**Predictive Modeling for Crystal Structure in Lithium-ion Battery Cathode Materials**

Group 12

CHEE 6397

Group Members: Hao Wang (2344601), Udom Ifeanyi (2220031), Kevin Tieu (1902070)

**Abstract**

This project explores the prediction of crystal systems in lithium-ion battery cathode materials using machine learning techniques. Utilizing a dataset from Kaggle, the study involves data analysis, including data distribution and correlation, and a gradient boost classification model. The gradient boost model demonstrated promising results with an accuracy of 0.669 after evaluated using five-fold cross validation, indicating its performance and potential for crystal structure prediction application.

1. **Background**

Lithium-ion batteries are pivotal in modern technology, with their performance significantly influenced by the properties of cathode materials. The crystal structure of these materials significantly affects their functionality. Predicting these structures is thus essential for advancing battery technology.

In lithium-ion battery (LIB) research, significant focus has been on cathode materials, given their critical role in battery performance. Recent studies have highlighted the importance of these materials' energy density, thermal stability, and charge rates [1], [2]. Research advancements have been made in exploring new cathode materials, showing promising results in battery capacity and life. The crystal structure of cathode materials is a key determinant in LIBs' performance. Studies indicate that the crystallography of these materials influences lithium-ion diffusion rates, directly impacting the efficiency of charging and discharging [3], [4]. Machine learning (ML) has emerged as a powerful tool for predicting the crystal structure of cathode materials [5], [6]. These methodologies, leveraging algorithms based on material composition and properties, have shown accuracy in structure prediction, marking a significant step forward in computational material science. However, the field continues to face challenges, particularly in the generalization of these predictive models across different types of cathode materials. Future research is geared towards developing more sophisticated ML models that can incorporate a broader array of variables for enhanced prediction accuracy [7].

1. **Dataset Description**

The dataset includes several information for each of the 339 cathode materials for lithium-ion batteries. The features of each data point include formation energy, energy above hull, band gap, number of sites, density, volume, band structure, and the material’s chemical composition. The label for this dataset is the crystal system of the cathode materials.

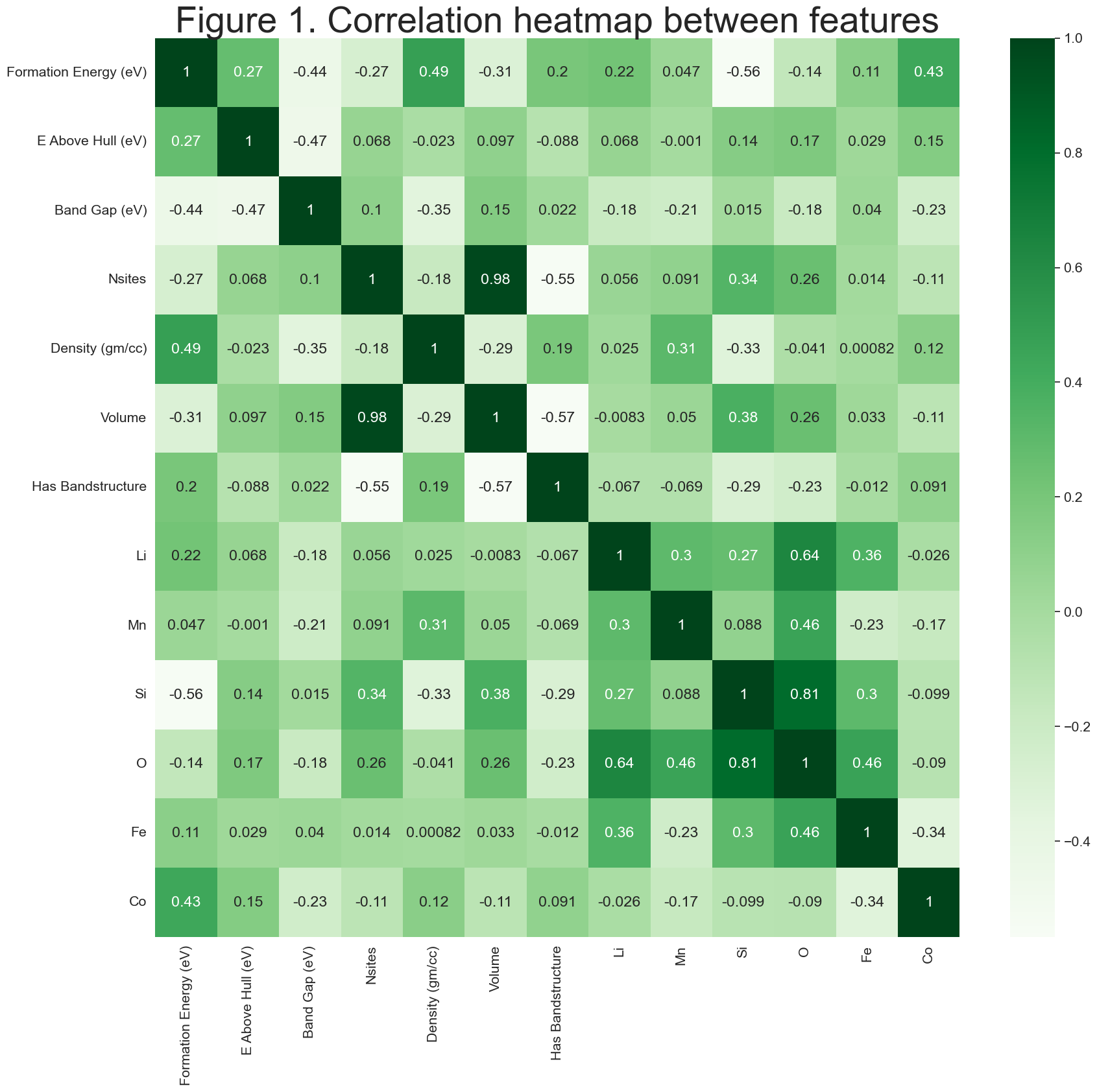
1. **Methodology**

Prior to data analysis, the dataset is filtered to eliminate missing or invalid data points. Then, the raw data is converted to manageable data structures. For example, the crystal systems, which are represented in string, are converted into integers using one-hot encoding. Also, chemical formulas for all cathode materials are broken down into their element compositions for better data analysis. In addition, some irrelevant information that does not have any effect on the labels are removed, such as materials ID.

Then, data analysis is conducted using several approaches. First, the correlations between features are investigated using a heat map. The map is used to indicate any correlation between each pair of features in the dataset. In addition, the data distribution of features in the dataset is evaluated to portrait a better picture of the dataset. This technique determines whether the dataset is balanced and biased toward a label or feature.

After the data analysis steps are completed, a decision tree-based gradient boost classification model is used to predict crystal system. The model is fine-tuned by optimizing the three hypermeters (number of weak learners, learning rate and maximum tree depth). The model’s performance is evaluated using five-fold cross-validation with the F1 score as the metrics.

1. **Results**
2. Correlation Analysis:



The extended heat map (Figure 1) provides additional insights into the correlations between different material properties. The volume and Nsites properties show a strong correlation of 0.98. Additionally, oxygen shows a strong connection with lithium and silicon (0.64 and 0.81, respectively). This means the lithium cathode structures would be likely to contain oxygen and silicon.

1. Data Distribution:

The sample distribution of features and labels are critical to understand the dataset’s properties. First, more than 80% of the materials in the dataset have bandstructure (as shown in figure 1).

A graph with a bar and a number of bars

Description automatically generated with medium confidenceA chart of a number of bars

Description automatically generated with medium confidenceFrom figure 3, the distribution of materials with different crystal systems is imbalanced in this dataset. The monoclinic crystal structure accounts for more than 40% while the triclinic structure only covers about 20%. This mild imbalance could result in the trained model being biased.

1. Classification Model Performance

The Gradient Boost classification model was employed to predict crystal structures. After fine-tuning, the model with 135 weak learners, a learning rate of 0.07 and the maximum tree depth of 5 shows the optimal performance. The model achieved an average 5-fold cross-validation accuracy of 0.669, indicating a strong predictive performance. Figure 4 shows the predicted structures versus the actual results from one training set with an accuracy of 0.644. Since the original dataset contains 40% of monoclinic materials, the model shows bias toward monoclinic predictions. About 48.4% of total false positive and negative results predicts monoclinic crystal system.

The feature importance of the model is also investigated. It shows volume and oxygen have the highest influence on the prediction, while bandstructure has little to no effect on the classification.

A graph of a number of different types of energy

Description automatically generated with medium confidenceA green squares with white text

Description automatically generated6. Key Takeaways:

The analysis highlights the potential of machine learning techniques, especially gradient boost classification, in predicting the crystal systems of lithium-ion battery cathode materials. The regression models and visualizations further deepen the understanding of the relationships between various material properties.

These results demonstrate the effectiveness of the chosen analytical methods and contribute significantly to the field of lithium-ion battery research. For the final report, you should insert each mentioned element (graphs, plots, histograms) directly from the Jupyter notebook outputs to provide visual and empirical evidence supporting these findings.

1. **Conclusion**

The study provides insights into the relationships between various properties of lithium-ion battery cathode materials. Future research could focus on refining models for more accurate predictions and exploring additional classification algorithms.

References

[1] M. S. Whittingham, "Lithium batteries and cathode materials," Chemical Reviews, vol. 104, no. 10, pp. 4271-4302, 2004.

[2] J. B. Goodenough and K. S. Park, "The Li-ion rechargeable battery: a perspective," Journal of the American Chemical Society, vol. 135, no. 4, pp. 1167-1176, 2013.

[3] M. Armand and J. M. Tarascon, "Building better batteries," Nature, vol. 451, no. 7179, pp. 652-657, 2008.

[4] Q. Yan et al., "Machine learning design of polymer dielectrics: machine learning of polymer dielectrics," Chemical Science, vol. 8, no. 3, pp. 1759-1765, 2017.

[5] B. Kang and G. Ceder, "Battery materials for ultrafast charging and discharging," Nature, vol. 458, no. 7235, pp. 190-193, 2009.

[6] Y. Sun, N. Liu, and Y. Cui, "Promises and challenges of nanomaterials for lithium-based rechargeable batteries," Nature Energy, vol. 1, no. 7, pp. 1-12, 2016.

[7] A. Jain et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Materials, vol. 1, no. 1, 011002, 2013.