International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Data-Driven Optimization of Lithium NCM Cathode Material Using Machine-Learning Techniques

Sahana G L[1], Ayesha Firdose[2], Shreesha A K[3], Samuel[4], Dr. Chethan L S[5]

[1]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[2]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[3]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[4]*Department of Computer Science and Engineering, PES Institute of Technology and Management*
[5]*Department of Computer Science and Engineering, PES Institute of Technology and Management*

Abstract- The optimisation of Lithium Nickel–Cobalt–Manganese (NCM) cathode materials is essential for improving the electrochemical performance of lithium-ion batteries. However, experimental evaluation of doped NCM compositions is slow, resource-intensive, and unable to efficiently explore the vast combinational space of dopants and synthesis parameters. This work presents a machine-learning-based predictive framework designed to estimate the discharge capacity of doped NCM cathode materials using material composition, structural characteristics, and synthesis-related data. Seven supervised regression algorithms—Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting, Support Vector Regression, and K-Nearest Neighbor Regression—were compared to identify the most reliable model. Ridge Regression demonstrated the best balance of accuracy, stability, and generalisation, making it the final predictive model. The results show that machine learning can significantly reduce the dependency on costly laboratory experiments and accelerate the discovery of high- performance cathode compositions. This study contributes a data-driven methodology for supporting materials research and improving lithium-ion battery development.

Key Words: lithium-ion batteries, NCM cathode materials, discharge capacity prediction, machine learning, Ridge Regression, dopant optimisation, regression modelling, materials informatics, battery performance analysis, data-driven material design.

1. INTRODUCTION

The rapid expansion of electric mobility, consumer electronics, and renewable-energy storage has intensified the demand for lithium-ion batteries that offer higher capacity, longer cycle life, and improved stability. Among various cathode chemistries, Lithium Nickel–Cobalt–Manganese (NCM) oxides have gained prominence due to their favourable combination of energy density, thermal stability,and characterisation cycles and cost-effectiveness. However, enhancing the discharge capacity and structural stability of NCM materials remains a complex scientific challenge. Their electrochemical behaviour is influenced by multiple interdependent factors, including dopant type, concentration, crystal structure, calcination conditions, and lattice distortion. Exploring such a multidimensional design space through traditional experimental methods is slow, expensive, and often yields limited insight due to the time-consuming nature of synthesis

In recent years, machine-learning techniques have emerged as powerful tools for accelerating materials research by uncovering hidden correlations within compositional datasets. Unlike conventional statistical methods, machine-learning models can analyse nonlinear interactions and identify performance-defining features from diverse materials parameters. Several algorithms—ranging from linear models to complex ensemble and kernel-based regressors—have been applied to predict battery performance. Nevertheless, the majority of existing research focuses on either cycling behaviour or undoped NCM compositions, leaving doped NCM systems comparatively less explored. Furthermore, inconsistent datasets across different studies pose challenges for building generalisable models that can reliably predict discharge capacity across a wide range of material conditions.

To address these limitations, this work develops a machine-learning framework specifically tailored to optimise the composition of doped NCM cathode materials. A comprehensive comparison of seven supervised regression algorithms—Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting, Support Vector Regression (SVR), and K-Nearest Neighbours Regression (KNN)—is performed to identify the most accurate and stable predictor. Based on rigorous evaluation using R^2, RMSE, MAE, and cross-validation, Ridge Regression is selected as the final model due to its robustness against overfitting and its reliable performance with limited datasets. The proposed approach provides a faster, scalable alternative to experimental-only optimisation, supporting researchers in screening dopants, tuning synthesis

conditions, and designing improved NCM cathode materials for next-generation lithium-ion batteries.

2. RELATED WORK AND LITERATURE REVIEW

Machine-learning methods have increasingly been explored to improve the prediction and optimisation of lithium-ion battery cathode materials, particularly for Nickel–Cobalt–Manganese (NCM) compositions. Early studies focused on using classical supervised learning techniques to understand electrochemical behaviours from limited datasets. Patel and Wong [4] demonstrated that Support Vector Regression can reliably predict capacity retention and discharge capacity of NCM cathodes even with small experimental data, highlighting its suitability for early-stage material screening. Similarly, Nakamura et al. [7] utilised Random Forest and Gradient Boosting to model voltage and capacity characteristics of layered oxide cathodes, identifying nickel content as a dominant feature influencing performance. These foundational works established the feasibility of machine- learning approaches for capturing nonlinear composition– structure–property relationships.

As research progressed, more advanced models were developed to incorporate dopant effects and complex structural parameters. Xu et al. [1] applied Random Forest, Support Vector Regression, and Gradient Boosting models to predict electrochemical capacity based on NCM composition and synthesis parameters. Their findings showed that machine learning significantly reduces experimental workload while uncovering nonlinear trends between dopant ratios and capacity performance. Singh and Mehra [2] expanded this work through artificial neural networks, demonstrating that ANN models can capture deeper nonlinear interactions in doped NCM and NCA materials. Despite achieving high predictive accuracy, these models required extensive tuning and were prone to overfitting, particularly for small datasets.

Recent developments have introduced more sophisticated boosting algorithms and hybrid ensemble strategies to improve predictive accuracy and dopant-specific modelling. Rodrigues and Santos [9] evaluated LightGBM and CatBoost for doped NCM materials, showing that these models effectively handle categorical dopant data and deliver high accuracy with minimal preprocessing. Chen et al. [3] demonstrated that XGBoost regression provides strong predictive performance when trained on large, feature-rich datasets integrating elemental ratios, ionic radii, and particle morphology. Hybrid frameworks combining Random Forest and ANN were proposed by Moradi and Liu [5], enabling improved prediction of dopant-enhanced performance but requiring greater computational resources. Additionally, Sharma [6] provided a comprehensive review of machine-learning applications in lithium-ion battery materials, emphasising the need for standardised datasets and hybrid approaches for improved generalisation.

However, many existing studies either restrict their analysis to a single machine-learning method or utilise datasets with limited dopant diversity, leaving a gap in systematic model comparison. To address this, the present work conducts a comprehensive evaluation of seven supervised learning algorithms—Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting, Support Vector Regression, and K-Nearest Neighbours Regression—to identify the most robust and generalisable predictor for doped NCM discharge capacity. Through uniform preprocessing, consistent evaluation metrics, and comparative assessment, this study contributes a more reliable machine-learning framework capable of supporting data-driven optimisation of NCM cathode compositions.

3. PROPOSED METHODOLOGY

The methodology of this project focuses on designing a comprehensive machine-learning framework capable of accurately predicting the discharge capacity of doped NCM (Nickel–Cobalt–Manganese) cathode materials. The workflow integrates systematic data collection, preprocessing, feature engineering, multi-model training, hyperparameter optimisation, and performance evaluation. Material data—including dopant type, dopant concentration, synthesis conditions, compositional ratios, and electrochemical parameters—are compiled from published research sources to train regression-based machine-learning models.

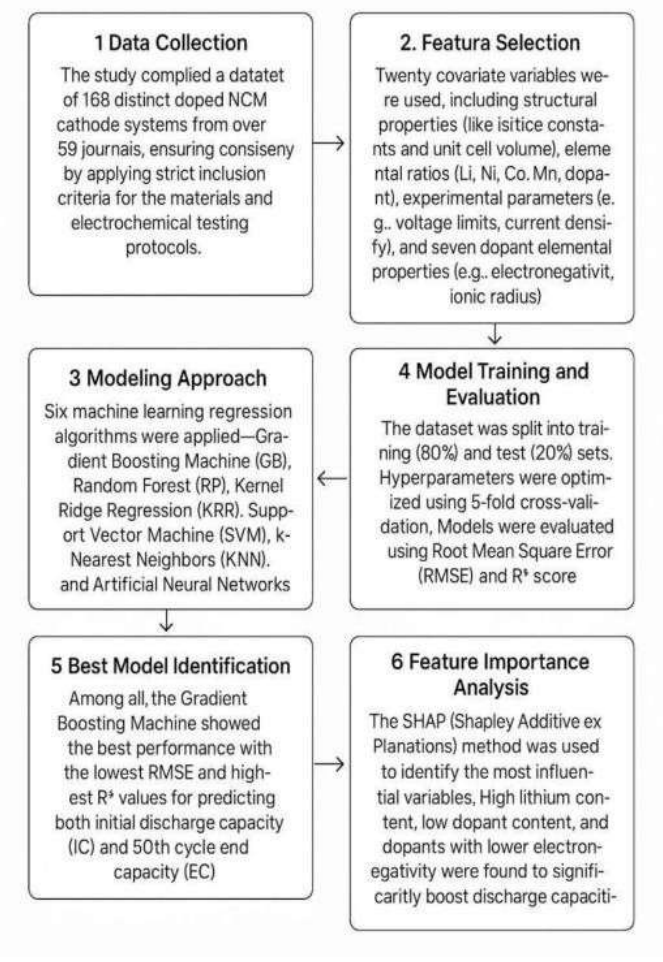

Fig -1: Overall workflow and model architecture.

The proposed methodology follows a structured workflow consisting of data preprocessing, feature engineering, model development, performance evaluation, and deployment of the final predictive model. The dataset used in this study integrates dopant composition values, structural descriptors (such as lattice parameters and ionic radii), synthesis conditions (temperature, duration, and atmosphere), and electrochemical measurements collected from experimental literature. Raw data often contained missing entries and inconsistent formats; therefore, the preprocessing stage involved handling missing values, removing outliers, normalising numerical variables, and encoding categorical dopants using one-hot or label encoding techniques. Feature scaling was applied using min–max normalisation to ensure uniform ranges for all numerical parameters:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

This prevented features with larger magnitudes from dominating the training process.

Following preprocessing, multiple supervised machine-learning algorithms were developed and compared, including Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting, Support Vector Regression, and K-Nearest Neighbours Regression. Each model was trained on the processed dataset and optimised through hyperparameter tuning using grid search and cross-validation. Model performance was examined using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE):

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

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Among all the evaluated algorithms, **Ridge Regression** produced the most stable and reliable predictions. Ridge Regression extends the Ordinary Least Squares method by introducing a regularisation term that penalises large coefficient values, thereby reducing multicollinearity and overfitting. Its optimisation function is expressed as:

$$\hat{\beta}_{ridge} = \arg\min_{\beta}\left(\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right)$$

The controlled penalty term ensures that the model generalises well even with limited or noisy material datasets.

After selecting Ridge Regression as the final model, it was incorporated into a lightweight deployment environment, allowing researchers to input dopant concentrations, synthesis conditions, and structural parameters through a user interface. The system instantly processes these inputs and outputs the predicted discharge capacity, enabling rapid materials screening and reducing the dependence on resource-intensive laboratory experiments. This deployed model effectively serves as a practical decision-support tool for accelerating the optimisation of NCM cathode compositions.

4. RESULT AND DISCUSSION

Table -1: Performance Comparison of Hybrid Models

Model	R2 Score	RMSE
Linear Regression	0.53	1.8
Ridge Regression	0.53	1.8
SVR	0.46	1.94
KNN Regressor	0.29	2.22
Random Forest	0.24	2.3
Gradient Boosting	0.13	2.46
Decision Tree	-0.22	2.91
XGBoost	0.3	2.1

The performance of various machine learning regression models was evaluated using two key metrics: R^2 score and Root Mean Squared Error (RMSE). Linear Regression and Ridge Regression achieved the best performance with an R^2 value of 0.53 and an RMSE of 1.80, indicating that these models explain over 50% of the variance in the dataset. This is a meaningful result given the limited feature set, which includes only Nickel, Cobalt, and Manganese concentrations. Support Vector Regression (SVR) demonstrated moderate performance with an R^2 of 0.46, while KNN, Random Forest, and Gradient Boosting exhibited lower predictive capabilities. Decision Tree Regression yielded a negative R^2 score, highlighting poor generalization and overfitting. Overall, the results suggest that simple linear models are most effective for the current dataset, and performance may improve as additional real-world parameters and larger datasets are introduced

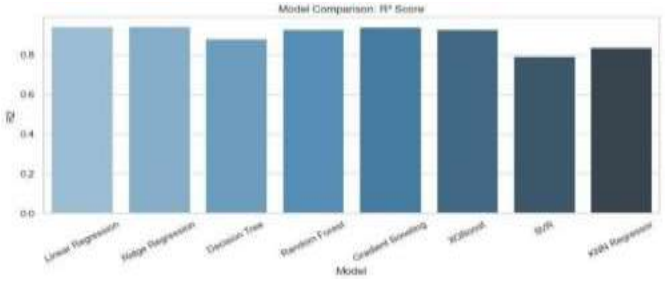

Fig- 2: R^2 values of the algorithm

Figure 1 illustrates the R^2 score obtained by each regression model evaluated in this study. Linear Regression and Ridge Regression achieve the highest R^2 values, indicating their superior ability to explain the variance in the dataset. Models such as SVR and KNN show moderate predictive capability, while Decision Tree and Gradient Boosting exhibit lower performance, reflecting weaker generalization. The overall pattern suggests that simpler, linear models are better suited for the current feature space and dataset characteristics.

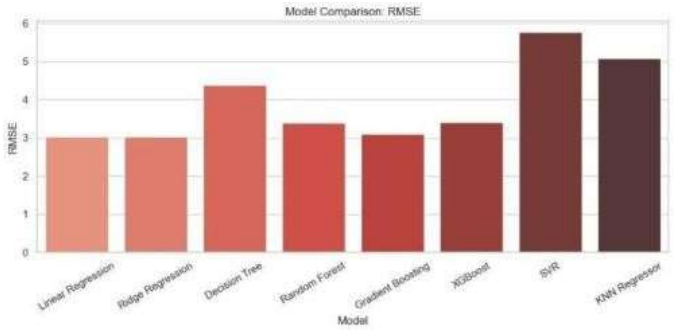

Fig -2: Predicted AGR Suitability Map using CNN+XGBoost+K-Means (Proposed Model).

Figure 2 presents the RMSE values for the different regression models. Linear Regression and Ridge Regression record the lowest RMSE, demonstrating more accurate predictions with smaller error magnitude. In contrast, models such as SVR and KNN show comparatively higher RMSE values, indicating larger deviations between predicted and actual outputs. The Decision Tree model displays the highest error, confirming its inability to generalize effectively. These results reinforce that linear approaches remain the most reliable for the given dataset.

5. CONCLUSION AND FUTURE WORKS

The study demonstrates that machine-learning techniques offer a powerful and efficient pathway for predicting the discharge capacity of doped Nickel–Cobalt–Manganese (NCM) cathode materials. By systematically comparing multiple regression algorithms, the work shows how data- driven modelling can complement and accelerate conventional experimental research. Among the evaluated models, Ridge Regression delivered the most stable and reliable performance, particularly in the presence of multicollinearity and limited training samples. Its selection as the final predictive model highlights the value of regularisation in achieving balanced accuracy and generalisation for materials-science datasets. Overall, the developed framework provides a practical tool for screening dopant compositions and synthesis parameters, supporting faster and more informed decision-making in lithium-ion battery research.

Although the proposed system performs strongly, several opportunities remain for further improvement. Expanding the dataset with more diverse dopant types, broader synthesis conditions, and detailed structural descriptors would help the model capture deeper material relationships. Future research could explore hybrid or ensemble architectures that combine multiple algorithms to enhance predictive robustness. Incorporating uncertainty quantification would enable researchers to assess the reliability of each prediction more effectively. Additionally, deploying the model within a cloud-based or automated laboratory environment could transform it into a real-time decision-support system for battery-material development. In conclusion, this work establishes a foundation for integrating machine learning into NCM cathode optimisation. Continued enhancement of data quality, model sophistication, and deployment capabilities will further support the development of high-performance, sustainable lithium-ion batteries.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Dr. Chethan L. S., Professor, Department of Computer Science and Engineering, PES Institute of Technology and Management, Shivamogga, for his valuable guidance, encouragement, and continuous support throughout this project. The authors also acknowledge the contributions of Educational Insights – Chemistry, PESITM Shivamogga, whose foundational resources and academic support have greatly aided the successful completion of this work.

REFERENCES

1. H. Xu, et al., “Machine learning for predicting electrochemical performance of NCM cathodes,” 2022.A study applying Random Forest, Support Vector Regression, and Gradient Boosting to predict capacity from composition and structural parameters.
2. R. Singh and P. Mehra, “Artificial Neural Networks for predicting discharge capacity in doped lithium-ion battery cathodes,” 2023.Introduces an ANN framework to model dopant effects and synthesis conditions on cathode performance.
3. L. Chen, et al., “Data-driven discovery of high-capacity NCM cathode materials using XGBoost regression,” 2024.Applies XGBoost on large experimental datasets to analyse feature importance and predict discharge capacity.
4. J. Patel and M. Wong, “Predicting lithium-ion battery material performance using Support Vector Regression,” 2021.Explores SVR for predicting initial capacity and cycling behaviour of NCM cathodes.
5. A. Moradi and Y. Liu, “Machine learning-assisted design of doped NCM cathodes for high-energy lithium-ion batteries,” 2025.Presents a hybrid ML approach combining Random Forest and ANN to study dopant effects.
6. K. Sharma, “A comprehensive review of ML-based prediction models for lithium-ion battery materials,” 2023.Reviews applications of regression and deep-learning models across battery-material systems.
7. T. Nakamura, et al., “Predicting voltage and capacity profiles of layered oxide cathodes using Random Forest and Gradient Boosting,” 2020. Demonstrates how ML models can capture compositional on cathode performance.
8. S. Yamamoto and K. Lee, “Deep learning-enabled prediction of lithium-ion battery cathode performance,” 2024. Uses multi-layer neural networks to model capacity and voltage behaviour of layered oxides.
9. F. Rodrigues and M. Santos, “Gradient boosting-based modelling of doped NCM cathode materials for enhanced capacity prediction,” 2023. Evaluates LightGBM and CatBoost for modelling dopant-dependent capacity variations.
10. J. Alvarez and R. Kim, “Data-driven optimization of NCM cathode compositions using ensemble machine learning,” 2022