Project Title

Battery Health Analytics - Predicting the Lifecycle and Health of Li-ion Batteries

Presented By:
Jai Kumar Gupta

Project: Battery Health Analytics
Date: June 27, 2025

Project Plan: Battery Health Analytics

This document outlines the updated status of the project, reflecting the completion of all data processing and modeling tasks

Phase 1: Data Processing (✅ Complete)

- **Task:** Loaded raw time-series data from 35 CS2_8 text files, parsed the data format, and consolidated the information into a usable structure.

Phase 2: Feature Engineering (✅ Complete)

- **Task:** Created a rich feature set to capture the physical degradation of the battery, including average/max temperature, average/min voltage, and cycle duration. Created the RUL and Health_State target variables.

Phase 3: Model Development & Evaluation (✅ Complete)

- **Step 5: Regression Task - Predicting RUL (✅ Complete)**
  - **Action:** Trained, tuned, and evaluated five different regression models. Performed an in-depth analysis of their performance and feature importances.
- **Step 6: Classification Task - Predicting Health State (✅ Complete)**
  - **Action:** Trained, tuned, and evaluated five different classification models. Analyzed their performance using classification reports, confusion matrices, and ROC/AUC curves.
- **Step 7: Compare Models and Summarize (✅ Complete)**
  - **Action:** Consolidated all results into a final report, comparing the performance of all models across both tasks and providing detailed insights.

- **Step 8: Finalize Project Report and Presentation (✅ Complete)**
-
  - **Goal:** To use the completed analyses and summaries to prepare the final deliverables for submission.
  - **Action Items:**
    1. Review and refine the **Final Project Report** document we've created to ensure it is clear, concise, and meets all submission guidelines.
    2. Use the **Presentation Slides** we've created as a basis for the final project presentation, practicing the delivery and ensuring all key findings are communicated effectively.
    3. Ensure the final Jupyter Notebook is clean, well-commented, and easily reproducible.

Project Overview

- **Objective:**
  - To predict the **Remaining Useful Life (RUL)** of a CS2_8 battery using regression models.
  - To classify the real-time **health state** ('Healthy', 'Moderate') of the battery using classification models.
- **Dataset:**
  - CALCE CS2_8 Battery Dataset.
  - Consists of raw time-series data from 35 charge-discharge cycles.
- **Approach:**
  - **Data Processing:** Parse and clean raw data from individual cycle files.
  - **Feature Engineering:** Create meaningful physical features (temperature, voltage, etc.).
  - **Modeling:** Train, tune, and evaluate a suite of machine learning models for both tasks.
  - **Analysis:** Compare model performance to identify the best approach for each task.

Methodology - Feature Engineering

The key to our success was transforming raw data into powerful predictive features.

- **Primary Health Indicator:**
  - `Capacity_mAh`: Calculated for each cycle to track degradation.
- **Engineered Physical Features:**
  - **Thermal Indicators:** `Avg_Temperature_C`, `Max_Temperature_C`
    - *Insight: An increase in temperature often signals higher internal resistance and cell aging.*
  - **Voltage Indicators:** `Avg_Voltage_mV`, `Min_Voltage_mV`
    - *Insight: A lower average voltage during discharge indicates a degraded state.*
  - **Time-Based Indicator:** `Cycle_Duration_s`
    - *Insight: A shorter discharge time corresponds to less available capacity.*

Task 1 Results - RUL Prediction (Regression)

Goal: To predict the exact number of cycles remaining.

Final Model Performance (After Tuning) Best Model:

The **Tuned Random Forest Regressor** was the clear winner, with an average prediction error of less than one cycle.

- **Key Insight:** The Random Forest model excelled because it successfully learned the complex, non-linear degradation patterns revealed by our engineered features—a capability the simpler models lacked.

Task 1 Insights - What Drives RUL?

Feature Importance for RUL Prediction

- The `Cycle` number is the most dominant predictor, as expected.
- Critically, **Average Temperature** emerged as the second most important feature.
- **Actionable Insight:** The strong correlation with average temperature confirms that thermal management is not just for safety; it is a critical factor in battery longevity. Our model validates that monitoring temperature is essential for accurate RUL forecasting.

Task 2 Results - Health State Classification

Goal: To classify each cycle as 'Healthy' or 'Moderate'.

Final Model Performance (After Tuning)

Best Models: The **Tuned SVM** and **Logistic Regression** both achieved flawless classification.

- **Key Insight:** The perfect scores indicate that our engineered features create a very clear and separable boundary between the different health states.

Task 2 Insights - What Defines Health State?

Visual and Feature Analysis

Confusion Matrices: The matrices for SVM and Logistic Regression were perfect, showing zero misclassifications.

ROC Curves: Both top models achieved a perfect Area Under the Curve (AUC) of 1.0, confirming their flawless ability to distinguish between classes.

Key Differentiator: The `Average Voltage` was the most decisive feature for classifying health state. This finding explains why even linear models achieved perfection: the voltage drop between 'Healthy' and 'Moderate' states is so distinct that it creates a clear, easily separable threshold.

Task 2 Insights - What Defines Health State?

Visual and Feature Analysis

Confusion Matrices: The matrices for SVM and Logistic Regression were perfect, showing zero misclassifications.

ROC Curves: Both top models achieved a perfect Area Under the Curve (AUC) of 1.0, confirming their flawless ability to distinguish between classes.

Key Differentiator: The `Average Voltage` was the most decisive feature for classifying health state. This finding explains why even linear models achieved perfection: the voltage drop between 'Healthy' and 'Moderate' states is so distinct that it creates a clear, easily separable threshold.

Executive Summary

This project successfully developed and evaluated a suite of machine learning models to predict the health and Remaining Useful Life (RUL) of a CS2_8 Li-ion battery. By engineering a set of robust physical features from raw time-series data—including temperature, voltage, and cycle duration—we were able to build highly accurate models for two distinct tasks: regression for RUL prediction and classification for health state assessment. The **Tuned Random Forest Regressor** was identified as the best model for predicting RUL, with an error of less than one cycle. For classification, the **Tuned Support Vector Machine (SVM)** achieved perfect accuracy in distinguishing between 'Healthy' and 'Moderate' battery states. The analysis confirms that a battery's thermal and voltage characteristics are key indicators of its degradation, providing a strong foundation for real-world battery management systems.

1. Introduction

The proliferation of electric vehicles (EVs) and portable electronics has made the accurate prediction of battery health and Remaining Useful Life (RUL) a critical field of study. An effective Battery Management System (BMS) that can forecast battery degradation is essential for ensuring reliability, safety, and user confidence. This project addresses this need by applying a comprehensive machine learning workflow to the CALCE CS2_8 battery dataset. The primary goals were to develop a model capable of accurately predicting the RUL in terms of remaining cycles and another model to classify the battery's current health state.

2. Methodology

The project followed a structured, multi-phase methodology, beginning with raw data and culminating in tuned, predictive models.

2.1. Data Processing and Feature Engineering

The initial phase involved processing 35 individual time-series files, each representing a single charge-discharge cycle. The key steps were:

- **Data Aggregation:** A script was developed to parse each file, handle header inconsistencies, and consolidate the key time-series data (Time, Voltage, Current, Temperature, Capacity).
- **Feature Creation:** For each cycle, a set of descriptive features was engineered from the discharge phase data to capture the physical state of the battery:
    - **Capacity_mAh:** The primary measure of battery health, representing the total discharged capacity for the cycle.

    - **Thermal Indicators:** Avg_Temperature_C and Max_Temperature_C.
    - **Voltage Indicators:** Avg_Voltage_mV and Min_Voltage_mV.
    - **Temporal Indicator:** Cycle_Duration_s.
- **Target Variable Creation:**
    - **RUL (for Regression):** Calculated as EOL - Cycle, where End of Life (EOL) was defined as the cycle where capacity first dropped below 80% of the initial capacity.
    - **Health State (for Classification):** A categorical feature created based on State of Health (SOH): 'Healthy' (SOH > 90%), 'Moderate' (80% <= SOH <= 90%), and 'Critical' (SOH < 80%).

3. Regression Analysis: Predicting RUL

The first modeling task focused on predicting the exact RUL.

3.1. Initial Model Performance

Models were first trained with their default scikit-learn parameters.

Model	RMSE	MAE	R-squared
Linear Regression	1.05e-14	9.64e-15	1.000
Random Forest Regressor	0.85	0.70	0.981
XGBoost Regressor	1.23	1.18	0.961
Neural Network (MLP)	2.56	2.15	0.832
SVR	**5.54**	**4.29**	**0.209**

- **Insight:** The default SVR performed very poorly, while the tree-based models (Random Forest, XGBoost) were highly effective from the start.

3.2. Final Performance (After Hyperparameter Tuning)

Model	Test R-squared	Test RMSE (cycles)
Random Forest (Tuned)	**0.986**	**0.73**
SVR (Tuned)	0.966	1.16
XGBoost (Tuned)	0.961	1.24

Neural Network (Tuned)	0.886	2.10

- **Insight:** Hyperparameter tuning provided a massive performance boost to the **SVR** model, raising its R^2 from 0.21 to 0.97. The **Tuned Random Forest** emerged as the best overall model, with an extremely low prediction error of only **0.73 cycles**.

3.3. Key Findings from Regression

- **Feature Importance:** The analysis revealed that Cycle was the most dominant predictor, followed by Avg_Temperature_C. This confirms that a battery's thermal signature is a critical indicator of its long-term health.
- **Visual Analysis:** The prediction plot for the Tuned Random Forest showed its superior ability to capture not just the primary downward trend of RUL but also the minor, real-world fluctuations, unlike simpler models.

4. Classification Analysis: Predicting Health State

The second task focused on classifying a cycle as 'Healthy' or 'Moderate'.

4.1. Initial Model Performance

Model	Accuracy	F1-Score
Logistic Regression	1.00	1.00
SVM	1.00	1.00
Decision Tree	0.91	0.91
Random Forest	0.91	0.91
KNN	0.73	0.74

4.2. Final Performance (After Hyperparameter Tuning)

Model	Accuracy	F1-Score	AUC
SVM (Tuned)	**1.00**	**1.00**	**1.00**
Logistic Regression	1.00	1.00	1.00
Random Forest (Tuned)	0.91	0.91	N/A

KNN (Tuned)	0.73	0.74	N/A

- **Insight:** Hyperparameter tuning did not significantly alter the top-performing models, as SVM and Logistic Regression already achieved perfect scores with their default settings.

4.3. Key Findings from Classification

- **Best Models:** The **Tuned SVM** and **Logistic Regression** are the best models for this task, both achieving flawless classification with perfect accuracy, F1-scores, and an Area Under the ROC Curve (AUC) of 1.00.
- **Feature Importance:** For classification, Avg_Voltage_mV was the second most important feature after Cycle. This provides a crucial insight: a drop in a battery's average operating voltage is a very clear signal that it has transitioned from a 'Healthy' to a 'Moderate' state, explaining why even linear models could find a perfect separating boundary.
- **Data Limitation:** A key finding was the severe class imbalance in the dataset. The test set contained no examples of the 'Critical' state, meaning we can only validate the models' ability to distinguish between 'Healthy' and 'Moderate' states.

5. Conclusion and Recommendations

This project successfully demonstrated that a data-driven approach using machine learning can yield highly accurate models for battery health monitoring. The process of engineering physically meaningful features from raw sensor data was validated as the most critical step for enabling high-performance modeling.

- **For Predicting Exact RUL:** The **Tuned Random Forest Regressor** is recommended. Its ability to model complex, non-linear relationships makes it ideal for precise RUL forecasting.
- **For Classifying Health State:** The **Tuned Support Vector Machine (SVM)** is recommended. Its perfect performance, combined with its computational efficiency, makes it a robust choice for real-time health state classification.

Analysis of Advanced Regression Results (Step 5)

This report provides a comprehensive analysis of the results from our advanced regression modeling, including hyperparameter tuning, feature importance, and a final model comparison.

1. Hyperparameter Tuning: The Impact of Optimization

The first step was to find the optimal settings for each of our complex models. The results were:

- **Random Forest:** Best with {'max_depth': None, 'min_samples_leaf': 1, 'n_estimators': 150}. This suggests the model benefits from having many deep trees to capture complex patterns.
- **SVR:** Best with {'C': 100, 'gamma': 0.1, 'kernel': 'rbf'}. This is a significant finding. The default C=1 was clearly not suitable. By increasing the C parameter, we allowed the model to create a more complex decision boundary, drastically improving its performance.
- **XGBoost:** Best with {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}. This points to a preference for deeper trees (max_depth=5) and a moderate learning rate.
- **Neural Network (MLP):** Best with {'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (50, 30)}. The model preferred the standard relu activation and a two-layer architecture.

Key Insight: Hyperparameter tuning had the most dramatic impact on the **SVR model**, transforming it from the worst-performing model into a strong contender. This highlights the importance of not relying on default model settings.

2. Final Model Performance Comparison

After tuning, we evaluated all models on the unseen test data. The final ranking is:

Model	Test R-squared	Test RMSE (cycles)
Linear Regression	1.000000	~0.00
Random Forest (Tuned)	**0.986175**	**0.73**
SVR (Tuned)	0.965586	1.16
XGBoost (Tuned)	0.960728	1.24
Neural Network (Tuned)	0.886317	2.10

Key Insights:

- **Best Overall Model:** The **Tuned Random Forest** is the champion. With an **R^2 of 0.986**, it can explain over 98.6% of the variance in the RUL. Its RMSE of **0.73** means its predictions are, on average, off by less than a single cycle, which is an exceptionally accurate result.
- **The SVR's Redemption:** After tuning, the SVR model's R^2 score jumped from a dismal 0.21 to a very impressive **0.966**. It is now the second-best model, proving that proper tuning is essential.
- **The Illusion of Perfection:** While Linear Regression has a perfect score, it's only because it found the simple RUL = EOL - Cycle formula. The Random Forest is superior because it learned this *plus* the more complex patterns from our other engineered features.

3. Feature Importance Analysis

This plot, generated from our best model (Tuned Random Forest), is one of the most valuable outputs of the entire project. It tells us *what* factors the model considered most important when predicting the battery's remaining life.

Key Insights:

1. **Cycle is King (Importance: ~0.67):** As expected, the current cycle number is the single most powerful predictor. This makes intuitive sense—the older a battery is, the less life it has left.
2. **Temperature is a Critical Secondary Indicator (Importance: ~0.20):** The Avg_Temperature_C is the second most important feature. This is a crucial finding, suggesting that as this battery degrades, its internal temperature during operation consistently changes in a predictable way.
3. **Other Features Matter:** Cycle_Duration_s, Max_Temperature_C, and the voltage features all contribute to the prediction, even if their individual importance is small. They provide the nuance that allows the Random Forest to outperform the simple Linear Regression.

4. Visualization of Tuned Models

- **Random Forest (Tuned):** The prediction line tracks the actual RUL almost perfectly, capturing both the main trend and the minor, real-world fluctuations. This is the visual proof of its superior performance.
- **SVR (Tuned):** The improvement is dramatic. The prediction line is no longer flat; it now follows the downward trend very closely and smoothly. It is an excellent fit, beaten only slightly by the Random Forest's ability to capture the finer details.

- **XGBoost & Neural Network (Tuned):** Both models show strong performance, with their prediction lines closely following the actual data. They exhibit slightly more variance and individual cycle errors than the Random Forest and SVR, which aligns with their slightly lower R^2 scores.

Overall Conclusion for Regression Task

Through advanced techniques, we have rigorously trained, tuned, and evaluated several models. The **Tuned Random Forest Regressor** is definitively the best model for this task, providing highly accurate and reliable RUL predictions.

Crucially, our feature importance analysis shows that while the cycle number is the dominant predictor, **thermal and voltage characteristics are key secondary indicators** that allow sophisticated models to outperform simple linear approaches.

Analysis of Advanced Classification Results (Step 6)

This report provides a comprehensive analysis of the results from our advanced classification modeling, which included hyperparameter tuning for all complex models and evaluation using confusion matrices and ROC/AUC curves.

1. Performance of Tuned Models

After performing a GridSearchCV to find the optimal settings for each model, the final evaluation on the unseen test data yielded the following performance:

Model	Accuracy	F1-Score
SVM (Tuned)	**1.00**	**1.00**
Logistic Regression	1.00	1.00
Random Forest (Tuned)	0.91	0.91
KNN (Tuned)	0.73	0.74

Key Insights:

- **Top Performers:** The **Tuned Support Vector Machine (SVM)** and the standard **Logistic Regression** both achieved perfect scores. They were able to flawlessly distinguish between 'Healthy' and 'Moderate' cycles on the test data.
- **The Power of Simplicity:** The fact that a relatively simple model like Logistic Regression performed perfectly suggests that the classes in our dataset are **linearly separable**. This means a straight line (or a simple hyperplane in higher dimensions) can be drawn to separate the 'Healthy' data points from the 'Moderate' ones in the feature space.
- **Strong but Imperfect:** The tuned Random Forest made only one error, showing it is also a very strong model for this task.
- **KNN Struggles:** Even after tuning, KNN remains the weakest model. This indicates its instance-based "voting" approach is less effective for this problem than the boundary-finding approaches of SVM and Logistic Regression.

2. Feature Importance for Classification

The feature importance plot from our Tuned Random Forest model tells us which factors were most decisive in classifying a cycle's health state.

Key Insights:

1. **Cycle is the Dominant Feature (Importance: ~0.35):** Just as with the regression task, the age of the battery is the most powerful classifier.
2. **Voltage is a Key Indicator (Importance: ~0.27):** Unlike in the regression task, Avg_Voltage_mV is the second most important feature here. This is a crucial insight: a drop in the average operating voltage during discharge is a very clear signal that the battery has transitioned from a 'Healthy' to a 'Moderate' state.
3. **Temperature Follows:** Max_Temperature_C and Avg_Temperature_C are the next most important features, confirming that thermal properties are strong indicators of battery degradation and health status.

3. Confusion Matrix and ROC Curve Analysis

These visualizations provide a deeper look into model performance beyond just the accuracy scores.

Confusion Matrices:

- **SVM & Logistic Regression:** Both models produced perfect confusion matrices with zero errors on the off-diagonal. They correctly classified all 8 'Healthy' and all 3 'Moderate' samples.
- **Random Forest:** This matrix shows the model's single error: it misclassified one 'Healthy' sample as 'Moderate'. This is a minor error, but it's what separates it from the perfect models.

Receiver Operating Characteristic (ROC) Curves:

(Note: This plot shows the curves for all models overlaid, with the Logistic Regression curve and its perfect AUC score highlighted as an example)

- **What it means:** The ROC curve plots the True Positive Rate against the False Positive Rate. A perfect model has a curve that goes straight up to the top-left corner. The Area Under the Curve (AUC) is a single number summarizing this: 1.0 is a perfect score.
- **Key Insight:** The plot shows that both **SVM and Logistic Regression have a perfect AUC of 1.00**. This is the strongest possible confirmation that these models can flawlessly distinguish between the 'Healthy' and 'Moderate' classes in our dataset. The other models have curves that are slightly below this perfect line, corresponding to their lower performance scores.

Overall Conclusion for Classification Task

The advanced classification task was highly successful and confirmed the high quality of our engineered features.

- **Best Models:** The **Tuned SVM** and **Logistic Regression** are the best models for this task, both achieving perfect classification scores.
- **Key Insight:** The problem of classifying battery health into 'Healthy' and 'Moderate' is linearly separable with the features we have engineered. A declining average voltage and rising temperature are the clearest indicators, after the cycle number, that a battery's health is changing.
- **Limitation:** As before, the analysis is limited by the lack of 'Critical' examples in the test set.