REPORT NO. 10644

Data Visualisation and Machine Learning for Truck Performance Prediction and Analysis

Results of research carried out as MRIWA Project M10644

at Electric Power Conversions Australia (EPCA)

by

Clayton Franklin

Date published: August 2026

DOI 10.71342/982904340372

Distributed by: MRIWA
1 Adelaide Terrace
Perth WA 6000
to which all enquiries should be addressed

Acknowledgements

MINISTER FOR MINES AND PETROLEUM

Hon Daniel Pastorelli MLA

CHIEF EXECUTIVE OFFICER, MINERALS RESEARCH INSTITUTE OF WESTERN AUSTRALIA

Nicole Roocke

LEAD INVESTIGATOR, ELECTRIC POWER CONVERSIONS AUSTRALIA

Clayton Franklin

AUTHORS:

Clayton Franklin

REFERENCE

The recommended reference for this publication is: Franklin, C. (2026) Data Visualisation and Machine Learning for Truck Performance Prediction and Analysis. MRIWA Report 10644, 45pp. DOI 10.71342/982904340372

DOI: 10.71342/982904340372

This report is available from the national Library of Australia

PARTICIPATING ORGANISATIONS

ABOUT THIS PUBLICATION

All work reported in this report was undertaken by Electric Power Conversions Australia (EPCA)

FUNDING ACKNOWLEDGEMENT

This research is funded by MRIWA (DOI 10.71342/430042142281) as part of Project M10644.

KEYWORDS AND TAGS

Predictive maintenance, battery electric mining trucks, predictive analytics, asset health monitoring, machine learning, data visualisation

DISCLAIMER

This report's scope statement, describing the basis on which the findings should be read, appears at the opening of the Executive Summary.

Published 2026 by the Minerals Research Institute of Western Australia

This report is published in digital format (PDF) and is available online at https://www.mriwa.wa.gov.au/research-projects/project-portfolio/

CONTACT

Minerals Research Institute of Western Australia

1 Adelaide Terrace

Perth WA 6000

+61 8 6180 4340

mail@mriwa.wa.gov.au

https://www.mriwa.wa.gov.au/

Executive Summary

This report documents work carried out on a single first-generation prototype vehicle operating at one site over a defined test period, using a training dataset drawn from a narrow range of operating conditions. The model performance results reported here characterise that prototype dataset and the models trained on it. They are not a characterisation of the E-777D production configuration, of EPCA's current monitoring and analytics platform, or of the performance of any other EPCA vehicle or deployment.

The global mining industry is under increasing pressure to reduce its carbon footprint and improve operational efficiency. Heavy haul trucks represent one of the largest consumers of diesel fuel in the sector, making them a prime target for electrification strategies. Electric Power Conversions Australia (EPCA) has pioneered a circular economy approach to retrofitting diesel-powered 100-tonne Caterpillar 777D haul trucks with full battery-electric powertrains, producing the EPCA E-777D electric haul truck.

This report presents the complete outcomes of MRIWA Project M10644, which developed and validated a data visualisation and machine learning platform for predicting and analysing the performance of EPCA's E-777D electric haul truck. The research was carried out across four project stages: (1) data scraping and storage organisation, (2) machine learning model development and initial validation, (3) data visualisation development, and (4) model validation and verification against new independent operational data.

Stage 1 established a robust automated data pipeline for ingesting, processing, and storing EPCA's truck telemetry data from its onboard systems into an AWS cloud environment, using Parquet file format and AWS Athena for efficient querying.

Stage 2 developed eleven machine learning models targeting key operational predictions, including state of charge at 5, 30, and 60 second horizons, time to charge, overheat risk, energy consumption, and chiller anomaly detection. The majority of models achieved R^2 values exceeding 0.99 on the held-out validation dataset, demonstrating strong predictive performance.

Stage 3 delivered a Microsoft Power BI dashboard directly linked to the AWS S3 and AWS Athena, enabling prediction and historical visualisation of all onboard truck performance data with multi-variable comparison and incremental data refresh.

Stage 4 completed the project by validating the machine learning models against a new independent operational dataset captured during live truck trials.

While predictive accuracy on the independent dataset fell short of operational requirements, the data pipeline and analytics platform provide a validated foundation for further model development as additional vehicles, conditions and data volumes are incorporated.

This project contributes to MRIWA's Research Priority Plan themes of "Infrastructure and Logistics - Data Driven Decisions," advancing the capability of the Western Australian mining sector to leverage data-driven technologies in support of sustainable and efficient electrified mining operations.

Table of Contents

Acknowledgements 2
Executive Summary 4
Terms, Abbreviations and Acronyms 6
List of Tables/Figures 7
1. Introduction 8
  1.1 Objectives of the research 9
  1.2 Scope 9
2 Methodology 10
  2.1 Data Scraping and Storage Organisation 10
  2.1.1 Data Sources and Structure 10
  2.1.2 Data Preprocessing Pipeline 12
  2.1.3 AWS Cloud Storage and Query Infrastructure 12
  2.2 Machine Learning Model Development and Initial Validation 15
  2.2.1 Target Variables and Feature Selection 15
  2.2.2 Feature Engineering 16
  2.2.3 Algorithm Selection 17
  2.2.4 Model Training and Validation 17
  2.3 Data Visualisation Development 18
  2.3.1 AWS Athena to Power BI Integration 18
  2.3.2 Data Transformation and Incremental Refresh 19
  2.3.3 Dashboard Functionality 19
  2.4 Model Validation and Verification 24
3 Results and Findings 26
  3.1 Data Pipeline 26
  3.2 Trial Results and Verification 26
4 Discussion 35
  4.1 Data Infrastructure and Pipeline 35
  4.2 Initial Machine Learning Model Performance 35
  4.3 Data Visualisation Platform 36
  4.4 Model Validation 36
5 Conclusions 37
6 Recommendations for further work 38
  6.1 Model Refinement and Retraining 38
  6.2 Data Infrastructure Expansion 38
  6.3 Visualisation Platform Enhancement 38
7 References 39
Appendix – Model Initial Key Features and Score 40

Terms, Abbreviations and Acronyms

Terms, Abbreviations, & Acronyms	Description
AWS	*Amazon Web Services*
AWS S3	*Amazon Web Services Simple Storage Service*
BTMS	*Battery Thermal Management System*
CSV	*Comma-Separated Values*
DCDC	*DC-to-DC Converter*
EPCA	*Electric Power Conversions Australia*
HMI	*Human Machine Interface*
HV	*High Voltage*
ICE	*Internal Combustion Engine*
LOOCV	*Leave-One-Out Cross-Validation*
LSTM	*Long Short-Term Memory*
LV	*Low Voltage*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
METS	*Mining, Equipment, Technology and Services*
ML	*Machine Learning*
MRIWA	*Minerals Research Institute of Western Australia*
ODBC	*Open Database Connectivity*
OEM	*Original Equipment Manufacturer*
Parquet	*Columnar storage file format optimised for big data*
PLC	*Programmable Logic Controller*
Power BI	*Microsoft Power Business Intelligence*
R^2	*Coefficient of Determination*
RMSE	*Root Mean Square Error*
SOC	*State of Charge*
SOE	*State of Energy*
SOH	*State of Health*
VCU	*Vehicle Control Unit*

List of Tables/Figures

Table 1: Test Session Lists 25

Table 2: Test Times and Conditions for the new Trials 26

Table 3: New Validation Results compared with Original Results....................................... 27

Table 4: Energy Consumption Per Trip From Session 1. Values are indexed to the measured consumption of trip 1 (trip 1 actual = 100). Measured consumption varied by less than 1.5% across the six trips. 33

Figure 1: Screenshot of Drive 1 Data Logged 2024-9-19-9-41-23...................................... 11

Figure 2: Parquet Vehicle File.. 13

Figure 3: Data Queried Via AWS Athena.. 14

Figure 4: Data being requested by Power BI from AWS Athena. 19

Figure 5: Dashboard Dataset All Auxiliary Data Displayed. 20

Figure 6: Dashboard Displaying Pack Current from Battery with Reduced Time Range..... 20

Figure 7: Dashboard Display Comparing Two Data Points. 21

Figure 8: Dashboard Display Comparing Different Temps from Different Systems............ 22

Figure 9: Dashboard Comparing Different Temps from Different Systems (Dataset Further Zoomed In).. 22

Figure 10: Dashboard Table and Graph Mode. 23

Figure 11: Graph Fullscreen Mode. 23

Figure 12: Bakers Hill Test Loop. Imagery ©2026 Google. Reproduced under Google Maps and Google Earth brand permissions. This image is third-party material and is not covered by the Creative Commons Attribution 4.0 International Licence applying to this publication..... 24

Figure 13: Charge Time Estimate Model Prediction for Test Session 1. 28

Figure 14: Time Till Charge Complete Model Prediction for Test Session 8 (Charging). 29

Figure 15: Overheating Risk for Test Session 1. 29

Figure 16: Total Tonnes Moved Before Battery Flat for Test Session 1. 30

Figure 17: Time to Battery Flat for Test Session 1. 30

Figure 18: Power Usage – 5s Ahead for Test Session 1. 31

Figure 19: SOC Prediction – 5 Seconds Ahead.. 32

Figure 20: SOC Prediction - 30 Seconds Ahead. 32

Figure 21: SOC Prediction - 60 Seconds Ahead. 33

Figure 22: Chiller Anomaly Prediction Model from Test Session 1..................................... 34

1. Introduction

The Western Australian mining industry is one of the most significant contributors to the State's economy, employing tens of thousands of workers and generating billions of dollars in exports annually (WA Government, 2026). However, the sector is also one of the largest consumers of diesel fuel in Australia. Heavy haul trucks, which transport ore from pit face to processing facilities across large open-cut mining operations, account for a substantial proportion of this consumption (Bao, Knights, Kizil, & Nehring, 2026). The resulting greenhouse gas emissions place the mining sector under pressure to identify and implement credible pathways to decarbonisation, consistent with the 2030 and 2050 net-zero commitments agreed upon under the 2016 Paris Agreement.

The electrification of heavy mining equipment represents one of the most promising avenues for reducing emissions in the sector. However, the transition from diesel to battery-electric vehicles presents significant engineering and operational challenges, particularly in relation to the reliability and predictability of performance in the extreme conditions typical of Australian mine sites.

A key barrier to broader industry adoption of electric haul trucks is the lack of mature data analytics and predictive intelligence capabilities. Unlike conventional diesel-powered trucks, which have decades of operational data and well-established maintenance regimes, battery electric vehicles in the mining context are a relatively new technology. Operators require confidence that the vehicles can perform reliably, that their state of charge can be accurately predicted, and that potential failures, particularly thermal events, can be anticipated and mitigated before they occur.

EPCA has directly addressed the hardware dimension of this challenge through its development of the E-777D electric haul truck, a fully retrofitted 100-tonne Caterpillar 777D fitted with a battery-electric powertrain. MRIWA Project M10644 addresses the data and intelligence aspects, developing the systems necessary to monitor, visualise, and predict the operational performance of this vehicle at scale.

1.1 Objectives of the research

The primary objective of this research project is to design, develop, and validate a machine learning and data visualisation platform for the EPCA E-777D 100-tonne electric haul truck. The platform is intended to:

- Enable ingestion, processing, and structured storage of all onboard telemetry data from the E-777D truck.
- Develop machine learning models capable of predicting key performance parameters, including state of charge, time to charge, overheat risk, energy consumption, and anomalous chiller behaviour.
- Provide an accessible, interactive data visualisation interface enabling operators and engineers to analyse truck performance data across time and across multiple onboard systems.
- Validate the developed models against independent operational data to assess performance and operational readiness for deployment in operational mining environments.

The research supports the broader industry goal of accelerating the adoption of battery-electric haul trucks in Western Australian mining operations by demonstrating that reliable, data-driven performance prediction is achievable with current technology.

1.2 Scope

The scope of this project encompasses four primary stages of work, aligned with the four project milestones defined in the MRIWA grant agreement:

- Stage 1: Data scraping, preprocessing, and storage organisation, including the development of an automated pipeline for ingesting EPCA E-777D telemetry data into an AWS cloud environment.
- Stage 2: Machine learning model development and initial validation, including the selection of appropriate algorithms, feature engineering, model training, and validation against a held-out portion of the training dataset.
- Stage 3: Data visualisation development, including the design and deployment of a Microsoft Power BI dashboard connected to the AWS data lake, enabling historical performance monitoring.
- Stage 4: Model validation and verification against a new operational dataset captured during live truck trials, determining model performance in real-world conditions and identifying any required refinements.

The project focuses on the EPCA E-777D truck operating at the Bakers Hill sand pit mine site in Western Australia. The machine learning models and visualisation platform are developed to be extensible to future truck deployments and additional datasets.

2 Methodology

2.1 Data Scraping and Storage Organisation

The first stage of the project addressed the challenge of transforming EPCA's raw onboard telemetry data into a structured, queryable, and cloud-hosted format suitable for machine learning and visualisation. The truck's onboard systems generate a continuous stream of data across seven separate subsystems, each recording values at one-second intervals from the moment the vehicle is powered on until it is powered off.

2.1.1 Data Sources and Structure

The seven onboard systems generating data are:

- Battery: Pack-level parameters covering charge state, cell-level voltages, thermal measurements, current, voltage and charging status.
- Auxiliary: Parameters for the auxiliary inverter and motor, covering thermal, electrical and speed measurements.
- Drive 1 & Drive 2: Inverter and motor parameters for each of the two main drives, covering thermal, electrical and rotational-speed measurements.
- Chiller: Parameters relating to the cooling system, covering fan and pump operation, coolant condition and power consumption.
- Charge Controller: Parameters relating to charging operations and charge control logic.
- Vehicle: Parameters from the broader vehicle system, including data from the original Caterpillar sensors, an inclinometer, the VCU, and other control boards.

For each system, two CSV files are generated per operational session: one recording input parameters sent to the system, and one recording output parameters returned by the system. Files are timestamped with the date and time of the recording session.

Figure 1 shows a sample of the raw CSV data logged from the Drive 1 system, before any processing takes place. It is included to illustrate the format and structure of the incoming data, not to convey specific readings. Each row represents one second of operation, with columns for module temperatures, coolant and hot spot temperatures, DC current and voltage, and other drive parameters. This is the starting point the preprocessing pipeline works from: raw, high-frequency, multi-column data that needs to be cleaned, structured, and converted before it can be queried or used in the machine learning models. The specific values in this snapshot carry no meaning outside the session they were recorded in and are not part of the technical evaluation.

Time	ModATem	ModBTem	ModCTem	GateDriTe	ContTemp	RTD1Tem	RTD2Tem	StallBurTe	CoolTemp	HotSpotTe	MotTemp	TorqShud	OilTemp	OilPres	AnIn4	AnIn5	AnIn6	DiIn1	DiIn2	DiIn3	DiIn4	DiIn5	DiIn6	DiIn7	DiIn8	MotAngEl	MotRPM	ElecOutFr	DeltResFi
19.09.2024 09:41:24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19.09.2024 09:41:25	19.8	20.1	19.7	17.4	14.5	300	300	44.8	11.2	11.2	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.9
19.09.2024 09:41:26	19.8	20.1	19.7	17.7	14.6	300	300	45.6	12.5	12.5	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	71.1	0	0	-70.9
19.09.2024 09:41:27	19.8	20.1	19.7	18	14.5	300	300	46.4	13.7	13.7	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.9	0	0	-70.8
19.09.2024 09:41:29	19.8	20.1	19.7	18.1	14.8	300	300	47	14.6	14.6	7.1	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.9	0	0	-70.8
19.09.2024 09:41:30	19.8	20.1	19.7	18.3	14.7	300	300	47.5	15.3	15.3	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.9	0	0	-70.8
19.09.2024 09:41:32	19.8	20.1	19.7	18.6	14.8	300	300	48	16.1	16.1	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.6	0	0	-70.5
19.09.2024 09:41:33	19.8	20.2	19.7	18.8	14.8	300	300	48.4	16.6	16.6	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.80001	0	0	-70.7
19.09.2024 09:41:34	19.8	20.1	19.7	18.8	14.9	300	300	48.7	17.1	17.1	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.7	0	0	-70.6
19.09.2024 09:41:36	19.8	20.1	19.7	19	15	300	300	49	17.5	17.5	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.5
19.09.2024 09:41:37	19.8	20.1	19.7	19.1	15.1	300	300	49.2	17.9	17.9	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.5
19.09.2024 09:41:38	19.8	20.1	19.7	19.3	15	300	300	49.4	18.2	18.2	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.6
19.09.2024 09:41:39	19.8	20.1	19.7	19.4	15.2	300	300	49.6	18.4	18.4	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.5
19.09.2024 09:41:40	19.8	20.1	19.7	19.6	15.1	300	300	49.7	18.6	18.6	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.5
19.09.2024 09:41:42	19.8	20.1	19.7	19.6	15.2	300	300	49.8	18.8	18.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.5
19.09.2024 09:41:43	19.8	20.1	19.7	19.8	15.2	300	300	49.9	19	19	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.80001	0	0	-70.5
19.09.2024 09:41:44	19.8	20.1	19.7	19.9	15.2	300	300	50	19.1	19.1	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.5
19.09.2024 09:41:46	19.8	20.1	19.7	20.1	15.4	300	300	50.1	19.2	19.2	6.5	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.5
19.09.2024 09:41:47	19.8	20.1	19.7	20.1	15.4	300	300	50.2	19.3	19.3	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.9	0	0	-70.5
19.09.2024 09:41:49	19.8	20.1	19.7	20.1	15.5	300	300	50.2	19.4	19.4	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.4
19.09.2024 09:41:50	19.8	20.1	19.7	20.3	15.3	300	300	50.3	19.5	19.5	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.4
19.09.2024 09:41:51	19.8	20.1	19.7	20.5	15.5	300	300	50.3	19.5	19.5	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.4
19.09.2024 09:41:52	19.8	20.1	19.7	20.6	15.4	300	300	50.4	19.6	19.6	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.5
19.09.2024 09:41:53	19.8	20.1	19.7	20.7	15.5	300	300	50.4	19.6	19.6	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.4
19.09.2024 09:41:55	19.8	20.1	19.7	20.7	15.6	300	300	50.4	19.7	19.7	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.4
19.09.2024 09:41:59	19.9	20.1	19.7	20.9	15.7	300	300	50.4	19.7	19.7	6.5	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.4
19.09.2024 09:42:00	19.8	20.1	19.7	21	15.8	300	300	50.5	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.4
19.09.2024 09:42:02	19.8	20.1	19.7	21.1	15.8	300	300	50.5	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.4
19.09.2024 09:42:03	19.8	20.1	19.7	21.1	15.9	300	300	50.5	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.4
19.09.2024 09:42:04	19.8	20.1	19.7	21.2	15.8	300	300	50.5	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.4
19.09.2024 09:42:05	19.8	20.1	19.7	21.3	15.9	300	300	50.4	19.8	19.8	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:06	19.8	20.1	19.7	21.3	15.9	300	300	50.3	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:08	19.8	20.1	19.7	21.4	16	300	300	50.3	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.3
19.09.2024 09:42:09	19.8	20.1	19.7	21.4	16	300	300	50.2	19.8	19.8	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.3
19.09.2024 09:42:10	19.8	20.1	19.7	21.5	15.9	300	300	50.2	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.3
19.09.2024 09:42:12	19.8	20.1	19.7	21.7	15.9	300	300	20.2	19.8	19.8	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:13	19.8	20.1	19.7	21.6	15.9	300	300	19.9	19.8	19.8	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:15	19.8	20.1	19.7	21.8	16.1	300	300	19.9	19.8	19.8	6.5	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.2	0	0	-70.3
19.09.2024 09:42:16	19.8	20.1	19.7	21.7	16	300	300	19.9	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.3
19.09.2024 09:42:17	19.8	20.1	19.7	21.8	16.1	300	300	19.9	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:18	19.8	20.1	19.7	22	16.2	300	300	19.9	19.9	19.9	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.3
19.09.2024 09:42:19	19.8	20.1	19.7	22.1	16.1	300	300	19.9	19.9	19.9	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:21	19.8	20.1	19.7	22.1	16.2	300	300	19.9	19.9	19.9	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.2
19.09.2024 09:42:22	19.6	19.8	19.7	22.2	16.3	300	300	19.9	19.8	19.8	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:23	19.3	19.5	19.4	22.2	16.1	300	300	19.8	19.8	19.8	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.3
19.09.2024 09:42:25	19.1	19.2	19.2	22.2	16.3	300	300	19.7	19.7	19.7	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.30001	0	0	-70.3
19.09.2024 09:42:26	18.9	19	19	22.2	16.3	300	300	19.6	19.6	19.6	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:28	18.8	18.9	18.8	22.2	16.2	300	300	19.5	19.5	19.5	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:29	18.7	18.9	18.7	22.3	16.3	300	300	19.4	19.4	19.4	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.4	0	0	-70.3
19.09.2024 09:42:30	18.6	18.8	18.6	22.4	16.3	300	300	19.3	19.3	19.3	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.5	0	0	-70.4
19.09.2024 09:42:31	18.7	19	18.7	22.9	16.3	300	300	19.3	19.2	19.2	6.5	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	69.6	0	0.1	-100
19.09.2024 09:42:32	18.8	19.2	18.9	23.2	16.3	300	300	19.2	19.1	19.1	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	69.5	0	0	-100.2
19.09.2024 09:42:34	18.9	19.3	19	23.2	16.4	300	300	19.2	19.1	19.1	6.6	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	70.9	0	0	0.8
19.09.2024 09:42:35	18.9	19.2	19	23.3	16.5	300	300	19.2	19.1	19.1	6.7	0	-25	0	0.16	0	0	0	0	0	0	0	1	0	0	68.9	0	0	-99.9

Figure 1: *Screenshot of Drive 1 Data logged 2024-9-19-9-41-23.*

2.1.2 Data Preprocessing Pipeline

An automated Python preprocessing pipeline was developed to prepare raw session data for storage. The pipeline performs file validation, timestamp extraction, schema-aware cleaning and structuring by subsystem, and conversion to Apache Parquet format. Parquet was selected for its columnar storage architecture, which substantially reduces file size and accelerates query performance for the large volumes of repeating, structured data generated by the truck (Amazon, n.d.). Implementation detail is withheld as commercially sensitive information and is available from EPCA on commercial terms.

Processing state and runtime status are logged to support resumption after interruption and to provide an audit trail. Figure 2 shows the final converted Parquet file from the data processing pipeline.

2.1.3 AWS Cloud Storage and Query Infrastructure

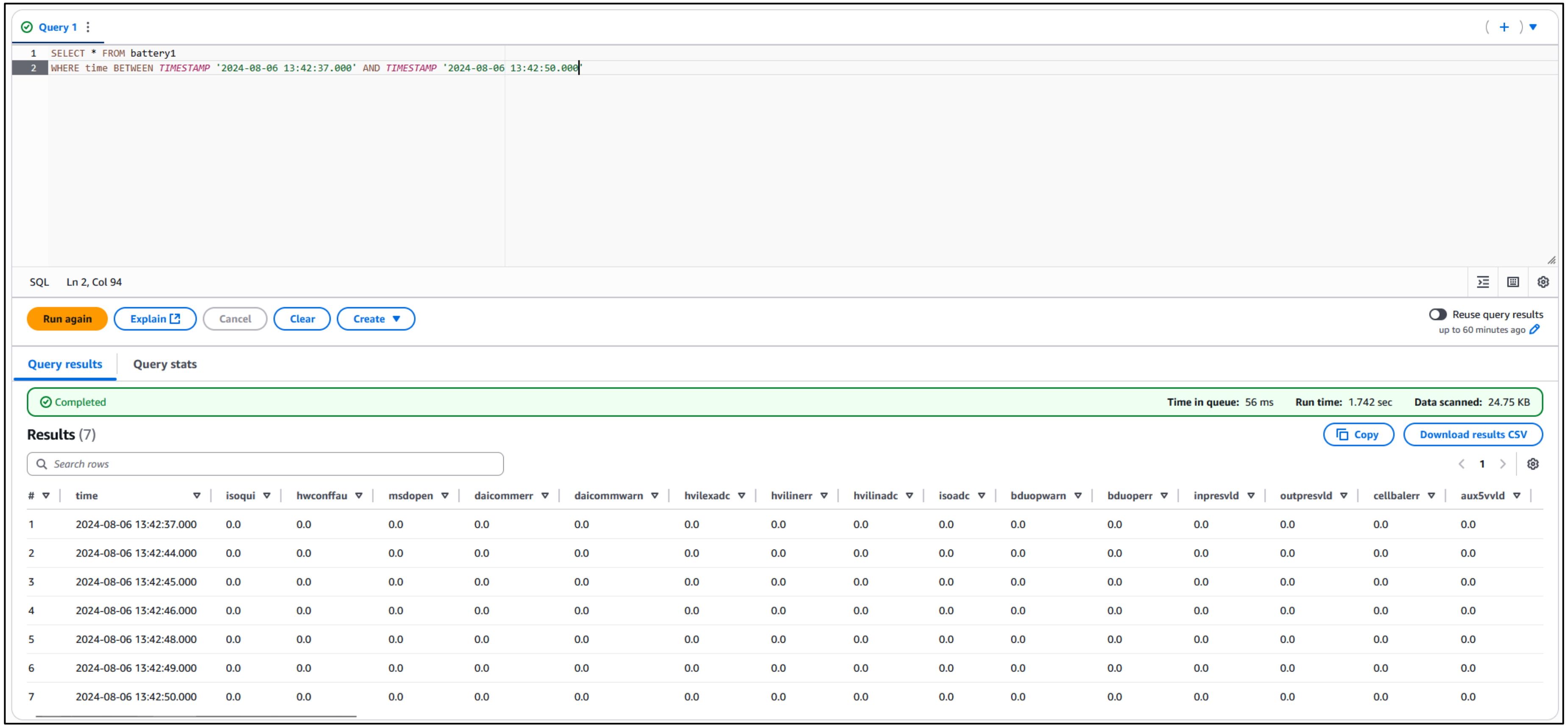
Processed Parquet files are uploaded to an AWS S3 bucket, organised by system type and partitioned by timestamp. Partitioning by time allows AWS Athena to skip irrelevant data partitions when processing time-bounded queries, significantly reducing query cost and latency (Amazon, n.d.). AWS Athena is configured to query the Parquet data directly in S3 without a separate database import step; see Figure 3, which demonstrates the usage of AWS Athena to search the created database.

vehicle.parquet

```
SELECT * FROM data
```

	time	plc_hb	24volt	vcumode	vehstate	whlspd	uptime	downtime	gearchngetime	bodypos	throttle	bodyup	brakeairpres	brakeoiltemp	downshifton	hoistposave	keyaccessvolt	ignition	parkbrakeleft	parkbrakeright
1	2024-10-10T09:39:43.000Z	0	0	0	0	0				0	0	0	0	0	0	0	0	0	0	0
2	2024-10-10T09:39:47.000Z	0	0	4	13	0				0	-0.06655	0	839	1528	1	3107	23877	0	417	312
3	2024-10-10T09:39:48.000Z	0	0	4	13	0				0	-0.0661	0	834	1528	1	3110	23895	0	420	315
4	2024-10-10T09:39:49.000Z	0	0	4	13	0				0	-0.06475	0	839	1531	1	3107	23858	0	417	315
5	2024-10-10T09:39:51.000Z	0	0	4	13	0				0	-0.0652	0	836	1528	1	3110	23858	0	420	315
6	2024-10-10T09:39:52.000Z	1	0	4	13	0				0	-0.0661	0	839	1528	1	3110	23802	0	420	312
7	2024-10-10T09:39:53.000Z	1	0	4	13	0				0	-0.06655	0	836	1531	1	3110	23830	0	420	315
8	2024-10-10T09:39:54.000Z	1	0	4	13	0				0	-0.0652	0	834	1528	1	3110	23793	0	420	315
9	2024-10-10T09:39:56.000Z	1	0	4	13	0				0	-0.0661	0	836	1528	1	3110	23821	0	420	315
10	2024-10-10T09:39:57.000Z	0	0	4	13	0				0	-0.06475	0	834	1528	1	3110	23821	0	417	315

Figure 2: *Parquet vehicle file.*



Query 1

```
SELECT * FROM battery1
WHERE time BETWEEN TIMESTAMP '2024-08-06 13:42:37.000' AND TIMESTAMP '2024-08-06 13:42:50.000'
```

SQL Ln 2, Col 94

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 56 ms Run time: 1.742 sec Data scanned: 24.75 KB

Results (7)

Copy Download results CSV

#	time	isoqui	hwconffau	msdopen	daicommerr	daicommwarn	hvilexadc	hvilinerr	hvilinadc	isoadc	bduopwarn	bduoperr	inpresvld	outpresvld	cellbalerr	aux5vvld
1	2024-08-06 13:42:37.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	2024-08-06 13:42:44.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	2024-08-06 13:42:45.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	2024-08-06 13:42:46.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	2024-08-06 13:42:48.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	2024-08-06 13:42:49.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	2024-08-06 13:42:50.000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3: *Data queried via AWS Athena*

.

2.2 Machine Learning Model Development and Initial Validation

Several machine learning models were developed, each targeting a specific operational prediction relevant to truck performance monitoring, safety, and operational planning.

2.2.1 Target Variables and Feature Selection

The key prediction targets were defined before constructing the models. The following are the decided prediction targets:

- Charge Time Estimate (Unit: Seconds): The amount of time it will take the vehicle to charge based on maximum charge capacity and current SOC. This is determined when the vehicle is in operation, not charging. The value increases from zero at the upper charge limit to a defined maximum at the lower limit of the usable SOC window.
- Time Till Charge Complete (Unit: Seconds): The amount of time required before the vehicle is fully charged. 0 seconds indicates the vehicle has reached 90% State of Charge. This metric is only applicable while the truck is charging.
- Overheat Risk (Binary Classification): Indicates if the vehicle is at risk of overheating if the vehicle were to continue operating at this level, potentially resulting in the vehicle shutting down. The classification is either normal operation or overheating risk.
- Tonnes Before Flat (Unit: Tonne): Determines the number of tonnes that can be moved before the vehicle needs to be charged. 0 tonnes indicates that the vehicle is at 10% SOC, the lower limit of the battery. The amount of tonnes it can move goes down linearly over time, regardless of the vehicle's bucket size.
- Time to Flat (Unit: Seconds): The amount of time required before the vehicle needs to be physically charged. 0 seconds indicates 10% SOC remaining, the limit for the battery before undervoltage errors occur. This does not include the time it takes to reach the charger.
- Power Usage - 5 seconds ahead (Unit: Watts): Predicts the power to be used by the vehicle 5 seconds in the future.
- State of Charge Prediction - 5 seconds ahead (Unit: %): Predicts the state of charge of the battery 5 seconds into the future.
- State of Charge Prediction - 30 seconds ahead (Unit: %): Predicts the state of charge of the battery 30 seconds into the future.
- State of Charge Prediction - 60 seconds ahead (Unit: %): Predicts the state of charge of the battery 60 seconds into the future.
- Energy Consumption per trip (Unit: kWh): Determines and predicts the energy consumption for any given trip. This trip is defined as the vehicle starting and reaching an arbitrarily defined destination.
- Chiller Anomaly Detection (Binary Classification): Detects abnormal chiller operation, categorising it as an anomaly when temperature and cooling rate are at a mismatch.

The variable set used for model training is withheld as commercially sensitive information and is available from EPCA on commercial terms.

2.2.2 Feature Engineering

In addition to the raw logged features, a number of derived features were engineered to improve model accuracy and capture temporal dynamics:

- Scaled features: Key variables such as state of charge, cell voltages, and inclination were scaled to percentage representations of their full operating range
- Time decomposition: The timestamp was decomposed into hour, minute, second, day of week, weekend indicator, and shift time to capture diurnal and shift-based operational patterns.
- State of Energy (SOE): A normalised measure of energy remaining in the pack.
- Estimated derived quantities: Operational planning metrics derived from pack energy state, including charge-time, runtime and haulage-capacity estimates.
- Instantaneous power and energy: Derived from logged electrical measurements at each timestep.
- Thermal score and cooling load: A composite score combining motor and hot spot temperatures with fan speed data, used as a derived risk indicator for the overheat detection model.
- Lagged and rolling features: Lagged values (1 and 2 timesteps back) and 3-second rolling averages for SOC, motor temperature, and pack voltage, capturing short-term temporal trends.
- Future-value features: Values from 5 and 10 seconds ahead for SOC were generated. These values were mainly used for forecasting and not included in future prediction models.

The full engineered feature set is withheld as commercially sensitive information and is available from EPCA on commercial terms.

2.2.3 Algorithm Selection

The majority of models use the Random Forest algorithm (mainly Regressor), selected for its ability to handle mixed numerical and categorical features, robustness to outliers and noisy sensor data, resistance to overfitting, and strong performance on large tabular datasets (Salman, Kalakech, & Steiti, 2024). All Random Forest models use a common estimator count and a fixed random seed for reproducibility; configuration values are withheld as commercially sensitive information and are available from EPCA on commercial terms.

Two models use alternative approaches:

- Overheat Risk: Random Forest Classifier, as the prediction target is binary (risk / no risk).
- Chiller Anomaly Detection: A Weighted LSTM Autoencoder, trained on normal operating data and using reconstruction error to identify anomalies. The model uses a fixed input window and feature weighting applied to emphasise chiller temperature and fan speed. Configuration values are withheld as commercially sensitive information and are available from EPCA on commercial terms.

Per-model configuration and feature sets are withheld as commercially sensitive information and are available from EPCA on commercial terms.

2.2.4 Model Training and Validation

Data was queried from the AWS Athena database and downloaded locally for model training. Each model's dataset was split into training and test subsets, with the test subset held out during training and used for initial validation of the models. The models all had the following characteristics from the dataset and cleaning:

- Pre-processing: Small files with less than ten seconds were removed, or zero data points were removed. Only information going out from devices was included (not information going to devices).
- Post-processing: Data points where the vehicle was in accessory mode were removed. Data points that experienced no data variation (static values) were disregarded.
- Dataset Size: 478 Sessions with various operational and charging sessions with a variety of lengths and loads.
- Dataset Split: 80% used for training, 20% used for validation. The split was done by time for the majority of models, with trips being used for the energy consumption per trip model.

Trained models and their validation results were exported and stored locally for use in the final stage of the project. The accuracy of the models and key features of the test subset can be found in the appendix to this report – Model Initial Key Features and Score. Additionally, the energy consumption per trip model used Leave-One-Out Cross-Validation (LOOCV) to account for the relatively small number of distinct trips in the training dataset.

While these initial scores are high, they were obtained on data drawn from the same operational sessions as the training set, and warrant careful scrutiny for overfitting. This is examined in Section 4.2.

- Power usage – 5s ahead model: The high MAPE (93.49%) and lower R^2 (0.8808) indicate that short-horizon power prediction is more sensitive to transient driving conditions not fully captured in the current feature set. This model also retained one future-shifted covariate (motor temperature 5 s ahead); its measured feature importance is negligible (Appendix), but its initial score should be read with that limitation in mind.
- Chiller anomaly detection: The recall of 0.5334 indicates the model identifies approximately half of genuine anomalies in the test set, while maintaining perfect precision. Stage 4 validation against new operational data with additional anomalous events was expected to further assess and refine this model.

2.3 Data Visualisation Development

Stage 3 developed an interactive data visualisation platform linking the AWS S3 to a Microsoft Power BI dashboard, enabling EPCA engineers and mine site operators to access, explore, and compare truck performance data across any time window and across all onboard systems.

2.3.1 AWS Athena to Power BI Integration

Linking Power BI to AWS Athena required the deployment of a Microsoft On-premises Data Gateway. The gateway acts as a secure bridge between the AWS cloud environment and Power BI, enabling near-live and scheduled data refresh. Multiple gateway instances were configured under a single gateway cluster to provide redundancy. An ODBC connection to Amazon Athena was configured on the gateway hosts using a managed data source definition and stored credentials. Figure 4 shows the data being retrieved from AWS S3 through the use of the newly created Data Gateway.

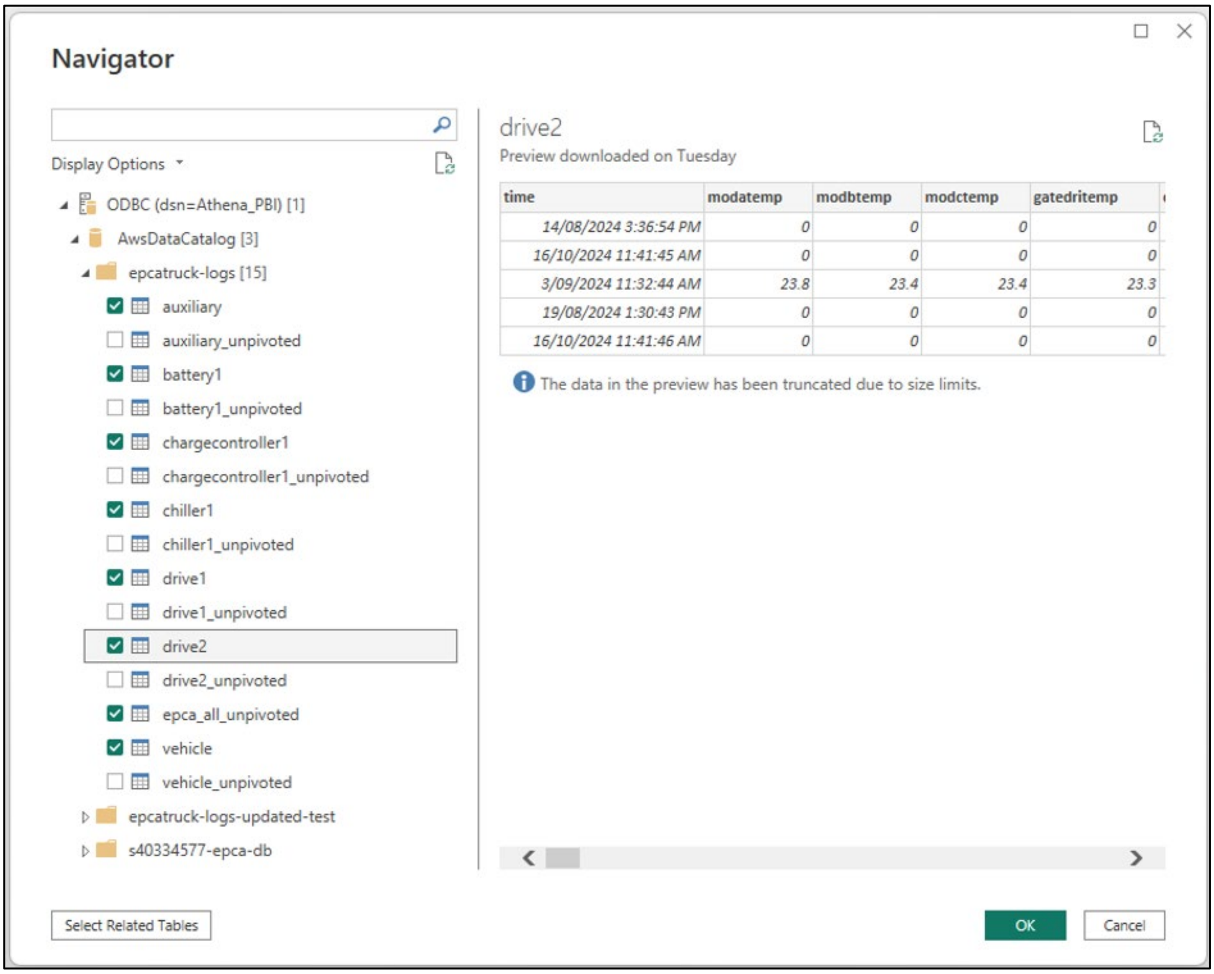


Figure 4: *Data being requested by Power BI from AWS Athena.*

2.3.2 Data Transformation and Incremental Refresh

Power Query code was written within Power BI to merge all individual system data tables into a single unified comparison table, retaining the key fields from each system required for visualisation. Incremental refresh was configured using RangeStart and RangeEnd parameters, ensuring that only new data within the incremental window is loaded from AWS Athena on each refresh cycle. The Tabular Editor external tool was used for post-publication modifications to the semantic model’s dataset structure.

2.3.3 Dashboard Functionality

The completed Power BI dashboard provides variable selection across all onboard systems, time-series graphs and tabular views with statistical summaries, adjustable time range filtering, multi-variable overlay for cross-system comparison, and a full-screen graph mode for detailed analysis. Figure 5 to Figure 11 show the Power BI dashboard in use. Note that the number of data points listed changes based on the zoom level for each graph.

Figure 5 shows the dashboard's default view, with all auxiliary system data displayed across the full recorded time range. This is the starting point a user sees when a dataset is first opened, before any variables or time windows have been selected. It illustrates the volume and density of data available for a single system.

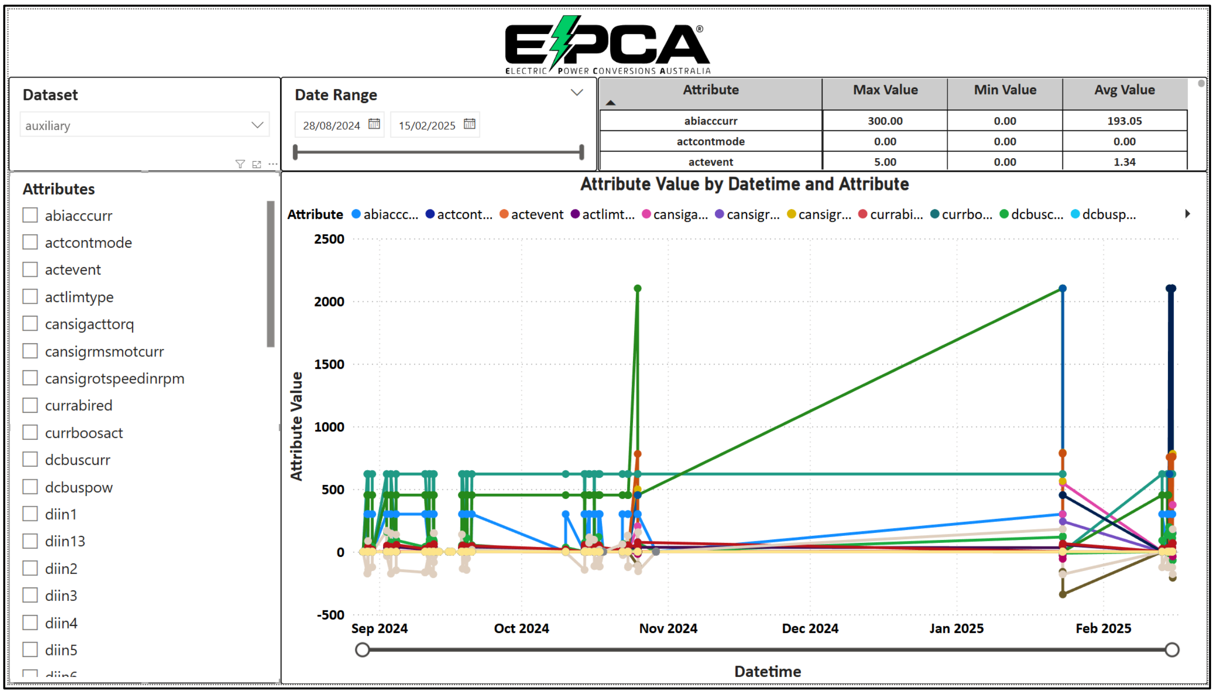


Figure 5: *Dashboard dataset - all auxiliary data displayed.*

Figure 6 demonstrates the adjustable time range filter. The full session data is narrowed to a shorter window using the data range selection, selecting only to show the 28/08/2024 – 13/10/2024 data points. This selection can also be used to show only one day as well.

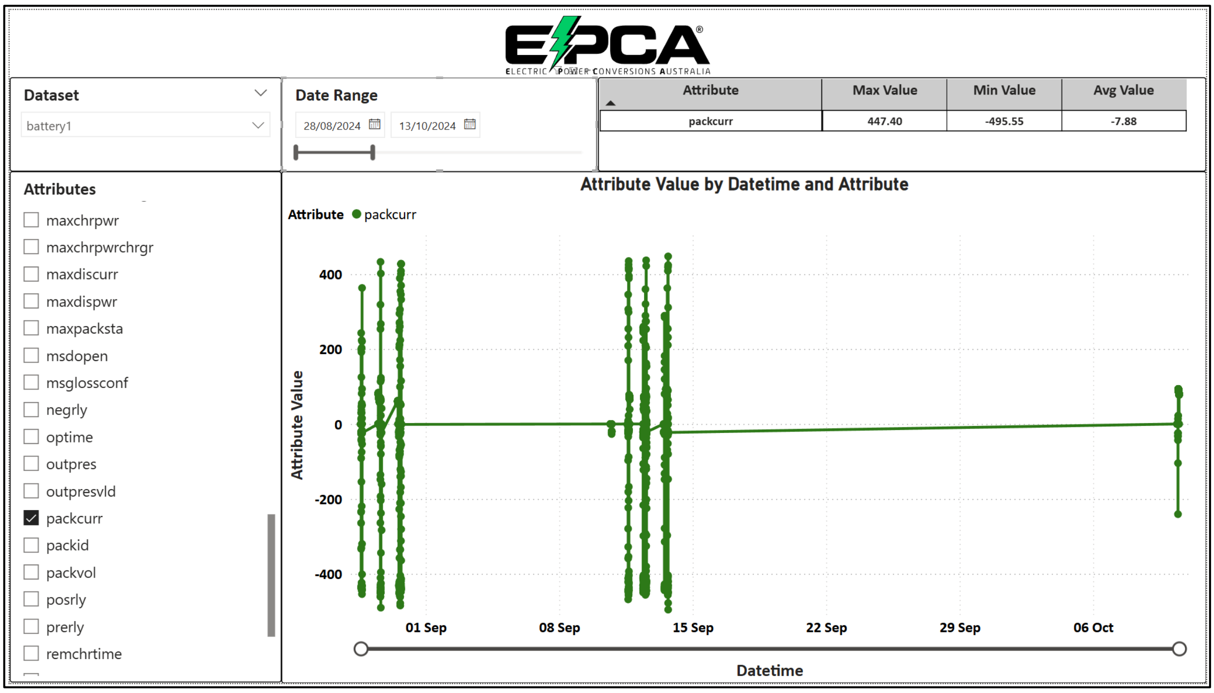


Figure 6: *Dashboard displaying pack current from battery with reduced time range.*

Figure 7 shows the multi-variable overlay function, with two attributes (battery status and humidity) plotted on the same chart and time axis. This allows a user to check whether two variables move together or diverge, without needing to export data or build a separate chart for each one.

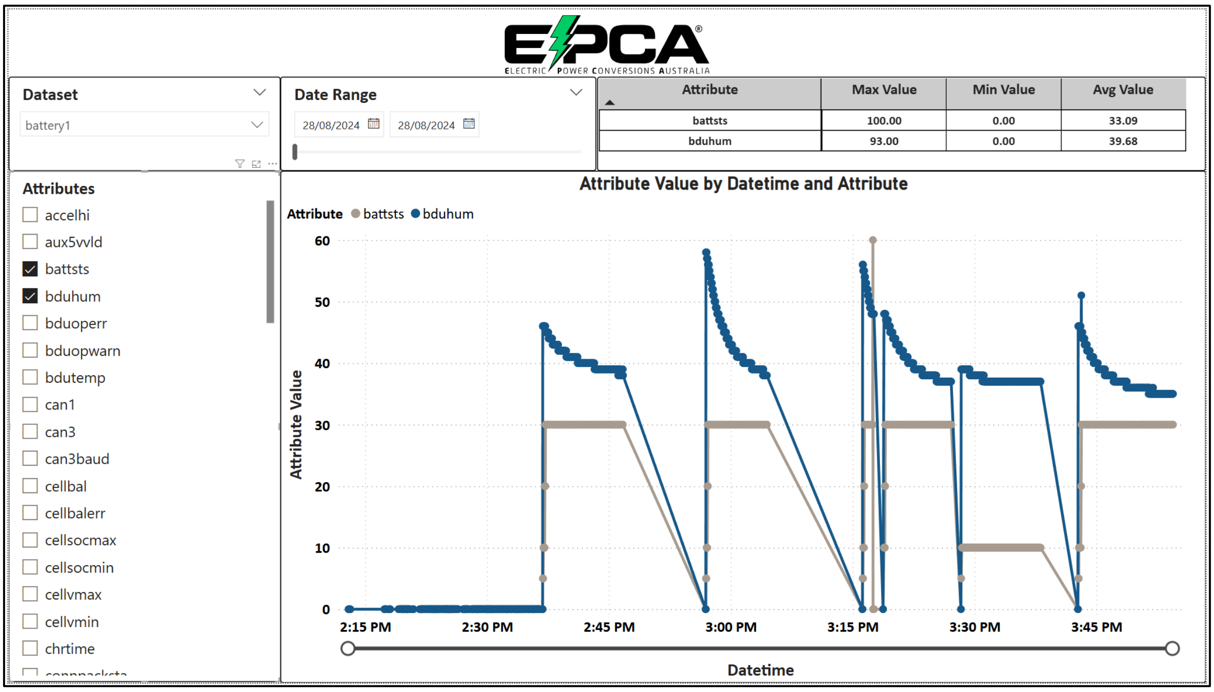


Figure 7: *Dashboard display comparing two data points.*

Figure 8 extends the overlay function across systems, comparing temperature readings from the auxiliary, drive, and chiller datasets on a single chart. This cross-system comparison is useful for diagnosing whether a temperature change in one system corresponds to a change elsewhere, such as a chiller response to rising drive temperatures. Figure 9 also shows the same dataset as Figure 8, zoomed in further to a window of only a few minutes. This illustrates the dashboard's range, from broad session-level trends down to second-by-second detail, using a secondary zoom level beneath the graph's x-axis.

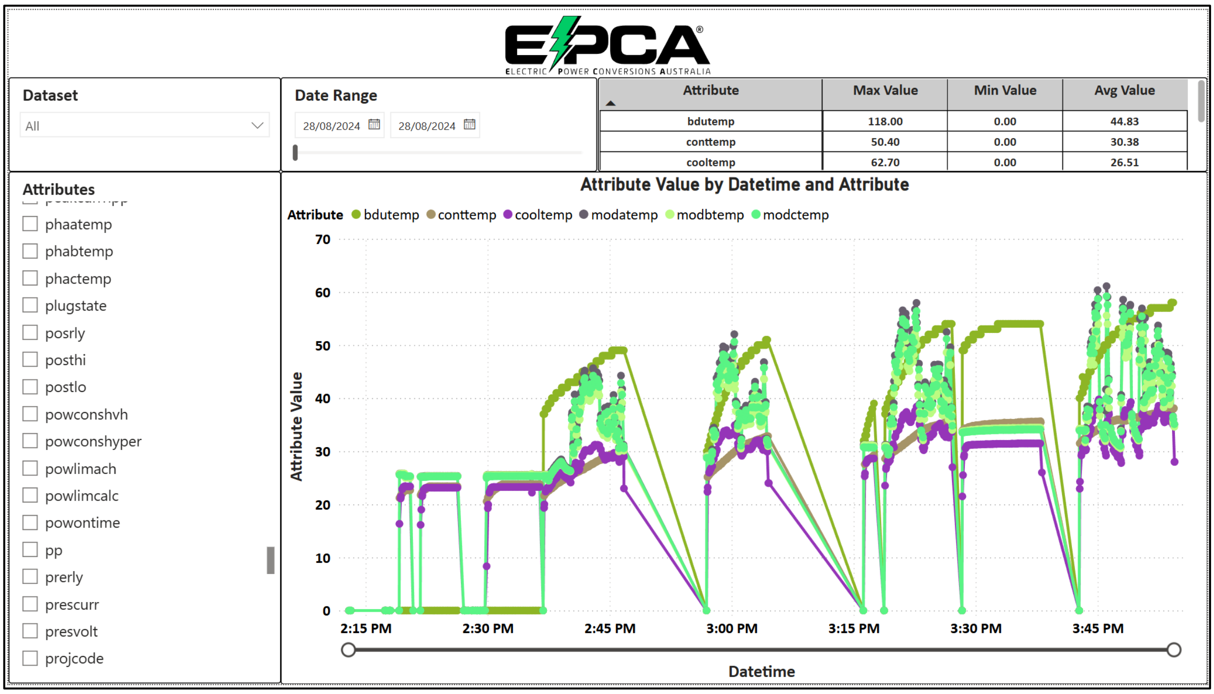


Figure 8: *Dashboard display comparing temps from different systems.*

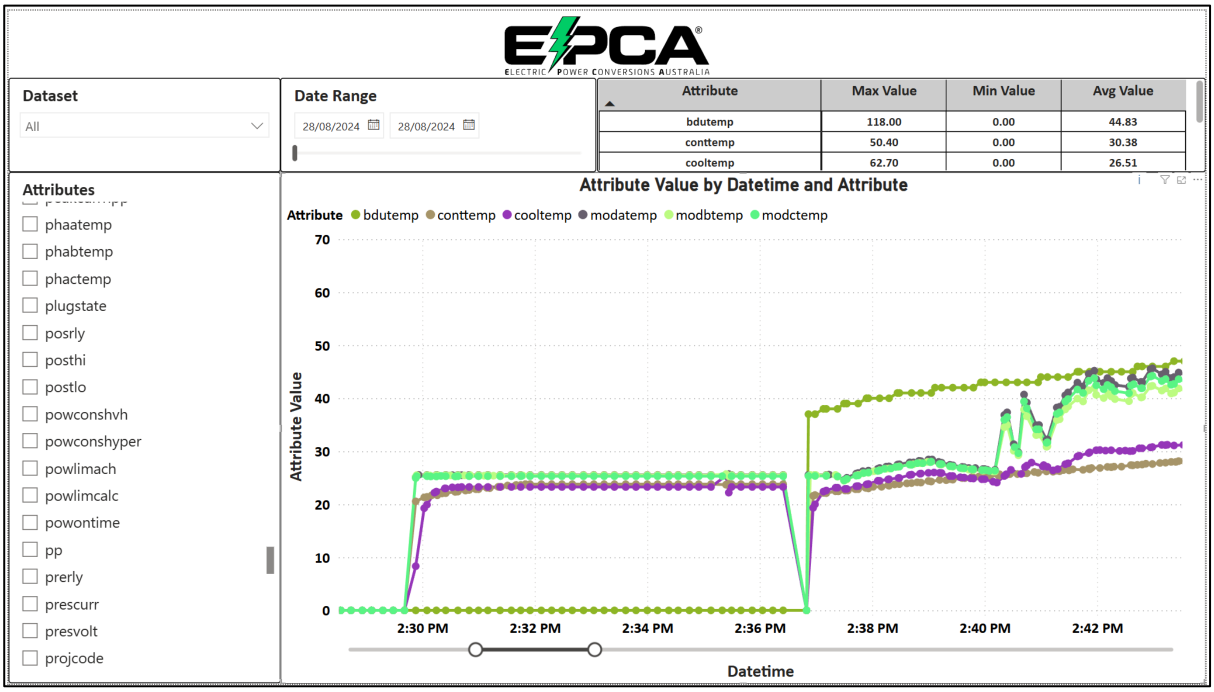
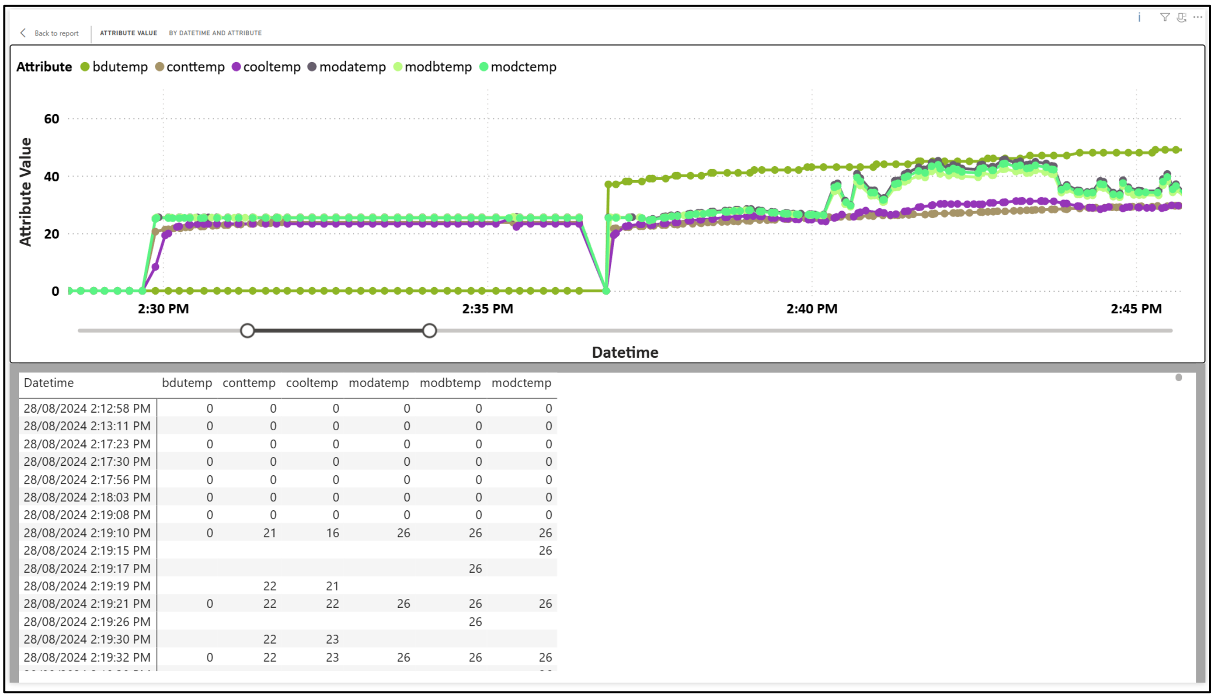


Figure 9: *Dashboard comparing temps from different systems (dataset further zoomed in).*

Figure 10 shows the table in graph mode, which displays the plotted chart alongside a data table listing the exact timestamped values behind it. This allows a user to move between a visual trend and the underlying figures directly, without switching views or exporting data separately. Figure 11 shows the graph in Fullscreen mode of the same data.



Datetime	bdutemp	conttemp	cooltemp	modatemp	modbtemp	modctemp
28/08/2024 2:12:58 PM	0	0	0	0	0	0
28/08/2024 2:13:11 PM	0	0	0	0	0	0
28/08/2024 2:17:23 PM	0	0	0	0	0	0
28/08/2024 2:17:30 PM	0	0	0	0	0	0
28/08/2024 2:17:56 PM	0	0	0	0	0	0
28/08/2024 2:18:03 PM	0	0	0	0	0	0
28/08/2024 2:19:08 PM	0	0	0	0	0	0
28/08/2024 2:19:10 PM	0	21	16	26	26	26
28/08/2024 2:19:15 PM						26
28/08/2024 2:19:17 PM					26	
28/08/2024 2:19:19 PM		22	21			
28/08/2024 2:19:21 PM	0	22	22	26	26	26
28/08/2024 2:19:26 PM					26	
28/08/2024 2:19:30 PM		22	23			
28/08/2024 2:19:32 PM	0	22	23	26	26	26

Figure 10: *Dashboard table and graph mode.*

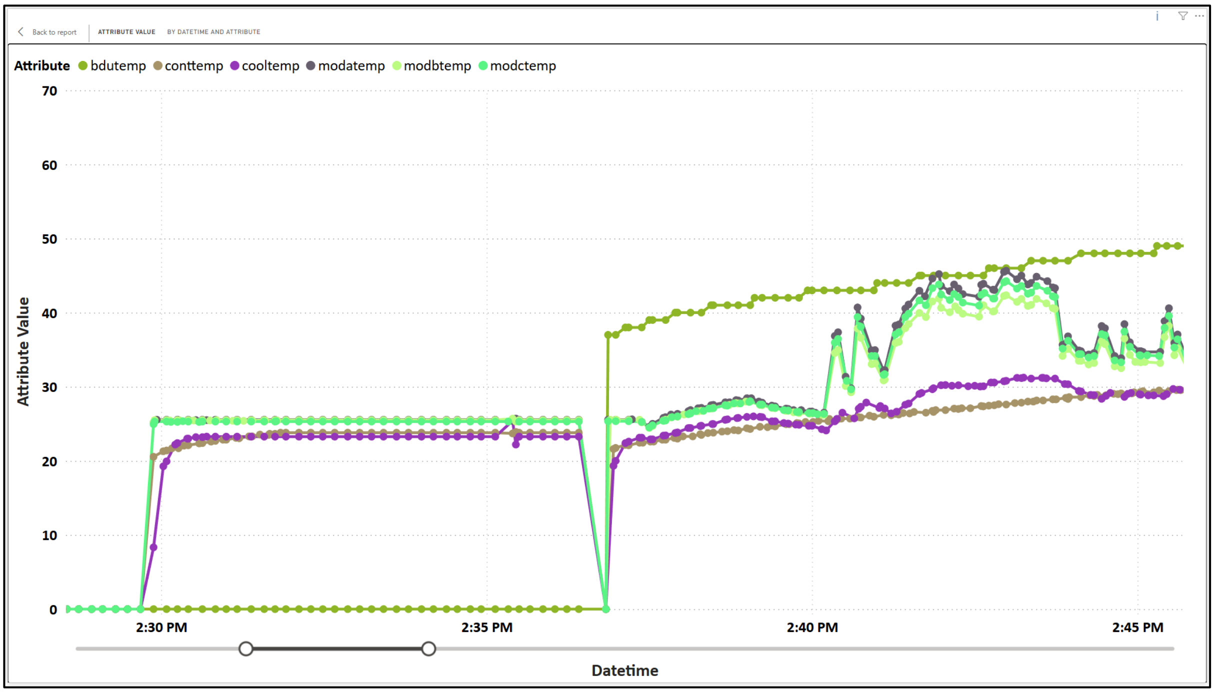


Figure 11: *Graph fullscreen mode.*

2.4 Model Validation and Verification

To validate the machine learning models developed, an independent dataset was captured during a new set of live truck operations at the Bakers Hill mine site. This final verification confirmed whether the model performance achieved on the held-out training dataset generalises to new operational conditions and identified any models requiring further tuning or retraining.

The validation dataset was collected during ten dedicated sessions at the Bakers Hill sand pit mine site: six haul-cycle sessions, one fault-induced chiller test session and three charging sessions. Each haul-cycle session involved the E-777D truck completing multiple loaded and unloaded haul cycles on the same circuit used during the original trials, covering a 1 km haul road with a gradient of approximately 7-15% depending on the position on the track. This route was the original test route where the original data was collected for training the model.

Figure 12: Bakers Hill test loop. *Imagery ©2026 Google. Reproduced under Google Maps and Google Earth brand permissions. This image is third-party material and is not covered by the Creative Commons Attribution 4.0 International Licence applying to this publication.*

Refer to Figure 12 for the test route. Sessions were conducted across a range of ambient temperatures to test model performance under varying thermal conditions, whilst Table 1 lists all the tests that were to be performed.

Table 1: *Test session list.*

Test Session	Conditions
1	6 Loaded Cycles (Up and Down the Hill)
2	6 Loaded Cycles (Up and Down the Hill)
3	6 Loaded Cycles (Up and Down the Hill)
4	6 Unloaded Cycles (Up and Down the Hill)
5	6 Unloaded Cycles (Up and Down the Hill)
6	6 Unloaded Cycles (Up and Down the Hill)
7	Chiller Pump Fault Induced Test
8	Charging
9	Charging
10	Charging

The final validation followed the same preprocessing pipeline established previously. New session data was converted from CSV to Parquet and uploaded to the AWS S3 buckets under a dedicated Stage 4 partition. AWS Athena queries were used to extract the new test data locally, and the same feature engineering transformations applied during initial training were applied to the new dataset to ensure consistency. The trained models were loaded from their saved export files and applied directly to the new test feature set without retraining, producing predictions for each target variable at each timestep. Predicted values were compared against actual measured values using the same metrics computed during the initial stage: R^2, MAE, and MAPE for regression models; accuracy, precision, recall, and F1 score for classification models.

3 Results and Findings

3.1 Data Pipeline

The data ingestion and storage pipeline was successfully developed and tested across all onboard data systems of the EPCA E-777D truck with the new validation dataset acquired. All available historical data from EPCA's operational trials was converted from CSV format to Parquet, uploaded to AWS S3, and registered with AWS Athena for querying. From there, the data was then put through the machine learning models for validation and was viewed from the EPCA dashboard. The final key outcomes achieved from the pipeline were the following:

- Automated preprocessing pipeline successfully processes all seven system data types, handling empty files, duplicate records, and timestamp extraction.
- All EPCA E-777D trial data successfully converted to parquet format and uploaded to AWS S3.
- AWS Athena queries validated and confirmed functional across all system partitions.
- Processing-state and runtime logs implemented for pipeline auditing
- Machine learning algorithm successfully queried and utilised new data added to AWS Athena to verify models.
- New data displayed on EPCA dashboard with machine learning model outputs.

The pipeline is extensible, allowing additional truck operational data to be ingested for ongoing validation and fleet-level analytics without modification to the core processing.

3.2 Trial Results and Verification

The data acquired during the new Trials were taken along the same track the original data was obtained from, as outlined in section 2.4. Table 2 shows that a total of approximately 46,920 seconds of operational and charging data was collected across the multiple sessions. Sessions cover loaded and unloaded haul cycles, a fault-induced chiller test, and three charging cycles, across ambient temperatures ranging from 28.2°C to 38.2°C.

Table 2: *Test Times and Conditions for the new Trials*

Test Session	Start Time	End Time	Ambient Average (°C)	Haul Cycle
1	09:14	10:20	30.4	6 Loaded Haul Cycles – nominal rated payload
2	14:00	15:15	36.7	6 Loaded Haul Cycles – nominal rated payload
3	09:08	10:05	28.2	6 Loaded Haul Cycles – nominal rated payload
4	14:12	15:30	32.8	6 Unloaded Haul Cycles
5	09:06	10:21	30.3	6 Unloaded Haul Cycles

Test Session	Start Time	End Time	Ambient Average (°C)	Haul Cycle
6	14:30	15:50	38.2	6 Unloaded Haul Cycles
7	16:20	16:30	37.4	Chiller Pump Fault-Induced Test
8	11:01	13:03	34.3	Charging – partial cycle to upper charge limit
9	11:30	13:20	30.4	Charging – partial cycle to upper charge limit
10	11:21	13:10	35.7	Charging – partial cycle to upper charge limit

Notable differences from the training dataset include two sessions conducted with high ambient temperatures exceeding 30°C that only had limited data in the training dataset, and one session in which a minor chiller pump intermittent fault was deliberately induced and logged before being rectified. These differences were considered important context when interpreting model performance against the new data. The following table shows the results obtained from the newly acquired dataset being determined by the current model, where all sessions were passed through the model, not just individual test sessions. Table 3 summarises the results and compares it with the original validation results from the test set.

Table 3: *New validation results compared with original results*

Model	Original Validation Results	New Validation Results
Charge Time Estimate	R^2: 0.9980 | MAE: 1,093 s | MAPE: 0.11%	R^2: 0.6814 | MAE: 12,440 s | MAPE: 18.73%
Time Till Charge Complete	R^2: 0.9748 | MAE: 53.1 s | MAPE: 21.78%	R^2: 0.7916 | MAE: 136.1 s | MAPE: 44.41%
Overheat Risk	Acc: 0.9972 | Prec: 0.958 | F1: 0.910	Acc: 0.6638 | Prec: 0.381 | F1: 0.320
Tonnes Before Flat	R^2: 0.9990 | MAE: 0.15 t | MAPE: 0.36%	R^2: 0.7241 | MAE: 8.94 t | MAPE: 22.81%
Time To Flat	R^2: 0.9989 | MAE: 0.11 s | MAPE: 0.36%	R^2: 0.6973 | MAE: 1,284 s | MAPE: 27.44%
Power Usage – 5s Ahead	R^2: 0.8808 | MAPE: 93.49%	R^2: 0.2214 | MAPE: 481.2%
SOC Prediction – 5s Ahead	R^2: 0.9937 | MAE: 0.283 % | MAPE: 0.54%	R^2: 0.8841 | MAE: 2.143 % | MAPE: 4.87%
SOC Prediction – 30s Ahead	R^2: 0.9937 | MAE: 0.321 % | MAPE: 0.77%	R^2: 0.7188 | MAE: 5.912 % | MAPE: 13.20%
SOC Prediction – 60s Ahead	R^2: 0.9911 | MAE: 0.389 % | MAPE: 0.81%	R^2: 0.5534 | MAE: 9.441 % | MAPE: 21.67%
Energy Consumption per Trip	R^2: 0.991	R^2: 0.8813
Chiller Anomaly Detection	Prec: 1.000 | Recall: 0.533 | F1: 0.686	Prec:0.412 | Recall: 0.214 | F1: 0.281

Every model shows a significant drop in accuracy; the SOC prediction – 5s ahead model has the best results of all the models, while power usage – 5s ahead, time to charge completion, and chiller anomaly detection show the largest declines. The following graphs are illustrative outputs showing each model's predictions against actual values from a single validation session — test session 1, except the time till charge complete model, which is shown for charging session 8.

The charge time estimate model in Figure 13 consistently has fluctuating values even during a lack of variability in the actual results. Although the general trend of upward change is still observed.

Figure 13: *Charge time estimate model prediction for test session 1.*

The time till charge completion model, shown in Figure 14 for charging session 8, shows similar results to the previous graph, where the trend line is observed from the predicted result, but the values themselves vary widely. The values seem to get more erratic towards the end of charging, but it's hard to tell as there should be no prediction below 0 seconds.

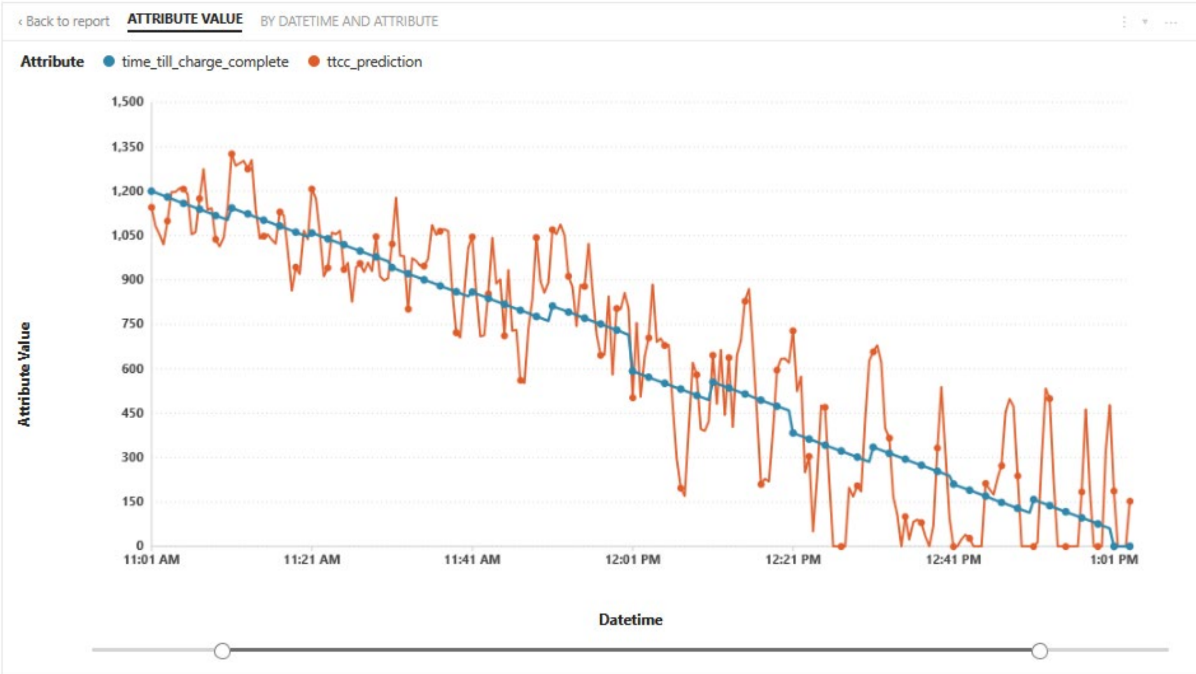


Figure 14: *Time Till Charge Complete model prediction for test session 8 (charging).*

Figure 15 shows the overheating risk model. It can be seen that the model overpredicts overheating risks significantly, indicating a systematic bias in the model's classification threshold under these conditions. There does not appear to be any trend regarding whether the overheat risk predictions match the actual overheat risks.

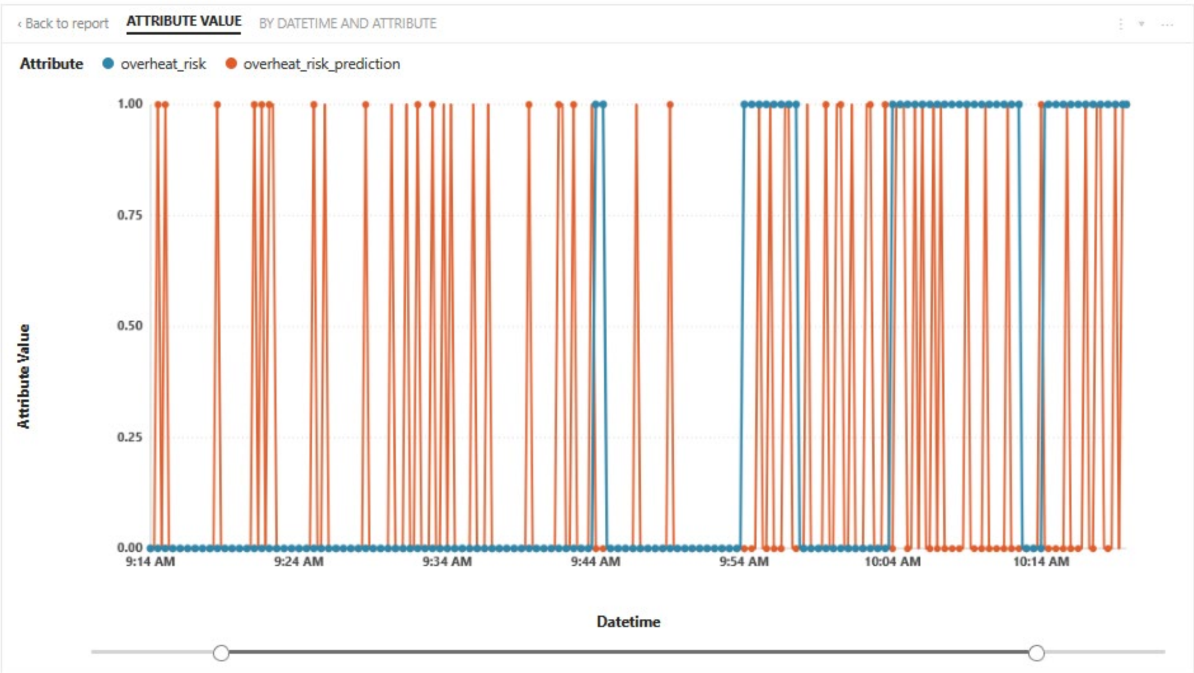


Figure 15: *Overheating risk for test session 1.*

Figure 16 shows the total tonnes moved before the battery flat, with predicted tonnage tracked against actual tonnage across six loaded haul cycles. The truck gains tonnes due to the regeneration of the vehicle. The two lines stay aligned throughout the session, showing a similar trend in the data but fail to maintain decent consistency.

Figure 16: *Total Tonnes Moved Before Battery Flat for test session 1.*

Figure 17 compares predicted and actual time remaining before the battery is depleted, over the course of test session 1. The downward trend is maintained, but the predicted value oscillates significantly more than the actual time to flat values.

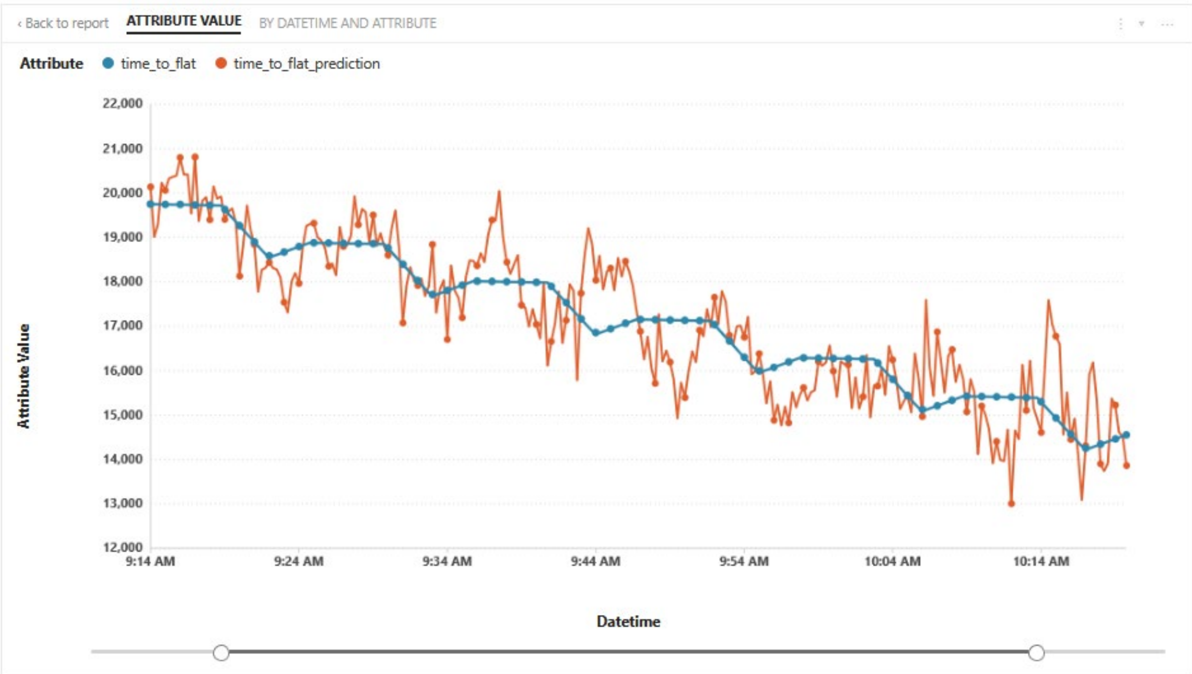


Figure 17: *Time to Battery Flat for test session 1.*

Figure 18 displays the power usage – 5s ahead model. The prediction line is visibly more irregular than the actual line, particularly during uphill and regenerative braking phases, where the model also gets the direction of power flow wrong. The general trend and movement of the data is still accurate.

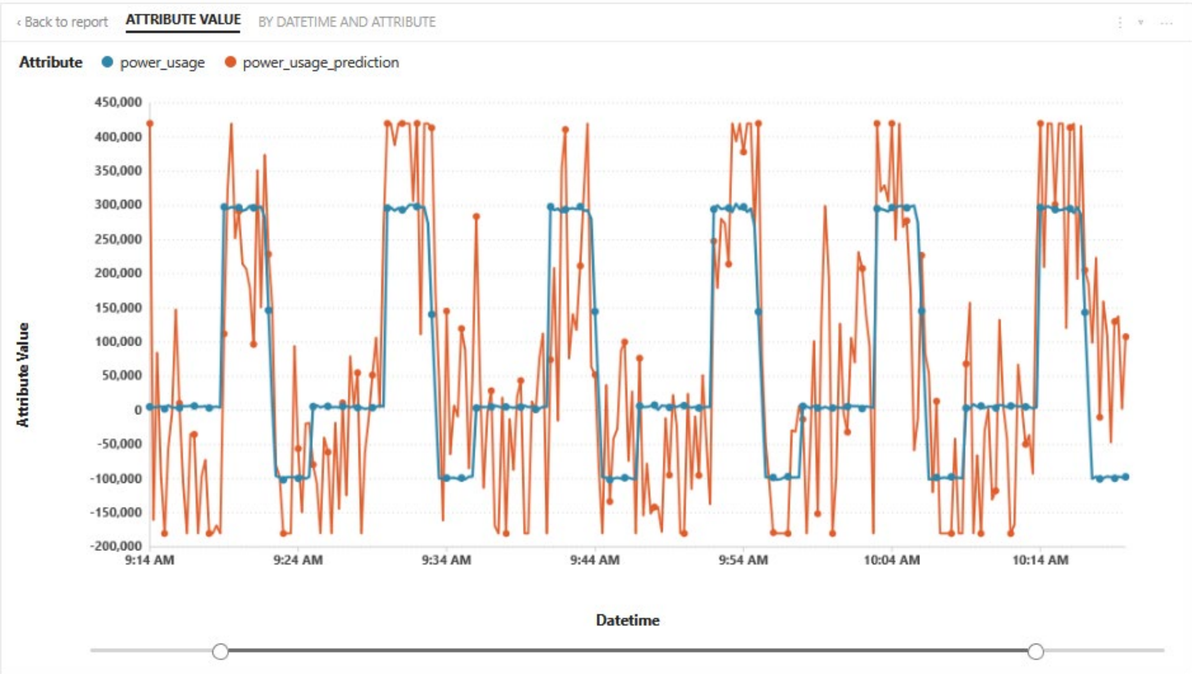


Figure 18: *Power Usage – 5s Ahead for test session 1.*

Figure 19 shows state of charge prediction – 5 seconds ahead model against the actual state of charge. The two lines closely track one another throughout the session compared to all the models. It mostly follows the trend of the actual values but overshoots and undershoots when the truck is going uphill and regenerating downhill.

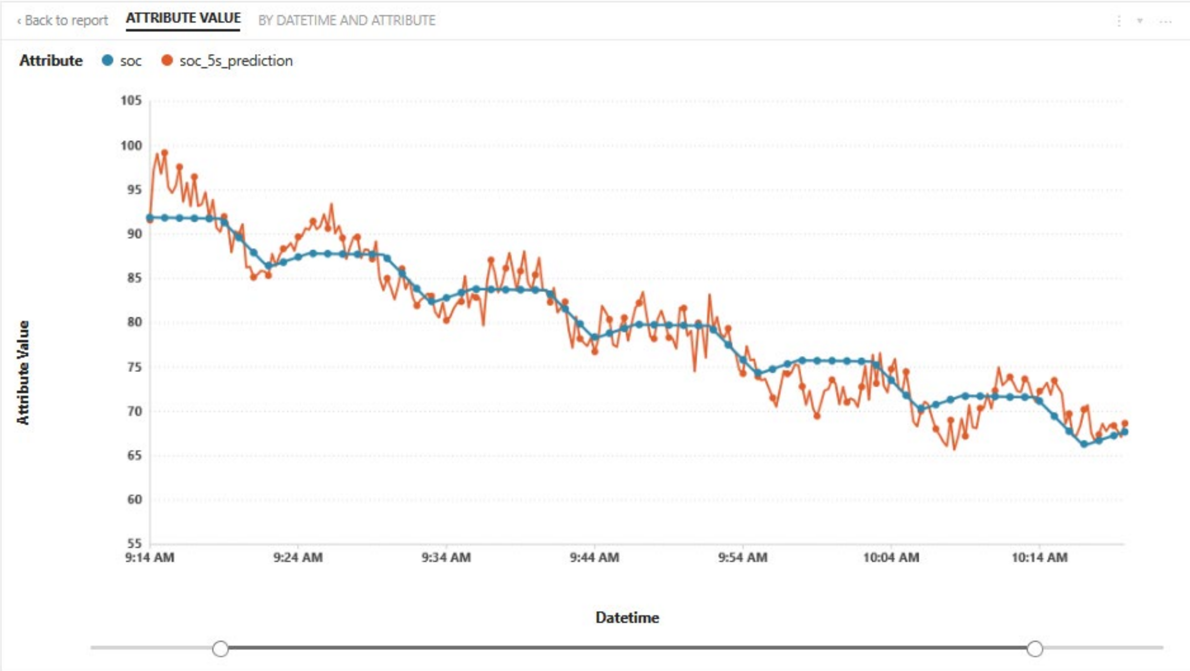


Figure 19: *SOC Prediction – 5 Seconds Ahead.*

Figure 20 shows that with SOC prediction - 30 seconds ahead model, the prediction shows slightly more deviation than the 5-second model, particularly during periods of rapid regeneration or discharge.

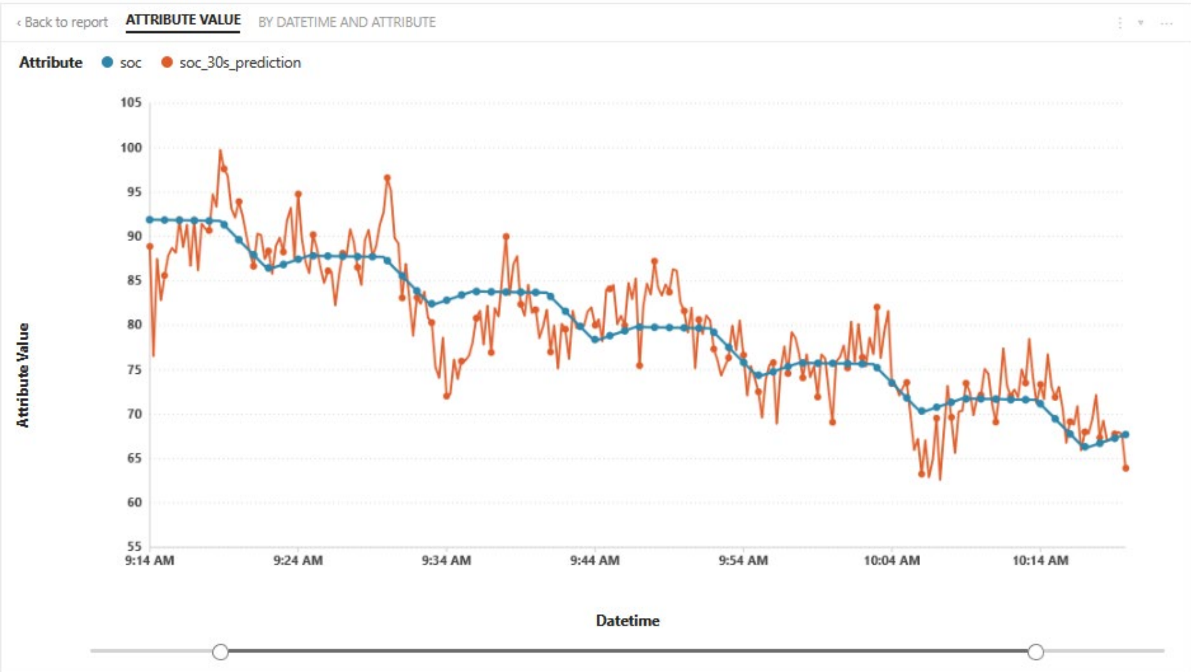


Figure 20: *SOC Prediction - 30 Seconds Ahead.*

Figure 21 displays the final SOC prediction – 60 seconds ahead; the prediction shows slightly more deviation than the 5sec and 30sec model. Deviation increases with this longer prediction horizon, which is expected with prediction error compounding over a longer time window.

Figure 21: *SOC Prediction - 60 Seconds Ahead.*

Table 4 shows predicted versus actual energy consumption for each of the six haul cycles in the test session 1. The model predicts energy for a full trip, based on it going uphill and downhill with a certain amount of load. The model's output reflects a full-loaded and return trip, which is still widely inconsistent compared to the actual measured consumption. Because the model was trained on loaded-plus-return trips and is validated here against one-way loaded cycles, this comparison reflects a validation-design mismatch as much as model generalisation, and should be read accordingly.

Table 4: *Energy Consumption Per Trip From Session 1. Values are indexed to the measured consumption of trip 1 (trip 1 actual = 100). Measured consumption varied by less than 1.5% across the six trips.*

Trip Number	Actual (indexed)	Predicted (indexed)
1	100.0	71.5
2	99.9	84.4
3	98.5	116.1
4	99.5	156.7
5	99.7	152.3
6	99.4	68.2

The final graph shown in Figure 22 is the chiller anomaly prediction model. There are only a few false positives compared to the actual chiller anomaly results. Granted, no anomalies actually occurred during test session 1.

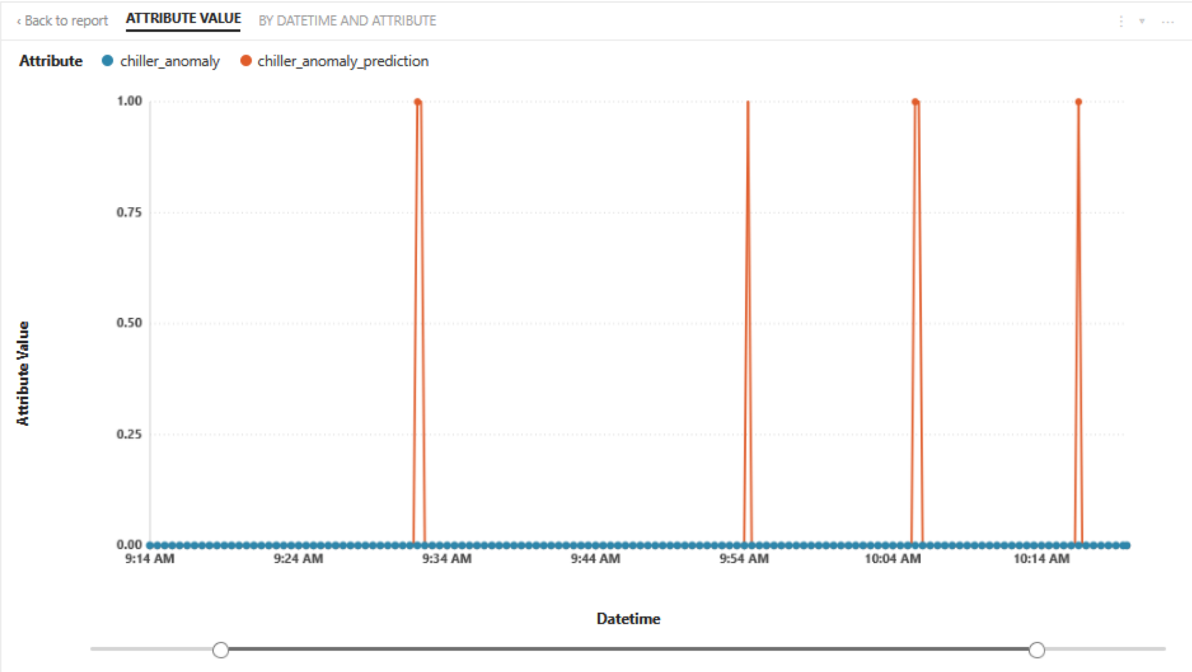


Figure 22: *Chiller Anomaly Prediction model from test session 1.*

4 Discussion

4.1 Data Infrastructure and Pipeline

The AWS-based data infrastructure developed provides a scalable and reliable foundation for EPCA's ongoing data analytics capability. The choice of parquet format appears well-suited to the volume and columnar structure of the truck's telemetry data, enabling querying via AWS Athena without a traditional relational database system. The automated preprocessing pipeline significantly reduced the manual effort previously required to prepare data for analysis and enables quick ingestion of new operational data. The partitioning strategy adopted for the S3 data lake ensures that query performance will be efficient (Amazon, n.d.), but as data volume grows with additional truck operating hours and future fleet deployments, the querying time and speed will need to be tested and potentially optimised.

4.2 Initial Machine Learning Model Performance

Across the prediction figures, the model output appears more irregular than the actual measured values. This is a result of how the predictions are generated rather than a reflection of irregular truck behaviour. The actual values are derived from direct sensor readings that follow the truck's real, continuous operating state. The model instead produces an independent prediction at each one-second timestep based on that moment's feature values (such as throttle, current, voltage, or incline), without reference to what it predicted a moment earlier (Salman, Kalakech, & Steiti, 2024). As a result, small fluctuations in the input data from one second to the next appear as visible noise in the predicted line, even during periods when the truck itself is behaving smoothly. This effect is more pronounced in models with lower accuracy scores, such as power usage - 5s ahead and time until charge complete, and less visible in models such as SOC prediction - 5s ahead, which rely on lagged and rolling features that give the model some awareness of the recent trend.

The high R^2 and low MAPE values achieved by the majority of models on the held-out validation dataset demonstrate that the random forest algorithm is well-suited to the structured, high-frequency tabular data generated by the E-777D truck. The feature engineering process helped in aiding some predictions, with the main contributing feature being cell voltage minimum scaled, being present highly in 5 of the major models (time before flat, tonnes before flat and all SOC predictions). The majority of the models appear to heavily rely on a singular value in the dataset, which is concerning as it potentially suggests overfitting to one value and not truly capturing the real-world trend. Although it should be noted that the majority of features are present for each of the models, only a handful of them would be relevant for its prediction; for example, the SOC prediction should have the cell voltage and current from the battery as the main variables looked at. The only model to deviate from the overreliance on one value is overheat risk, which is most likely due to overheat risk needing to take into account multiple temperatures of a variety of devices.

4.3 Data Visualisation Platform

The Power BI dashboard provides EPCA's engineering team with visuals for understanding truck operational behaviour and trends in the data. Prior to this project, analysis of truck performance data required manual extraction and processing of individual CSV files. The dashboard eliminates this barrier, enabling engineers to visualise any combination of variables from any operational session within a reasonable amount of time. The incremental refresh capability ensures the dashboard remains current as new data is ingested, and it will become increasingly important for performance analysis as the data volume grows over time.

4.4 Model Validation

The validation results reveal a consistent and significant pattern of model performance degradation when applied to the new live operational data. While the models performed strongly on the held-out portion of the initial training data, a dataset drawn from a similar set of operational sessions and conditions as the training data itself, several key differences in the dataset exposed the limits of the current models' generalisation capability.

One of the most significant sources of degradation appears to be the limited diversity of the training dataset in the modelling approach. The training data was collected across a relatively narrow range of operating conditions: moderate ambient temperatures, consistent haul cycle profiles and limitations of the initial prototype machine. The new validation dataset introduced conditions that were either absent or severely underrepresented in training, such as high ambient temperatures and fast charging. Models that rely heavily on thermal features, particularly the power usage – 5s ahead and overheat risk models, were most affected, as the relationship between thermal inputs and the target variable changes meaningfully under high-ambient conditions. The drastic change in accuracy of the models from the validation test set and the initial test set heavily suggests that the model was overfitted to only work on the initial test set and not represent real-world changes.

The SOC prediction - 5s ahead model was the only model to maintain acceptable performance across all test sessions, which is consistent with the model's near-instantaneous prediction horizon and its reliance on cell voltage that is relatively invariant to the majority of operating conditions (except under heavy load or fast charging). Against the project's committed accuracy target of state-of-charge prediction within ±5%, the 5-second-ahead model met the target on the independent validation data (MAE 2.14%), while the 30-second and 60-second horizons did not (MAE 5.91% and 9.44% respectively). The longer-horizon SOC prediction models (30s and 60s) showed progressively greater degradation, reflecting the compounding of prediction error over longer time windows and greater sensitivity to data not represented in training.

The results are not fit for unsupervised operational use in their current form, confirming that models trained on a narrow range of operating conditions do not generalise reliably to the fuller range encountered in service. Additionally, post-tuning of the models after testing with the new validation set would help improve prediction accuracy. The majority of models require retraining on a more diverse and representative dataset before their predictions can be relied upon across the full range of conditions encountered at an active mine site.

5 Conclusions

This report presents the complete outcomes of MRIWA Project M10644, which set out to develop and validate a data visualisation and machine learning platform for the EPCA E-777D 100-tonne electric haul truck.

Stage 1 successfully established an automated, cloud-based data pipeline capable of ingesting, processing, and storing all onboard telemetry data from the E-777D truck in a structured and queryable format. The pipeline is adaptable and extensible to accommodate future operational data from additional truck deployments.

Stage 2 developed eleven machine learning models targeting key operational performance parameters. The majority of models achieved R^2 values exceeding 0.99 on the held-out validation dataset, demonstrating strong predictive capability across SOC prediction, operational planning, and safety monitoring applications.

Stage 3 delivered a fully functional Power BI dashboard directly linked to the AWS data lake, providing EPCA's engineering team with a powerful and accessible tool for historical performance monitoring across all onboard systems.

Stage 4 completed the project with validation of the machine learning models against an independent operational dataset from live truck trials. The validation characterised the generalisation performance of the models under real operating conditions and confirmed that the data infrastructure, pipeline and visualisation platform operated reliably throughout the trials.

The Stage 4 validation against live operational data from Bakers Hill confirmed that the models, while demonstrating strong performance on initial training data, require further development and a substantially expanded dataset before they can be considered ready for unsupervised operational deployment. This outcome is consistent with expectations for a first-generation machine learning platform applied to a novel vehicle type with limited operational history.

The project has successfully established the complete technical foundation - data infrastructure, feature engineering, model architecture, and visualisation - required to close this gap efficiently as the EPCA E-777D accumulates additional operating hours. The findings of this project provide the Western Australian mining industry with a clear, evidence-based roadmap for deploying data-driven performance prediction in battery-electric haul truck operations, and demonstrate that the technical barriers to doing so are tractable with continued systematic data collection and model development.

6 Recommendations for further work

Based on the findings from all four project stages, the following recommendations are made for further development of the platform and the broader electric truck program:

6.1 Model Refinement and Retraining

- Continuous retraining: As the truck accumulates additional operating hours and as more trucks are deployed, models should be periodically retrained on the expanded dataset to maintain and improve prediction accuracy across evolving operational profiles.
- Expand training dataset diversity: Collecting data across a wider range of operating conditions, specifically targeting peak summer ambient temperatures (>40°C), partial and interrupted charging cycles, varied payload weights, and a curated set of documented equipment fault events for anomaly model training, should result in more accurate models.

6.2 Data Infrastructure Expansion

- Fleet-level integration: The data pipeline and model architecture could be extended to accommodate data from additional EPCA truck deployments, with truck-specific identifiers and site-specific parameters incorporated into the data schema for fleet-level analytics.
- Real-time inference: Future development should explore integration and development of real-time model inference into the truck's VCU or a connected edge device, enabling live performance alerts and decision support for operators in the field.

6.3 Visualisation Platform Enhancement

- Predictive overlay: The Power BI dashboard should be extended to display machine learning model predictions alongside actual measured values, enabling operators to compare predicted and actual performance in near-real-time.
- Alert and notification system: A threshold-based alert mechanism should be integrated into the dashboard, notifying operators when the overheat risk or chiller anomaly models exceed defined confidence thresholds.
- Mobile accessibility: Future iterations of the dashboard should be optimised for mobile device access, enabling mine site operators to monitor truck performance from the field.

7 References

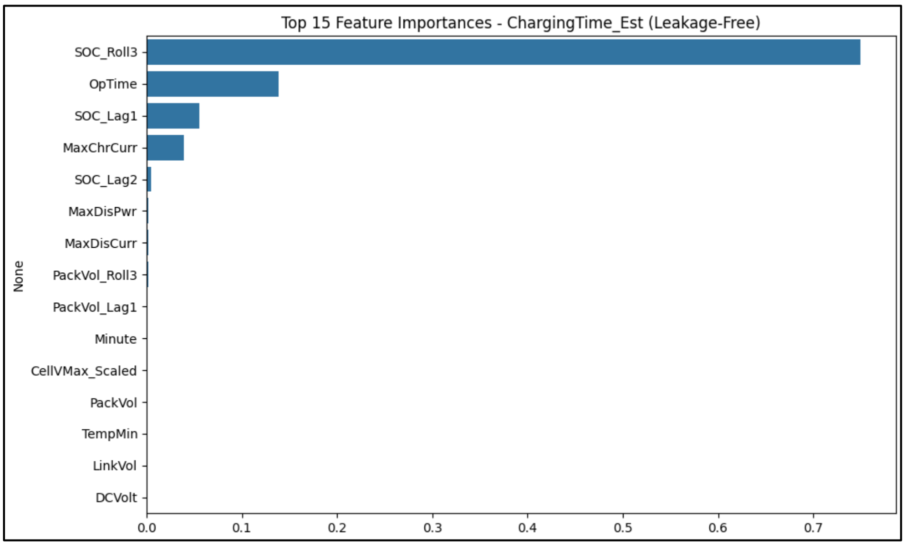
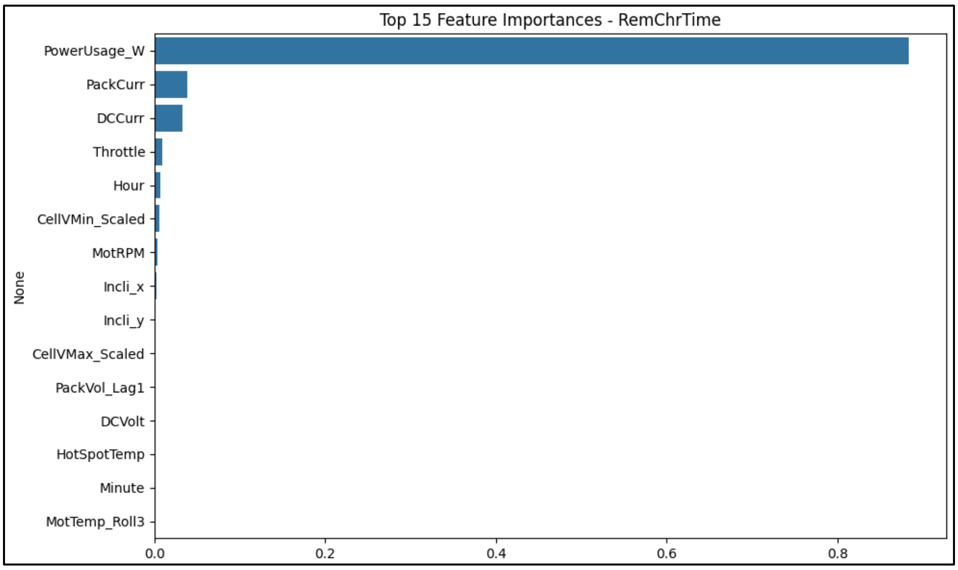
Amazon. (n.d.). *Optimize Data*. Retrieved from AWS: https://docs.aws.amazon.com/athena/latest/ug/performance-tuning-data-optimization-techniques.html

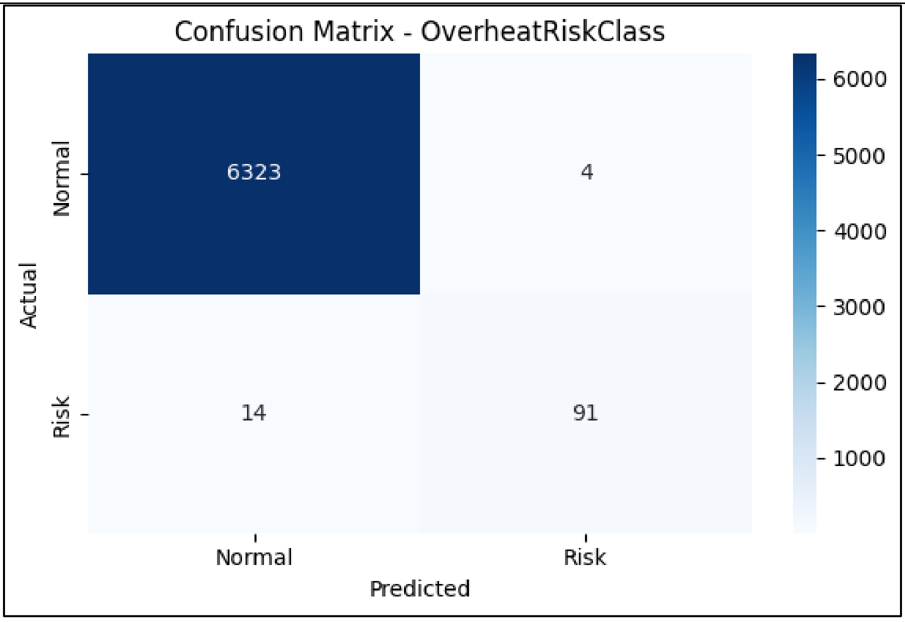
Bao, H., Knights, P., Kizil, M., & Nehring, M. (2026). Comparative assessment of CO2 emissions from mining haulage truck alternatives to achieve 2030 climate targets. *Resources Policy*. doi:10.1016/j.resourpol.2025.105826

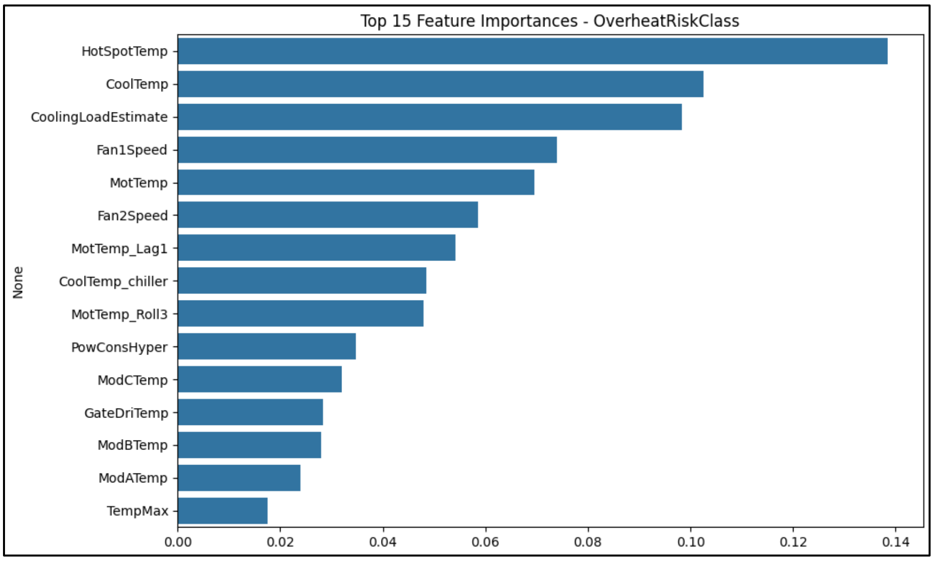
Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 69-79. doi:10.58496/BJML/2024/007

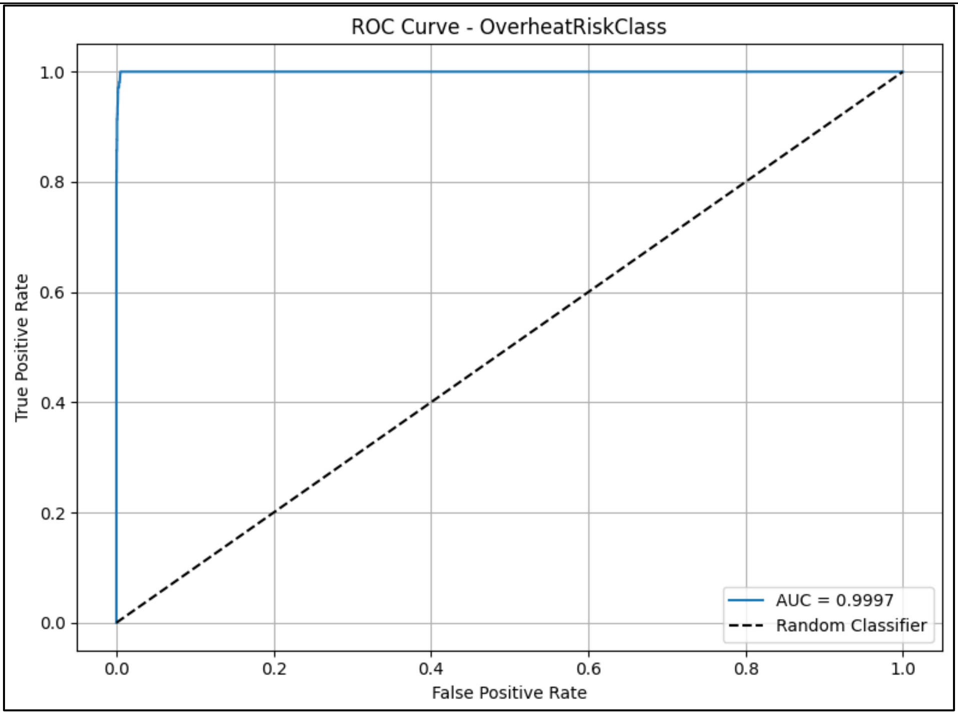
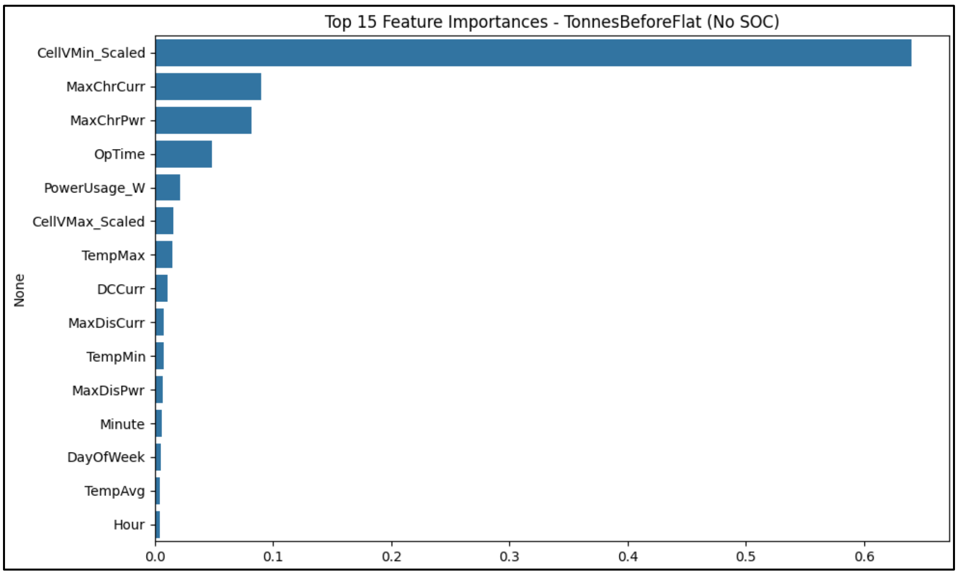
WA Government. (2026, May 28). *Economic Indicators*. Retrieved from Government of Western Australia: https://www.wa.gov.au/organisation/department-of-mines-petroleum-and-exploration/economic-indicators

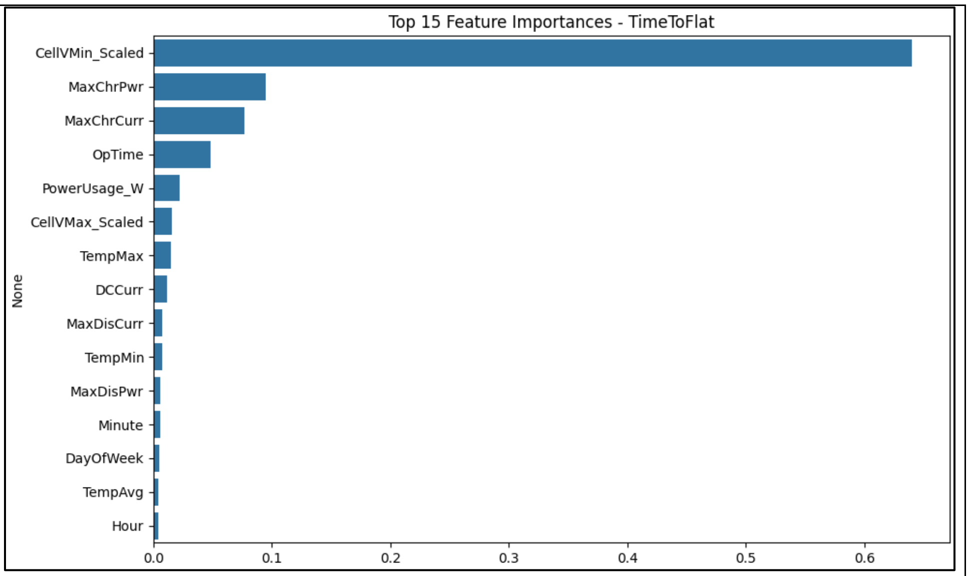
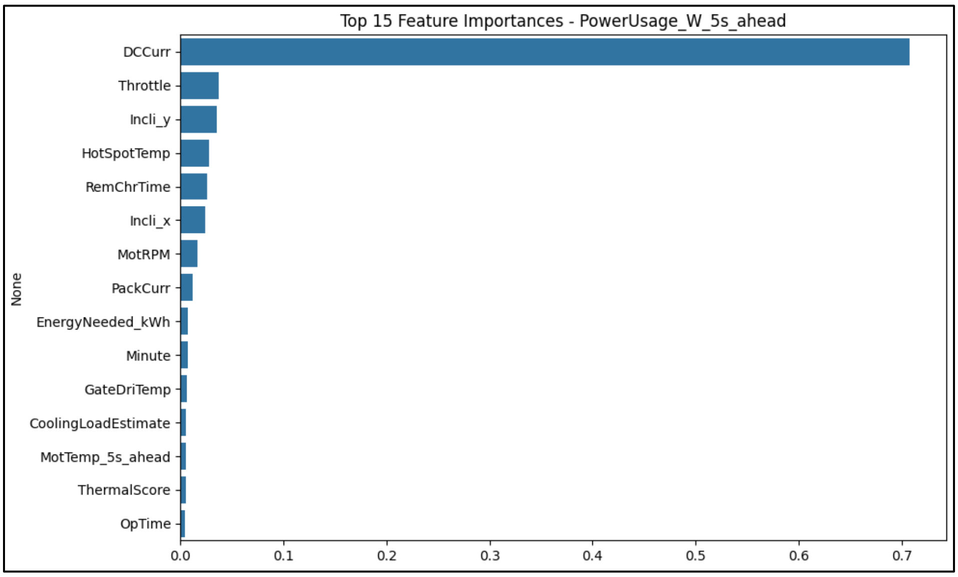
Appendix – Model Initial Key Features and Score

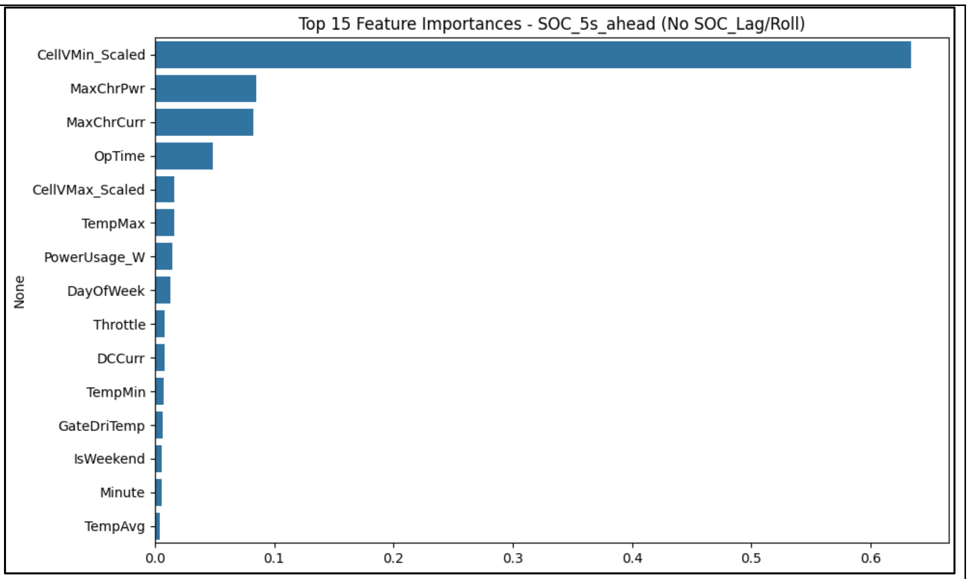
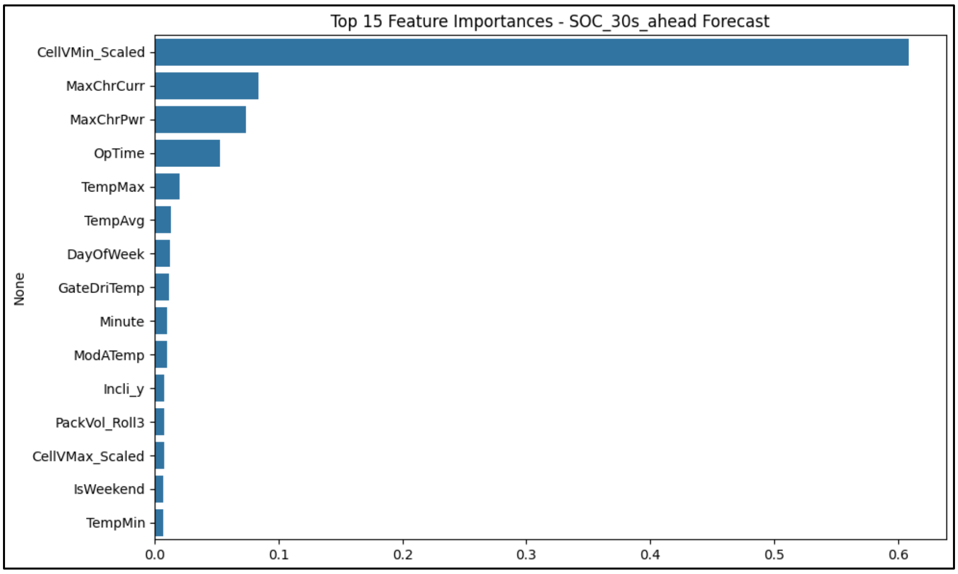
<table>
<tr><th>Model</th><th>Validation Score + Key Features</th></tr>
<tr><td>Charge Time Estimate</td><td>R^2: 0.9980
MAE: 1093.52 sec
MAPE: 0.11%

</td></tr>
<tr><td>Time Till Charge Complete</td><td>R^2 (RemChrTime): 0.9748
MAE: 53.14 sec
MAPE: 21.78%

</td></tr>
<tr><td>Overheat Risk</td><td>Accuracy: 0.9972
F1 Score: 0.9100</td></tr>
</table>

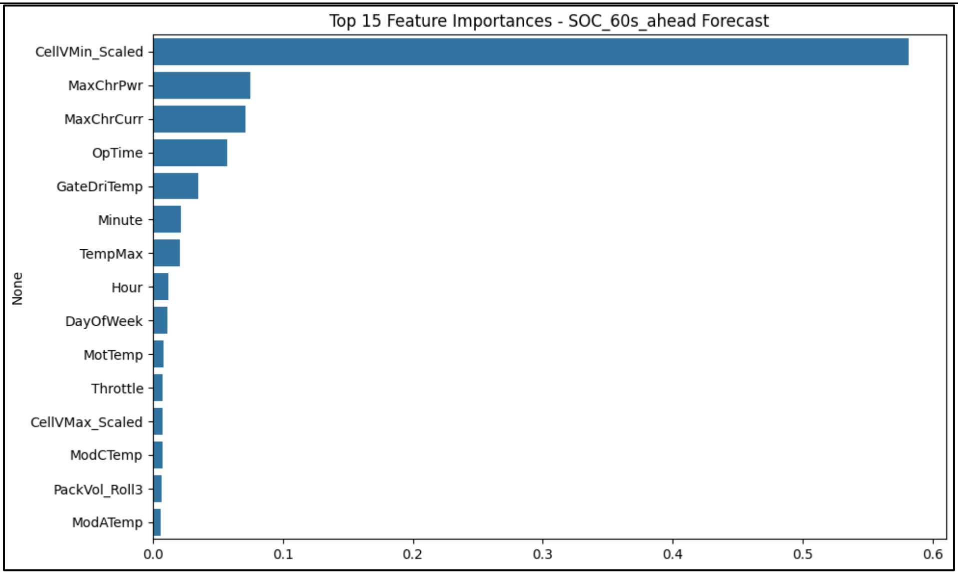

Confusion Matrix - OverheatRiskClass
Actual
Normal
Risk
6323
4
14
91
Normal
Risk
Predicted
6000
5000
4000
3000
2000
1000

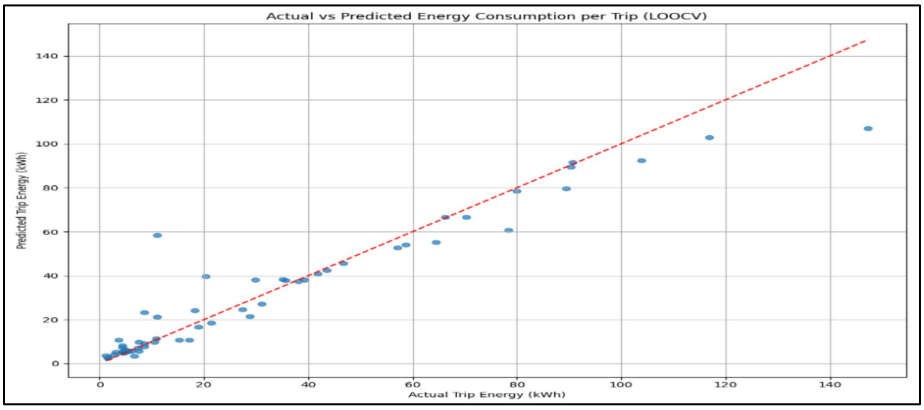

Top 15 Feature Importances - OverheatRiskClass
HotSpotTemp
CoolTemp
CoolingLoadEstimate
Fan1Speed
MotTemp
Fan2Speed
MotTemp_Lag1
CoolTemp_chiller
MotTemp_Roll3
PowConsHyper
ModCTemp
GateDriTemp
ModBTemp
ModATemp
TempMax
None
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

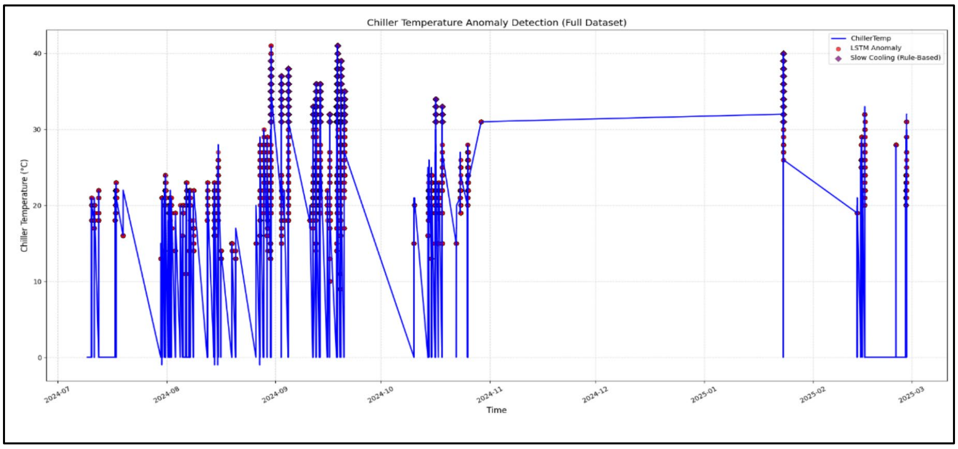
	
Tonnes before Flat	R^2 (no SOC): 0.9990 MAE: 0.15 tonnes MAPE: 0.36% 
Time to Flat	R^2 (TimeToFlat): 0.9989 MAE: 0.11 sec MAPE: 0.36%

	
Power Usage - 5s Ahead	R^2 (PowerUsage_W_5s_ahead): 0.8808 MAPE: 93.49% 
State of Charge Prediction – 5s Ahead	R^2 (no SOC_lags): 0.9937 MAE: 0.2832 MAPE: 0.54%

<table>
<tr><td></td><td>
</td></tr>
<tr><td>State of Charge Prediction – 30s Ahead</td><td>R^2 (SOC_30s_ahead): 0.9937
MAE: 0.3212
MAPE: 0.77%

</td></tr>
<tr><td>State of Charge Prediction – 60s Ahead</td><td>R^2 (SOC_60s_ahead): 0.9911
MAE: 0.3886
MAPE: 0.81%</td></tr>
</table>

<table>
<tr>
<td></td>
<td>
</td>
</tr>
<tr>
<td>Energy Consumption</td>
<td>R^2: 0.991


</td>
</tr>
<tr>
<td>Chiller Anomalies</td>
<td>Precision: 1.000
Recall: 0.5334
F1 Score: 0.6857


</td>
</tr>
</table>