Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3347~3355
e-ISSN: 2548-6861

Ensemble Tree-Based Machine Learning for Predicting Volumetric Change in Lithium-Based Battery Electrode Materials

Arsenio Farrell Winoto [1]*, Gustina Alfa Trisnapradika [2]*
* Department of Informatics Engineering, Universitas Dian Nuswantoro
arseniow777@gmail.com [1], gustina.alfa@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2026-06-18
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Battery electrode materials,
Ensemble tree models,
Hyperparameter tuning,
Machine learning,
Volumetric change prediction.

ABSTRACT

Volumetric instability in lithium-based electrode materials remains a persistent challenge in electric vehicle battery development, as identifying stable material combinations through conventional laboratory methods is both time-consuming and resource-intensive. This study develops and compares three ensemble tree-based machine learning models, namely Random Forest, XGBoost, and CatBoost, to predict the maximum volume change percentage of lithium-based electrode materials. A dataset of 52,503 samples was constructed by integrating electrode pair data with structural and electronic features from the Materials Project API, enriched with compositional descriptors extracted using the Matminer Magpie preset. Each model underwent baseline evaluation followed by hyperparameter tuning using Optuna with Bayesian optimization over 100 trials, assessed using RMSE, MAE, and R^2. All three models achieved R^2 test above 0.98, with Random Forest yielding the best performance at RMSE of 17.1713, MAE of 4.0954, and R^2 test of 0.9900. SHAP analysis identified density discharge as the most determinant predictor across all models, reflecting its physicochemical role in representing the final crystal structure state following lithium intercalation. These findings confirm that ensemble tree-based models offer a reliable and efficient alternative to wet laboratory experimentation for lithium-based electrode material discovery.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Reducing greenhouse gas emissions has become increasingly critical as the proliferation of fossil fuel-powered vehicles continues to be a primary contributor to global greenhouse gas emissions [1]. According to the International Energy Agency (IEA), the transportation sector accounted for 7.98 gigatons, approximately 21.7% of total global energy-related CO_2 emissions in 2023, reflecting a continued upward trend from 7.73 Gt recorded in 2021 [2]. Consequently, the adoption of electric vehicles (EVs) has emerged as one of the most viable pathways toward achieving net zero emissions, as transportation electrification represents the single most significant contributor to road traffic carbon emission reduction [3]. Nevertheless, the successful large-scale deployment of EVs is fundamentally contingent upon battery quality, as batteries inevitably undergo capacity degradation with prolonged cycling [4]. Identifying optimal electrode material combinations thus becomes essential to improving battery longevity. Conventional wet laboratory experiments, however, are inherently costly and low-throughput, relying on extensive trial and error procedures, machine learning therefore presents a more efficient alternative by substantially reducing operational costs while accelerating the material discovery process [5].

Prior work in machine learning-driven battery research has been conducted notably by Moses et al. [6]. who constructed a database of charged and discharged electrode pairs comprising more than 190,000 entries and applied deep neural network models to predict average voltage and percentage volume change across the dataset. However, their models were evaluated without per-ion specialization, leaving predictive accuracy for lithium-based electrode materials uncharacterized despite lithium being the dominant working ion in commercial EV batteries, and no prior study has specifically applied ensemble tree-based models with

Bayesian hyperparameter optimization to this subset. Among all active ion species, lithium possesses the lowest reduction potential and smallest ionic radius, endowing lithium-based batteries with superior gravimetric capacity, volumetric capacity, and power density [7]. Ensemble tree models were further selected as they consistently outperform deep learning on tabular data[8] and other machine learning algorithms with a p-value below 0.001[9], making them well suited for the heterogeneous tabular dataset employed here. This study therefore presents a novel contribution by being the first to systematically develop, compare, and optimize ensemble tree-based models specifically for volumetric change prediction within the lithium working ion subset.

This study addresses prediction of the maximum volume change percentage as lithium ions intercalate and deintercalate from electrode crystal structures, a parameter of critical importance as volumetric variations induce mechanical stress, structural deterioration, and capacity loss[4]. Irreversible volume changes also serve as a direct indicator of battery State of Health, making accurate prediction essential for preventing premature cell failure[10]. The developed models are expected to serve as a faster and more cost-effective material screening tool relative to conventional laboratory experimentation, contributing to more durable EV batteries and representing a computational effort toward reducing greenhouse gas emissions globally.

II. METHOD

This study was conducted through literature review, data collection, preprocessing, and model development with Optuna-based hyperparameter tuning [11]. The overall research workflow is illustrated in Figure 1.

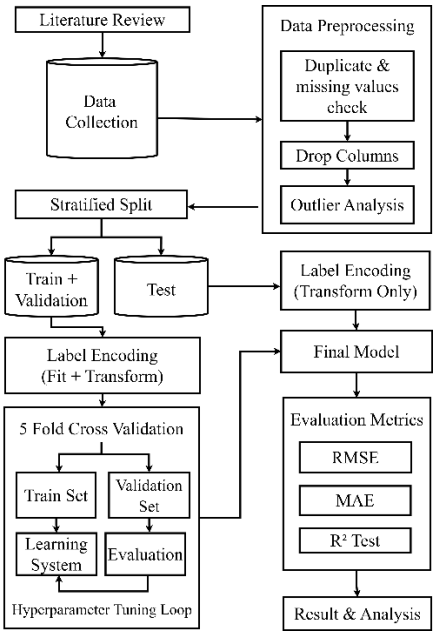

Figure 1. Research Workflow Diagram

A. Data Collection

The dataset used in this study was sourced from [6], which constructed a database of paired battery electrode materials comprising charged and discharged electrode pairs, with attributes covering material pair identifiers, working ion species, average voltage, and maximum delta volume percentage, encompassing all working ion species. From this database, only samples with lithium as the working ion were retained, yielding the lithium-specific subset used in this study. All unique material IDs derived from the charge and discharge identifier columns were subsequently submitted to the Materials Project API via MPRester to retrieve structural and electronic properties for each material, specifically band gap, density, space group, crystal system, and crystal structure. Data retrieval was performed in batches of 50 material IDs per request, with a retry mechanism of up to five attempts and exponential back-off delay to handle connection failures, and raw retrieved data were cached locally as individual JSON files per material ID to avoid redundant API calls.

Element-based compositional features were then extracted from the retrieved crystal structures using the Matminer ElementProperty featurizer with the Magpie preset, producing 37 compositional descriptors per material, including electronegativity statistics, covalent radius statistics, Mendeleev number statistics, atomic mass statistics, and stoichiometric norm values, computed independently for the charged and discharged conditions. Additionally, two derived compositional features, namely the number of elements and the number of metal ions, were computed directly from the crystal structure composition and appended to the dataset. The resulting dataset comprised 52,503 rows and 87 columns, with feature groups summarized in Table I.

TABLE I
DATASET FEATURE GROUP

Feature Group	Description	Total
Material Identity and property	id charge, id discharge, working ion, average voltage	4
Structural & Electronic (charge)	band gap, density, space group, crystal system	4
Structural & Electronic (discharge)	band gap, density, space group, crystal system	4
Magpie Features (charge)	electronegativity mean, covalent radius range, mendeleev number deviation, atomic mass mean, etc.	37
Magpie Features (discharge)	electronegativity mean, covalent radius range, mendeleev number deviation, atomic mass mean, etc.	37
Target	maximum delta volume percentage	1
Total		87

B. Data Preprocessing

Prior to model training, the dataset underwent a series of preprocessing steps. A duplication check confirmed that no duplicate rows were present, and a missing value inspection established that the dataset was complete with no missing entries, eliminating the need for imputation. Several columns were subsequently removed, namely the material identity columns comprising id charge, id discharge, and working ion, which carry no predictive value, as well as the average voltage column, which does not constitute the target variable used in this study. This process reduced the total number of columns from 87 to 83 columns. The overall preprocessing pipeline is summarized in Table II.

TABLE II
DATA PREPROCESSING PIPELINE

Preprocessing Step	Output
Duplicate and missing values check	No duplicates or missing values found
Drop identity columns and average voltage	83 features retained
Stratified train-test split (80:20)	Train: 42,002 / Test: 10,501
Label encoding on the crystal system (fit on train only)	7 classes per condition

All preprocessing steps successfully produced a clean dataset ready for model training. Prior to data splitting, a target distribution analysis was conducted, revealing a highly right-skewed distribution with a skewness value of 15.14 and kurtosis of 360.15, spanning a range from 0.0002 to 7,325.67%. These extreme values were retained because they represent materials that intrinsically exhibit large volumetric changes, and their removal would risk discarding scientifically relevant material information [12]. Data splitting was subsequently performed at an 80:20 ratio, yielding a training set of 42,002 samples and a test set of 10,501 samples, using decile bin-based stratification in which the target variable was discretized into ten quantile bins prior to splitting to ensure a consistent target distribution proportion between both subsets. The training set was further partitioned into five folds for cross-validation, whereby each fold utilized approximately 33,600 samples for training and 8,400 samples for validation in a rotating fashion, ensuring that every training sample served as a validation sample exactly once across the five iterations. Both subsets exhibited consistent distribution patterns, with a training mean of 37.49 and a test mean of 38.45, confirming that stratified splitting successfully preserved proportional target representation across both partitions with no data leakage detected between subsets. The final preprocessing step involved label encoding of the categorical crystal system features for the charged and discharged conditions, comprising seven classes including Cubic, Hexagonal, Monoclinic, Orthorhombic, Tetragonal, Triclinic, and Trigonal. The encoder was fitted exclusively on the training data and subsequently applied to the test data to prevent data leakage [13]. Furthermore, an explicit assertion was applied to verify that no sample overlap existed between the training and test subsets, confirming zero data leakage between partitions and ensuring the integrity of the evaluation protocol.

C. Model Development

This study developed three ensemble tree models, namely Random Forest, XGBoost, and CatBoost. The three algorithms were selected based on ensemble mechanism diversity, covering the bagging and boosting paradigms. Random Forest represents the bagging paradigm, while XGBoost and CatBoost represent boosting-based ensembles with distinct mechanisms whereby XGBoost employs second-order gradient optimization and CatBoost applies ordered boosting to reduce prediction shift.

LightGBM was excluded as its histogram-based gradient boosting mechanism is algorithmically redundant with XGBoost and would not contribute meaningful methodological diversity. Graph Neural Networks were excluded as they require graph-structured crystal representations with atoms as nodes and chemical bonds as edges, whereas the dataset used in this study is tabular and does not include such representations, placing GNN application outside the scope of this study. All models were trained on the preprocessed training data and evaluated on the separate test set through two developmental stages, comprising baseline model evaluation followed by hyperparameter tuning.

In the first stage, all models were executed using their respective default parameters without modification. In the second stage, as hyperparameter selection significantly influences the predictive performance of tree-based models [14], hyperparameter tuning was performed using the Optuna framework with Bayesian optimization based on the Tree structured Parzen Estimator method over 100 trials per model. The objective function minimized was the mean RMSE from 5 Fold Cross-validation on the training data as this approach ensures hyperparameters are evaluated across multiple distinctive data subsets, enabling more robust generalization performance [15]. Early stopping with 50 rounds was applied to XGBoost and CatBoost to prevent overfitting.

Collectively, these measures constitute a multi-layered overfitting prevention strategy. Stratified splitting ensures that the target distribution is proportionally represented in both subsets, five-fold cross-validation as the tuning objective ensures that hyperparameters are evaluated across five distinct and non-overlapping data partitions rather than a single validation subset, early stopping prevents excessive iteration beyond the point of generalization improvement, and fitting the label encoder exclusively on training data eliminates the possibility of target distribution leakage into the preprocessing stage. The search space defined for each model is presented in Table III.

TABLE III
HYPERPARAMETER SEARCH SPACE

Model	Parameter	Range
Random Forest	N estimators	100-500
	Max depth	5-25

	Min samples split	2–15
	Min samples leaf	1-8
	Max features	sqrt, log2, 0.2-0.9
	Bootstrap	True/False
	Max samples	0.5-1.0 (if bootstrap true)
	Min impurity decrease	0.0-0.05
XGBoost	N estimators	50-1500
	Learning rate	0.005-0.35 (log)
	Max depth	3-10
	Subsample	0.5-1.0
	Colsample by tree	0.5-1.0
	Colsample by level	0.5-1.0
	Reg alpha	0.0-5.0
	Reg lambda	0.0-5.0
	Min child weight	1-10
	Gamma	0.0-5.0
CatBoost	Iterations	200-1000
	Learning rate	0.01-0.1 (log)
	Depth	4-8
	12 leaf reg	0.01-5.0 (log)
	Bagging temperature	0.0-1.0
	Random strength	0.01-5.0 (log)
	Border count	64-200

Each model was assigned a search space designed according to its algorithmic characteristics. Random Forest extends the search over max features to a continuous float range and treats bootstrap as a tunable parameter. For XGBoost, an initial trial was enqueued using default parameter values to ensure that Optuna recognizes the already-performative parameter region before exploring the broader search space. The optimal hyperparameter configuration identified by Optuna for each model is presented in Table IV, enabling full replication of the tuning results reported in this study.

TABLE IV
BEST HYPERPARAMETER CONFIGURATION

Model	Parameter	Value
Random Forest	N estimators	204
	Max depth	23
	Min samples split	3
	Min samples leaf	1
	Max features	0.7046
	Bootstrap	False
	Max samples	N/A (bootstrap disabled)
	Min impurity decrease	0.01443
XGBoost	N estimators	963
	Learning rate	0.03164
	Max depth	9
	Subsample	0.9132
	Colsample by tree	0.9912
	Colsample by level	0.8866
	Reg alpha	4.336
	Reg lambda	4.043
	Min child weight	2
	Gamma	3.077
CatBoost	Iterations	334
	Learning rate	0.06523
	Depth	8
	12 leaf reg	0.04418
	Bagging temperature	0.26089
	Random strength	1.42175
	Border count	182

Upon obtaining the optimal hyperparameter combination, each model was retrained using the entire training set with the optimal parameters and evaluated on the test set.

D. Evaluation Metrics

Model performance was evaluated using three complementary metrics, namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), selected on the basis of their complementary diagnostic value given the characteristics of the target distribution in this study. RMSE was selected because the target variable exhibits a highly right-skewed distribution with extreme outliers reaching up to 7,325.67%, and its quadratic penalty on large errors makes it particularly sensitive to model failure on materials with extreme volumetric change, which are precisely the cases most critical to detect. MAE was selected as a complementary metric because its linear penalty structure renders it more robust to outliers, providing a representative measure of average prediction error across the majority of samples exhibiting low to moderate volumetric change. R^2 was selected to quantify the proportion of target variance explained by the model as a whole, providing a scale-independent measure of overall fit [16]. The simultaneous use of all three metrics ensures that model performance is assessed from three non-redundant perspectives, capturing sensitivity to extremes, average accuracy on common samples, and overall explanatory power respectively. These metrics are formulated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

Where $\hat{y}_i$ is the predicted value, y_i is the actual value, $\bar{y}$ is the mean of the actual values, and n is the number of samples.

III. RESULT AND DISCUSSION

A. Exploratory Data Analysis

Exploratory data analysis was conducted to characterize the dataset and target variable distribution prior to modeling. The dataset encompasses electrode material pairs retrieved from the Materials Project database without restriction to any particular chemical family, with compositional diversity

captured through 37 Magpie descriptors per condition computed from the crystal structure composition of each material, encoding elemental statistics including atomic mass, electronegativity, covalent radius, Mendeleev number, valence electron configurations, and stoichiometric norm values. Two derived features, namely the number of constituent elements and the number of metal ions per formula unit, were further computed directly from the crystal structure composition. The dataset exclusively comprises lithium as the working ion and spans seven crystal systems, namely Cubic, Hexagonal, Monoclinic, Orthorhombic, Tetragonal, Triclinic, and Trigonal, across both charged and discharged conditions, with the distribution across crystal systems being non-uniform, reflecting the inherently unequal crystallographic representation of lithium intercalation compounds in the Materials Project database. The target variable, maximum delta volume percentage, spans from 0.0002% to 7,325.67% and exhibits a highly right-skewed distribution with a skewness of 15.14 and kurtosis of 360.15, indicating that the majority of electrode pairs undergo moderate volumetric change while a minority of high-deformation materials produce the observed distributional skew, a range that reflects the physical diversity of lithium-based electrode materials rather than a data quality artifact. As illustrated in Figure 2, this distributional characteristic is further confirmed by substantial deviation from the theoretical normal quantile line across the upper tail region.

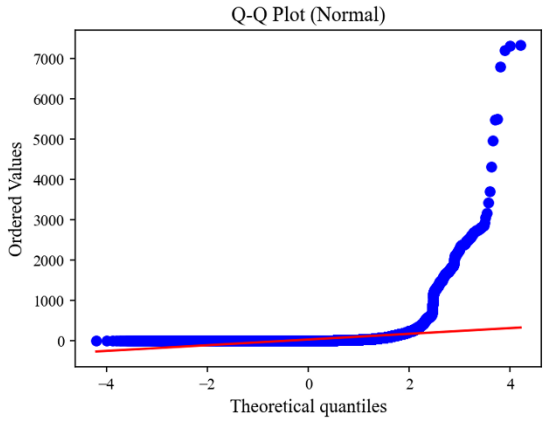

Figure 2. Q-Q Plot

Figure 2 shows that the target distribution deviates substantially from normality. Stratified splitting was subsequently applied to ensure a consistent target distribution proportion between the training and test sets, yielding a training mean of 37.49 and a test mean of 38.45, which confirms that decile bin-based stratification successfully preserved proportional target representation across both subsets, thereby preventing evaluation bias attributable to distributional imbalance.

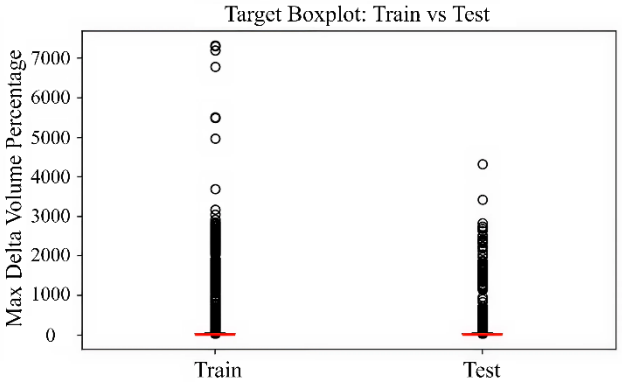

Figure 3. Target Distribution Boxplot

Figure 3 further corroborates this finding, as both subsets display comparable interquartile ranges and outlier distributions, providing additional visual evidence that the stratification procedure effectively maintained distributional equivalence between partitions.

B. Baseline and Tuned Model Performance

Hyperparameter optimization was performed exclusively on the training set through five-fold cross-validation, keeping the test set completely unseen throughout the process. Table V presents the mean train and validation performance across five folds for each tuned model, where RF, XGB, and CB denote Random Forest, XGBoost, and CatBoost, respectively.

TABLE V
TUNED MODEL PERFORMANCE ACROSS FIVE-FOLD CROSS-VALIDATION

Metric	Partition	RF	XGB	CB
RMSE	Train	4.0278	1.8251	5.1160
	Validation	22.7142	22.9502	21.1290
MAE	Train	2.4959	1.2498	3.2540
	Validation	4.4062	3.0583	4.5162
R^2	Train	0.9995	0.9999	0.9992
	Validation	0.9825	0.9803	0.9826

CatBoost achieved the lowest validation RMSE at 21.1290, followed by Random Forest at 22.7142 and XGBoost at 22.9502. XGBoost recorded the lowest validation MAE at 3.0583, while Random Forest and CatBoost achieved comparable R^2 validation values of 0.9825 and 0.9826

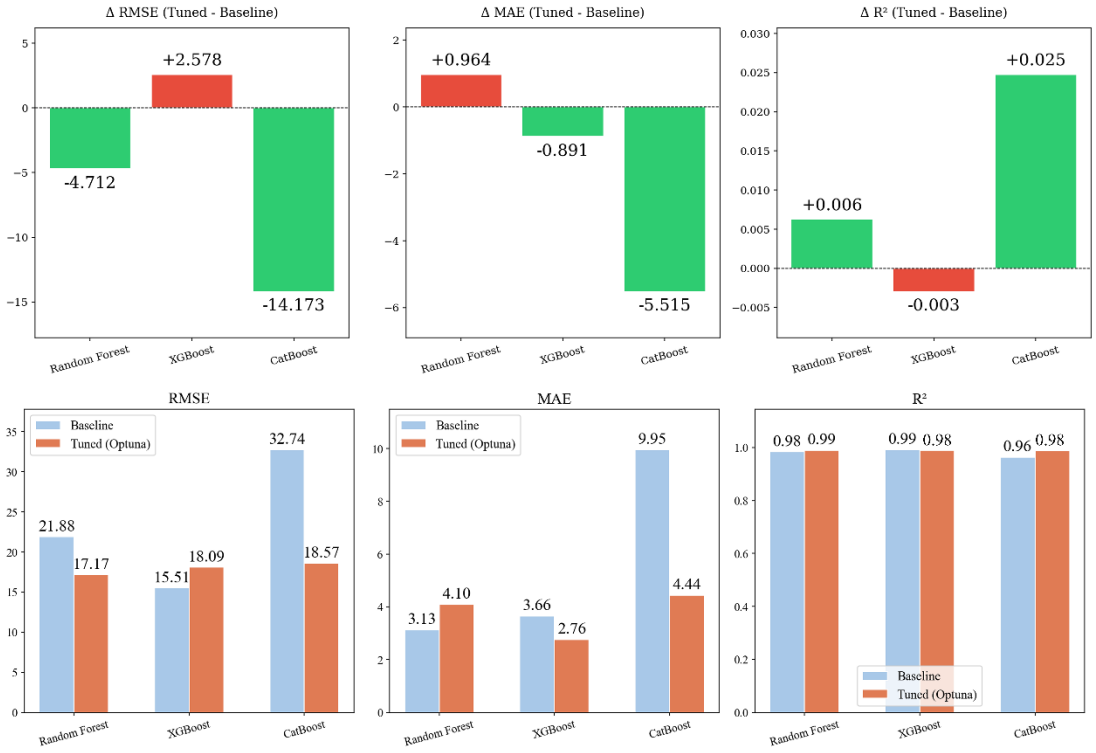

Figure 4. Baseline vs. Tuned Model Performance Comparison Across RMSE, MAE, and R²

respectively. The consistent performance across all five held-out validation folds confirms that the tuned hyperparameter configurations generalize beyond the training partition rather than overfitting to any single fold.

Each tuned model was subsequently evaluated on the held-out test set, with results presented in Table VI, where the test set had not been accessed at any point during training or optimization.

TABLE VI
BASELINE AND TUNED MODEL PERFORMANCE ON TEST SET

Metric	Stage	RF	XGB	CB
RMSE	Baseline	21.8832	15.5085	32.7447
	Tuned	17.1713	18.0868	18.5721
MAE	Baseline	3.1316	3.6555	9.9548
	Tuned	4.0954	2.7649	4.4399
R² Test	Baseline	0.9837	0.9918	0.9635
	Tuned	0.9900	0.9889	0.9883

Tuning improved test set performance to varying degrees, with CatBoost and Random Forest showing the largest RMSE reductions. XGBoost showed an unusual pattern, with RMSE increasing from 15.5085 to 18.0868 while MAE decreased from 3.6555 to 2.7649, suggesting that optimizing for cross-validation RMSE led the model to prioritize common samples over extreme values. The consistent alignment between CV and test results across all models confirms these improvements reflect genuine generalization rather than favorable data partitioning. Figure 4 illustrates this through metric deltas and baseline-tuned comparisons across RMSE, MAE, and R², with CatBoost showing the largest RMSE delta (-14.173) followed by Random Forest (-4.712), while XGBoost's positive RMSE delta (+2.578) alongside a negative MAE delta (-0.891) visually confirms the same tradeoff, consistent across all models except XGBoost on RMSE.

C. *Model Evaluation and Comparison*

All three tuned models recorded R² test values above 0.98, confirming strong explanatory capability across the target range. Random Forest achieved the best overall performance with the lowest RMSE of 17.1713 and highest R² test of 0.9900, while XGBoost recorded the lowest MAE of 2.7649, reflecting superior accuracy on common samples despite a marginally higher RMSE of 18.0868. CatBoost ranked third by RMSE at 18.5721 while still achieving a competitive R² test of 0.9883, indicating that all three models converge toward comparable overall explanatory power. Inter-model differences in R² test are relatively narrow, spanning 0.9883 to 0.9900, whereas RMSE differences are more pronounced across the three models.

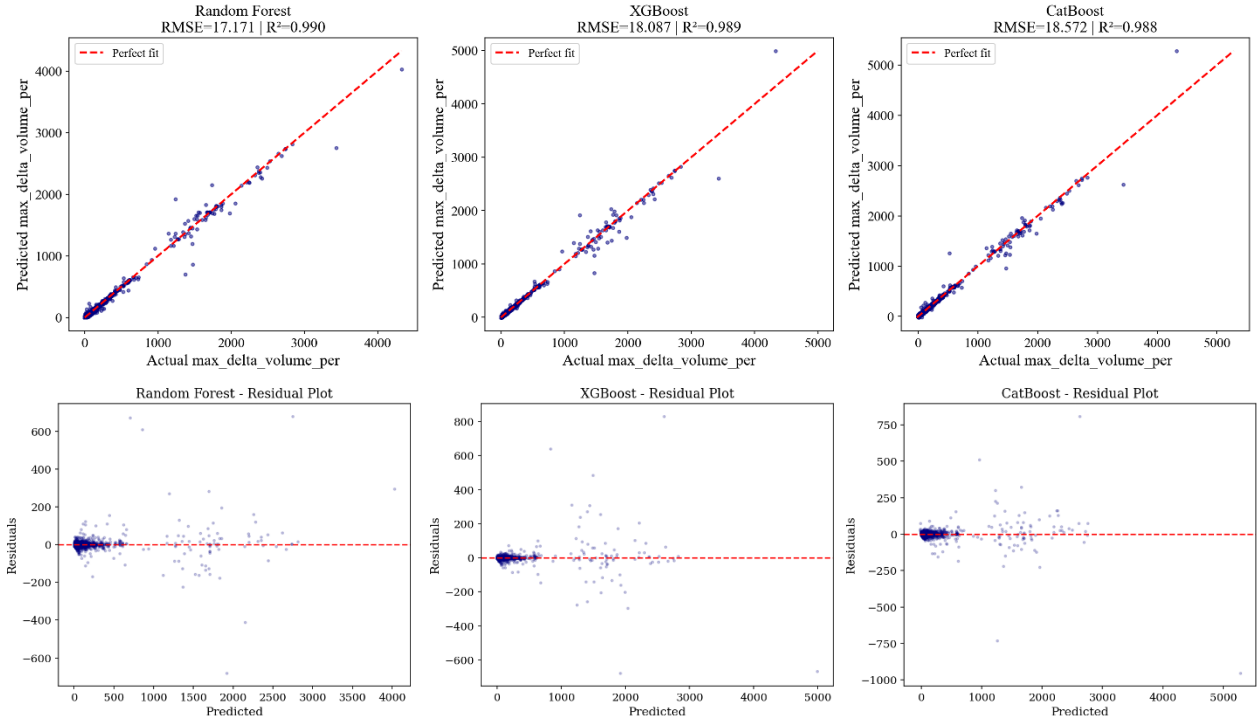

Figure 5. Actual vs Predicted and Residual Plot for All Tuned Models

Figure 5 shows that all three models exhibit predicted values concentrated closely along the perfect fit line across the low to moderate target value range. Deviation from the perfect fit line becomes apparent at higher target values, though all three models maintain relatively consistent point density along the reference line across the full target range, further corroborating the quantitative metric results presented above.

Residual analysis further confirms this pattern. All models produce residuals relatively randomly distributed around zero across the low to moderate predicted value range, though residual variance increases at higher predicted values across all models, reflecting heteroskedasticity consistent with the right-skewed target distribution. Random Forest, XGBoost, and CatBoost demonstrate contained residual dispersion, corroborating their superior quantitative performance.

The strong predictive performance achieved without feature normalization or outlier removal reflects the recursive partitioning nature of tree-based models, which maintain consistently high performance regardless of feature scaling [17], and demonstrate robustness on non-normally distributed data [18].

To assess whether inter-model differences are statistically meaningful, Wilcoxon signed-rank tests were conducted on per-fold RMSE scores for all model pairs, with Cohen's d as a complementary effect size measure, and the results are presented in Table VII.

TABLE VII
WILCOXON SIGNED-RANK TEST RESULTS

Pair	W	p-value	Cohen's d
Random Forest vs XGBoost	7.0	1.0000	-0.038
Random Forest vs CatBoost	4.0	0.4375	0.217
CatBoost vs XGBoost	1.0	0.1250	-0.807

No pair achieved statistical significance at the p-value below 0.05 threshold due to the structural limitation of the Wilcoxon signed-rank test at n equal to 5, making Cohen's d the primary indicator of practical differences. Random Forest versus XGBoost yielded d equal to negative 0.038, confirming negligible practical difference between the two models. CatBoost versus XGBoost yielded d equal to negative 0.807, indicating a medium to large practical advantage of CatBoost over XGBoost in cross-validation RMSE, while Random Forest versus CatBoost yielded d equal to 0.217, indicating a small practical difference between the two.

Five-fold cross-validation conducted entirely on the training data confirms that reported performance reflects genuine learned patterns rather than overfitting to a single test partition, with consistent results across all folds despite elevated RMSE in fold three attributable to a higher concentration of extreme volumetric change samples. External validation on independent material compositions or experimental data was not conducted, which constitutes a recognized limitation and a direction for future research.

XGBoost required the shortest tuning duration at 588.33 seconds owing to GPU acceleration via CUDA, followed by

CatBoost at 1,980.70 seconds due to its ordered boosting mechanism, and Random Forest at 7,410.40 seconds reflecting the absence of GPU support across 100 trials with five-fold cross-validation.

D. Feature Importance and SHAP Analysis

SHAP-based feature importance analysis was conducted to interpret the contribution of individual features to the predictions of all three models, as SHAP provides a theoretically grounded framework rooted in game theory that consistently quantifies each feature's contribution to model output [19][20]. The SHAP bar plot presents the mean absolute SHAP value per feature, representing the global contribution of each feature to the model output. Given the high dimensionality of the dataset comprising 83 features, the analysis focuses on the top five most influential features per model to provide a concise yet informative representation of the dominant predictors driving each model's predictions.

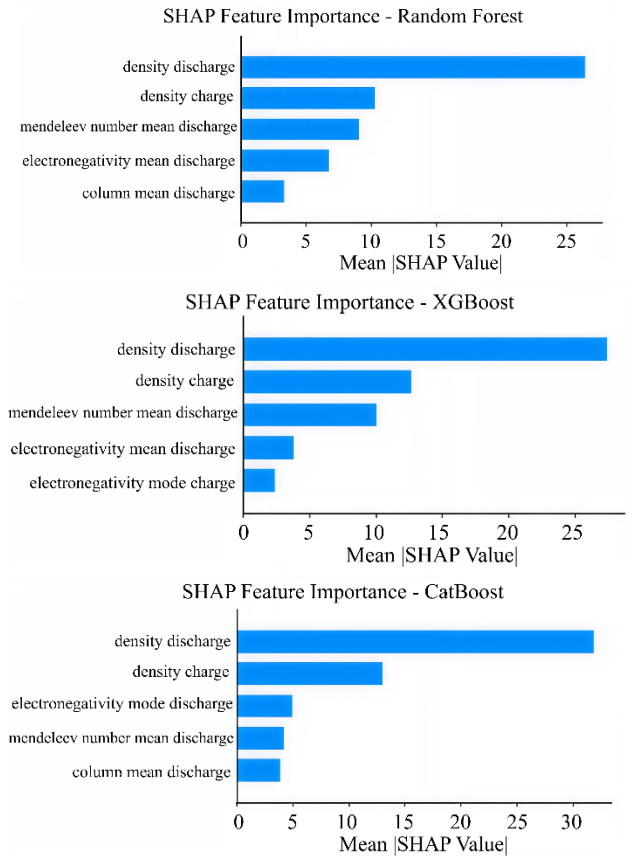

Figure 6. Feature Importance

Figure 6 shows that density discharge consistently ranked first across all three models, followed by density charge in second position, providing strong evidence that structural properties under discharge conditions constitute the most determinant predictors of volumetric change magnitude. Random Forest, XGBoost, and CatBoost demonstrate a distributed pattern in which Magpie features including mendeleev number mean discharge and electronegativity mean discharge also contribute significantly. This is physicochemically justified as discharge density directly represents the final crystal structure state following lithium intercalation, rendering the density differential between charge and discharge conditions the primary indicator of volumetric deformation magnitude. The convergence of importance rankings across algorithmically distinct models further confirms that these features reflect genuine physicochemical relationships rather than model-specific artifacts.

Beyond the density features, several Magpie-derived descriptors consistently emerge as significant contributors. Mendeleev number mean discharge encodes periodic table positioning governing crystal lattice flexibility during intercalation, electronegativity mean discharge captures ionic bond strength between lithium and the host structure directly influencing deformation magnitude, column mean discharge modulates electronic interactions determining structural response to intercalation, and valence orbital p electrons discharge governs bonding directionality and lattice deformability. The consistent emergence of these discharge-state descriptors across all three models confirms that the compositional and electronic characteristics of the fully intercalated material state are the primary physicochemical determinants of volumetric deformation in lithium-based electrode pairs.

IV. CONCLUSION

This study developed and evaluated three ensemble tree-based machine learning models, namely Random Forest, XGBoost, and CatBoost, for predicting the maximum volume change percentage of lithium-based electrode materials using a dataset of 52,503 samples enriched with structural, electronic, and compositional features from the Materials Project API and Matminer Magpie preset. All three models recorded R^2 test values above 0.98, with Random Forest emerging as the best-performing model following Optuna-based Bayesian hyperparameter tuning, achieving RMSE of 17.1713, MAE of 4.0954, and R^2 test of 0.9900. SHAP analysis consistently identified density discharge as the most determinant predictor across all models, providing physicochemical evidence that the structural state of electrode materials under discharge conditions is the primary driver of volumetric deformation magnitude. The primary limitation of this study is that all models were trained and evaluated exclusively on computationally derived data from the Materials Project without external experimental validation, though five-fold cross-validation results confirm that reported performance reflects genuinely learned patterns rather than favorable test set partitioning.

These findings confirm that ensemble tree-based models serve as computationally efficient and reliable screening tools for lithium-based electrode material discovery, offering a viable alternative to conventional laboratory experimentation. From a scientific standpoint, the identification of density discharge as the primary deformation predictor provides actionable design guidance, as candidates exhibiting minimal

density differential between charge and discharge states indicate greater mechanical stability and extended cycle life. From an industrial perspective, the Random Forest model can be directly applied to screen large candidate pools from the Materials Project prior to synthesis, prioritizing mechanically stable materials and eliminating high-deformation candidates early, thereby reducing experimental burden and accelerating the development of more durable electric vehicle batteries.

REFERENCES

[1] M. F. Ng, Y. Sun, and Z. W. Seh, "Machine learning-inspired battery material innovation," *Energy Adv.*, vol. 2, no. 4, pp. 449–464, 2023, doi: 10.1039/d3ya00040k.

[2] E. Ndhlovu, D. Mhlanga, and B. Duri, "Correction: Decarbonising urban transport: an overview of electric vehicles, public transport, and sustainable infrastructure in achieving net-zero emissions (Discover Global Society, (2025), 3, 1, (53), 10.1007/s44282-025-00201-9)," *Discov. Glob. Soc.*, vol. 3, no. 1, 2025, doi: 10.1007/s44282-025-00267-5.

[3] Z. Gao *et al.*, "Electric vehicle lifecycle carbon emission reduction: A review," *Carbon Neutralization*, vol. 2, no. 5, pp. 528–550, 2023, doi: 10.1002/cnl2.81.

[4] T. Rahman and T. Alharbi, "Exploring Lithium-Ion Battery Degradation: A Concise Review of Critical Factors, Impacts, Data-Driven Degradation Estimation Techniques, and Sustainable Directions for Energy Storage Systems," *Batteries*, vol. 10, no. 7, 2024, doi: 10.3390/batteries10070220.

[5] L. Tang, P. Leung, Q. Xu, and C. Flox, "Machine Learning Orchestrating the Materials Discovery and Performance Optimization of Redox Flow Battery," *ChemElectroChem*, vol. 11, no. 15, pp. 1–17, 2024, doi: 10.1002/celc.202400024.

[6] I. A. Moses, V. Barone, and J. E. Peralta, "Accelerating the discovery of battery electrode materials through data mining and deep learning models," *J. Power Sources*, vol. 546, no. August, p. 231977, 2022, doi: 10.1016/j.jpowsour.2022.231977.

[7] N. Bhandari, G. J. Martis, and S. L. Gaonkar, "Advancements in lithium-ion batteries: sustainability and market impact," *Discov. Mater.*, vol. 5, no. 1, 2025, doi: 10.1007/s43939-025-00291-x.

[8] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, 2022.

[9] S. Uddin and H. Lu, "Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data," *PLoS One*, vol. 19, no. 4 April, pp. 1–12, 2024, doi: 10.1371/journal.pone.0301541.

[10] G. Bree, H. Hao, Z. Stoeva, and C. T. John Low, "Monitoring state of charge and volume expansion in lithium-ion batteries: an approach using surface mounted thin-film graphene sensors," *RSC Adv.*, vol. 13, no. 10, pp. 7045–7054, 2023, doi: 10.1039/d2ra07572e.

[11] T. Kee and W. K. O. Ho, "Optimizing Machine Learning Models for Urban Sciences: A Comparative Analysis of Hyperparameter Tuning Methods," *Urban Sci.*, vol. 9, no. 9, 2025, doi: 10.3390/urbansci9090348.

[12] Y. Ma, P. Xu, M. Li, X. Ji, W. Zhao, and W. Lu, "The mastery of details in the workflow of materials machine learning," *npj Comput. Mater.*, vol. 10, no. 1, pp. 1–17, 2024, doi: 10.1038/s41524-024-01331-5.

[13] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[14] J. P. Lai, Y. L. Lin, H. C. Lin, C. Y. Shih, Y. P. Wang, and P. F. Pai, "Tree-Based Machine Learning Models with Optuna in Predicting Impedance Values for Circuit Analysis," *Micromachines*, vol. 14, no. 2, 2023, doi: 10.3390/mi14020265.

[15] P. Heidari and A. Milan, "Combining K-fold cross validation with bayesian hyperparameter optimization for accuracy enhancement of land cover and land use classification," *Sci. Rep.*, vol. 15, no. 1, pp. 1–16, 2025, doi: 10.1038/s41598-025-23336-w.

[16] A. B. Hassanat *et al.*, "A Novel Outlier-Robust Accuracy Measure for Machine Learning Regression Using a Non-Convex Distance Metric," *Mathematics*, vol. 12, no. 22, pp. 1–20, 2024, doi: 10.3390/math12223623.

[17] J. M. H. Pinheiro *et al.*, "The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks," *IEEE Access*, vol. 13, no. November, pp. 199903–199931, 2025, doi: 10.1109/ACCESS.2025.3635541.

[18] C. Okolie *et al.*, "Assessment of explainable tree-based ensemble algorithms for the enhancement of Copernicus digital elevation model in agricultural lands," *Int. J. Image Data Fusion*, vol. 15, no. 4, pp. 430–460, 2024, doi: 10.1080/19479832.2024.2329563.

[19] F. Qayyum, M. A. Khan, D. H. Kim, H. Ko , and G. A. Ryu, "Explainable AI for Material Property Prediction Based on Energy Cloud: A Shapley-Driven Approach," *Materials (Basel).*, vol. 16, no. 23, pp. 1–24, 2023, doi: 10.3390/ma16237322.

[20] H. Qin, Y. Zhang, Z. Guo, S. Wang, D. Zhao, and Y. Xue, "Prediction of Bandgap in Lithium-Ion Battery Materials Based on Explainable Boosting Machine Learning Techniques," *Materials (Basel).*, vol. 17, no. 24, 2024, doi: 10.3390/ma17246217.