Quest Journals
Journal of Software Engineering and Simulation
Volume 9 ~ Issue 10 (2023) pp: 11-21
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Predicting Crystal System of Cathode Materials in Lithium-Ion Batteries Using Machine Learning Models

Tyler Fu
Princeton International School of Mathematics and Science, NJ, USA

ABSTRACT: *The crystal system of a lithium-ion battery cathode can significantly impact its chemical properties. This study aimed to use data from The Materials Project to build a machine learning model to predict the crystal structure of a cathode. Statistical tests demonstrated a strong correlation between the crystal structure of a substance and its chemical properties. To develop this model, the dataset was randomly divided into the training set, which contains 80% of the data and the remaining 20% for testing, and an XGBoost decision tree model was then trained to predict the three major types of crystal structures (monoclinic, orthorhombic, and triclinic) of cathodes. Remarkably, the model achieved a prediction accuracy of 94%, surpassing previously reported benchmarks of 75% in another study. This research establishes the feasibility of predicting the crystal structure based on the chemical properties of materials. Additionally, the study not only identifies key features for accurate prediction but also enriches understanding of the relationship between crystal systems and different cathode materials.*

KEYWORDS: *Lithium-ion cathodes, battery, crystal system, machine learning, encoding, XGBoost*

R*eceived 26 Oct., 2023;* ***R****evised 06 Nov., 2023;* ***A****ccepted 08 Nov., 2023* © *The author(s) 2023.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

Lithium-ion (Li-ion) batteries consist of three major parts: cathode, anode, and electrolyte. Each element engages in a redox reaction, in which certain reactants acquire electrons while others lose them. Functionally, the cathode acts as the oxidizing agent, seeking to grab electrons; conversely, the anode is the reducing agent, aiming to release electrons. The electrolyte, on the other hand, allows lithium ions to move between the cathode and anode. During the discharge process of a lithium-ion battery, positively charged lithium ions (Li^+) travel from the anode to the cathode through the electrolyte [1]. The anode oxidizes lithium into lithium ions, which then bind to the cathode [2]. Simultaneously, electrons traverse from the cathode to the anode via a circuit, creating the flow of electric current. This study focuses on the structural aspects of the cathode within Li-ion batteries [1].

Li-ion batteries have many different types of cathodes that usually are crystals. A crystal is a repeating arrangement of atoms, and the smallest arrangement of atoms that can repeatedly produce a crystal structure is called a "unit cell". There are 14 basic unit cells called Bravais lattices, falling within 7 main primitive crystal systems, shown in Table 3. The specific crystal system depends on factors like the distance between the corners of the unit cell and the angles between the edges of the unit cell [3]. Lithium ions can easily bind or unbind themselves from a crystal structure through the process of intercalation. Intercalation is one of the reasons lithium-ion batteries can discharge and recharge many times (Layered Structures and Intercalation Reactions). The aim of lithium-battery cathodes is to have the lowest reduction potential possible – a substance's tendency to get reduced – while maximizing the reduction potential of the anode because the difference in redox energies determines the voltage of the battery [4].

The crystal system of a cathode significantly influences its electrochemical properties, directly impacting battery performance, such as capacity and voltage. The ability to accurately predict the crystal system is instrumental in estimating cathode performance for specific applications. Leveraging the capabilities of machine learning to handle complex data patterns and make accurate predictions, this study sets out to achieve its primary objective: to develop a machine-learning model capable of accurately predicting the crystal system of a lithium-ion battery cathode based on the characteristics of the cathode material. The dataset employed in this study originates from The Materials Project, an open web-based platform with access to the physical and chemical properties of many materials.

II. METHOD

2.1 Data

The dataset employed in this research originates from the Materials Project. The dataset comprises 339 records corresponding to distinct cathode materials and contains 11 variables. These variables include material ID, chemical formula, space group, formation energy, energy above hull (eabovehull), band gap, number of sites (nsites), density, volume, hasbandstructure, and crystal system of each cathode material. Among the 11 variables, material ID was excluded from the analysis because it does not contribute to the substance's properties. There was no evident missing data, and duplicated rows were thoroughly checked before and after removing the "material ID" column. Table 1 provides an overview of the remaining 10 columns within the dataset, along with their respective descriptions. Additionally, Table 2 presents randomly selected five rows from the dataset to offer a snapshot of the nature of the data.

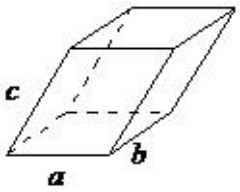
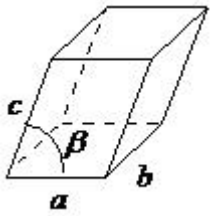
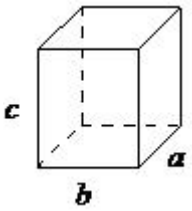
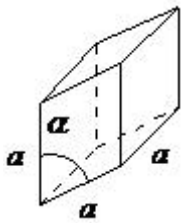
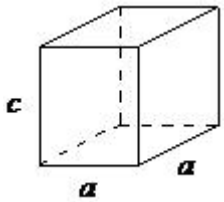
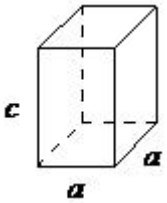
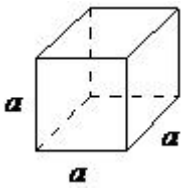
Table 1: Description of Variables in the Dataset

Formula	The chemical formula of the crystal, which is the ratio of the different elements in one unit cell
Spacegroup	Spacegroup notation describes the symmetry of an infinitely repeating lattice. The first letter is the Bravais lattice symbol. The numbers after the first letter describe the screw axis symmetry: rotation and translation. The first number, n, means rotation of 360 degrees divided by n, and the second number means translation by that number of unit cells. The last letter describes the glide plane symmetry: translation and reflection. The letter depends on the direction of the translation [5].
FormationEnergy	The change in energy (electron volts or eV) when one unit cell of the substance is formed
EAboveHull	Energy above hull is the formation energy difference between a compound and its most stable form, in electron volts (eV) per atom [6] and [7]. Energy above the hull is useful in this project because many of the materials in the dataset have the same formula but different structures.
BandGap	The band gap is the minimum amount of energy (eV) to excite one electron from the valence orbital to the next highest unoccupied orbital, called the conduction orbital. The smaller the bandgap gets, the more conductive the substance is.
Nsites	The number of atoms inside one unit cell, including fractions of atoms shared with other unit cells.
Density	Density of the substance in grams per cubic centimeter.
Volume	Volume of a unit cell in cubic angstroms.
HasBandstructure	A band is a filled orbital in the atom. A band structure is a graph showing the allowed orbitals in a substance, including the conduction orbital (mentioned in the band gap) [8]. Some substances do not have an allowed conduction band, so they do not have a band structure [9].
Crystalsystem	The crystal system is the shape of a unit cell. There are only three crystal systems in the data, monoclinic, triclinic, and orthorhombic, out of the seven main crystal systems shown in Table 3.

Table 2: Random Selection of Five Rows from the Dataset

Material Id	Formula	Space group	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Nsites	Density (gm/cc)	Volume	Has Bandstructure	Crystal System
mp-849394	Li2MnSiO4	Pc	-2.699	0.006	3.462	16	2.993	178.513	TRUE	monoclinic
mp-762762	LiFe2(SiO4)2	P1	-2.426	0.114	0	39	2.753	547.911	FALSE	triclinic
mp-762828	LiMnSiO4	Pna21	-2.623	0.054	0.11	84	3.55	864.216	FALSE	orthorhombic
mp-566680	Li2MnSiO4	P21nm	-2.705	0	3.052	16	3.039	175.842	TRUE	orthorhombic
mp-767709	Li2Mn3Si3O10	C2/c	-2.747	0.016	2.578	36	3.334	421.286	TRUE	monoclinic

Table 3: Seven Main Primitive Crystal Systems

System	Lengths and Angles	Unit Cell Shape
triclinic	a≠b≠c and ∠bc≠∠ca≠∠ab	
monoclinic	a≠b≠c, ∠bc=∠ab = 90°, and ∠ca ≠ 90°	
orthorhombic	a≠b≠c and ∠bc=∠ca=∠ab = 90°	
rhombohedral	a=a=a and ∠aa ≠ 90°	
hexagonal	a=a≠c, ∠ca = 90°, and ∠aa = 120°	
tetragonal	a=a≠c and ∠ca=∠aa = 90°	
cubic	a=a=a and ∠aa = 90°	

Source: Chemistry LibreTexts

2.2 Exploratory data analysis

Exploratory data analysis was the initial step to gain an understanding of the data. First, the number of unique variables and the number of entries for each variable were identified, and then the variables were divided into classes of categorical and numerical variables. In this dataset, the categorical variables are formula, spacegroup, hasbandstructure, and crystal system; however, hasbandstucture was coded to 0 and 1, representing false and true, respectively. The numerical variables are formation energy, volume, nsites, volume, energy above hull (eabovehull), and band gap. For each numerical variable, summary statistics were calculated, including the mean, standard deviation, minimum, maximum, 1st quartile, median, and 3rd quartile. This information is presented in Table 4. Similarly, for each categorical variable, the number of each unique value and its corresponding frequency was calculated. This information is also shown in Table 4. As part of the analysis, histograms were generated to compare the three crystal systems and hasbandstructure, depicted in Figures 1 and 2. Figure 1 delineates that 65 crystals do not have a bandstructure, while 274 crystals have a bandstructure. The distribution of 139 monoclinic, 128 orthorhombic, and 72 triclinic crystals is also exhibited. Figure 2 provides visual representations of formation energy, eabovehull, bandgap, and hasbandstructure color-coded by the crystal system.

	count	mean	std	min	25%	50%	75%	max
Formation Energy (eV)	339.0	-2.616950	0.183809	-2.985	-2.7575	-2.605	-2.5255	-2.012
E Above Hull (eV)	339.0	0.058215	0.030363	0.000	0.0355	0.062	0.0815	0.190
Band Gap (eV)	339.0	2.079740	1.087968	0.000	1.2655	2.499	2.9680	3.823
Nsites	339.0	38.837758	23.133142	10.000	26.0000	31.000	52.0000	132.000
Density (gm/cc)	339.0	2.984003	0.353968	2.200	2.7605	2.947	3.1060	4.201
Volume	339.0	467.765619	292.674559	122.581	286.3815	358.537	601.6965	1518.850
Has Bandstructure	339.0	0.808260	0.394252	0.000	1.0000	1.000	1.0000	1.000

Table 4: Summary Statistics of Numerical Variables and Categorical Variables

	count	unique	top	freq
Formula	339	114	LiFeSiO4	42
Spacegroup	339	44	P1	72
Crystal System	339	3	monoclinic	139

Figure 1: Count Plot for Band Structure and Crystal System

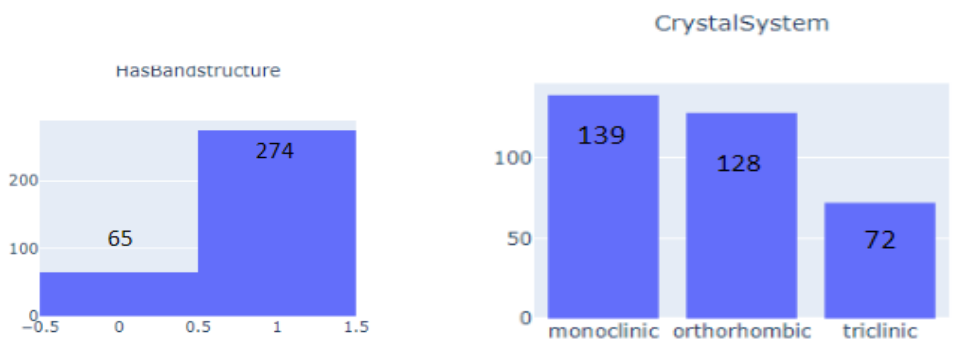
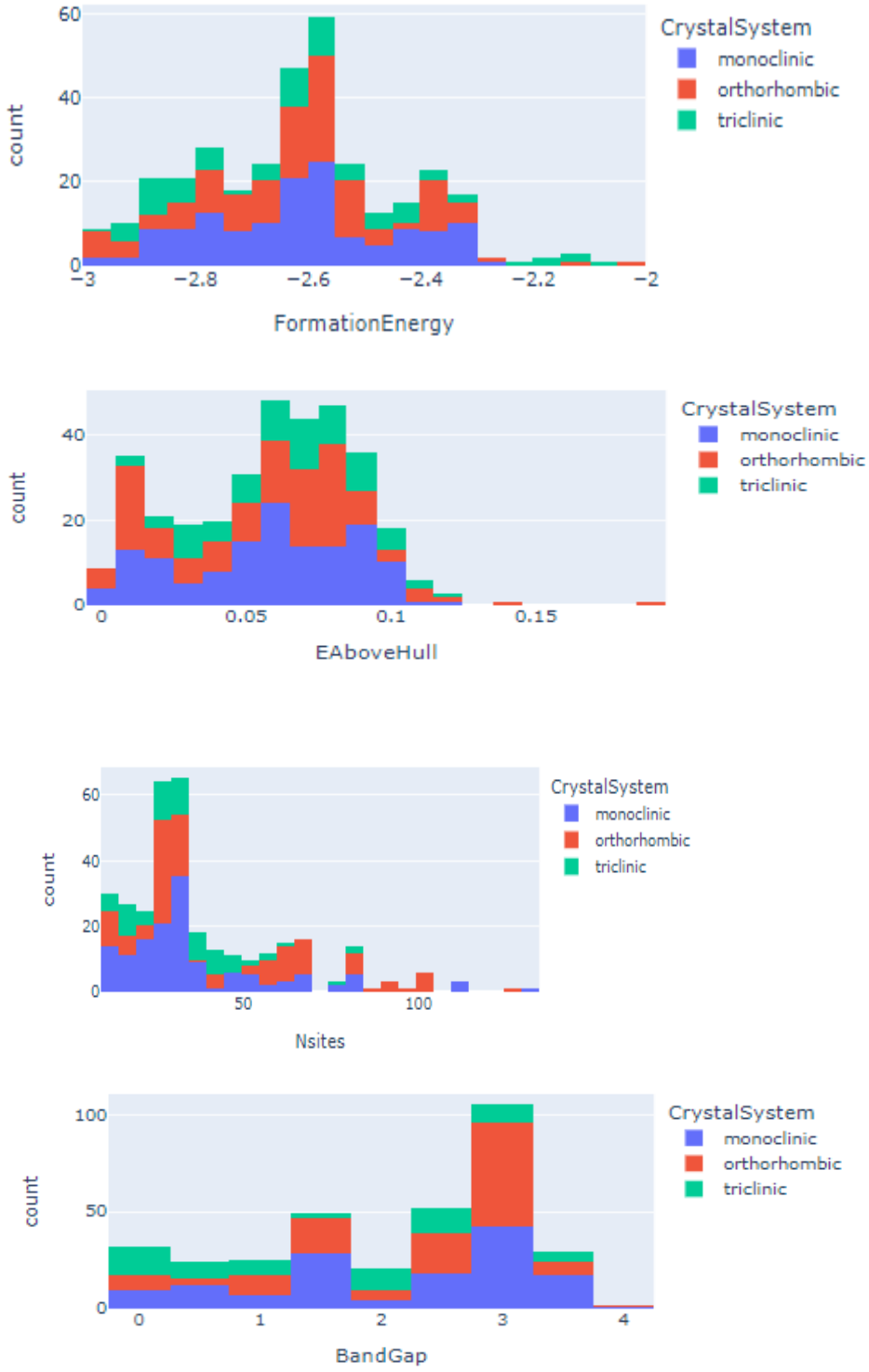
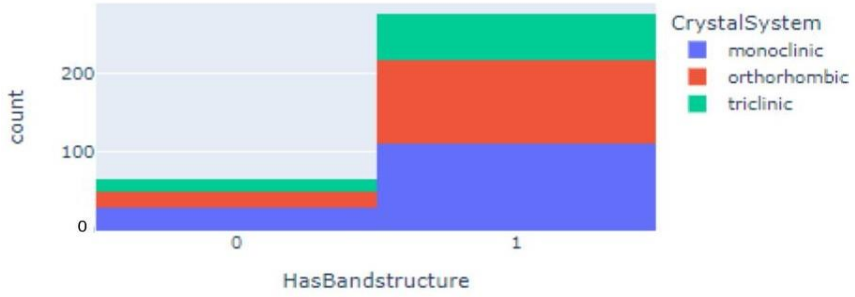


Figure 2: Histograms of Numerical Variables Color-Coded by Crystal System





This comprehensive exploratory analysis offers a foundational insight into the dataset's characteristics and sets the stage for subsequent analytic steps.

2.3 Statistical Tests

First, the Pearson correlation method tested the linear correlation between the numerical variables. The correlation coefficient ranges from -1, by which two variables make a straight line with a negative slope, indicating a negative linear relationship, to 1, by which two variables make a straight line with a positive slope, indicating a positive linear relationship. Figure 3 shows a correlation heatmap highlighting variable pairs with correlation coefficients above 0.4 or below -0.4 [10].

Next, a one-way analysis of variance [11] test was performed for each dependent variable on each independent variable one at a time. ANOVA compares the variance between and within the groups to determine if there is any statistically significant difference among the means of levels of independent variables. The variance between groups is represented by the ratio of the sum of squares between groups and degrees of freedom. The sum of squares between groups measures the variability between the groups, which is how far away each number in a group is from the overall mean of all the groups combined. The degrees of freedom for the sum of squares are the number of independent pieces of information in the test, and for the sum of squares between groups, the degrees of freedom is one less than the total number of groups in the test. The variance within variables is the ratio of the sum of squares within and degrees of freedom. The ratio of the variance between and variance within is called the F-value. The p-value signifies the probability of getting an F-value that is the same or more extreme than the observed F-value. The ANOVA test was conducted on each categorical variable compared with all the numerical variables, and the resulting p-values are recorded in Table 5. If the p-value was less than 0.05, the null hypothesis was rejected. The null hypothesis assumes no differences between the variables, and any difference observed is due to random chance.

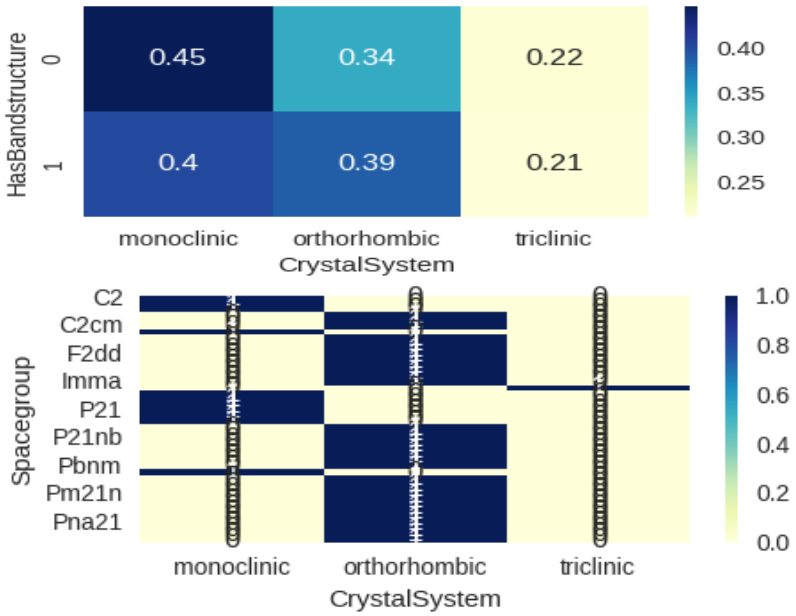
Lastly, Figure 4 displays two contingency tables comparing band structure and spacegroup against the three crystal systems. Each cell within the contingency table shows the percent of the specific combination of the two compared categories.

Figure 3: Correlation Heatmap Between Numerical Variables

Table 5: P-Values from Various One-Way ANOVA Models

	Crystal System	Spacegroup	Hasbandstructure	Formula
Formation Energy	0.98242	9.427552e-11	0.000244	1.058187e-190
Energy Above Hull	0.11273	4.196027e-08	0.105841	5.690134e-23
Band Gap	0.000073	0.000004	0.693063	1.690479e-64
Nsites	0.000265	3.623385e-22	2.838366e-28	0.000102
Density	0.015524	3.793224e-13	0.000488	1.601994e-20
Volume	0.001541	5.157082e-26	3.818104e-30	0.000004
Hasbandstructure	0.748037	0.000092	0	0.000014

Figure 4 : Contingency Table of Band Structure and Spacegroup vs Crystal System



In the correlation heatmap illustrated in Figure 3, nsites strongly correlates with volume with a coefficient of 0.98, and density and formation energy were weakly correlated. Moreover, band gap was negatively correlated with both formation energy and energy above hull; band structure was negatively correlated with nsites and volume. In Table 5, with the exception of crystal system versus formation energy and band structure versus band gap, most p-values lie below 0.05, which means they are all somewhat correlated. In the contingency tables displayed in Figure 4, the second plot is interesting because it shows that the crystal system strongly influences the spacegroup of a crystal because 100% of a certain spacegroup falls entirely in one crystal system.

2.4 Preprocessing

To prepare the data for the machine learning model, each chemical formula, spacegroup, and crystal system was encoded to a numerical number. Label encoding was applied for the output variable crystal system. Given the substantial cardinality of spacegroup and chemical formula, meaning numerous unique values, they were encoded differently. Spacegroup was frequency encoded, and a number was assigned based on the frequency of the certain spacegroup, and the formula was encoded using the CatBoost encoder.

2.5 Principal Component Analysis

Principal component analysis (PCA) was adopted to condense the number of variables by combining a large number of original variables into a few new ones called principal components. Before initiating principal component analysis, all the data were normalized, ensuring that each column had a mean equal to zero and a standard deviation equal to one. At first, the data were reduced to two principal components. The proportion of variance explained, which is how well a principal component "explains" the dataset, was calculated for the two principal components [12]. The proportion of variance explained by the first principal component was about 31.45%, and the proportion of variance explained by the second principal component was approximately 20.56%, totaling 52.01%. After that, the data were reduced to three principal components; the proportion of variance explained by the new third principal component was about 15.70%, totaling 67.71%, which was better. However, the principal components were not used in the final machine learning model to preserve as much information as possible.

2.6 Machine Learning Model

The machine learning model employed in this study utilized the XGBoost classifier, a decision tree-based algorithm in Python [13]. A decision tree entails a series of nodes and branches collectively classifying crystal systems. XGBoost leverages a feature called boosting, which creates decision trees sequentially so that each subsequent decision tree is designed to rectify errors made in the previous decision tree. Each decision tree contributes a bit to the prediction, so all the decision trees working together create an impactful machine learning model. XGBoost classifier is one of the most efficient machine learning algorithms, often outperforming most other models, such as logistic regression [14].

Before starting the machine learning, the output variable, the crystal system, was separated from the input variables. Then, 80% of the data were chosen randomly to be part of the training set, and the remaining 20% were for testing. Minmax scaler was adopted to normalize the training set, and the testing set was normalized individually. Hence, the mean was equal to zero, and the standard deviation was equal to one. After normalizing the data, an XGBoost classifier was trained on the scaled training data using the default parameters of XGBoost. Then, the machine learning model was tested on the testing data.

After testing the model, a confusion matrix and a classification report were generated. A confusion matrix illustrates how many crystal systems the model predicted correctly and incorrectly, as shown in Figure 5 [15]. In the classification report, precision is the ability of the classifier to not label a positive case as negative; recall is the ability to predict positive cases correctly; f1-score is the weighted average of precision and recall; support is the actual number of positive cases; and accuracy is the percentage of true predictions [16]. Table 6 summarizes the classification report. Figure 6 presents a plot of the feature/variable importance, which is the frequency a feature was used in the decision tree. A variable with high cardinality can affect the feature importance plot. Thus, Figure 7 offers a different plot of feature importance based on the gain for each variable, reflecting their individual contributions to the model. Lastly, Figure 8 illustrates the initial four rows of the decision tree model.

III. RESULTS

Confusion matrix of prediction is presented in Figure 5, and classification report is presented in Table 6. Feature importance based on variable count and based on gain are presented in Figure 7 and 8.

Figure 5: Confusion Matrix of Crystal System Predictions

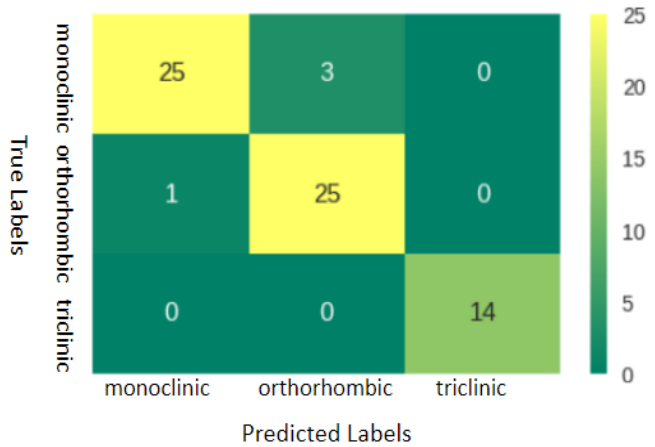


Table 6: Classification Report of Crystal System Predictions

	precision	recall	f1-score	support
monoclinic (0)	0.96	0.89	0.93	28
orthorhombic (1)	0.89	0.96	0.93	26
triclinic (2)	1.00	1.00	1.00	14
accuracy			0.94	68
macro average/ unweighted average	0.95	0.95	0.95	68
weighted average	0.94	0.94	0.94	68

Figure 6: Feature Importance Based on Variable Count

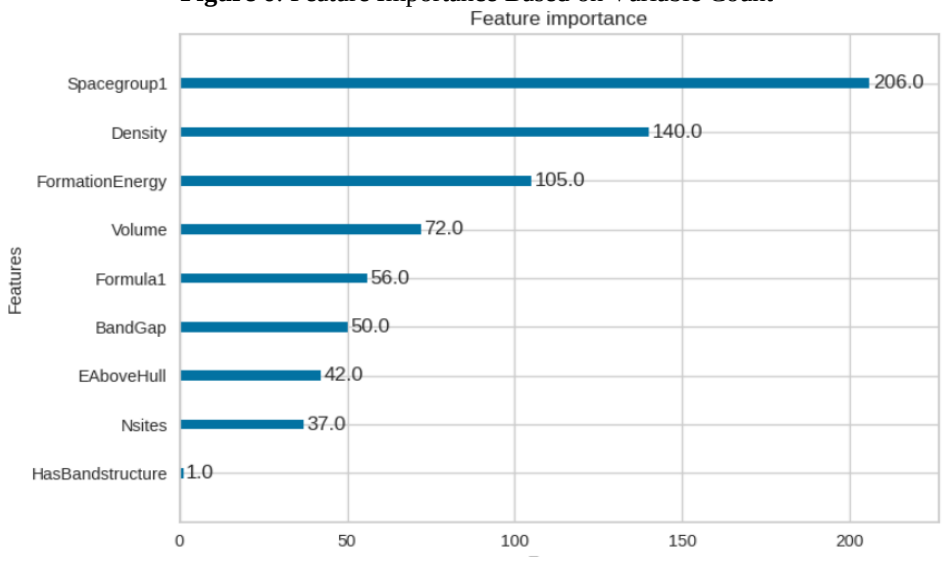


Figure 7: Feature Importance Based on Gain

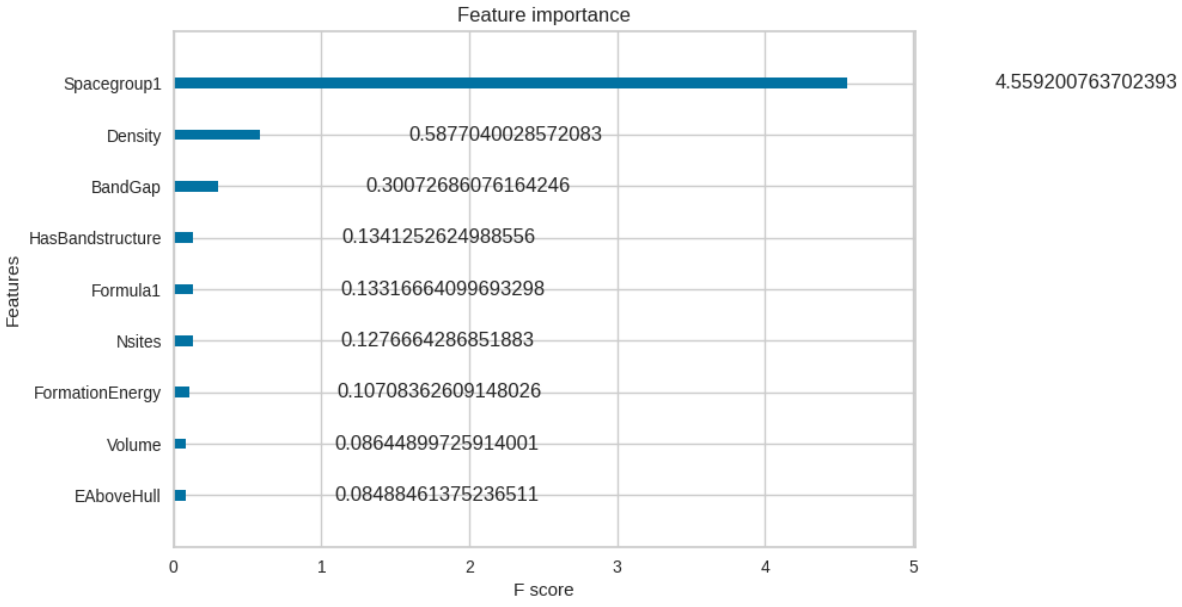


Figure 8: The First Four Rows of the Decision Tree

3.1 Discussion of Results

Based on the confusion matrix in Figure 5 and the classification report in Table 6, the model's performance metrics range from 0.89 to 1.00. Notably, the model achieved an accuracy of 94%, surpassing the performance of a different study that achieved 75% accuracy [17].

The model excelled in predicting triclinic crystal systems, scoring 1.00 across all performance metrics related to triclinic crystals. This could suggest that the model is overfitting, although the performance metrics for other crystal systems exhibited lower scores. Unfortunately, there was insufficient data for cross-validation tests to confirm or refute if the model was overfitting.

According to the feature important plots, the most important feature was spacegroup. Spacegroup, which describes the symmetry of a crystal lattice, is probably very closely related to the crystal system. The close relationship between spacegroup and the crystal system was also seen in the contingency table (Figure 4), where a spacegroup corresponded uniquely to only one type of crystal system, or in other words, two different crystal systems could not have the same spacegroup. Additionally, density, band gap, formation energy, volume, and nsites were important features in predicting the crystal structure.

One limitation of this study was that encoding high cardinality categorical variables using a CatBoost encoder could induce overfitting because it may greatly increase the data's complexity. Comprehensive data can make it easier for the machine learning model to memorize the training data than to discern meaningful patterns that better represent real-world scenarios. A future solution to mitigate this is to incorporate more extensive datasets with additional information about lithium-ion battery cathodes. More data will help boost the accuracy of the machine learning model and prevent overfitting because there will be more possibilities, and meanwhile, allow cross-validation tests to reduce overfitting. Furthermore, employing an alternative ensemble method in XGBoost or a different machine learning model could improve performance.

IV. CONCLUSION

In this study, a machine learning model was constructed to classify whether lithium-ion battery cathodes had a monoclinic, orthorhombic, or triclinic crystal system using data from the Materials Project. First, critical statistical values were extracted during exploratory data analysis, and correlation and analysis of variance were tested. Subsequently, all the data were encoded using CatBoost and label encoders. Then, the dataset was split into training and testing groups and normalized using a minmax scaler. An XGBoost classifier with default parameters was fitted to the training data and was tested with the test data. As a result, the XGBoost model performed exceptionally well, yielding an overall 94% accuracy rate. The model improves existing machine learning methods for predicting the crystal system of lithium-ion battery cathodes, and it holds the potential to be incorporated into real-world applications, offering valuable insights and facilitating informed decisions within the realm of lithium-ion battery cathode research and development.

REFERENCES

[1]. https://www.energy.gov/energysaver/articles/how-lithium-ion-batteries-work
[2]. https://www.chem.tamu.edu/rgroup/marcetta/chem362/HW/2017%20Student%20Posters/Lithium%20Ion%20Batteries.pdf
[3]. Raja, P. and Barron, A. Physical Methods and Nano Science. LibreTexts
[4]. Manthiram, A. A reflection on lithium-ion battery cathode chemistry. Nature Communications **11**, 1550 (2020).
[5]. https://www.chemistry.mcmaster.ca/~xman/cccw18/files/Crystal_Systems_and_Space_Groups_coloured.pdf
[6]. Bartel, C. Review of Computational Approaches to Predict the Thermodynamic Stability of Inorganic Solids. Journal of Material Science, 2022 (published online).
[7]. Liu, M., et. al. Spinel Compounds as Multivalent Battery Cathodes: A Systematic Evaluation Based on *Ab Initio* Calculations. Journal of Energy and Environmental Science, Issue 3, 2015.
[8]. Clark, J. Band Structure. LibreTexts.
[9]. https://www.mrl.ucsb.edu/~seshadri/2018_218/Bands-k-space.pdf
[10]. Correlation Coefficient Review. Khan Academy
[11]. Analysis of Variance. Khan Academy

[12]. Brems, Matt. A One-Stop Shop for Principal Component Analysis. Medium, Towards Data Science, 2022
[13]. Morde, V. XGBoost Algorithm: Long May She Reign! Medium, Towards Data Science, 2019
[14]. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboos
[15]. Mohajon, J. Confusion Matrix for Your Multi-Class Machine Learning Model. Medium, Towards Data Science, 2021
[16]. Kohli, S. Understanding a Classification Report For Your Machine Learning Model. Medium, Towards Data Science, 2019
[17]. Shandiz, M. and Gauvin, R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. Computational Materials Science, Volume 117, 2016, 270-278.