Problem Statement:

The goal is to develop models or algorithms capable of estimating the State of Health (SoH) of batteries based on historical usage data. Solutions that require minimal data for accurate SoH estimation or that perform well under varying environmental conditions will be given additional consideration.

Objectives:

1. Develop machine learning or data-driven models to accurately estimate battery SoH.
2. Implement physics-based models for SoH estimation.
3. Predict the Remaining Useful Life (RUL) of the battery.
4. Design user-friendly visualization tools for real-time SoH monitoring.

Introduction to Battery State of Health (SoH) Prediction:

As the globe moves toward renewable energy and electrification of transportation, battery reliability and longevity have become essential considerations. Batteries, particularly lithium-ion (Li-ion) batteries, are commonly used in electric vehicles (EVs), energy storage systems (ESS), consumer electronics, and other applications because to their high energy density, efficiency, and extended cycle life. However, batteries, like all energy storage devices, degrade over time, resulting in reduced capacity, power output, and overall performance. This degradation has a direct impact on the battery's State of Health (SoH), a crucial indicator used to assess its ability to function as expected in comparison to its initial condition.

What is State of Health (SoH)?

SoH is a measure of a battery's health compared to its initial state when it was brand new. It is typically expressed as a percentage and reflects the battery's remaining capacity and ability to function efficiently.

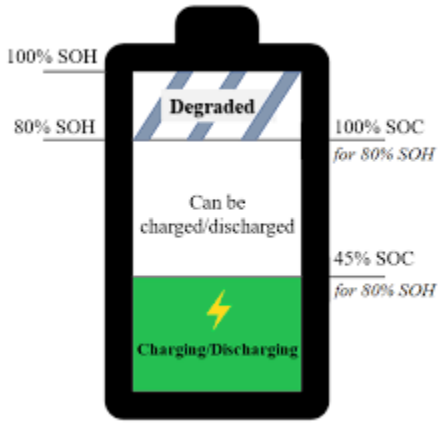
- **SoH = 100%**: The battery is in perfect health, delivering full rated capacity and power.
- **SoH < 100%**: The battery has degraded, and its capacity or performance is lower than its original specification.
- **SoH < 80%**: In many applications, this is considered the threshold at which a battery is no longer suitable for normal use (e.g., in EVs, the vehicle range may become unacceptable).

Mathematically, SoH can be defined as:

SoH = Current Capacity/Nominal Capacity × 100

Where:

- **Current Capacity** is the amount of charge the battery can hold in its degraded state.
- **Nominal Capacity** is the rated capacity when the battery was new.



What Makes SoH Prediction Vital?

For battery-powered devices to operate in a safe, effective, and cost-effective manner, accurate SoH prediction is essential. The following are some implications of knowing SoH:

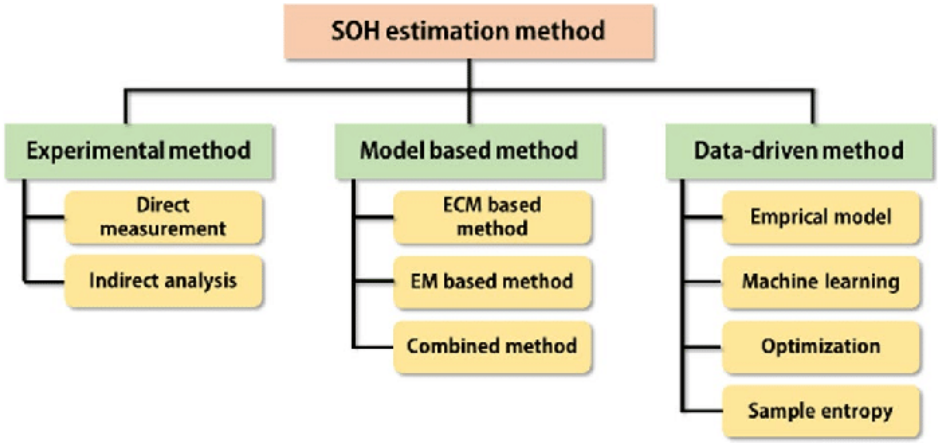
1. **Safety:** As batteries age, they may be more vulnerable to catastrophic failures such thermal runaway, short circuits, and overheating. By alerting users when a battery is getting close to dangerous circumstances, monitoring SoH helps avoid these kinds of failures.

2. **Performance Optimization:** A high degree of battery performance is needed for many applications. Users can maintain optimal performance, particularly in energy-demanding industries like electric automobiles, by adjusting operations based on the SoH.

3. **Extending Battery Life:** To extend the battery's lifespan, proper temperature regulation, charge/discharge cycles, and usage patterns can all be implemented with the aid of SoH monitoring.

4. **Maintenance and Replacement**: Predicting SoH can help plan for preventive maintenance or battery replacement, reducing unexpected downtime and repair costs. **End-of-Life (EoL) Prediction**: SoH is closely tied to the prediction of the battery's **Remaining Useful Life (RUL)**. As the SoH drops, RUL predictions become essential for estimating how long the battery will continue to provide acceptable performance before needing replacement.

To address the challenge of predicting battery State of Health (SoH), various approaches can be employed, each offering distinct advantages and challenges. Here's a **summary of the main methods:**



1. Electrochemical Models

These models simulate the internal electrochemical behavior of batteries, including processes such as ion movement, charge transfer, and chemical reactions.

- **Pros**: They offer **high accuracy** and a **clear interpretation** of battery performance since they are grounded in the fundamental chemistry of the system.
- **Cons**: Implementation is complex, and the models require significant **computational resources** and **detailed knowledge** of battery materials.
- **Example**: Pseudo-two-dimensional (P2D) model.

2. Empirical Models

Empirical approaches are based on experimental data and derive relationships between observable battery characteristics, such as voltage and capacity, to estimate SoH.

- **Pros**: **Simple**, **quick** to implement, and largely **data-driven**.
- **Cons**: May **not be as accurate under varying conditions** and often **lacks general applicability** across different battery types.
- **Example**: Capacity fade models that correlate cycle count or voltage trends with SoH.

3. Machine Learning Models

Machine learning techniques leverage extensive datasets, which may include variables like voltage, current, and temperature, to capture complex, non-linear relationships for SoH estimation.

- **Pros**: Capable of **managing complex scenarios** and large datasets, making them **well-suited for real-time applications**.
- **Cons**: They often **require large amounts of labeled data** for training and can be **less interpretable** than traditional models.
- **Examples**: Neural networks like CNNs, RNNs, and Transformer models.

4. Statistical Data-Driven Models

Statistical models use techniques to find relationships between input parameters and SoH through fitting methods.

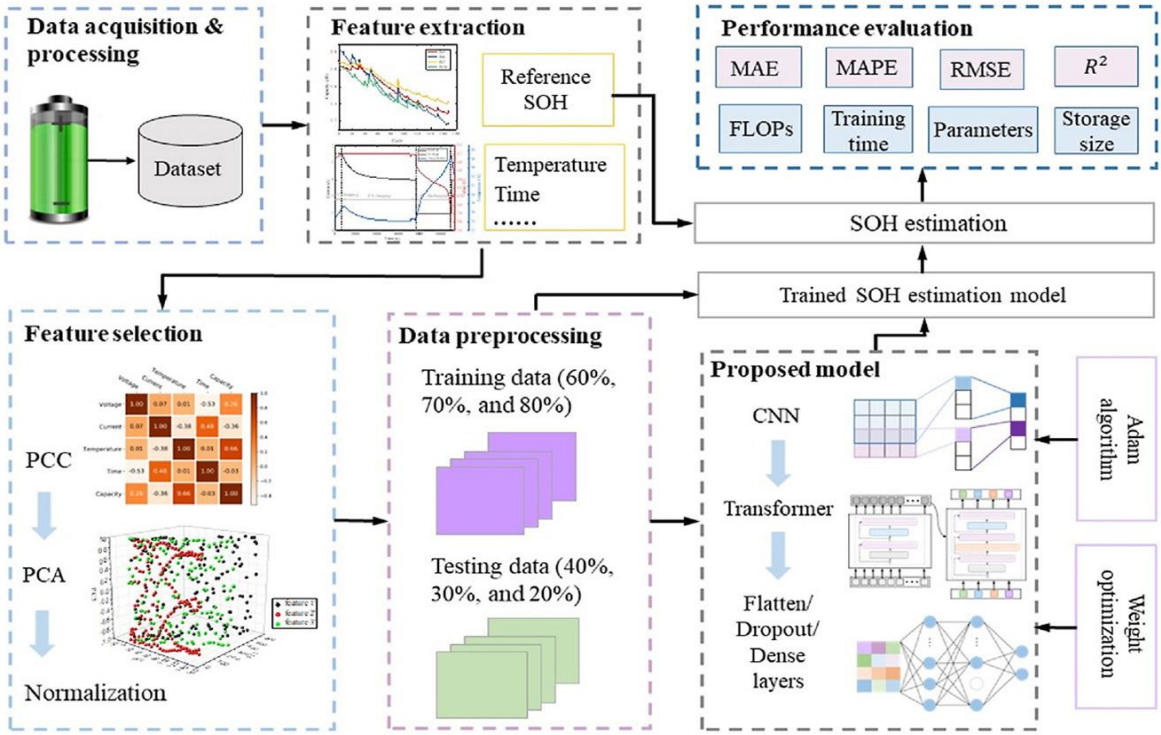
- **Pros**: These methods are **less computationally demanding** and offer more **straightforward interpretation**.
- **Cons**: They may **lack the precision** of more advanced machine learning or physics-based models.
- **Example**: Kalman filters or regression-based models.

5. Hybrid Models

These combine elements from physics-based electrochemical models with data-driven techniques like machine learning, merging the interpretability of electrochemical processes with the adaptability and scalability of data-driven approaches.

- **Pros**: **High accuracy**, **better generalization**, and **balance between interpretability and real-world adaptability**.
- **Cons**: **More complex** to develop and **require careful integration** of different methodologies.
- **Example**: Hybrid models combining electrochemical principles with machine learning algorithms like deep learning, like **PINNs** (Physics-Informed Neural Networks)

Brief Overview of Our Approach:



The objective of this project is to predict the **State of Health (SoH)** of lithium-ion batteries by leveraging a combination of traditional feature extraction and machine learning models. The process begins with data acquisition from the **NASA Battery Dataset**, followed by feature extraction from current and voltage curves. We apply **Pearson correlation** analysis to select the most relevant features for SoH prediction. Finally, a machine learning pipeline using a **CNN-Transformer model** is employed, with the **convolutional neural network (CNN) capturing localized patterns** in the data and the **transformer model generalizing across different batteries**. This combination helps achieve accurate and robust SoH estimation.

Data Engineering Pipeline:

Data Acquisition and Preprocessing

The data is sourced from the **NASA Ames Prognostics Center of Excellence** battery dataset. For this study, all battery datasets(except 3rd and 5th) from the NASA Battery dataset are utilized. The dataset contains variables like **voltage, current, temperature, capacity**, and others over multiple cycles.

- **Preprocessing**: The raw data was stored in .mat format, which required conversion into structured **CSV** files for ease of handling. This conversion process ensures that we have accessible and labeled datasets for further processing.
- Our Code notebook details the code for reading and restructuring the .mat data, allowing consistent formatting across all the battery datasets.

Feature Extraction:

Feature extraction is crucial to capture the behavior of the battery as it degrades over time. From the provided dataset, we extract several features based on current and voltage curves. These features are:

1. Mean Voltage and Current

The average voltage and current during charge and discharge cycles serve as key indicators of a battery's performance. A **declining mean voltage** typically signals **capacity fade**, often linked to **side reactions** like **electrolyte decomposition** or **lithium plating**. Similarly, the **average current** during discharge reflects the **battery's load capacity**; any fluctuations may indicate issues within the battery's internal structure or rising resistance.

2. Standard Deviation

Standard deviation measures the variability of voltage and current readings over time. **High variability** in voltage can suggest **instability in electrochemical** reactions or **significant changes** in **internal resistance**. This could stem from separator degradation, electrode fatigue, or shifts in electrolyte composition. Tracking **standard deviation** helps gauge **operational consistency**, with increased instability often pointing to deteriorating health.

3. Kurtosis

Kurtosis assesses the "**tailedness**" of voltage and current distributions. A **high kurtosis** value indicates the **presence of extreme values** in the dataset, which may reflect **abnormal operating conditions or transient events**, such as **voltage spikes** during high loads. This metric is particularly useful for **identifying potential safety issues**, such as **thermal runaway or short circuits**, thus providing insights into **degradation patterns**.

4. Skewness

Skewness evaluates the **asymmetry** of the voltage and current distributions. A **positively skewed voltage distribution** often signifies a prevalence of low voltage readings, hinting at **lithium-ion depletion or increased internal resistance**. Conversely, a **negatively skewed current distribution** may indicate **frequent high-current pulses**, potentially **accelerating aging processes like lithium plating**. Understanding skewness helps **detect imbalances in charge and discharge cycles** that impact battery performance.

5. Charging Time

Charging time directly reflects **how efficiently a battery can accept energy**. As batteries age, increased internal resistance—caused by **solid-electrolyte interphase (SEI)** growth, **electrode degradation, or electrolyte depletion**—can lead to **longer charging times**. An increase in charging time across cycles signals declining efficiency and warrants a closer look at the battery's internal conditions.

6. Accumulated Charge

Accumulated charge represents the **total energy cycled** through the battery during its lifespan. This metric is vital for evaluating **effective capacity utilization**. Discrepancies between accumulated charge and expected performance can indicate internal inefficiencies, such as **heightened resistance** or **irreversible loss of active material**. Monitoring this feature contextualizes capacity loss and energy delivery issues over time.

7. Curve Slope

The slope of the voltage-current curve during charge and discharge reflects the **battery's responsiveness to changing current demands**. A **steeper slope** indicates **lower internal resistance**, facilitating efficient energy transfer. Over time, a **flattening slope** suggests **increased resistance** due to factors like **electrode passivation** or **electrolyte degradation**. Analyzing the slope provides insights into the battery's capability to manage dynamic load changes, which is crucial for applications like electric vehicles.

8. Curve Entropy

Curve entropy **quantifies the disorder in the voltage and current signals**, shedding light on the system's complexity. **Higher entropy** values often indicate **greater irregularity in charge and discharge cycles**, which may be associated with **unpredictable degradation behaviors**. A significant **increase in entropy**, particularly in voltage signals, can signify **declining electrochemical stability or heightened thermal effects**. Monitoring entropy thus serves as a diagnostic tool for identifying subtle shifts in battery health.

All the features are extracted by taking one whole cycle data at once, so a single datapoint of feature is computed from one whole cycle data

Below are the features extracted from all battery datasets:

	cycle	mean_voltage	mean_current	std_voltage	std_current	...	slope_voltage	slope_current	entropy_voltage	entropy_current	SoH
0	1	3.556946	-1.990533	0.226595	0.202011	...	-0.000193	-0.000033	-4.039001	-39.336811	100.000000
1	2	3.561476	-1.990278	0.232008	0.202426	...	-0.000195	-0.000033	-4.009105	-39.367648	99.498985
2	3	3.566752	-1.989947	0.224124	0.202994	...	-0.000192	-0.000033	-4.215857	-39.388808	98.918547
3	4	3.568795	-1.989601	0.221551	0.203501	...	-0.000192	-0.000033	-4.427441	-39.368330	98.916498
4	5	3.563971	-1.989548	0.235643	0.203319	...	-0.000199	-0.000033	-3.662594	-39.410114	98.289755
...	...	...	...	...	...	...	...	...	...	...	...
1394	52	3.065651	-1.991047	0.423930	0.208020	...	-0.000596	-0.000054	5.081981	-39.320559	93.636929
1395	53	3.087680	-1.990491	0.409291	0.210270	...	-0.000585	-0.000056	5.600099	-39.258813	97.899449
1396	54	3.080080	-1.990194	0.411531	0.210894	...	-0.000586	-0.000056	5.591191	-39.303709	97.946004
1397	55	3.063813	-1.990608	0.423972	0.209126	...	-0.000599	-0.000055	5.326688	-39.262819	94.493922
1398	56	4.060790	-0.671836	0.224683	1.160515	...	-0.017762	-0.091747	-285.798364	-30.836173	0.000000

1399 rows × 16 columns

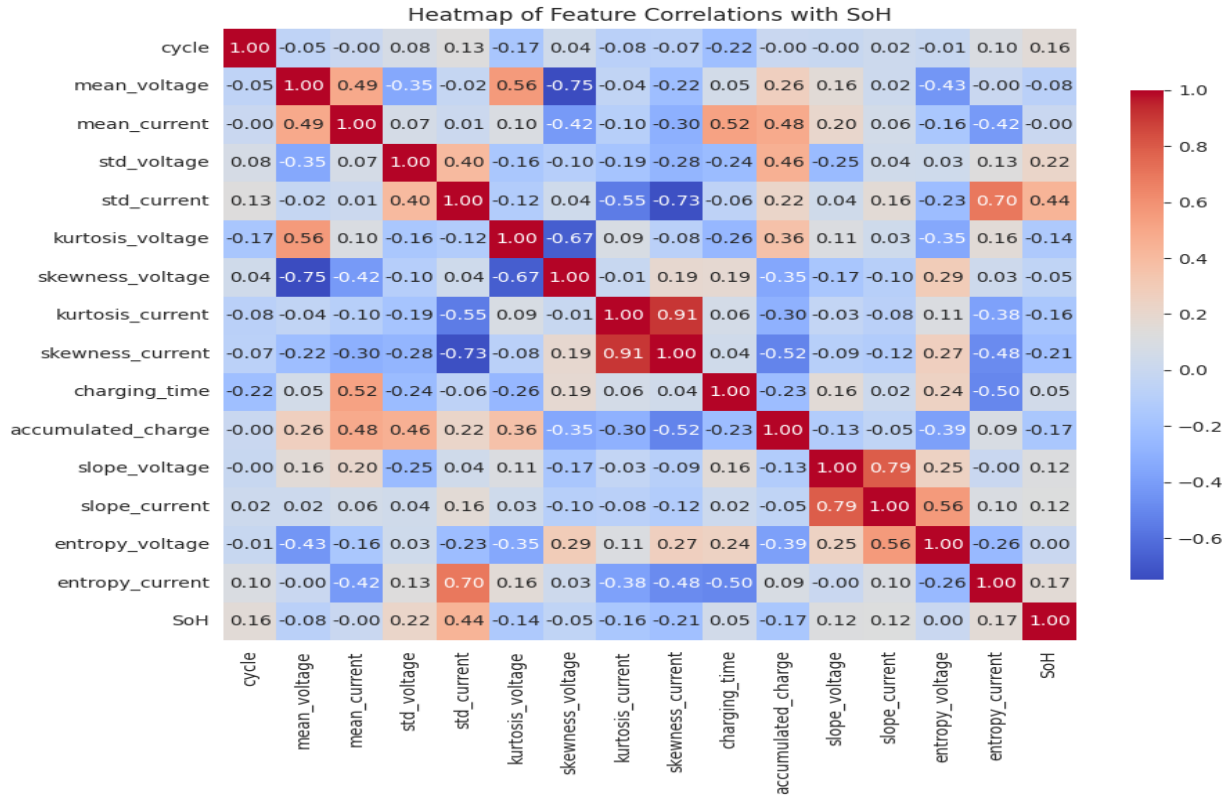
Feature Selection:

After feature extraction, **Pearson Correlation Coefficient (PCC)** is used to select the most relevant features correlated with SoH. PCC measures the linear correlation between two variables and is given by:

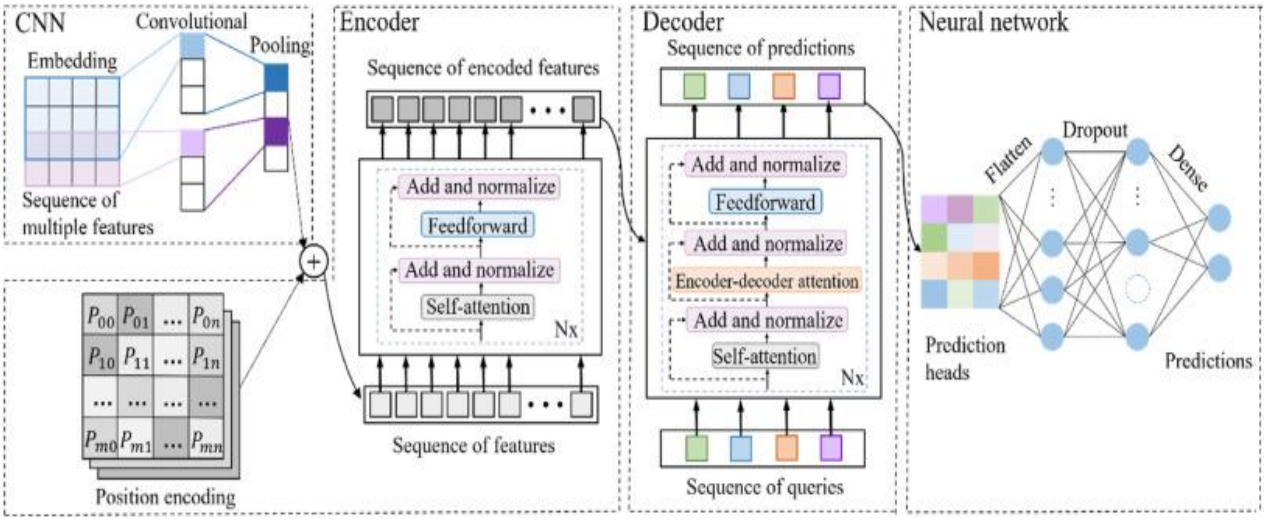
$$PCC = n(\sum xy) - (\sum x)(\sum y) / [n\sum x2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]$$

- **Significance**: Features with higher correlation to SoH are selected for the machine learning model. This reduces dimensionality and removes irrelevant data, improving the model's performance.

```
Pearson Correlation Coefficients with SoH (Sorted):
std_current           0.436634
std_voltage           0.215422
entropy_current       0.165371
cycle                 0.157552
slope_current         0.123725
slope_voltage         0.117525
charging_time         0.047937
entropy_voltage       0.000579
mean_current         -0.001670
skewness_voltage     -0.050192
mean_voltage         -0.079309
kurtosis_voltage     -0.143431
kurtosis_current     -0.160655
accumulated_charge   -0.166153
skewness_current     -0.211226
```



Model: CNN-Transformer:



The selected features are then used to train a **CNN-Transformer model**, a hybrid architecture combining convolutional layers and transformer blocks. The CNN is adept at capturing **local features** from the input data, while the transformer helps model **long-range dependencies** across the sequence of battery cycles.

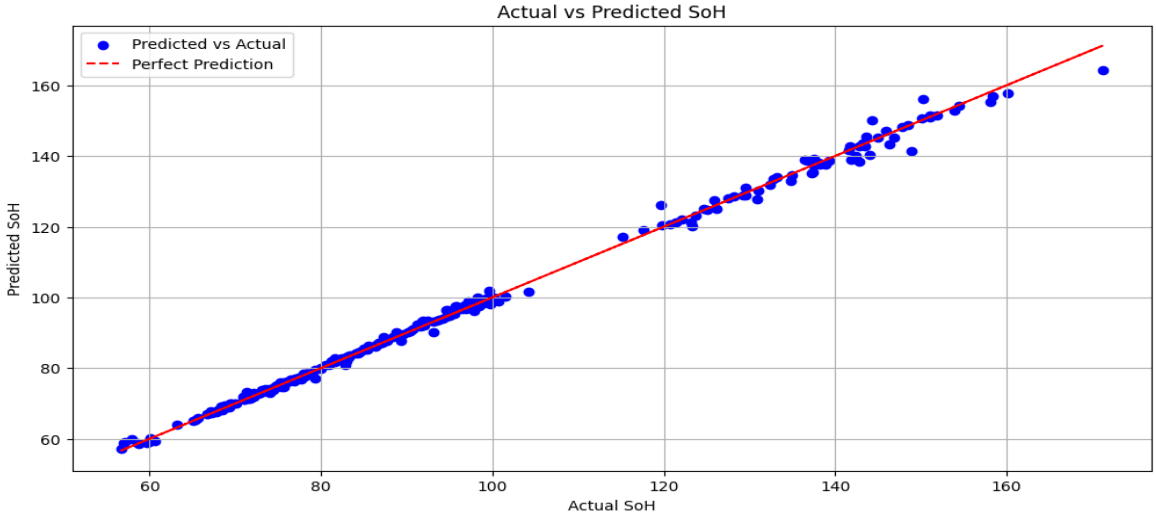
- **CNN Layers**: The CNN component extracts hierarchical local features from the time-series data, identifying localized patterns in voltage, current, and temperature variations over time.
- **Transformer Layers**: The transformer's attention mechanism captures the global relationships between different data points, enhancing the model's ability to generalize across different batteries.
    - It uses **positional encodings** to retain the temporal information of the battery cycles.

According to the research, this architecture offers the advantages of both CNN's locality-preserving properties and transformer's ability to model long-term dependencies efficiently.

This model is trained on the combined dataset of the batteries and validated using metrics such as **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE), Root Mean Square Error (RMSE)**, and **R-squared (R^2)** to assess prediction accuracy.

Final Model:

In our final model, we have performed the same steps as mentioned above and also crafted the **CNN-Transformers model**. In Addition, just for simplicity and **Comparison** purposes, we have also used a rather simple and computationally light model, the **ExtraTreesRegressor**. We combined all the data from all the batteries, **normalized** the features using **MinMaxScaler**, then performed **PCC Analysis**, took the **top 7 features** for further training, split the data into training and testing, then **fitted** the model on the training set, and at the end **tested** on the test set. Here are the results for the prediction of **ExtraTreesRegressor**:



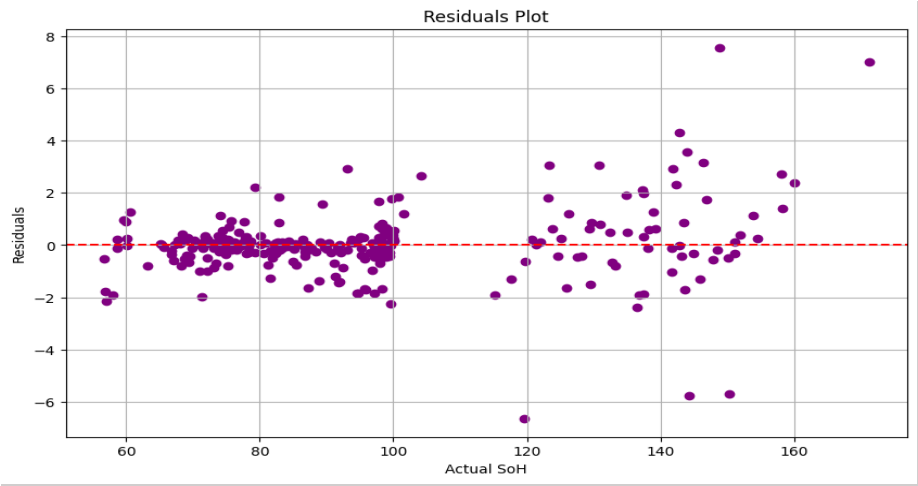
Results (ExtraTreesRegressor):

- **Mean Absolute Error (MAE): 0.74**
- **Mean Squared Error (MSE): 1.73**
- **Root Mean Squared Error (RMSE): 1.32**
- **R-squared (R^2): 1.00**
- **Accuracy Score: 0.9928433504713539**

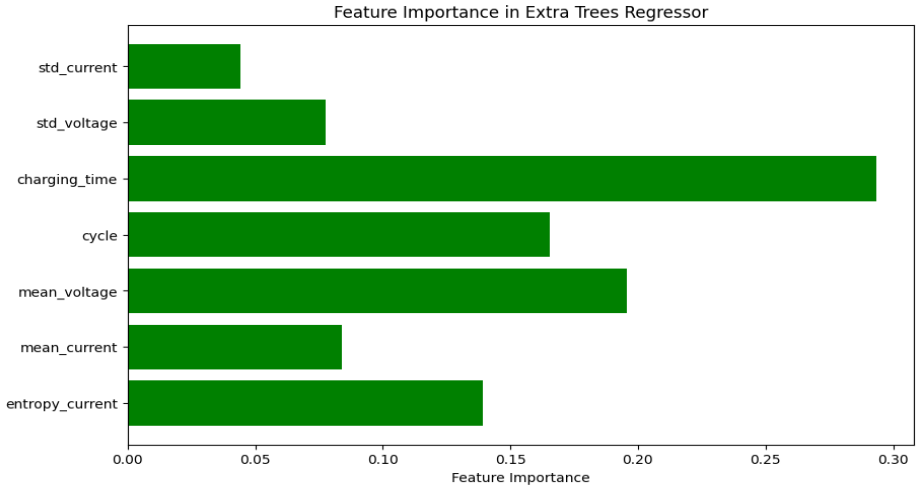
Overall, the results were pretty good, suggesting that the feature extraction was useful and relevant features following the physics-chemistry based principles of the working of a battery were extracted. Also, the overall pipeline was perfectly implemented to get impressive results.

More plots for visualization:

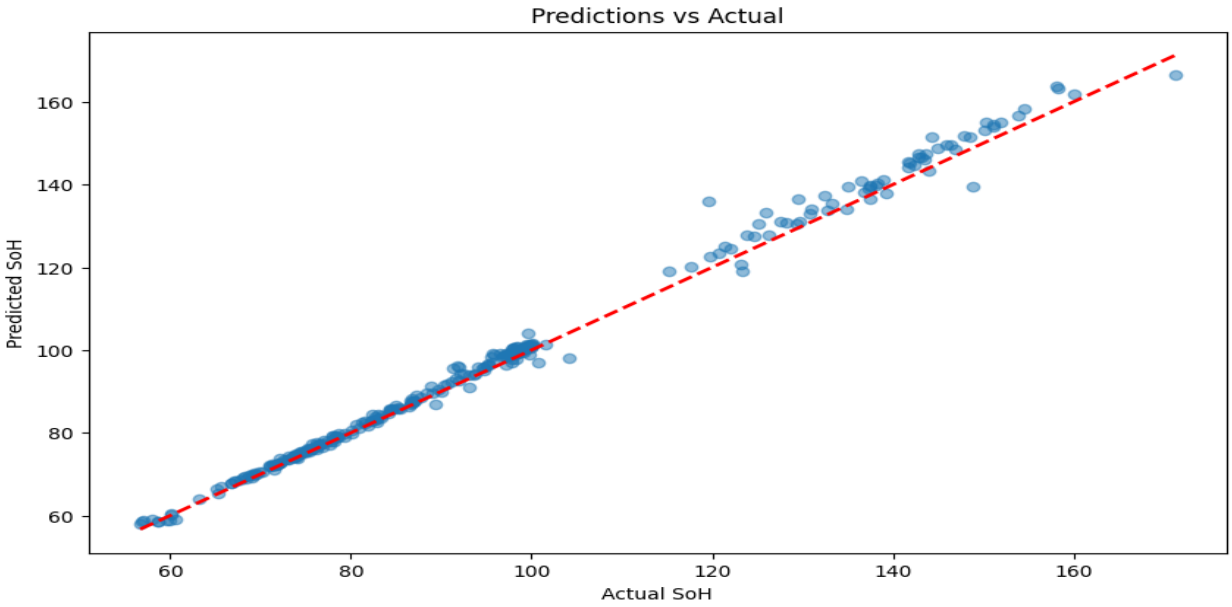
Residual Error Visualization:



Feature Importance Visualization:



Here are the results for the prediction of **CNN-Transformers model**:

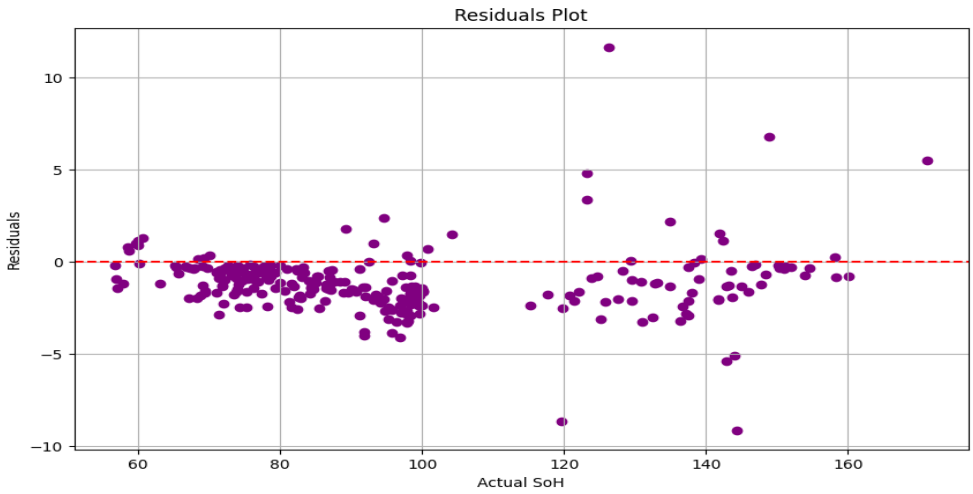


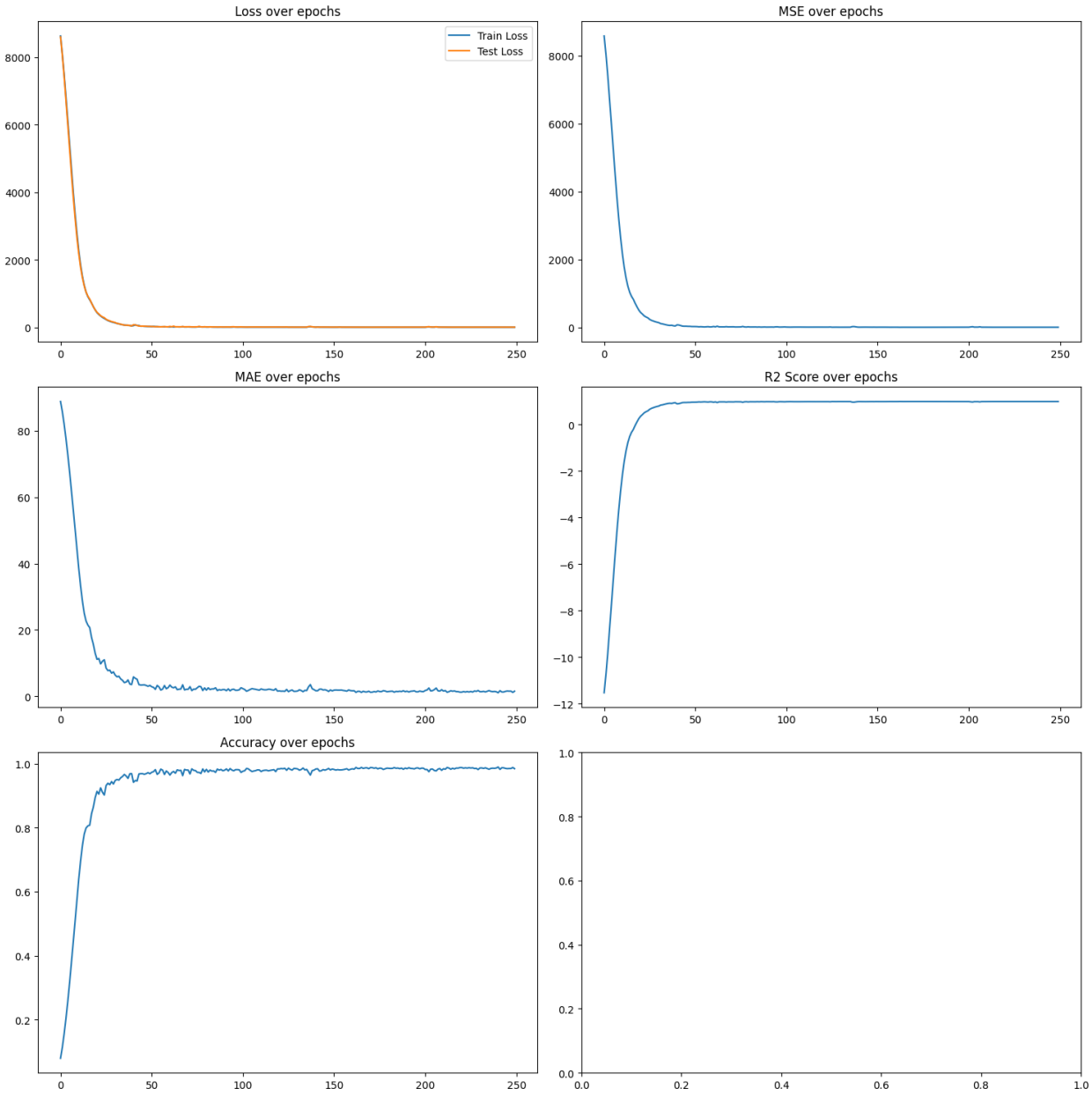
Results (CNN-Transformers model):

- **MSE: 4.1651**
- **MAE: 1.5177**
- **Root Mean Squared Error (RMSE): 2.04**
- **R2: 0.9939**
- **Accuracy: 0.9843**

More plots for visualization:

Residual Error Visualization:





The final results on the testing set containing a variety of different types data from various datasets was really good using the CNN-Transformers. But it the evaluation metrics of the ExtraTreesRegressor was a slightly better than the CNN-transformers with an **R2 of 1.00** which may show that the **extra trees model may be showing overfitting** on the type of data it was trained on, whereas the CNN-transformers model with an **R2 of 0.9939** is showing **good generalization capability**, with **potential scalability** on new unseen datasets.

Further Improvements:

First is we can use **lesser amount of data but take data from each type of battery dataset**, enable the model to **learn diverse battery datasets**, with **limited data** to **save resources**.

Second thing is we can **decrease the amount of features** we are using **by employing PCA**, so that we can **retain maximum feature variance** and hence **maximum feature information**, with **minimizing the number of features used and thus the computational cost**, such that any model ranging from simple random forest or ExtraTrees to complex CNN-Transformer model can learn their best with minimal computational cost and **maximizing accuracy, improving feasibility and real-world scaling so that we can have reasonable real world applications.**

References:

The overall understanding, idea and approach to the problem was taken from the following research papers:

Battery State of Health Estimate Strategies: From Data Analysis to End-Cloud Collaborative Framework (mdpi.com) → (this explains all possible methods and approaches to the problem and discusses the results of few of the methods)

A novel state-of-health estimation for the lithium-ion battery using a convolutional neural network and transformer model - ScienceDirect → (This research paper details the implementation of a CNN-Transformer architecture on the NASA battery dataset and discusses the results in-depth)

https://www.nature.com/articles/s41467-024-48779-z → (from this research paper we got the ideas for what kind of features to be extracted from the dataset for maximum relevance and importance with our problem statement of SoH prediction)

https://www.kaggle.com/code/rajeevsharma993/battery-health-nasa-dataset → (from this Kaggle notebook we got the code snippet for setting up our data pipeline for structuring the .mat files into usable csv files (the '**load_data**' function in the code))