Received: 25 November 2025 | Revised: 17 April 2026 | Accepted: 24 June 2026 | Published online: 30 July 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–15
DOI: 10.47852/bonviewJDSIS62028409

BON VIEW PUBLISHING

Comparative Study of Machine Learning Algorithms for Crystal Structure Predictive Modeling of Li-ion Battery Cathodes

Mega Novita[1,*], Alok Singh Chauhan[2,*], Rizky Muliani Dwi Ujianti[3], Sigit Ristanto[4], Bambang Agus Herlambang[5], Kazuyoshi Ogasawara[6], Chong Geng Ma[7], Mikhail Brik[7, 8, 9, 10] and Michal Piasecki[11]

[1] *Postgraduate Program of Natural Science Education, Universitas PGRI Semarang, Indonesia*

[2] *School of Computer Applications and Technology, Galgotias University, India*

[3] *Department of Food Technology, Universitas PGRI Semarang, Indonesia*

[4] *Department of Physics Education, Universitas PGRI Semarang, Indonesia*

[5] *Department of Informatics, Universitas PGRI Semarang, Indonesia*

[6] *Department of Chemistry, Kwansei Gakuin University, Japan*

[7] *School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, China*

[8] *Centre of Excellence for Photoconversion, University of Belgrade, Serbia*

[9] *Academy of Romanian Scientists, Romania*

[10] *Institute of Physics, University of Tartu, Estonia*

[11] *Department of Theoretical Physics, Jan Dlugosz University in Czestochowa, Poland*

Abstract: Identification of the crystal structure of the cathodes of Li-ion batteries is important to improve energy storage technologies. This research has three goals: (1) to determine the important parameters to predict the crystal systems (monoclinic, orthorhombic, and triclinic); (2) to evaluate different machine learning algorithms to make the correct prediction; and (3) to give insights on how to optimize materials used in batteries. We have studied a Kaggle dataset comprising a number of physicochemical descriptors such as chemical formula, energy above the convex hull, band gap, number of atomic sites in the crystal, density, and volume. We evaluate and compare systematically the predictive performance of 10 machine learning algorithms, including Random Forest, Support Vector Machine, Logistic Regression, K-Nearest Neighbor, Decision Tree, Gradient Boosting, XGBoost, Naive Bayes, AdaBoost, and CatBoost. The final model with the best predictive performance resulted from the optimized stacking model that achieved an accuracy of 72.06% and an Receiver Operating Characteristic–Area Under the Curve (ROC–AUC) of 0.8438 after removing leaky features and applying advanced validation methods. These results are consistent with past research that reveals that ensemble learning methods can be used to extract intricate trends in materials information. Nevertheless, this study is based on one dataset with a rather small sample size, which can impact the broader applicability of the findings.

Keywords: machine learning classification, ensemble learning, Li-ion batteries, crystal systems, battery cathodes

***Corresponding author:** Mega Novita, Postgraduate Program of Natural Science Education, Universitas PGRI Semarang, Indonesia. Email: novita@upgris.ac.id and Alok Singh Chauhan, School of Computer Applications and Technology, Galgotias University, India. Email: alok.chauhan@galgotiasuniversity.edu.in

1. Introduction

The performance of cathodes in lithium-ion (Li-ion) batteries is a factor that defines their performance, durability, and efficiency to a large extent [1–3]. The performance of Li-ion batteries is dependent on the cathode materials, which include lithium cobalt oxide and lithium iron phosphate combined with lithium nickel-manganese-cobalt-based materials, which support reversible Li-ion transport during the charging and discharging process. The increased utilization of electric cars and portable electronic devices has greatly increased the need to develop high-performance energy storage systems [4]. To ensure such advancements, more powerful predictive modeling methods are crucial toward maximizing cathode materials. High-level models make possible a more precise prediction of material characteristics and optimize battery operation. The monoclinic, orthorhombic, and triclinic crystal systems are significant in Li-ion battery cathodes due to their unique structural features and effects on Li-ion transport and stability [5, 6]. These differences play an important role in determining electrochemical performance in battery materials. Triclinic crystal systems, however, have the lowest symmetry and are typically linked with more complex and less stable structures. This decreased symmetry can restrict Li-ion mobility and can influence electrochemical performance. Nevertheless, even with their distinctive structural arrangements, they may offer alternative routes to the process of ion transport and, thus, provide the possibility of the development of novel battery materials.

Machine learning (ML) and artificial intelligence (AI) have become extensively used in data-driven approaches to predict and optimize the properties of battery materials in the field of materials science [7–11]. These methods allow analyzing complex data efficiently and enhance the knowledge of the relations between material structure and properties [12–16]. Models have been created by researchers that can be used to accurately predict the behavior of a material based on structural and electronic properties. The complexity of monoclinic, orthorhombic, and triclinic crystal systems however poses serious challenges to predictive modeling. Moreover, the existing research also has various limitations such as a small dataset, incomplete comparison between ML models, and a lack of validation strategies. Such difficulties can influence the generalization and reliability of the models. To address these issues, various machine learning algorithms have been explored to allow data-driven predictive modeling of materials in the field of battery materials, with the objective of improving the design and performance of battery materials. The authors have already studied numerous representatives of materials science and computational modeling, especially in the area of energy materials [17]. Our studies since 2011 have focused on the development of laser- and optics-based materials, with theoretical analysis being done using the theoretical framework of Discrete Variational-Xalpha and Discrete Variational Multi-Electron. The research that we conducted on red phosphorus, which is further subdivided into oxide-based and fluoride-based categories, has resulted in important publications and insights. Our research on fluoride-based phosphorus has led to major publications in the highest-ranking journals, including studies of the multiple forms of the transition metal ions and the light-absorbing properties of various doped materials. Recently, our research has been involved in the study of the effects of pressure on rubies, the optical properties of $Li_2TiO_3:Mn^{4+}$ and Cr^{3+}, and the $4f^n$-$4f^{n-1}5d$ transitions of Ln^{3+} ions in various host crystals [18]. Moreover, our work explores the application of ML and deep learning methods to identify novel red phosphor materials, pointing to the potential of these advanced methods [19, 20].

Although significant advances have been made, several challenges remain in ensuring the reliability and generalization of machine learning models, particularly when limited datasets and information leakage are involved. Unlike past researchers, this work takes a leakage-free ML pipeline along with stratified cross-validation to achieve more reliable and generalizable results. This study investigates the accuracy of different ML algorithms in predicting crystal systems in Li-ion battery cathodes, such as monoclinic, orthorhombic, and triclinic crystals. The aims are to find important predictive characteristics, compare the behavior of different algorithms, and give preliminary comments for future research. They applied a variety of ML models on a dataset of structural and electronic features including a set of models of Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, Logistic Regression, Gradient Boosting, XGBoost, Naive Bayes, CatBoost, and AdaBoost. Ensemble methods are identified as a result of having the ability to model complex relationships within the dataset. The current work helps to comprehend the applications of ML solutions to classifying crystalline systems and gives a deeper insight that can be used in future research in the field of optimization of battery materials.

2. Literature Review

2.1. Machine learning in materials science

In recent years, data-driven approaches have significantly contributed to the growing use of ML–based approaches in materials science, particularly for modeling complex and nonlinear interactions between material properties and performance in materials discovery. The traditional experimental and computational approaches can be time-consuming and resource-intensive, particularly when dealing with big and complicated data.

Conversely, ML methods allow more effective analysis of data, which allows researchers to discover significant patterns and contribute to the faster discovery and optimization of materials. According to past studies, ML approaches can be successful in predicting various material properties, including structural stability, electronic properties, and performance-related energy consumption. These models can produce precise predictions that help in making decisions about the design of materials by making use of features such as formation energy, band gap, and atomic configurations [21]. These strategies can be particularly useful in situations where the nonlinear dependence between variables is difficult to model using traditional techniques.

Applications of ML and AI in materials science have undergone a significant evolution throughout the years. It has used a variety of algorithms, beginning with simple linear models and more complex ensemble learning algorithms, in the hope of improving predictive performance and model robustness. Of these, the ensemble methods have received special focus since they are able to discover complex patterns in high-dimensional data through a combination of a number of base learners.

Even though considerable advances have been achieved, the overall predictive performance of ML models is limited by factors such as data quality, feature engineering, and dataset size. Generalization and reliability in model prediction can be influenced by a lack of data and the intricacy of material structures. Thus, further research remains to be aimed at the enhancement of the performance of models, their increased interpretability, and the development of more robust methods of materials prediction.

2.2. Machine learning for Li-ion batteries

The use of ML in the study of Li-ion batteries has become one of the most widely discussed research areas due to its potential to enhance the performance of the battery and speed up the process of discovering new materials. Li-ion batteries find common application in a variety of applications such as electric vehicle applications as well as in portable electronic devices, the effectiveness, durability, and efficiency of Li-ion batteries being largely influenced by the properties of cathode materials. As a result, understanding and optimizing these materials has become a key focus in recent research.

Past research has utilized ML methods in different parts of Li-ion battery systems, such as estimating state of charge, predicting battery health, and analyzing material properties. Specifically, important properties including formation energy, band gap, and structural stability, which are critical to assess the suitability of cathode materials, have been predicted using ML models [22]. The methods allow quicker and more economical analysis than conventional experimental methods.

A number of ML algorithms have been put into practice in this field, including classical algorithms like Support Vector Machine and Logistic Regression as well as more modern algorithms such as Gradient Boosting, Random Forest, and XGBoost. Ensemble learning techniques have been shown to perform better than the other methods in that they can be used to model nonlinear relationships among battery material datasets. These models fuse a group of learners to enhance predictive performance and robustness, which makes them particularly suitable in the scenario of high-dimensional and heterogeneous data.

Although the results are promising, there are still difficulties in implementing ML in Li-ion battery studies. A lot of studies are based on a small amount of data, and this can affect the generalizability of the models. Moreover, variation in feature representation and evaluation measures may cause discrepancies in the performance reported. As such, further studies are needed to create more comprehensive and reliable ML-based models in predicting and optimizing Li-ion battery materials.

2.3. Crystal structure prediction in battery materials

The crystal structure is important in determining the overall electrochemical behavior of Li-ion battery cathodes. Various crystal systems, which include monoclinic, orthorhombic, and triclinic crystal systems, have different structural characteristics that affect Li-ion mobility, stability, and overall battery efficiency. Such structural differences have a direct influence on the ion diffusion pathways, electronic properties, and material durability during the charge/discharge processes.

Past records have emphasized that orthorhombic crystal structures tend to be linked with good electrochemical performance as a result of the balanced structural stability and efficient Li-ion transportation [18]. Conversely, monoclinic structures offer greater flexibility of the atomic structure but can decrease ion mobility because of lower symmetry. The triclinic systems are typically associated with more complicated and less stable structures, which may limit their usefulness in applications of high-performance batteries.

Although the crystal structures are important, predicting crystal systems using the material properties is not an easy task. The relationships between formation energy, band gap, atomic sites, density, and volume are nonlinear and extremely complicated. Traditional approaches may not easily find such interactions, and this is especially true in cases where big and high-dimensional data are involved.

Previous studies have explored the use of ML methods in order to overcome these obstacles. ML models are able to identify more complex patterns in the data and have more accurate predictions about crystal systems provided a range of material characteristics. However, predicting crystal structures of different materials fairly using a range of ML algorithms has yet to be applied to a study that involves conducting a thorough comparison of a number of models.

2.4. Research gap

Although ML applications in materials science and in the study of Li-ion batteries are rapidly developed, a number of challenges are present. Previous research has demonstrated the possibilities of ML methods when it comes to predicting material properties and optimizing battery performance [17]. However, a lot of these studies are limited in scope and related to particular algorithms or particular aspects of battery systems without providing a more comprehensive comparison of multiple ML models.

Moreover, the complexity of crystal structures, including monoclinic, orthorhombic, and triclinic, is a major issue with predictive modeling [18]. The interactions of structural and electronic properties are very nonlinear and can hardly be described by the methods used in the past. Even though it is not a new study that explores the use of ML to predict materials, the classification of various crystal systems with the help of different ML algorithms is an area that is still under-explored.

The other critical limitation comes in regard to the availability of data and validation strategies. Many of the current studies rely on small datasets and fail to use strong validation methods, including cross-validation, which can impact the reliability and the capability of the results to generalize [19]. In addition, the research on the combination of extensive feature analysis and model comparison in order to gain a better understanding of the underlying factors that influence the crystal system prediction is still lacking. Unlike other studies, this paper uses a leakage-free ML pipeline with stratified cross-validation to obtain more reliable and generalizable findings.

3. Methods

Data on the physicochemical properties of Li-ion silicate cathodes were obtained by extracting the appropriate data from an openly available dataset retrieved from Kaggle.com. This database contains a set of variables including material identifiers, elemental composition, key physicochemical properties such as formation energy, energy above the convex hull, band gap, and the number of atoms at each atomic site, material density, unit cell volume, and crystal structure information. It is important to note that this study is completely founded on one publicly available dataset, and no external validation dataset was used.

Thorough preprocessing was conducted as part of the data preparation process to analyze the data with ML. This entailed the systematic elimination of irrelevant data and the accurate handling of any absent data. The numerical features were subjected to normalization processes to standardize the scales of such features in a bid to eliminate any potential deviations in the magnitudes of such features and to enable the robust training of models. The dataset was subsequently stratified into separate training and test sets [23] after preprocessing the data. To be reproducible, the

random seed was fixed when dividing the data and training the model. Not only did this approach optimize the model training, but it also enabled deep validation and evaluation, maintaining the reliability and applicability of the trained models. Moreover, feature selection was carried out to decide and isolate the most important features to predict the crystal system classification.

Moreover, the Spacegroup feature was eliminated in all experiments since it directly encodes crystallographic symmetry and introduces data leakage. Including this feature may lead to overly optimistic results. It was therefore omitted to make sure that the model only learns using physically meaningful descriptors. The chosen features in this study were these descriptors of formation energy, energy above the convex hull, band gap, atomic site count (Nsites), density, and volume because the variables directly relate to the physical and structural properties of the materials. The aim of such a process was to increase predictive accuracy by focusing on the most significant features needed to classify. The dataset was processed with techniques such as principal component analysis (PCA) and feature selection based on the model in order to extract meaningful information. A group of ML algorithms was chosen to be compared, and they included various methodologies in a wide range of complexity. Included in this analysis were Logistic Regression [24], Decision Tree [25, 26], Random Forest [27], K-Nearest Neighbor [28, 29], Support Vector Machine [30], Gradient Boosting [31], XGBoost [32, 33], Naive Bayes [34, 35], AdaBoost [36–38], and CatBoost [39, 40].

Table 1 summarizes the ML models used in this study. The models were fitted with optimized parameters that were obtained after RandomizedSearchCV to enhance predictive accuracy. Stratified 5-fold cross-validation was used to improve model robustness and generalization and was incorporated into

Table 1
Summary of machine learning models and their mathematical expressions

No.	Machine learning model	Description	Mathematical expression
1	Logistic Regression	Linear model for binary classification based on the logistic function (sigmoid)	$P\left(Y=\frac{1}{X}\right)=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$ Logistic Regression evaluates the probability $P\left(Y=\frac{1}{X}\right)$ that a given instance X is part of class 1. The logistic function $1+e^{-z}$ transforms the input z into a value within the range of 0 to 1, enabling probabilistic interpretation. To determine the probability estimate, the model computes probabilities by applying the logistic function to a linear combination of input variables X
2	Decision Tree	Constructing a tree-based model for classification or regression involves recursively splitting the feature space	Decision rule at node t: $X_j \leq s$, where X_j is the feature and s is the split point
3	Random Forest	Training on bootstrapped samples from the dataset builds a collection of decision trees. In classification scenarios, the model determines the final class label based on a majority voting scheme, whereas for regression tasks, the final output is computed as the average of individual predictions	Random Forest prediction: $\hat{y}=mode(y_1, y_2, \ldots, y_n)$
4	K-Nearest Neighbor (KNN)	The most common class among an instance's K closest neighbors in the dataset is determined by an instance-based learning technique	KNN prediction: $\hat{y}=mode(y_1, y_2, \ldots, y_K)$
5	Support Vector Machine	The model determines the optimal hyperplane by maximizing the separation between data points identified as support vectors in the feature space	$w \cdot x+b = 0$. is the hyperplane equation. When the input vector is x, where w represents the weight vector, and b denotes the bias term
6	Gradient Boosting	The ensemble approach constructs models sequentially, where each new model aims to correct the errors of the previous one	The gradient descent update rule is given by $\theta^{(t+1)}=\theta^{(t)}-\eta\nabla L\theta^{(t)}$. the model parameters at iteration t are represented by $\theta^{(t)}$, the learning rate is denoted by η, the gradient of the loss function L with respect to the model parameters is ∇L

(*Continued*)

Table 1
(*Continued*)

No.	Machine learning model	Description	Mathematical expression
7	XGBoost	Optimized implementation of gradient boosting with regularization and parallelization	Loss function: $Loss = \sum_{i-1}^{n} L(y_i, \hat{y}_i) + \lambda\Omega(f)$. λ represents the regularization parameter, and $\Omega(f)$ denotes the regularization term that imposes a penalty on the complexity of the model
8	Naive Bayes	A classifier that functions according to Bayes' theorem, presuming that the features are independent of the class	Naive Bayes classifier: $\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$
9	AdaBoost	Ensemble method combining numerous weak learners by assigning weights to training examples	AdaBoost prediction: $\hat{y} = sign(\sum_{t-1}^{T} \alpha_t h_t(x))$. $\alpha_t$is the weight assigned to weak learner $h_t(x)$
10	CatBoost	Gradient boosting algorithm designed for categorical features, optimizing training speed and memory usage	CatBoost prediction is given by $\hat{y} = \sum_{k-1}^{K} sign(f_k(x))$, where $f_k(x)$ is the prediction of the k -th base learner

the training process. A complete ML pipeline was created, which includes train-test splitting, balancing of classes by the Synthetic Minority Over-sampling Technique (SMOTE), optimization of the ML pipeline by RandomizedSearchCV, and stratified cross-validation. Leakage-prone properties like Space-group were explicitly ignored in order to ensure that the model only learns on physically meaningful descriptors. Moreover, SMOTE was used to reduce the imbalance of the classes in the dataset. Moreover, a stacking ensemble method was adopted to enhance classification performance. The stacking method combined several base learners, such as XGBoost, Gradient Boosting, Random Forest, and CatBoost, whereas the meta-learner was Logistic Regression. In this way, the model is able to capture different patterns and enhance the overall predictive performance. All experiments were done in Python with popular libraries, such as scikit-learn, pandas, and NumPy, which guarantee reproducibility of the results. The data that were utilized in the current research were retrieved from an open-source dataset (Kaggle).

Table 1 summarizes the ML models used in this study, including their descriptions and corresponding mathematical formulations. All models were trained and evaluated on a dataset designed for crystal system classification. During the training phase, model parameters were adjusted to improve their ability to accurately classify different crystal structures.

Each experiment was carried out using Python and Jupyter Notebook environment, with libraries scikit-learn, XGBoost, CatBoost, SHapley Additive exPlanations (SHAP), and imbalanced-learn. Data preprocessing techniques were used, such as imputation, scaling, and class balancing using SMOTE. The accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and SHAP interpretability analysis were used to assess model performance.

Performance evaluation was conducted using a separate test dataset, considering metrics such as accuracy, precision, recall, F1-score, and ROC-AUC [41, 42]. This approach allows for a thorough assessment of model performance behavior across all crystal classes.

To improve model robustness and minimize overfitting, hyperparameter optimization was applied to selected high-performing models using RandomizedSearchCV combined with stratified 5-fold cross-validation.

The randomized search procedure investigated several key hyperparameters, including tree depth, number of estimators, learning rate, and regularization strength, to achieve an optimal trade-off between model complexity and generalization.

Accuracy is a fundamental evaluation metric that represents the proportion of correctly predicted instances relative to the total number of predictions. It can be expressed as follows:

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Prediction} \tag{1}$$

Accuracy provides a general measure of model performance, but it may be unreliable for imbalanced datasets. Therefore, additional metrics are required for a more complete evaluation.

The precision of the model is the ability to distinguish genuine positive cases from all the cases it labels as positive. It is calculated as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

When there is a significant cost associated with false positives, precision becomes very crucial. For example, misdiagnosing a healthy individual as having a condition might result in needless stress and therapy in the medical diagnostics field. Precision emphasizes the model's ability to eliminate false positives by assisting us in calculating the percentage of positive predictions that are correct.

Recall, sometimes referred to as sensitivity, evaluates how well the model can locate all pertinent examples within a given category. The formula is:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negatives} \tag{3}$$

Regardless of false positives, the focus of this statistic is on how successfully the model detects positive cases. When missing a

positive case could have major negative effects, high recall is essential. For example, a security threat may be the subject of dire consequences if it is not detected or if an important illness diagnosis is missed. Recall is therefore a part of the consideration of recording the highest number of true positive cases in the model.

Recall and precision are combined together, and that's why the F1-score is an equitable, comprehensive evaluation of the model's performance. It is estimated by:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

The F1-score is appropriate when it is necessary to trade off between recall and precision, particularly when dealing with imbalanced data. The F1-score is a better measure of the model's effectiveness because it considers the errors made by the model in predicting the positive class as well as the errors made in predicting the negative class. It can be used to test the model's performance in a case that requires recall and/or precision, ensuring the model can correct relevant situations and avoid false alarms.

ROC-AUC is an indication of how well the model is able to separate different classes at different thresholds. The receiver operating characteristic (ROC) curve is a plot of the true positive rate as a function of the false positive rate, or the percentage of correct positive identifications and the percentage of incorrect negative identifications over all the different thresholds. The area under the curve (AUC) is a general indication of how well the model performs:

$$\text{AUC} = \int_0^1 \text{TPR}\,(x)\, d\text{FPR}(x) \tag{5}$$

The higher the AUC value, the more effective the model's separation of instances. The closer it gets to 1, the better the model is able to distinguish between the instances. If the AUC is close to 0.5, then the model is not very useful and is not much better than random guesses and needs improvement. This statistic can be used to easily compare the trade-offs between sensitivity (being able to identify positive cases) and specificity (being able to identify negative cases) at different thresholds. It lets us know whether the model is capable of distinguishing well between the different classes and provides us with a full understanding of the model's capabilities.

Multiple performance metrics are important for assessing the performance of ML models. They argue either in general terms if the model is or isn't effective or, more specifically, with respect to features of the model such as recall, precision, and class separation. These signals can be used to communicate to stakeholders what they have selected, tuned, and applied to the models. The process ends with the improvement of the classification of crystal systems and the advancement in materials science, assisting researchers and engineers in developing novel and efficient materials and technologies that could be beneficial in different areas. Besides, these measures will enhance communication between data scientists and domain experts, which will make sure that the models developed are appropriate for the domain-specific necessities and problems.

4. Experimental Setup

Transparency, reusability, and reliable assessment of ML models for predicting the crystal systems of Li-ion battery cathode materials were ensured by the experimental workflow of this study. The experiments were performed in a Python-based Jupyter Notebook environment with popular scientific computing and ML libraries such as NumPy, Pandas, Scikit-learn, XGBoost, CatBoost, SHAP, and imbalanced-learn. A fixed random seed value of 42 was used when splitting the datasets, cross-validating, and training the models to ensure the results are reproducible. The experimental setup involved data preprocessing, feature selection, hyperparameter optimization, model training, and interpretability analysis.

This research was conducted using a dataset readily available on the Kaggle platform that includes physicochemical and crystallographic properties of Li-ion battery cathode materials. A total of 339 samples and 11 attributes of monoclinic, orthorhombic, and triclinic crystal systems were used in the dataset. The following parameters were selected as key features: formation energy, energy above hull, band gap, density, crystal volume, and number of atomic sites (Nsites). The "Spacegroup" attribute was purposely omitted from all experiments since it directly represents crystallographic symmetry information and could cause data leakage. By restricting only physically meaningful material descriptors, these models learned only from the material descriptors and not specific structural labels.

A full preprocessing process was done before the model was trained. Irrelevant attributes and missing values were eliminated systematically, and numerical attributes were scaled to ensure stability of training and uniform distribution of data. The data were then stratified and split into training and testing datasets. The distribution of classes was balanced using the SMOTE, which increases the number of minority class instances to make the model more robust. Furthermore, feature engineering and dimensionality reduction techniques such as PCA and model-based feature selection were used to select the most informative descriptors for classification of crystal systems. The final dataset split had an 80:20 ratio for training and testing and was performed using a fixed random seed to guarantee experimental consistency.

We compared 10 ML algorithms: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine, Gradient Boosting, XGBoost, Naive Bayes, AdaBoost, and CatBoost. Furthermore, a stacking model was developed by combining XGBoost, Random Forest, Gradient Boosting, and CatBoost as base learners and Logistic Regression as the meta-learner. The performance prediction and overfitting were minimized by performing hyperparameter optimization with RandomizedSearchCV integrated with stratified 5-fold cross-validation. The optimized stacking model performed best among all the evaluated models in terms of classification.

The main prediction model was an optimized stacking ensemble model, as shown in Figure 1. The base classifiers in a stacking ensemble model were optimized as follows: XGBoost, Random Forest, Gradient Boosting, and CatBoost were chosen as the base classifiers, and Logistic Regression was selected as the meta-learner. RandomizedSearchCV was used for hyperparameter tuning. Stratified cross-validation was used for better model generalization in the optimization process. The parameters of the algorithms were systematically optimized for best classification performance. For the XGBoost classifier, the number of estimators used was 302, the max depth was 7, and the learning rate was 0.0879. Random Forest model used 406 trees, and trees are max depth of 17 and a min split size of 8. CatBoost had 400 iterations, depth 8, and a learning rate of 0.05, while Gradient Boosting had 300 estimators, depth 7, and a learning rate of 0.1. The maximum number of iterations was set to 3000, and the regularization parameter (C) of Logistic Regression was set

Figure 1
Proposed optimized stacking ensemble framework

```
from sklearn.ensemble import StackingClassifier, GradientBoostingClassifier, RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from catboost import CatBoostClassifier
from imblearn.pipeline import Pipeline as ImbPipeline
from imblearn.over_sampling import SMOTE
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score

# Define the enhanced stacking model with base estimators and tuned parameters
stack_model_enhanced = StackingClassifier(estimators=[
     ('xgb', XGBClassifier(n_estimators=302, max_depth=7, learning_rate=0.0879, eval_metric='logloss', random_state=42)),
     ('rf', RandomForestClassifier(n_estimators=406, max_depth=17, min_samples_split=8, random_state=42, n_jobs=-1)),
     ('cat', CatBoostClassifier(iterations=400, depth=8, learning_rate=0.05, verbose=0, random_state=42)),
     ('gb', GradientBoostingClassifier(n_estimators=300, max_depth=7, learning_rate=0.1, random_state=42))],
   final_estimator=LogisticRegression(max_iter=3000, C=0.849), cv=5, n_jobs=-1, passthrough=True)

# Create the full pipeline with preprocessing and SMOTE
stack_pipe_enhanced = ImbPipeline([('prep', preprocessor), ('smote', SMOTE(random_state=42)), ('model', stack_model_enhanced)])
stack_pipe_enhanced.fit(X_train, y_train)

# Make predictions and evaluate the model
y_pred_enhanced = stack_pipe_enhanced.predict(X_test)
y_pred_proba_enhanced = stack_pipe_enhanced.predict_proba(X_test)
acc_enhanced = accuracy_score(y_test, y_pred_enhanced)
prec_enhanced = precision_score(y_test, y_pred_enhanced, average='weighted', zero_division=0)
rec_enhanced = recall_score(y_test, y_pred_enhanced, average='weighted', zero_division=0)
f1_enhanced = f1_score(y_test, y_pred_enhanced, average='weighted', zero_division=0)
roc_auc_enhanced = roc_auc_score(y_test, y_pred_proba_enhanced, average='weighted', multi_class='ovr')
```

to 0.849. Considering the class imbalance, SMOTE was applied before model training.

The performance of the model was assessed with several statistical measures that are commonly used in the field of classification: accuracy, precision, recall, F1, and ROC-AUC to evaluate the performance of the model in classification. A confusion matrix analysis and SHAP-based interpretability analysis were also conducted to investigate the behavior of classification and feature contributions. SHAP was used for the analysis of model interpretability. To quantify the contribution of individual features, SHAP summary plots were generated for selecting individual features to classify crystal systems. It was found that the most influential descriptors to predict crystal systems are band gap, energy above hull, formation energy, and volume.

The whole implementation code, the trained model configurations, and the experimental notebook were prepared in the Jupyter Notebook format to facilitate reproducibility and future research, and the dataset and source code were shared publicly on a GitHub repository.

5. Result and Discussion

The first step in this study was to import a database of different properties of Li-ion battery cathodes. The database is of great value for constructing a predictive model to classify crystal systems of cathode materials. Efficient data processing and analysis were carried out with the pandas library. Since the dataset was preprocessed, there were 339 rows and 11 columns. The columns are the different properties of the materials, including Formula, Materials Id, Band Gap (eV), Formation Energy (eV), Volume, Density (gm/cc), Nsites, EAbove Hull (eV), Has Band Structure, and Crystal System.

A representative subset of data is shown in Table 2 and information about some of the main physicochemical properties of Li-ion battery cathode materials for the various crystal systems. The samples chosen are from different classes of materials to provide a good balance of the sample set. This collection contains materials with and without an associated band structure, to give a balanced representation. Parameters like "Formation Energy (eV)" and "Band Gap (eV)" enable the understanding of the

Table 2
Dataset sample of Li-ion battery cathode crystal structure

No.	Materials Id	Formula	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Nsites	Density (gm/cc)	Volume	Has Band Structure	Crystal system
1	mp-849394	Li_2MnSiO_4	–2.699	0.006	3.462	16	2.993	178.513	TRUE	Monoclinic
2	mp-850949	$LiMnSiO_4$	–2.65	0.027	1.052	28	3.507	291.575	FALSE	Monoclinic
3	mp-566680	Li_2MnSiO_4	–2.705	0	3.052	16	3.039	175.842	TRUE	Orthorhombic
4	mp-761776	$LiMn(SiO_3)_2$	–2.824	0.036	0.037	80	3.343	850.626	FALSE	Orthorhombic
5	mp-767711	$Li_2Mn_2Si_4O_{11}$	–2.904	0.012	0.904	38	3.098	441.742	TRUE	Triclinic
6	mp-868673	$Li_{15}Mn_{15}SiO_{32}$	–2.194	0.028	0.311	63	4.093	595.723	FALSE	Triclinic

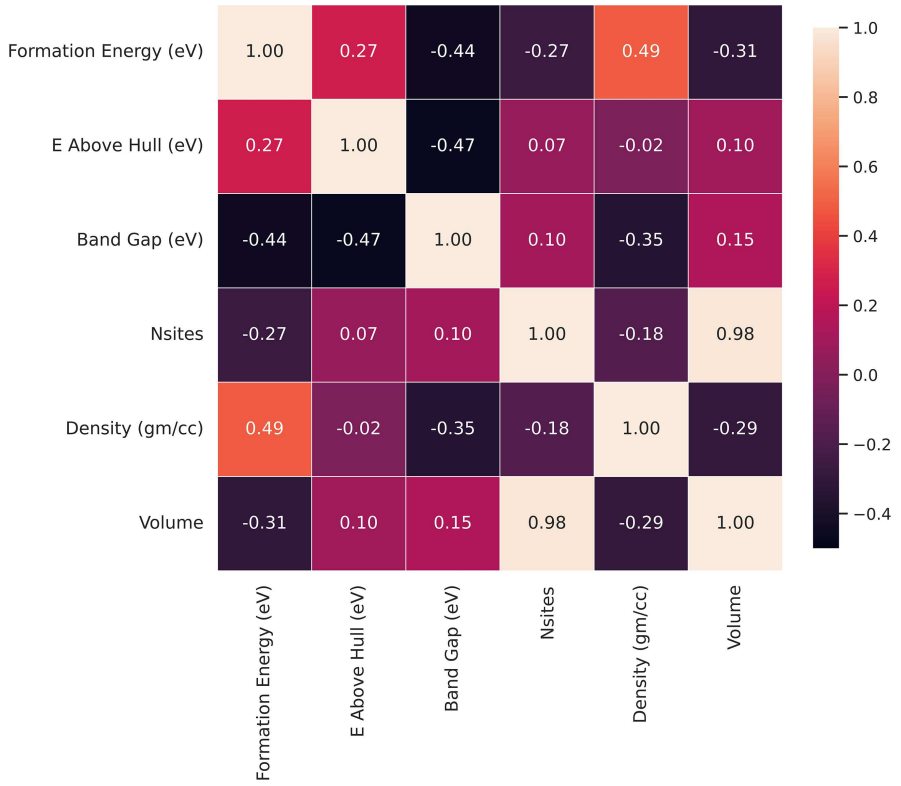
stability and electronic properties of materials, which are important to determine their efficiency as cathode material for Li-ion batteries. A few of the possible crystal structures that could be encountered in the cathodes of Li-ion batteries are shown in the table, and six sample observations are inserted as a snapshot of the data. The data consists of 339 entries of 3 different crystal systems. The dataset consists of 339 samples belonging to three crystal systems. The monoclinic class contains 139 samples, of which 110 have an associated band structure and 29 do not. The orthorhombic class contains 128 samples, of which 106 have an associated band structure and 22 do not. The triclinic class contains 72 samples, of which 58 have an associated band structure and 14 do not. This extensive dataset contains a wide range of data for studying and forecasting properties of a variety of Li-ion battery cathode materials.

The rows of the table are different materials that are used in Li-ion batteries, with each row having a unique "Materials Id", and its chemical formula is used to describe it. The column "Materials Id" is just to identify each material and is not used in calculations. The "Formation Energy (eV)" column shows the energy needed to form the material from its constituent elements, and the lower the energy cost, the more stable the material. The term eV means the difference in energy between the energy of a material and the energy of the most stable (lowest energy) elemental form of that material. Smaller "EAbove Hull" values mean that the compound is more stable and that the compound is more likely to stay together as a compound as opposed to breaking down into its elements. The value of the "Band Gap (eV)" column is the energy that it takes to move from the valence band to the conduction band. The value is significant to determining the electrical conductivity of the material since it is the energy barrier to reach for electrical conduction. Nsites is a number corresponding to the number of atomic sites within the unit cell of the crystal structure of the material, and "Density" (in grams per cubic centimeter) is the mass of material per unit volume. "Volume" refers to the overall volume of the unit cell. In addition, the "Has Band Structure" column indicates whether or not the material has an electronic band structure. Lastly, the "Crystal System" column indicates the crystal structure of the materials. The data were split in the ratio of 80:20, with 80% of the data assigned to a training set and 20% to a test set. To increase the strength of the evaluation, a stratified 5-fold cross-validation was employed in this study. Moreover, interpretability analysis was performed using SHAP to gain insights into the contribution of the individual features to the output of the model [43, 44].

The observations underscore the need for the application of ML methods to represent the complex, nonlinear relationship between battery material data.

The correlation matrix of the key physicochemical descriptors with respect to the crystal system classification of Li-ion battery cathodes is given in Figure 2. The results show that the number of atomic sites (Nsites) and crystal volume are very strongly correlated (r = 0.98) with materials that have larger unit cells having more atomic sites. However, moderate negative correlations in band gap and formation energy (−0.44) and band gap and energy above hull (−0.47) were found, indicating possible interactions between the electronic and thermodynamic properties of materials.

Figure 2
Correlation matrix heat map



The results also show that there is a moderate positive correlation (0.49) between density and formation energy and a weak to moderate negative correlation between volume and density (−0.29) and between volume and formation energy (−0.31). Overall, the results of the correlation analysis indicate that the relationships between the structural and electronic descriptors are complex and partially nonlinear. The results validate the application of ML techniques that are able to learn complex interactions among features and predict the crystal system. The results support the application of ML techniques that can learn complex interactions between features and can predict the crystal system.

Figure 3 shows the distribution and relationships between important features in the dataset. In general, there are no apparent strong linear relationships between most of the variables, except for a positive trend between Nsites and volume. The distributions of the crystal systems are quite overlapping, so that there is no simple separation of the classes by means of individual features. The findings indicate that there is a need for more sophisticated ML models to extract intricate patterns from the data.

To explore further relationships between some of the key features, correlation analysis was carried out in addition to the distribution analysis. The results show a high positive correlation between Nsites and volume, suggesting that materials with larger unit cells generally contain a greater number of atomic sites.

Moreover, there were moderate negative correlations between the formation energy, energy above hull, and the band gap, suggesting possible interactions between the thermodynamic stability and electronic properties. These results also demonstrate the complexity of feature interactions and the importance of considering the data's nonlinear relationships with advanced ML models.

The success of the stacking model is a measure of the extent to which its base models can identify complementary patterns

Figure 3
Data distribution across variables

in the data. The XGBoost, Random Forest, Gradient Boosting, and CatBoost models are able to identify various nonlinear relationships, and the Logistic Regression meta-learner combines their results into a more robust and generalized forecast.

The optimized hyperparameters for the stacking model are summarized in Table 3. All the ML models are compared with respect to their performance, and a comparison of their performance is displayed in Table 4.

A confusion matrix analysis was used to assess the classification accuracy of the proposed optimized stacking model. The confusion matrix of the optimized stacking model is shown in Figure 4.

The performance of the model on classification is good, especially for the orthorhombic class that has the largest number of correct classifications. There is some overlap in feature characteristics between the monoclinic and triclinic crystal systems, and the complexity of these low-symmetry crystal structures is observed, with some misclassifications. The results obtained indicate that the model is capable of representing the overall trend, but the separation of similar crystal systems is still difficult.

The results of all the ML models evaluated and used in the improved ML pipeline are shown in Table 4. The performance results show that the optimized stacking classifier outperforms the other classifiers, achieving an accuracy of 72.06% and a ROC-AUC of 0.8438, which highlights its ability to learn intricate patterns in the data. The advantage of ensemble-based methods over individual models is clearly demonstrated in the individual models, where XGBoost and Gradient Boosting showed good predictive power, followed by Random Forest and CatBoost.

However, simpler models like AdaBoost, Support Vector Machine, Decision Tree, Logistic Regression, K-Nearest Neighbor, and Naive Bayes exhibited relatively poor performance. These differences highlight the value of ensemble approaches for capturing the intricacies of the data, particularly in nonlinear scenarios.

To evaluate the performance of each class of crystal, micro-averaged precision, recall, and F1-score were computed and complemented by accuracy. The usage of SMOTE and stratified cross-validation led to increases in the sensitivity of the minority classes and helped to build more reliable models.

The overall accuracy (72.06%) is a scientifically meaningful result since the model now does not use crystallographic leakage but only chemistry-derived descriptors. With the small number of data points and the removal of features that are symmetric, the model shows some interesting predictive ability and is not trivial. The aim of this extensive comparison is to have all the ML models in Table 1 evaluated in the experimental analysis in the same way. This updated result also takes into account data leakage, cross-validation, hyperparameter tuning, and class balancing and gives a more realistic and reliable result.

The ROC curves of the optimized stacking model are given in Figure 5. The results show that ensemble-based models, especially stacking, work well in the modeling of complex relationships in battery material data. The evaluation results show that the model performs well in each class. The model shows good discriminative power in all crystal systems—triclinic class with AUC 0.95, followed by orthorhombic and monoclinic crystal classes. The results indicate that the model is able to distinguish different crystal structures fairly well.

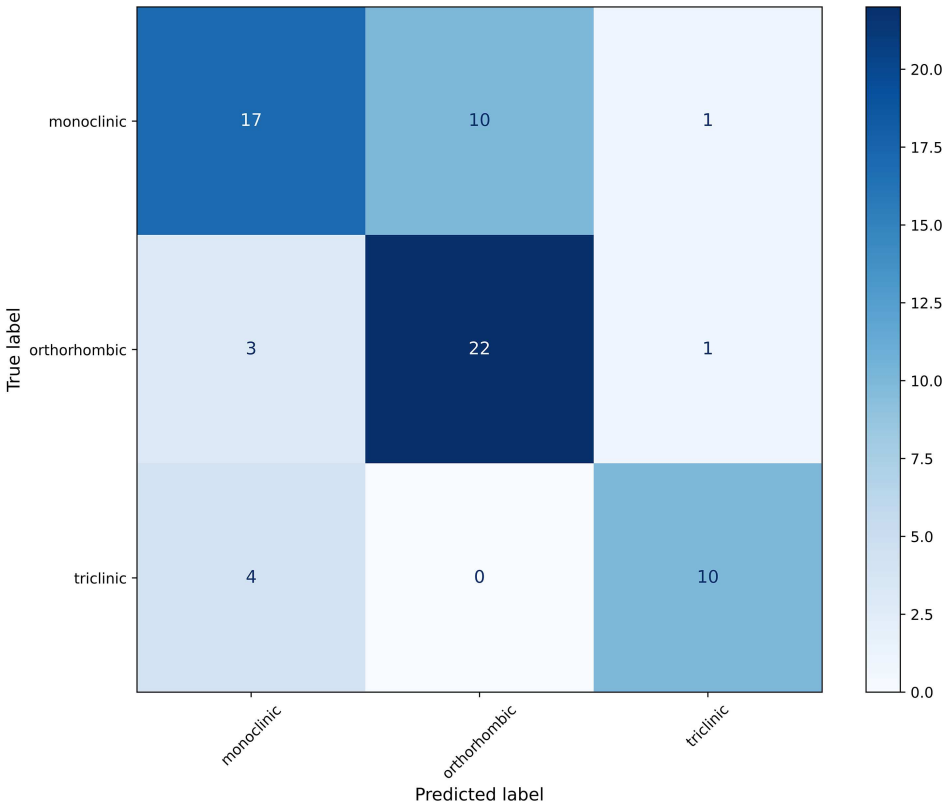
Table 3
Optimized hyperparameters for tuned stacking model

Model	Hyperparameters
XGBoost	n_estimators = 302, max_depth = 7, learning_rate = 0.0879
Random Forest	n_estimators = 406, max_depth = 17, min_samples_split = 8
CatBoost	iterations = 400, depth = 8, learning_rate = 0.05
Gradient Boosting	n_estimators = 300, max_depth = 7, learning_rate = 0.1
Logistic Regression (Meta Learner)	C = 0.849, max_iter = 3000

Table 4
Performance metrics for all machine learning models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Optimized Stacking	0.7206	0.7220	0.7206	0.7169	0.8438
XGBoost	0.6812	0.7059	0.6700	0.6776	0.8360
Gradient Boosting	0.6694	0.6926	0.6540	0.6636	0.8411
Random Forest	0.6576	0.6698	0.6606	0.6600	0.8281
CatBoost	0.6517	0.6807	0.6480	0.6564	0.8275
AdaBoost	0.6176	0.6218	0.6176	0.6181	0.7622
Decision Tree	0.6176	0.6183	0.6176	0.6169	0.6939
K-Nearest Neighbor	0.5735	0.5733	0.5735	0.5684	0.7333
Support Vector Machine	0.5294	0.5135	0.5294	0.5171	0.7052
Logistic Regression	0.4706	0.4558	0.4706	0.4612	0.6723
Naive Bayes	0.3529	0.4811	0.3529	0.3273	0.5611

Figure 4
Confusion matrix of optimized stacking model



In order to improve readability, we used SHAP analysis to obtain the results. The most influential features that predict the crystal systems are band gap, energy above hull, formation energy, and volume, as depicted in Figure 6. These features contribute the most to the model output and thus play a critical role in battery material classification. The dominance of band gap, formation energy, and energy above hull is physically meaningful because they are electronic stability and thermodynamic feasibility and they are closely connected with the formation of crystal structure.

Structurally, orthorhombic is a favorable crystal structure and consistent with previous studies for Li-ion battery applications; that is, it is structurally rigid yet has efficient Li-ion mobility [36–38]. Orthorhombic materials like $CsPbI_3$ and $LiFePO_4$ have been found to possess strong lithium storage capability and cycle stability [43]. The geometric structure they have is one in which their three unequal axes are joined at right angles, which allows the ions to transport efficiently and minimizes degradation of the structure for repeated charge–discharge cycles. This structural arrangement is in turn responsible for stable electrochemical properties, and this has made orthorhombic phases attractive for high-performance and long-life battery applications.

Monoclinic, which has two right angles and one oblique angle, can have more flexible lithium arrangements but may also have less symmetry to inhibit ion flow. This can have a negative impact on the electrochemical performance. The triclinic system, in which all the axes and angles are unequal, usually has poor ion conduction and less stability, which are less desirable characteristics for high-performance batteries. Hence, the Li-ion cathode is typically in an orthorhombic crystal configuration, which provides an attractive combination of capacity, stability, and electrochemical efficiency factors essential to applications that demand reliable, long-life energy storage.

The results of this study are comparable with those of the previous work in the field of ML applications to battery materials. A number of studies have shown that ensemble-based models constructed with Random Forest, Gradient Boosting, and XGBoost tend to outperform individual models because they are capable of detecting complex nonlinear relationships in material properties. For instance, recent research on battery health estimation and materials prediction has demonstrated that ensemble-based methods offer better prediction accuracy and robustness than individual methods [43]. Other studies on battery material prediction also claim good results through ML models, though the accuracy level is usually dependent on the amount of data, its representation, and how it is validated. Thus, the results of the performance obtained are more realistic predictions of actual performance, but direct comparisons among the studies should be treated with some caution because of differences in the characteristics of the data and the methods used for assessing performance.

With the relatively small number of data points being used, and the very good accuracy achieved by several models, the danger of overfitting should be taken into account. Overfitting occurs when a model is too well known to the training data, picks up noise, and thus fails to generalize to new data. But it should be noted that these results need to be treated with care. Although these advances have been made, there are some limitations to this study. The model may also be limited in its ability to generalize to future data due to the small number of instances in the dataset. Cross-validation was used to ensure robustness, but further testing with larger and more diverse datasets is needed [43]. In addition,

Figure 5
Multi-class ROC curve of crystal system classification for optimized stacking model

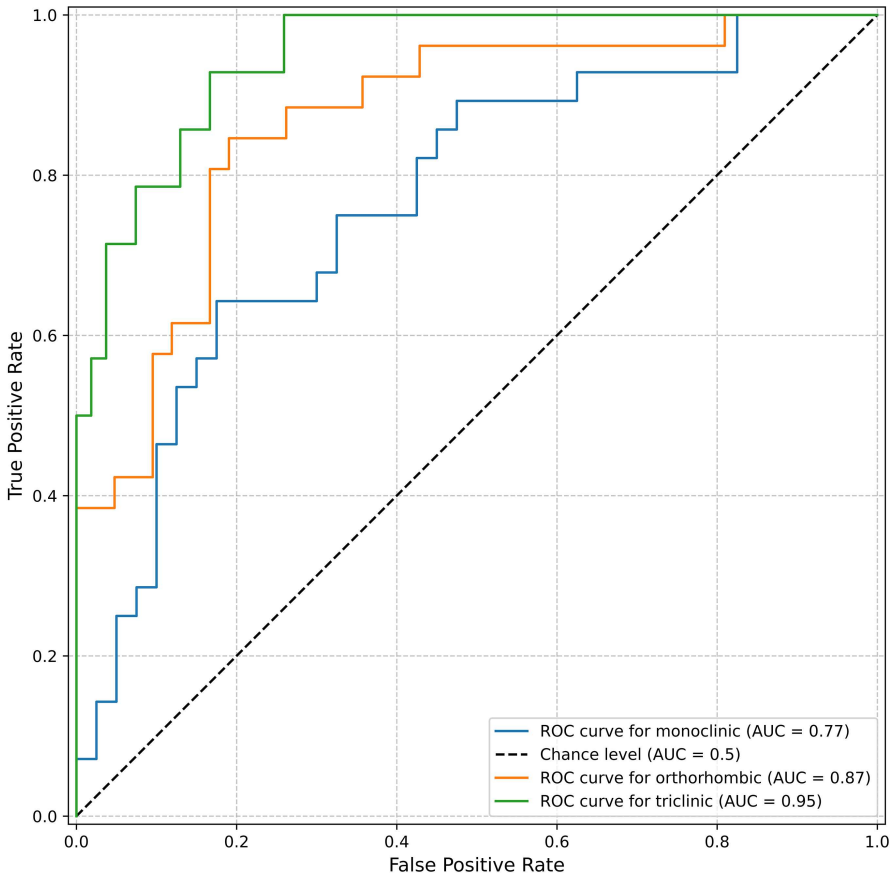
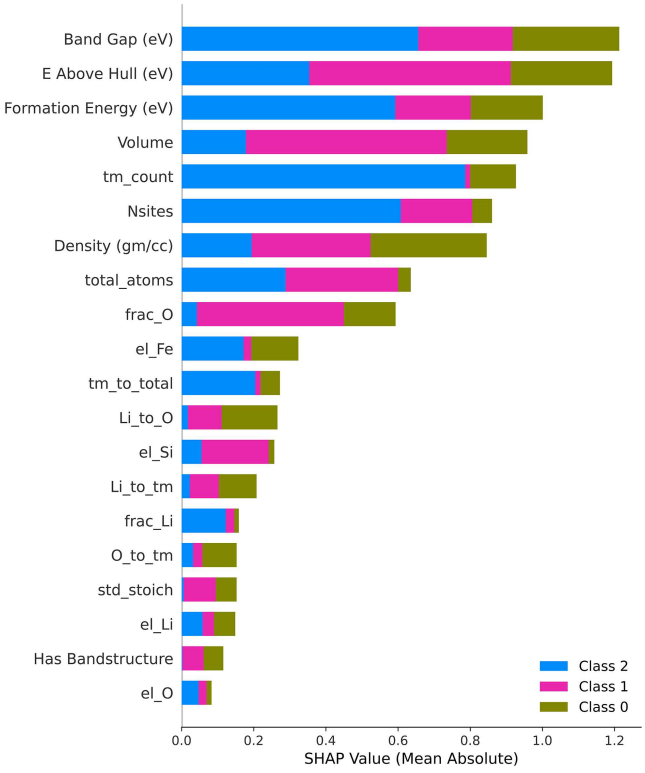


Figure 6
SHAP summary plot of feature contributions to crystal system prediction



the binary value of the feature "Has Band Structure" can lead to bias because of the availability of data rather than material properties. In the future, more sophisticated feature engineering and deep learning techniques could be used.

6. Conclusion

The crystal systems of the Li-ion battery cathodes were analyzed and classified using ML methods with the support of an important dataset of structural and electronic properties of these materials. The results of exploratory analysis, with a correlation heatmap and scatter plot matrix, highlighted a number of significant relationships between features, such as the high positive correlation (0.98) between the number of atomic sites (Nsites) and crystal volume and significant negative correlations between formation energy, energy above hull, and band gap. The results show that the complex relation between electrical properties and structural stability is relevant for battery performance.

The performance comparison of 10 ML algorithms revealed that ensemble model methods like Gradient Boosting, Random Forest, CatBoost, and XGBoost performed better than the single algorithms. The results show that the optimized stacking ensemble classifier demonstrated better predictive performance than other ML models. The same results were obtained in previous works, which revealed that ensemble learning methods not only can enhance the prediction accuracy and stability of the ensemble but also can improve the predictive accuracy by fusing multiple models and decrease the overfitting phenomenon. As for materials, it was discovered that it was most beneficial to have an

orthorhombic crystal system like those of previous studies, which have shown that this crystal form is stable and has good ion transport properties for Li-ion battery applications. Contrary to the previous studies, which were mainly based on material characteristics, the data-driven ML techniques employed in this study are capable of learning the complex feature interactions and enhancing the classification of crystal systems. This emphasizes the value of ML in conjunction with conventional materials science analysis.

Results provide several opportunities for additional studies. First, the electrochemical properties of orthorhombic structures should be considered as they are preferred. Second, the predictive modeling ability should be further improved using ensemble ML methods. Lastly, with respect to the design and optimization of cathode materials, the volume of the crystal, formation energy, and the band gap need to be considered.

The results have great potential but are not widely generalizable and are based on a small sample size, which may create problems for the generalizability of the type of models used. The scope of this study can be extended in the future by increasing the amount of data, optimizing feature engineering, using cross-validation techniques, and analyzing interpretability techniques such as SHAP. The results show the potential of ML methods in the fields of improving the classification and understanding of the crystal system of Li-ion battery cathodes and in developing more efficient and reliable Li-ion battery materials.

From an application perspective, the proposed leakage-free machine learning framework can assist researchers in rapidly predicting crystal systems from composition-derived descriptors, thereby reducing the need for extensive experimental screening during the early stages of cathode material discovery. This workflow also provides a reproducible and leakage-free framework for future crystal system prediction studies.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/divyansh22/crystal-system-properties-for-liion-batteries/data. All source code, model implementation, and experimental workflow are available from the GitHub repository at https://github.com/alokchauhan-collab/Crystal-Structure-Predictive-Modeling-of-Li-ion-Battery-Cathodes.

Author Contribution Statement

Mega Novita: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Alok Singh Chauhan:** Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing, Visualization, Supervision. **Rizky Muliani Dwi Ujianti:** Software, Investigation, Data curation, Visualization. **Sigit Ristanto:** Methodology, Validation, Writing – review & editing, Project administration. **Bambang Agus Herlambang:** Writing – review & editing. **Kazuyoshi Ogasawara:** Resources, Writing – review & editing. **Chong Geng Ma:** Resources. **Mikhail Brik:** Resources, Writing – review & editing, Supervision. **Michal Piasecki:** Resources, Writing – review & editing.

References

[1] Olabi, A. G., Abbas, Q., Shinde, P. A., & Abdelkareem, M. A. (2023). Rechargeable batteries: Technological advancement, challenges, current and emerging applications. *Energy*, *266*, 126408. https://doi.org/10.1016/j.energy.2022.126408

[2] Jo, S., Han, J., Seo, S., Kwon, O.-S., Choi, S., Zhang, J., . . . , & Lim, J. (2023). Solid-state reaction heterogeneity during calcination of lithium-ion battery cathode. *Advanced Materials*, *35*(10), 2207076. https://doi.org/10.1002/adma.202207076

[3] Manthiram, A. (2020). A reflection on lithium-ion battery cathode chemistry. *Nature Communications*, *11*(1), 1550. https://doi.org/10.1038/s41467-020-15355-0

[4] Jia, H., Cheng, Z., Tan, C., Yang, Y., Shi, Y., Zhou, Y., . . . , & Li, L. (2026). Application of machine learning to rechargeable aqueous zinc-ion batteries: Advancements and prospects. *Journal of Energy Storage*, *154*, 121180. https://doi.org/10.1016/j.est.2026.121180

[5] Xu, C., Du, W., Zhai, C., Wang, T., Zhao, C., Liang, H., & Li, S. (2026). Magnesium air batteries in terms of machine learning, anode alloying, electrolyte and air-cathode - A review. *Journal of Magnesium and Alloys*, *15*, 101989. https://doi.org/10.1016/j.jma.2026.101989

[6] Catillo, M., Buonocore, F., D'Onofrio, S., Giusepponi, S., Ferlito, S., Marchio, S., & Celino, M. (2026). Machine learning for redox potential prediction and application on a Na-ion battery system. *Solid State Sciences*, *177*, 108336. https://doi.org/10.1016/j.solidstatesciences.2026.108336

[7] Alzamer, H., Jaafreh, R., Kim, J.-G., & Hamad, K. (2025). Artificial intelligence and Li ion batteries: Basics and breakthroughs in electrolyte materials discovery. *Crystals*, *15*(2), 114. https://doi.org/10.3390/cryst15020114

[8] Orhadahwe, T. A., Oladunni, A. A., & Onadeko, O. O. (2024). Application of machine learning tools to material science: A mini-review. *Physical Science & Biophysics Journal*, *8*(1), 000269. https://doi.org/10.23880/psbj-16000269

[9] Chong, S. S., Ng, Y. S., Wang, H.-Q., & Zheng, J.-C. (2024). Advances of machine learning in materials science: Ideas and techniques. *Frontiers of Physics*, *19*(1), 13501. https://doi.org/10.1007/s11467-023-1325-z

[10] Xu, P., Ji, X., Li, M., & Lu, W. (2023). Small data machine learning in materials science. *npj Computational Materials*, *9*(1), 42. https://doi.org/10.1038/s41524-023-01000-z

[11] Yu, J., Jin, L., Ji, Y., & Li, Y. (2025). Data-driven composition-only machine learning for high-performance solid-state electrolytes. *Materials Chemistry Frontiers*, *9*(19), 2871–2878. https://doi.org/10.1039/D5QM00438A

[12] Liu, Z., Pan, J., & Liu, Q. (2023). Jīqì xuéxí zài shèjì gāo xìngnéng lǐ diànchí zhèngjí cáiliào yǔ diànjiězhì zhōng de yìngyòng [Application of machine learning in the design of cathode materials and electrolytes for high-performance lithium batteries]. *Progress in Chemistry*, *35*(4), 577–592. https://doi.org/10.7536/PC220937

[13] Zazoum, B. (2023). Lithium-ion battery state of charge prediction based on machine learning approach. *Energy Reports*, *9*, 1152–1158. https://doi.org/10.1016/j.egyr.2023.03.091

[14] Zhang, L., Shen, Z., Sajadi, S. M., Prabuwono, A. S., Mahmoud, M. Z., Cheraghian, G., & El Din, E. M. T. (2022). The machine learning in lithium-ion batteries: A review. *Engineering Analysis with Boundary Elements*, *141*, 1–16. https://doi.org/10.1016/j.enganabound.2022.04.035

[15] Valizadeh, A., & Amirhosseini, M. H. (2024). Machine learning in lithium-ion battery: Applications, challenges, and future trends. *SN Computer Science*, *5*(6), 717. https://doi.org/10.1007/s42979-024-03046-2

[16] Chhetri, M., & Martirosyan, K. S. (2025). Utilizing machine learning to predict the charge storage capability of lithium-ion battery materials. *Eurasian Chemico-Technological Journal*, *27*(1), 3–11. https://doi.org/10.18321/ectj1651

[17] Novita, M., Ristanto, S., Saptaningrum, E., Supriyadi, S., Marlina, D., Rondonuwu, F. S., . . . , & Brik, M. G. (2023). Study on local-structure symmetrization of K_2XF_6 crystals doped with Mn^{4+} ions by first-principles calculations. *Materials*, *16*(11), 4046. https://doi.org/10.3390/ma16114046

[18] Novita, M., Marlina, D., Ogasawara, K., Seok, K. J., & Soo, K. Y. (2021). Study on the optical luminescence properties of Li_2Tio_3: Mn^{4+} and Cr^{3+}. *Chemistry Letters*, *50*(1), 52–56. https://doi.org/10.1246/cl.200596

[19] Novita, M., Chauhan, A. S., Ujianti, R. M. D., Marlina, D., Kusumo, H., Anwar, M. T., . . . , & Brik, M. G. (2024). Exploring deep learning and machine learning for novel red phosphor materials. *Journal of Luminescence*, *269*, 120476. https://doi.org/10.1016/j.jlumin.2024.120476

[20] Lee, Y., Jeon, B., & Lim, S. (2026). A comparative study on machine learning models for estimating cathode material compositions in black powders using prompt gamma-ray spectra. *Nuclear Engineering and Technology*, *58*(4), 104069. https://doi.org/10.1016/j.net.2025.104069

[21] Li, W., Erickson, E. M., & Manthiram, A. (2020). High-nickel layered oxide cathodes for lithium-based automotive batteries. *Nature Energy*, *5*(1), 26–34. https://doi.org/10.1038/s41560-019-0513-0

[22] Urban, A., Seo, D.-H., & Ceder, G. (2016). Computational understanding of Li-ion batteries. *npj Computational Materials*, *2*(1), 16002. https://doi.org/10.1038/npjcompumats.2016.2

[23] Hastie, T., Tibshirani, R., James, G., & Witten, D. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). USA: Springer.

[24] El Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). *Machine learning for practical decision making: A multidisciplinary perspective with applications from healthcare, engineering and business analytics*. Switzerland: Springer. https://doi.org/10.1007/978-3-031-16990-8

[25] Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

[26] Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: From efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, *6*, 1124553. https://doi.org/10.3389/frai.2023.1124553

[27] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007

[28] Zhang, S. (2022). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4663–4675. https://doi.org/10.1109/TKDE.2021.3049250

[29] Cunningham, P., & Delany, S. J. (2022). k-Nearest Neighbor Classifiers - A tutorial. *ACM Computing Surveys*, *54*(6), 128. https://doi.org/10.1145/3459665

[30] Jabardi, M. (2025). Support vector machines: Theory, algorithms, and applications. *Infocommunications Journal*, *17*(1), 66–75. http://doi.org/10.36244/ICJ.2025.1.8

[31] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[32] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6), 15501329221106935. https://doi.org/10.1177/15501329221106935

[33] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[34] Theodoridis, S. (2025). *Machine learning: From the classics to deep networks, transformers, and diffusion models* (3rd ed.). USA: Academic Press. https://doi.org/10.1016/C2023-0-51257-1

[35] Huang, Z., Tong, F., Chen, G., Chen, X., Xu, X., Mu, Z., . . . , & Gu, X. (2026). Prediction of Lithium-ion battery states via combination of implantable sensors and machine learning. *Journal of Energy Storage*, *147*, 120010. https://doi.org/10.1016/j.est.2025.120010

[36] Feng, X., Hua, Y., Zou, J., Jia, S., Ji, J., Xing, Y., . . . , & Liao, J. (2022). Intelligible models for healthcare: Predicting the probability of 6-month unfavorable outcome in patients with ischemic stroke. *Neuroinformatics*, *20*(3), 575–585. https://doi.org/10.1007/s12021-021-09535-6

[37] Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence*, *94*, 103770. https://doi.org/10.1016/j.engappai.2020.103770

[38] Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An efficient AdaBoost algorithm with the multiple thresholds classification. *Applied Sciences*, *12*(12), 5872. https://doi.org/10.3390/app12125872

[39] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, *7*(1), 94. https://doi.org/10.1186/s40537-020-00369-8

[40] Meng, K., Bai, K., & Sun, S. (2024). Artificial intelligence driven design of cathode materials for sodium-ion batteries using graph deep learning method. *Journal of Energy Storage*, *101*, 113809. https://doi.org/10.1016/j.est.2024.113809

[41] Powers, D. (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

[42] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[43] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

[44] Zhang, H., Gao, J., Shi, Z., Gao, B., Wang, X., Wu, J., . . . , & Han, Y. (2026). High-throughput screening of efficient ionic liquid additives for aqueous zinc-ion batteries assisted by machine learning. *Journal of Energy Storage*, *154*, 121278. https://doi.org/10.1016/j.est.2026.121278

How to Cite: Novita, M., Chauhan, A. S., Ujianti, R. M. D., Ristanto, S., Herlambang, B. A., Ogasawara, K., …, & Piasecki, M. (2026). Comparative Study of Machine Learning Algorithms for Crystal Structure Predictive Modeling of Li-ion Battery Cathodes. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62028409