Milestone 1 Report

1. Introduction

Milestone 1 focused on understanding the CMAPSS turbofan engine dataset, preprocessing it, performing exploratory data analysis (EDA), and preparing the foundation for building an accurate Remaining Useful Life (RUL) prediction model. The emphasis in this stage was on obtaining clarity about the data characteristics, identifying relevant features, cleaning the dataset, and visualizing engine behavior to support further model development.

---

2. Work Completed in Milestone 1

2.1 Dataset Understanding

The CMAPSS dataset contains multivariate sensor readings collected from multiple turbofan engines under different operating conditions and fault modes. Four datasets (FD001–FD004) were considered, each with varying complexity in terms of operating conditions and failure modes. Each engine runs until failure, allowing true RUL labeling.

Key components explored:

- cycle
- three operational settings
- 21 sensor measurements
- dataset-level differences (FD001–FD004)

This understanding was necessary to decide how to treat each dataset individually or in a combined form.

2.2 Data Loading and Cleaning

The raw NASA CMAPSS `.txt` files were loaded and structured into DataFrames. Cleaning steps included:

- Assigning meaningful column names
- Removing irrelevant sensors (those with constant values across cycles)
- Handling missing or inconsistent values
- Adding dataset identifiers to distinguish FD001–FD004 when combining them

This resulted in a clean, uniform dataset ready for feature engineering and modeling.

2.3 Exploratory Data Analysis (EDA)

EDA was carried out to understand engine degradation patterns and sensor behavior.

Key analysis performed:

- Distribution of sensor values across cycles
- Identification of stable and highly variable sensors
- Understanding dataset differences
- Visualizing how RUL decreases with cycles for engines in FD001–FD004

Sensor behavior plots showed that some sensors remain stable throughout engine life, while others exhibit gradual drift or fluctuations. This helped in identifying which sensors are likely to be more informative for RUL prediction.

2.4 RUL Computation

For each engine, true RUL was computed using:

```
RUL = max_cycle_for_engine − current_cycle
```

This produced the ground-truth target variable necessary for supervised learning.

2.5 Sensor vs RUL Relationship Visualization

Multiple plots were generated showing the relationship between RUL and selected sensor values across engines in FD001–FD004. These plots demonstrated:

- RUL always decreases linearly
- Some sensors (such as sensor_2) show minimal drift
- Different datasets have different sensor behaviors depending on complexity

This provided insight into which sensors correlate with degradation and which sensors do not.

2.6 Dataset Consolidation Plan

A strategy was finalized to combine all four datasets:

- Standardize all datasets to the same feature format
- Add dataset-specific one-hot encoded flags (ds_FD001, ds_FD002, etc.)
- Ensure consistent scaling across all datasets
- Prepare a unified feature set for model training

This ensures the model can learn general degradation patterns across multiple operating conditions.

---

3. Summary of Milestone 1 Outcomes

Milestone 1 successfully delivered:

- Complete understanding and structuring of the CMAPSS dataset
- Cleaned and consolidated dataset inputs
- Exploratory visualizations highlighting sensor behaviors and RUL trends
- Feature extraction and identification of meaningful variables

- A clear strategy for training combined models using FD001–FD004

The foundation is now in place for designing, training, and optimizing the RUL prediction model.

---

4. Work Planned for Milestone 2

Milestone 2 will focus on model building, optimization, and validation. The following tasks are planned:

4.1 Windowing and Sequence Preparation

Time-series sequences will be created using sliding windows (e.g., sequence length = 30), essential for feeding data into LSTM/GRU-based models.

4.2 Feature Scaling and Normalization

All engine sensor data will be standardized using MinMaxScaler or StandardScaler to ensure consistency across datasets.

4.3 Model Architecture Development

Multiple deep learning architectures will be implemented and evaluated:

- LSTM
- Bi-LSTM
- CNN-LSTM
- GRU
- Hybrid models

The selected architecture will be tuned for best performance.

4.4 Cross-Dataset Training

A unified model will be trained on combined FD001–FD004 data using one-hot encoded dataset identifiers.

4.5 Model Evaluation

The trained model will be evaluated using:

- MAE
- MSE
- RMSE
- RUL prediction plots
- Engine-wise evaluation

This ensures the model generalizes across multiple engine conditions.

4.6 Visualization of Model Performance

Training curves, loss graphs, and prediction comparison plots will be generated to interpret model behavior.

---

5. Conclusion

Milestone 1 established a strong foundation by cleaning, analyzing, and understanding the CMAPSS turbofan engine data. The insights obtained directly support the model-building activities in Milestone 2, where advanced deep learning techniques will be applied for accurate RUL prediction.