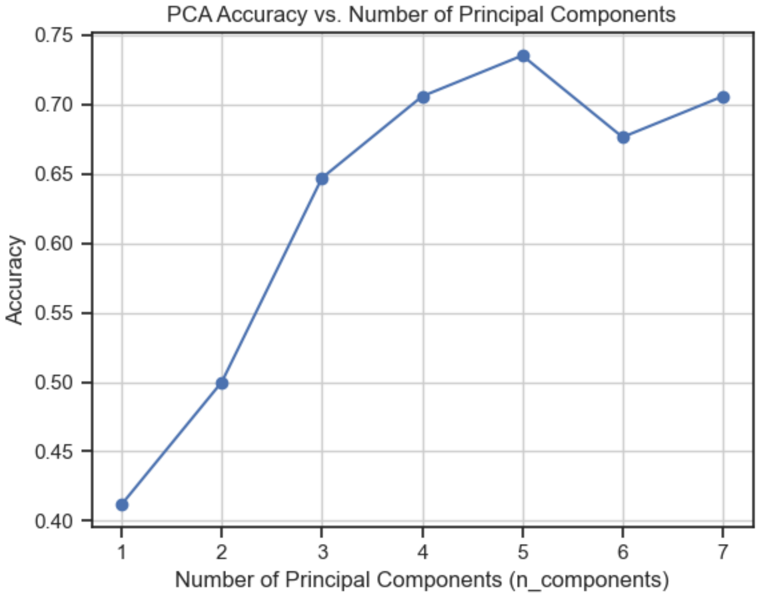

Figure 6. Model Accuracy with PCA numbers

4.5. KNN

Finally, K-Nearest Neighbor (KNN) algorithm is used to classify different features by measuring the distance between them and then it is classified as LAZY LEARNING. Take an instance, if a sample is in the range of features and most of the nearest sample belongs to a feature, then the sample will be attributed to that feature as well. Therefore, the accuracy is difficult to be guaranteed under the condition that multiple features are not significant. In the actual validation, the accuracy was only 56%, and for the best k-fold a lookup was performed with a value of 1.

4.6. Accuracy

In the following table, for the accuracy of the algorithms used in this research is summarized. In the classifications models for the crystal system, if PCA=5 is added to the random forest, the accuracy reaches 74%. This is quite high compared to the algorithm of KNN and Decision tree alone (See Table 2).

Table 2. Classification accuracy of models

	Decision Tree	Random Forest	XGBoost	Random Forest with PCA	KNN@k=1
Accuracy	0.51	0.63	0.52	0.74	0.56

4.7. The Effect of Training Data

In the last module, the number of training poles data for the algorithm was explored and the accuracy of the model was seen to be different for different numbers of training. For example, the highest accuracy of the model in the random forest algorithm added by PCA occurs at 60% training volume. For XGboost, on the other hand, it occurs at 90% training data. This can be explained by data overfitting with higher dimensional data complexity (See Figure 7).

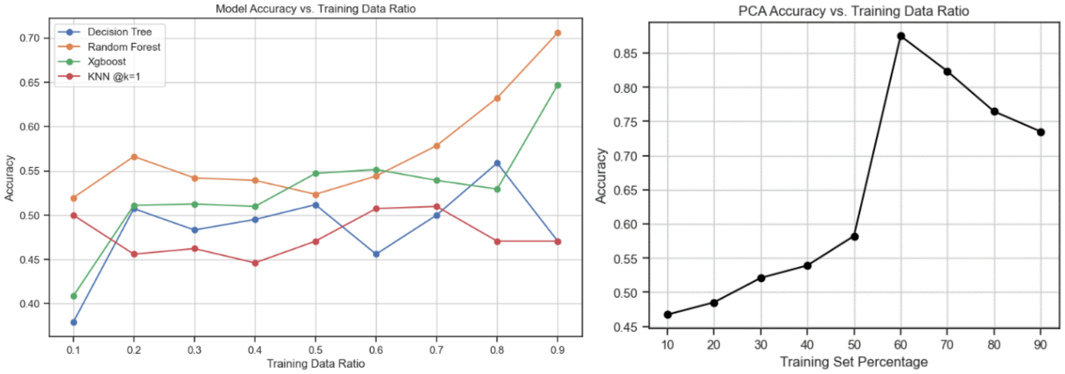

Figure 7. Training data ratio influence on model accuracy

5. Conclusion and future work

In summary, the core focus of this research is the classification of battery cathode materials. Initially, a multidimensional analysis of the data was conducted, and the most informative features were selected for classification prediction, with corresponding interpretations provided for each dataset. Subsequently, in the realm of model exploration, the latest classification algorithms were introduced, such as XGBoost and PCA, and comparing them to traditional models. All analytical findings are presented through graphical visualizations, enhancing the depth of understanding of the classification outcomes.

In this research process, the classification of the crystal system was tested with 5 different models, and the corresponding accuracy was obtained after processing the data for various physicochemical features. Through the learning and research of these 5 classification models which mentioned above, crystal classification of cathode material for battery becomes clear, and parameter optimization of K-fold, learning rate and depth et.al in the algorithm is achieved. The adjustment of the parameters of these algorithms also helped a lot in understanding the whole model mechanism, which made the accuracy of the classification of the crystal system also improved.

Finally, the random forest with PCA=5 is the most accurate for classification, reaching to 74% accuracy, indicating the usefulness of adjusting the parameters and data pre-processing efforts.

Adjustment for training data is also one of the key measures when the dataset is constant. However, there is limited understanding of the visualization and theory of the modeling process, and in the future, more details of the model need to be analyzed to improve the overall accuracy of the classification.

The algorithms for these classifications have helped screen potential candidates suitable as battery cathode materials faster and enable better optimization of the model with more quality data in the future. The AI algorithms that aid with material design are evolving and more work needs to be done in the future for the data and models that need to be developed. At the same time, there is a need to be more aware of the technical needs of current industrial developments, for example, in the battery industry, how the first charge and discharge of a battery is deployed to activate the battery can also be predicted and fitted using machine learning. Focusing on the needs of the industry can lead to a wider and more valuable application of machine learning.

References

[1] Joshi, R. P., Eickholt, J., Li, L., Fornari, M., Barone, V., & Peralta, J. E. (2019). Machine learning the voltage of electrode materials in metal-ion batteries. ACS applied materials & interfaces, 11(20), 18494-18503.

[2] Ling, C. (2022). A review of the recent progress in battery informatics. npj Computational Materials, 8(1), 33.

[3] Fujimura, K., Seko, A., Koyama, Y., et al. (2013). Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. Advanced Energy Materials, 3(8), 980-985.

[4] Luo, Z., Yang, X., Wang, Y., Liu, W., Liu, S., Zhu, Y., ... & Deng, Y. (2020). A survey of artificial intelligence techniques applied in energy storage materials R&D. Frontiers in Energy Research, 8, 116.

[5] Nitta, N., Wu, F., Lee, J. T., & Yushin, G. (2015). Li-ion battery materials: present and future. Materials today, 18(5), 252-264.

[6] Lv, D., Bai, J., Zhang, P., Wu, S., Li, Y., Wen, W., ... & Yang, Y. (2013). Understanding the high capacity of Li2FeSiO4: in situ XRD/XANES study combined with first-principles calculations. Chemistry of Materials, 25(10), 2014-2020.

[7] Casanova, R., Saldana, S., Chew, E. Y., Danis, R. P., Greven, C. M., & Ambrosius, W. T. (2014). Application of random forests methods to diabetic retinopathy classification analyses. PLOS one, 9(6), e98587.

[8] Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 52.

[9] Chen, R. C., & Liao, C. Y. (2018, April). Deep learning to predict user rating in imbalance classification data incorporating ensemble methods. In 2018 IEEE International Conference on Applied System Invention (ICASI) (pp. 200-203). IEEE.

[10] Motooka, T., & Uda, T. (2020). Multiscale modeling methods. In Handbook of silicon based MEMS materials and technologies (pp. 249-261). Elsevier.

[11] Dasari, D., & Varma, P. S. (2022). Data Cleaning Techniques Using Python. Technology, 1(1), 11-21.

[12] Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

[13] Souvignier, B., Wondratschek, H., Aroyo, M. I., Chapuis, G., & Glazer, A. M. (2015). Space groups and their descriptions.