CS 453/553 Project Report

Nabil Abdel-Rahman
Hritvik Jekki Venkateshwarulu
Daniel Kutnyi

***Abstract*—This paper presents a study on battery materials property prediction. We explore the dataset presented and apply various data mining/machine learning techniques to analyze this dataset and predict properties such as Capacity, Voltage, Efficiency, Energy, and Conductivity of the battery based on the compound formula. Our results demonstrate that it is possible to achieve a certain degree of predictive efficiency using machine methods.**

I. Introduction and Background

A. General Background

The rapid advancement of battery technology is crucial for the development of energy storage solutions, particularly in electric vehicles, renewable energy integration, and portable electronics. The performance of a battery is significantly influenced by the properties of its materials, such as capacity, voltage, efficiency, energy density, and conductivity. Predicting these properties accurately can accelerate the discovery of new battery materials and enhance existing ones. Data mining and machine learning techniques offer a powerful means of analyzing complex datasets and extracting meaningful insights to optimize material selection and performance.

B. Specific Problem

Despite significant research, predicting battery material properties remains challenging due to the complex relationships between chemical composition and electrochemical behavior. Traditional methods, such as experimental testing and theoretical calculations, are time-consuming and expensive. Therefore, leveraging data-driven approaches to predict material properties efficiently is essential for innovation in battery technology.

II. Motivation and Objective

A. Problem Statement

Existing methods for battery material property prediction rely heavily on empirical and theoretical models, which often require extensive experimental validation. These methods are not only resource-intensive but also limited in scalability. Additionally, data scarcity and the high-dimensional nature of battery material properties make accurate prediction difficult using conventional techniques.

B. Contribution and Novelty

This study applies machine learning techniques to predict key battery material properties based on compound formulas. Specifically, we utilize Decision Tree, Random Forest, Fully Connected Neural Networks, and Transformer models to analyze the dataset. By leveraging these advanced data mining techniques, we aim to:

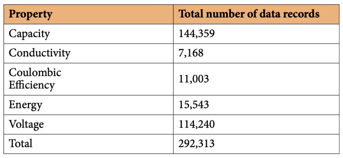
- Develop predictive models for battery properties.
- Identify the most influential features for property prediction.
- Compare the effectiveness of different machine learning algorithms in this domain.

III. Data Collection and Analysis

A. Dataset

The dataset used in this study was generated by mining textual data from 229,061 scientific papers using ChemDataExtractor [1] (version 1.5), specifically adapted for battery research. The extraction process involved natural language processing (NLP) techniques to identify and extract chemical-property relationships from scientific texts. Named entity recognition (NER) was employed to detect chemical compounds, while relation extraction algorithms linked these compounds to their corresponding battery performance metrics. The dataset was constructed by parsing all scientific papers that contained the word "battery," ensuring comprehensive coverage of relevant research. It includes 292,313 data records, capturing 214,617 unique chemical-property relationships among 17,354 unique chemicals and their properties, including capacity, voltage, conductivity, Coulombic efficiency, and energy. Approximately 117,403 data points represent multivariate data series. The dataset is publicly available, providing a representative overview of battery material information extracted from literature, along with a Graphical User Interface (GUI) to facilitate usage.

B. Dataset Introduction and Analysis



Property	Total number of data records
Capacity	144,359
Conductivity	7,168
Coulombic Efficiency	11,003
Energy	15,543
Voltage	114,240
Total	292,313

Fig. 1. Number of data records for each property

The records distribution by different properties can be found in the figure 1 and property distribution in the figure 2.

The dataset extraction algorithm achieved an overall precision of 80%, as verified against a subset of data with known correct records. However, occasional inaccuracies were

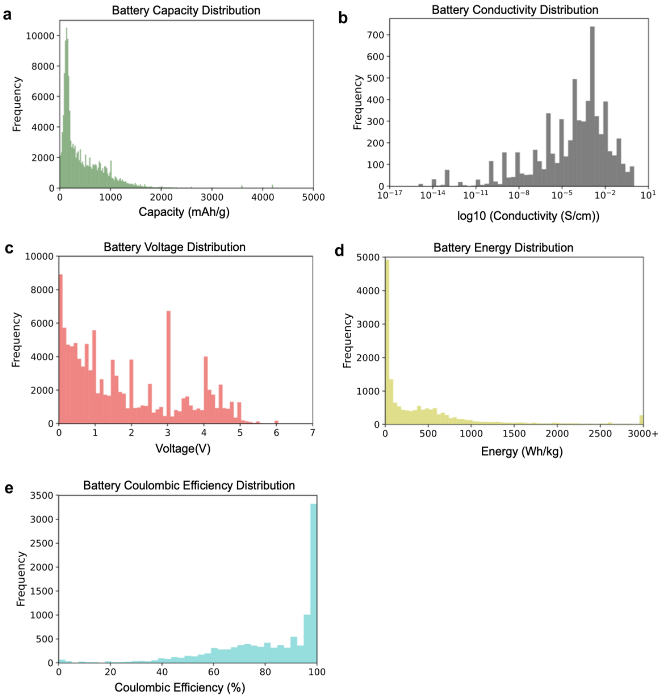


Fig. 2. Property value distribution

identified. To enhance data reliability, future analyses can focus on records with correctness values labeled as either ’none’ or ’true’ 3.

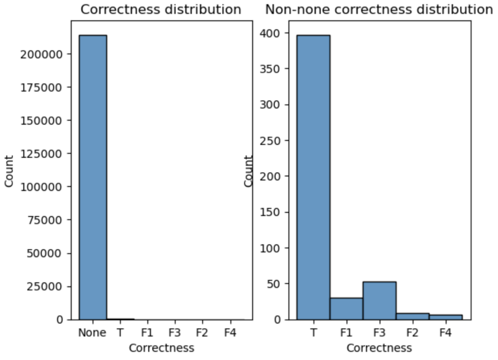


Fig. 3. Correctness Distribution

To further refine the dataset, we examined the ’warning’ field, which marks questionable values. The database authors introduced warning flags—L, R, and S—to identify values that may be extreme, potentially irrelevant to battery materials, or part of a data series. Removing flagged data would improve the database’s overall precision to approximately 85%, with gains of 4.2% and 0.8% for the L and R flags, respectively.

Additionally, we performed an outlier analysis to identify extreme values within the dataset. A boxplot and violin visualization were used to assess value distributions for key properties 4.

Outliers were detected using the 1.5 × IQR criterion:

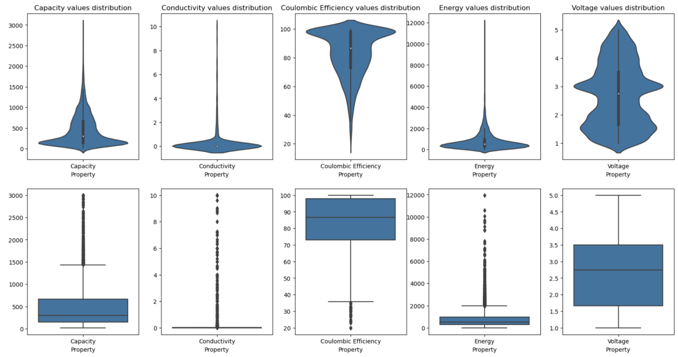


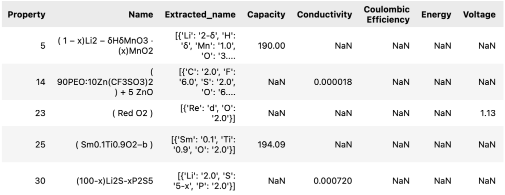
Fig. 4. Value Distribution Boxplots

- Capacity: 413 outliers
- Conductivity: 367 outliers
- Coulombic Efficiency: 39 outliers
- Energy: 682 outliers
- Voltage: 0 outliers

Total Outlier Rows: 1,428

After detecting and removing these outliers, the dataset was refined for subsequent modeling and analysis.

The original dataset contained separate records for each chemical-property combination. To prepare the data for modeling, we performed a pivot operation, consolidating multiple property records into a single entry per chemical compound. This process resulted in a refined dataset containing 3027 unique chemical records, each with a complete set of associated properties. The structure of the dataset at this point is presented in the figure 5.



Property	Name	Extracted_name	Capacity	Conductivity	Coulombic Efficiency	Energy	Voltage
5	(1 − x)Li2 − δHδMnO3 · (x)MnO2	[{'Li': '2-δ', 'H': 'δ', 'Mn': '1.0', 'O': '3....	190.00	NaN	NaN	NaN	NaN
14	(90PEO:10Zn(CF3SO3)2) + 5 ZnO	[{'C': '2.0', 'F': '6.0', 'S': '2.0', 'O': '6....	NaN	0.000018	NaN	NaN	NaN
23	(Red O2)	[{'Re': 'd', 'O': '2.0'}]	NaN	NaN	NaN	NaN	1.13
25	(Sm0.1Ti0.9O2−b)	[{'Sm': '0.1', 'Ti': '0.9', 'O': '2.0'}]	194.09	NaN	NaN	NaN	NaN
30	(100-x)Li2S-xP2S5	[{'Li': '2.0', 'S': '5-x', 'P': '2.0'}]	NaN	0.000720	NaN	NaN	NaN

Fig. 5. Final dataset structure after preprocessing

We further enriched the dataset by decomposing each chemical formula to calculate additional features representing the fractional composition of chemical elements within each compound. For instance, for the compound $(BiO)_4CO_3(OH)_2$, the elemental fractions calculated are Bi_fraction = 4, C_fraction = 1, O_fraction = 7, Li_fraction = 0, etc. This feature engineering step was implemented using the Mendeleev Python library.

Finally, we conducted an analysis of null value proportions in the finalized dataset 6.

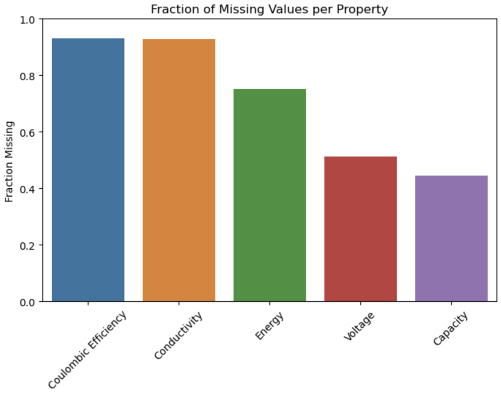


Fig. 6. Null values distribution

IV. Method

A. Algorithm Design and Implementation

Explain the data mining/machine learning algorithms used and how they were implemented.

1) Decision Tree: The Decision Tree algorithm was applied to predict five different battery material properties: Capacity, Conductivity, Coulombic Efficiency, Energy, and Voltage. Decision Trees work by recursively splitting the dataset based on the most significant feature at each node, creating a hierarchy of decision rules that optimize for minimal prediction error.

Each Decision Tree model was trained using the scikit-learn library in Python. The dataset was split into an 80% training set and a 20% test set. The models were tuned by adjusting hyperparameters such as the maximum tree depth and minimum samples per leaf to prevent overfitting.

2) Random Forest Regressor: The Random Forest model effectively captures complex non-linear interactions between variables, making it particularly suitable for predicting battery material properties, which are often characterized by intricate relationships. Additionally, Random Forest provides valuable insights by ranking the importance of each feature, thereby identifying the most influential chemical elements and compound attributes contributing to battery performance. Its inherent robustness to overfitting and ability to handle high-dimensional data further enhance its applicability to this task.

3) Fully Connected Neural Net: A neural network architecture was designed as a sequential model with multiple dense layers to learn the relationships between material compositions and battery properties. The architecture underwent several iterations, with the initial version containing three hidden layers [128, 64, 32] neurons and an enhanced version with two hidden layers [64, 32] to reduce overfitting. All hidden layers utilized ReLU activation functions with linear activation for the output regression layer. To address overfitting concerns, we implemented comprehensive regularization: batch normalization after each hidden layer, dropout (0.3 after the first layer, 0.2 after subsequent layers), L2 regularization (coefficient 0.001) on all hidden layers, and early stopping monitoring validation loss with a patience of 30 epochs. Optimization strategy employed the Adam optimizer with exponential learning rate decay, starting at 0.001 with a decay rate of 0.9 every 1000 steps. Mean Squared Error as the loss function, batch size of 32 samples, and 20% validation split. Architecture parameters were tailored for each target variable based on data characteristics:

```
architectures = {
’Capacity’: {
’layers’: [128, 64, 32],
’dropout’: [0.2, 0.2],
’l2’: 0.0005
},
’Energy’: {
’layers’: [64, 32],
’dropout’: [0.3],
’l2’: 0.001
},
’Voltage’: {
’layers’: [128, 64, 32],
’dropout’: [0.2, 0.2],
’l2’: 0.0005
}
}
```

V. Experiment and Discussion

A. Experimentation

1) Decision Tree: Using our algorithm design and implementation, we successfully generated 5 decision trees, one for each target. To assess model performance of the Decision Tree models, we analyzed the prediction performance on the test dataset and evaluated Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R² score. The results are summarized in Table I, which presents MAE, MSE, and R² scores for each property.

TABLE I
Decision Tree Model Performance for Battery Properties

Target	**MAE**	**MSE**	**R² Score**
Capacity	135.38	43342.03	0.309
Conductivity	0.0049	0.000061	-0.600
Coulombic Efficiency	11.64	237.42	-0.484
Energy	286.43	157947.60	-0.102
Voltage	0.7496	0.8318	-0.158

Among the five trained models, only the Capacity prediction model achieved a positive R² score, indicating a meaningful correlation between input features and the target variable. The complete Decision Tree structure for Capacity prediction is illustrated in Figure 7.

The decision tree model for capacity demonstrated the best performance, achieving an R² score of 0.309, indicating a moderate correlation between the input features and the target variable. The remaining four models exhibited negative R² scores, suggesting that they performed worse than a simple mean-based prediction.

Regardless, the MAE and MSE for all five targets are relatively high in proportion to their ranges throughout the

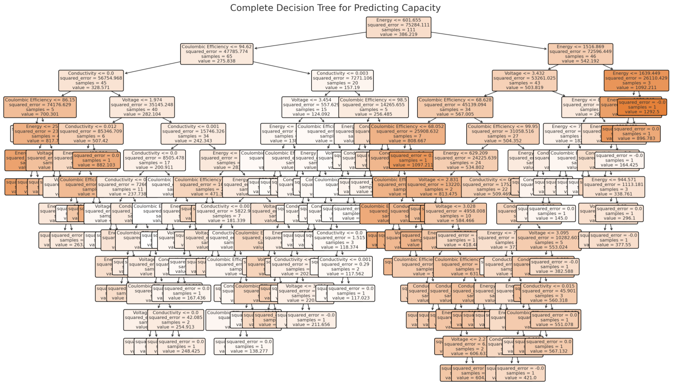
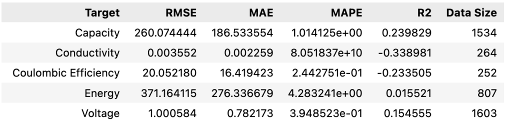

Complete Decision Tree for Predicting Capacity
Energy <= 601.655
squared_error = 75284.111
samples = 111
value = 386.219

Fig. 7. Complete Decision Tree for Predicting Capacity

dataset, indicating that the models struggle to capture accurate values to be applied towards our data.

Feature importance analysis revealed that Columbic Efficiency, Conductivity, Energy, and Voltage were the most significant predictors for Capacity. The Decision Tree visualization in Figure 7 provides a detailed breakdown of the model's decision-making process.

2) Random Forest Regressor: For each battery property, the dataset entries containing non-null values were divided into training and testing subsets with an 80-20 split. The scikit-learn Random Forest regressor was trained with 1000 decision trees, and the following evaluation metrics were computed on the test set: Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The results of these evaluations are summarized in the accompanying table 8.



Target	RMSE	MAE	MAPE	R2	Data Size
Capacity	260.074444	186.533554	1.014125e+00	0.239829	1534
Conductivity	0.003552	0.002259	8.051837e+10	-0.338981	264
Coulombic Efficiency	20.052180	16.419423	2.442751e-01	-0.233505	252
Energy	371.164115	276.336679	4.283241e+00	0.015521	807
Voltage	1.000584	0.782173	3.948523e-01	0.154555	1603

Fig. 8. Random Forest Results

Furthermore, feature importance scores from the Random Forest model were analyzed to identify the most influential variables 9. Notably, elemental fractions such as oxygen and lithium emerged as significant predictors for battery property estimation.

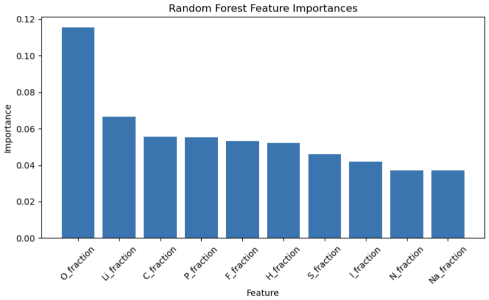


Fig. 9. Random Forest Feature Importance

3) Fully Connected Neural Net: Neural network models were initially aimed to train five targets: Capacity, Energy, and Voltage, Columbic efficiency, and Conductivity, but due to a lack of enough data after processing and cleaning, Columbic efficiency and conductivity were removed. The model trained separately for each target property (Capacity, Energy, and Voltage) using standardized inputs and outputs. Conducted two experimental rounds: initial models using elemental compositions with basic engineered features, and enhanced models incorporating chemistry-informed features like electronegativity differences and element interactions. Each model trained for up to 300 epochs with early stopping to prevent overfitting. The performance varied across target properties, with modest improvements from the enhanced models: Performance comparison of models across target properties:

Target	Initial		Enhanced		ΔR^2
	R^2	RMSE	R^2	RMSE	
Capacity	0.003	297.8	0.091	284.4	+0.088
Energy	-0.124	396.5	-0.107	393.5	+0.017
Voltage	0.130	1.015	0.194	0.977	+0.064

Feature importance analysis revealed that neural networks relied on specific features: for Voltage prediction, Li_fraction (0.458), F_fraction (0.351), and electronegativity difference (0.278); for Capacity prediction, Si_fraction (0.600), Na_fraction (0.586), and max_electronegativity (-0.317); for Energy prediction, Ti_fraction (0.299) and Na_fraction (0.255).

B. Discussion

The models faced significant challenges:

- Severe data sparsity (47-92% missing values)
- Limited samples (211-1282 depending on target)
- Weak input-output correlations
- High dimensionality (over 100 features)

Despite optimizations and feature engineering, these limitations restricted predictive performance, highlighting the need for larger, more complete datasets for effective battery material property prediction.

VI. Conclusion

In conclusion, this study demonstrates the possibility of employing advanced data mining and machine learning techniques (Decision Tree, Random Forest, and Fully Connected Neural Networks) to predict battery material properties such as Capacity, Energy, and Voltage based on chemical composition. The analysis highlights elemental composition and chemistry-informed features as crucial predictors of battery performance. Despite challenges such as data sparsity, limited sample size, and high dimensionality, the models provided moderate predictive power, particularly in capacity and voltage estimation. Future work should focus on expanding the dataset size and completeness, incorporating additional chemical insights, and exploring more sophisticated modeling approaches to enhance prediction accuracy and facilitate the discovery of innovative battery materials.

References

[1] J. Huang S., Cole, “A database of battery materials auto-generated using chemdataextractor,” *Scientific Data*, vol. 7, no. 260, 2020.