# Causal Graph-Based Anomaly Detection for Battery Modules in Electric Heavy-Duty Vehicles

Yuantao Fan[1], Carlos Camacho[2], Sepideh Pashami[1] and Sławomir Nowaczyk[1]

[1] *Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Sweden*
*{firstname.lastname}@hh.se*
[2] *Volvo Group Truck Technology, Göteborg, Sweden*
*carlos.camacho@volvo.com*

## Abstract

Heavy-duty battery electric vehicles rely on large and complex energy storage systems (ESS), composed of multiple battery modules, whose individual health and reliability are critical to vehicle performance and safety. This study applies an unsupervised anomaly detection framework, COSMO (Consensus Self-Organizing Models), to a naturalistic real-world dataset collected during routine operations of in-service heavy-duty vehicles. We extend the baseline COSMO by incorporating causal discovery algorithms to help detect early signs of faults in ESS across heterogeneous missions and external conditions. On-board sensors data is collected as a multivariate time series, including information such as voltage, current, temperature, state of charge, etc. Given the wide range of applications of heavy-duty vehicles, these signals typically exhibit extreme variability even under normal operation, making anomaly detection challenging.

Causal graph discovery allows us to acquire latent structures that capture the underlying relationships among these influential features. The resulting learned causal graphs, for each battery module, serve as a more consistent representation that captures each battery module's usage and behavior over time. Since battery modules within the same ESS are expected to behave similarly under comparable operating conditions, COSMO models them as a homogeneous group. We then mark as anomalous modules that are identified to exhibit causal graph representations deviating markedly from the consensus.

## 1. Introduction

The ongoing transition from fossil fuel-driven mobility solutions to electromobility is a key enabler of more sustainable transportation, particularly in the heavy-duty (HD) sector, where reducing emissions has a significant societal impact. At the core of electric trucks lies the energy storage system (ESS), which is not only key to operational efficiency but also safety critical. Failures in battery packs can lead to costly unplanned downtime, accelerated degradation, and sometimes even more severe consequences, making predictive maintenance and health management strategies essential for battery electric vehicles (BEVs).

This study evaluates causal graphs as model representations for anomaly detection. Our goal is to assess whether causal structures learned from sensor data and encoded in the form of graphs can be helpful in indicating early signs of impending component faults, deterioration, or anomalies in the ESS of HD BEVs. The dataset we use includes both vehicle usage features and dedicated measurements for each battery pack in ESS, with two documented cases of battery replacement providing rare but valuable ground truth for validation.

In a series of studies (Fan, Nowaczyk, & Rögnvaldsson, 2015; Fan, Nowaczyk, Antonelo, et al., 2016), we developed and employed a framework for anomaly detection, namely the Consensus Self-Organizing Models (COSMO). This approach is essentially based on the idea of "wisdom of the crowd:" assuming that the majority of units in a homogeneous population are healthy. By comparison, individual units that deviate from the majority can be considered abnormal. Two types of anomaly scoring features are computed to indicate potential faults: distance-to-peers features, which capture dissimilarity relative to the fleet average, and deviation-level features, which quantify how likely an individual unit is to be significantly deviating from its peers (accounting for the normal variation within the homogeneous group).

In particular, COSMO has been shown to work very well for relatively small fleets of vehicles, for example, buses operating in a single city (Rögnvaldsson, Nowaczyk, Byttner, Prytz, & Svensson, 2018). However, it struggles with highly heterogeneous populations, and thus, in this study, we set out to analyze data collected worldwide by employing causal discovery algorithms to learn directed dependencies among the CAN signals. Causal graphs offer an interpretable representation of system behavior, making it possible to trace how changes in one signal propagate through the system, thus assisting in enabling explainable predictive maintenance (Pashami et al., 2023). This is particularly valuable for conducting root cause analysis, making diagnosis more efficient, and identifying early equipment deterioration.

In this work, we investigate three pairwise causal discovery algorithms: the Additive Noise Model (ANM) (Hoyer, Janzing, Mooij, Peters, & Schölkopf, 2008), the Conditional Distribution Similarity statistic (CDS) (Fonollosa, 2016), and Information-Geometric Causal Inference (IGCI) (Daniusis et al., 2012). To compare two different causal structures, we adopt a weighted Structural Hamming Distance (wSHD) incorporating domain knowledge, where high edge weights emphasize causal links between battery parameters, while edges involving only vehicle usage signals are assigned lower importance.

The contribution of this study are threefold: i) we apply and investigate three causal discovery algorithms to learn causal structure between CAN signals for anomaly detection; ii) we propose a weighted SHD, based on relevancy of the feature to component of interest, to measure the distances between learn causal graphs; iii) we integrate causal graphs into the COSMO framework and evaluate their performance on a real-world dataset from heavy-duty battery electric vehicles in detecting early signs of battery faults that ultimately lead to pack replacements.

## 2. RELATED WORK

Majority of the anomaly detection studies for batteries are prediction-based, or reconstruction-based methods (Dong & Lin, 2021; Li, Wang, Xu, Wu, & Li, 2024; Bhaskar et al., 2023; Xiong, Sun, Yu, & Sun, 2020). These approaches typically rely on self-supervised learning to construct a reference model approximates the behavior of a healthy system, and anomalies are flagged when the observed behavior deviates significantly from reference model's predictions. In contrast, only a small fraction of the works, e.g. (Shin, Lee, & Kim, 2023), explore anomaly detection from a causality-driven perspective, leveraging unsupervised learning to flag anomalies based on comparing structural dependencies captured from different samples. Moreover, causality-based methods are inherently more explainable, which also provide valuable insights for, e.g., fault diagnosis, and root cause analysis, by revealing how features influence one other and by identifying changes that emerge in the underlying causal structure.

Previous development on COSMO framework mainly focuses on analyzing univariate sensor data for anomaly detection. The only study to incorporate multivariate sensor data is by Rögnvaldsson, in their work (Rögnvaldsson et al., 2018), which exploits pairwise relations between signals to identify model settings with a higher likelihood of discovery new knowledge, or anomalous occurrences.

## 3. PROBLEM STATEMENT

The multivariate time series dataset used in this study was collected from the CAN bus of heavy-duty battery electric trucks, comprising approximately 1,500 trips logged. Twelve types of signals are included: seven related to the operating conditions and attributes of the battery packs, and the remaining five signals reflecting vehicle usage. Due to the high costs associated with battery failures, the OEM has implemented a preventive maintenance strategy, which means that actual failure examples are very rare. The dataset contains only two confirmed cases of battery replacements due to increased risk or fault. In both cases, one of the six battery packs was replaced. In the first case, the replacement was conducted following several fault code instances indicating a risk of imminent abnormal thermal event. In the second case, heat had developed on one of the busbars, which led to the decision to replace the affected battery pack.

Throughout this paper, we use the following notation. Let the multivariate time series $\mathbf{x}$ of each trip $l$ and vehicle $v$ be denoted by

$$X = \{\, x^i_{v,l,t} \ \mid \ t = 1, 2, \dots, T(v),\ i = 1, 2, \dots, K \,\}, \quad (1)$$

where $x^i_{v,l,t}$ is the value of the $i^{\text{th}}$ feature at time $t$ for trip $l$ of vehicle $v$, and $T(v)$ denotes the end-of-life of the battery pack that was subsequently replaced in the workshop. The objective of our method is to identify deviations in the data that occur prior to failure. Due to the lack of better information, we assume that the time of failure coincides with the time of repair. Ideally, any deviations should be detected sufficiently early to allow for maintenance actions, such as inspection or repair, to be scheduled proactively. For simplicity, we assume a constant period of interest prior to repair, which we refer to as the *prediction horizon* (PH). For a battery replacement carried out on vehicle $v$ at time $\tau$, we define the set of samples that should be labeled as *faulty* to be:

$$F_v = \{\, \epsilon^v(t) \ : \ \tau - \text{PH} \leq t \leq \tau \,\}. \quad (2)$$

The corresponding set of *healthy* samples is then given by $H_v = \overline{F_v}$. The task of anomaly detection is to assign, through a scoring function $f : \mathbf{x} \mapsto \delta$, higher scores to faulty samples than to healthy samples, i.e., for as many samples as possible $\delta_{F_v} > \delta_{H_v}$.

## 4. METHOD

The COSMO approach is particularly well-suited for environments that involve fleets of similar equipment, but where defining a clear baseline of "normal" behavior is difficult, whether due to varying operating conditions or differences in use. The core idea is to "understand" the typical level of variability in the data across the fleet, and then compare one unit against its peers in relation to this baseline. In overly heterogeneous operations, however, this degree of variability often becomes highly contextual, which is something COSMO typically fails to capture, as it assumes a homogeneous population of samples undergoing the same, or very similar, underlying process.

The method proceeds by first extracting relevant signal characteristics from the multivariate time series using a selected model. By comparing the learned representations with an appropriate similarity metric, deviations are detected following a "wisdom of the crowds" principle, i.e. individual sample deviates from the majority can be identified as an anomaly.

### 4.1. Causal Inferences

In this study, therefore, in the first step, we learn causal relationships between the CAN signals using causal inference algorithms, as a means to reduce the effect of mission characteristics and external conditions on the peer-to-peer distance measures. The learned graph, or its equivalent adjacency matrix, is utilized as the model representation of the given trip. Three causal graph discovery algorithms were selected and investigated: the additive noise model (ANM), the Conditional Distribution Similarity Statistic (CDS), and the Information Geometric Causal Inference (IGCI) method. ANM is one of the most widely used approaches for pairwise causal inference. Its key principle is to test whether the data fit an additive noise formulation in one direction, while the model can be rejected in the reverse direction. CDS computes the variability of rescaled $y$ (or $x$) values after binning along $x$ (or $y$). A lower variability suggests the direction $x \to y$ (or $y \to x$). IGCI is a pairwise causal discovery method that assumes minimal noise and an invertible causal mechanism, exploiting inherent asymmetries to infer causal direction.

Given a multivariate time series $\mathbf{x}$, we represent its causal structure as a directed graph $G = (V, E)$, where each vertex $v \in V$ corresponds to a signal (feature) and each directed edge $(v_i \to v_j) \in E$ reflects a causal relationship from feature $i$ to feature $j$. For simplicity, the graph is encoded by its adjacency matrix $A \in \{0,1\}^{|V|\times|V|}$, where $A_{ij} = 1$ indicates the presence of a causal link from $v_i$ to $v_j$, and $A_{ij} = 0$ otherwise.

### 4.2. Distance Metric Comparing Causal Graphs

The next step of COSMO is to compare two representations. In this work, structural Hamming distance (SHD) (Dhanakshirur[1], Laumann, & Park, 2024) and a weighted version of it, were used for computing the dissimilarity between two causal graphs. The SHD between two causal graphs counts the number of edge additions, deletions, or reversals required to transform one graph into the other. Let $A, B \in \{0,1\}^{|V|\times|V|}$ denote the adjacency matrices of two directed graphs with the same set of vertices. The SHD can be computed as $\mathrm{SHD}(A, B) = \sum_{i \neq j} \mathbf{1}(A_{ij} \neq B_{ij})$, where $\mathbf{1}(\cdot)$ is the indicator function. Equivalently, SHD counts the number of entries that differ between $A$ and $B$, thereby capturing mismatches in both edge presence and orientation.

To account for the varying importance of signals with respect to a specific component fault under investigation, we also propose to use a weighted version of SHD, denoted throughout the text as wSHD, given by

$$\mathrm{wSHD}(A, B; W) = \sum_{i \neq j} W_{ij}\, \mathbf{1}(A_{ij} \neq B_{ij})\,. \tag{3}$$

In practice, the weight mask $W_{ij}$ is constructed based on the type of signals connected by the edge: edges between two battery-related parameters are assigned the highest weight, edges between battery and vehicle usage parameters are assigned medium weight, and edges solely between vehicle usage parameters are assigned the lowest weight. This ensures that discrepancies on edges most relevant to the investigated battery fault contribute most strongly to the overall distance. The standard SHD is recovered as a special case when $W$ is a mask of ones.

### 4.3. COSMO Framework

Within the homogeneous group, learned representations of each sample are compared in pairs. For a given crowd of $n$ samples, a distance matrix comprises pairwise distances between all samples:

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix} \tag{4}$$

where elements on the primary diagonal are equal to zero, i.e., distances $d_{ij} = d_{ji}$. In this study, we set $n$ equal to the number of battery packs in ESS, as they constitute a homogeneous group.

Deviation detection is performed based on the most central pattern (row with the lowest sum total), which reflects the normal behavior of the group and is denoted by $c$. The set of distances from this central sample $c$ to the other samples $m$ is then used as an empirical distribution. The z-score of sample $m$ is then estimated as the number of samples in the empirical distribution that is further away from the most central pattern $c$:

$$z(m) = \frac{|\{i = 1, ..., N : d_{i,c} > d_{m,c}\}|}{N}. \tag{5}$$

where $|\cdot|$ denotes the cardinality of the set. The null hypothesis is that all samples are drawn from the same distribution,

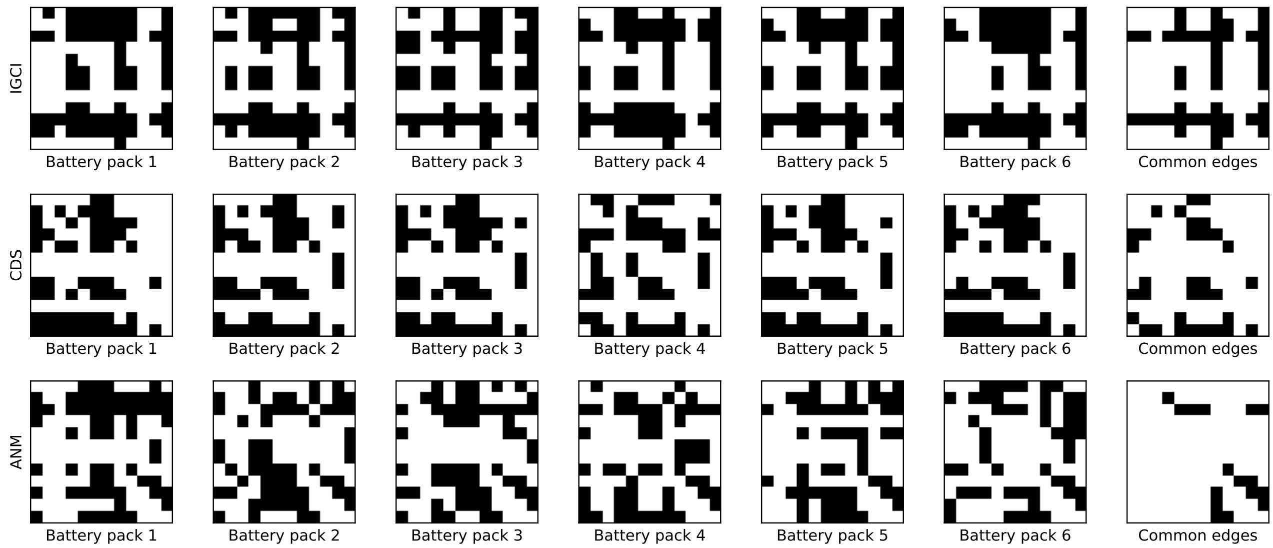


Figure 1. Learned causal graphs (i.e., adjacency matrix with dark cell corresponding to causal relationship between two features) for each of the battery packs using three causal discovery algorithms (IGCI, CDS, ANM) from a three example trips.

in which case the z-scores should be uniformly distributed between zero and one. This hypothesis is tested by comparing the average z-score over a certain period with the value expected from a uniform distribution. The deviation level is computed using the negative logarithm of the one-sided p-value:

$$Deviation\ level(\bar{z}) = -\log_{10}\left[\Phi\left(\frac{\bar{z}-0.5}{\sigma_n}\right)\right], \qquad (6)$$

where $\Phi(\cdot)$ is the normal cumulative distribution function, $\bar{z}$ is the average of the z-scores, $\sigma_n = (12n)^{-1/2}$, and $n$ is the number of valid days during a predetermined period, e.g., 30 day period. All trips/trajectories within this period were included for computing the the deviation level.

In this study, we investigate two types of features generated by COSMO for anomaly detection: deviation-level features and distance-to-peers features. The latter are computed as the mean distance of a unit to all other units within the homogeneous group.

## 5. EXPERIMENT RESULTS

### 5.1. Learning Causal Structure

The time-series dataset was segmented into approximately 1,500 trips, each lasting between 30 minutes and 3 hours. For every trip, six causal graphs were learned, one for each battery pack in the ESS, using three different causal discovery algorithms. To compute the causal graphs, we employed the Causal Discovery Toolbox library (Kalainathan & Goudet, 2019). The graphs were constructed from both battery parameters and vehicle usage features. Battery parameters are specific to individual battery packs, whereas vehicle usage features are shared across all packs. Figure 1 presents three sets of causal graphs (represented as adjacency matrices) obtained using ANM, CDS, and IGCI across three example trips. Each dark cell in the matrices denotes a directed causal relationship from one signal to another. As illustrated, the causal graphs of the six battery packs within the same trip (on each row) exhibit a consistent structure, sharing several common causal relationships, with a few differing edges. These variations will later be assessed using the Structural Hamming Distance (SHD) for anomaly detection.

### 5.2. COSMO Features

Figure 2 and Figure 3 illustrate the two types of COSMO features computed for detecting anomalies. Figure 2 shows the distance-to-peers feature, which reflects the deviation of each battery pack from the fleet average. As seen in the figure, the distance feature for battery pack four exhibits significant deviations some time before the repair was performed, suggesting potential for use in predictive maintenance service. Figure 3 presents the deviation-level feature, which clearly indicates a strong and persistent deviation in the 4th pack. This deviation lasted for some time and ultimately led to a service intervention in which the 4th battery pack was replaced. No further repairs to the ESS of this vehicle were required, and the deviation-level features of the remaining packs did not show deviations of comparable magnitude.

### 5.3. Performance Comparison

Table 1 summarizes the performance, measured in terms of average precision (AP) and area under the receiver operating characteristic curve (AUROC), across different approaches.

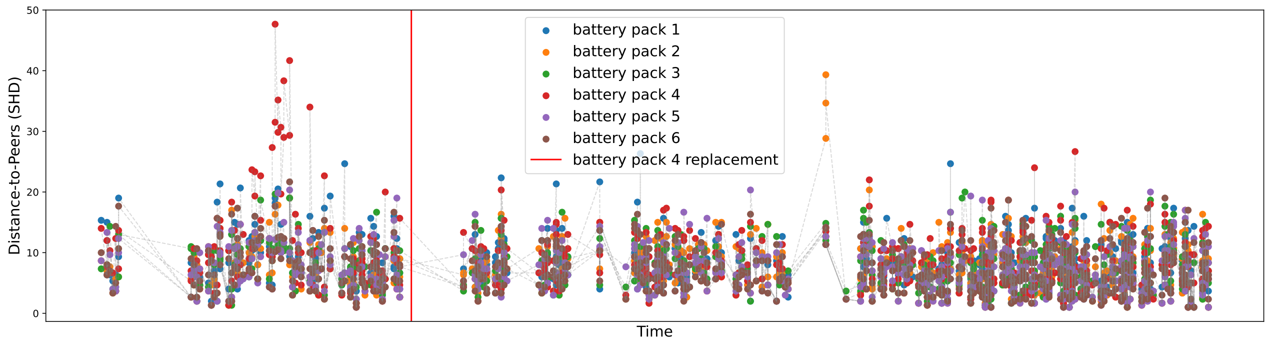


Figure 2. Distance-to-Peers feature of 6 battery packs computed based on IGCI and SHD; Red vertical line corresponding to replacement of battery pack 4, i.e. the first repair case.

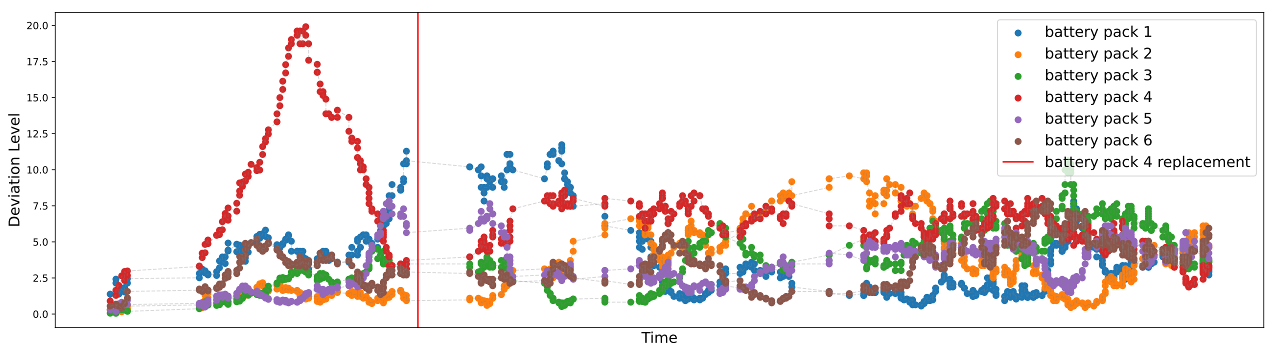


Figure 3. Deviation level, i.e. p-values, computed based on distance-to-peers features shown in Figure. 2

The comparison includes the choice of causal discovery algorithms, the distance metrics used between causal graphs, and a baseline against a well-known conventional method, Isolation Forest. The isolation forest, employed using the scikit-learn library (Pedregosa et al., 2011), were trained on aggregated values of 360 data trajectories collected from fault-free battery packs, with 100 estimators and provided with the true contamination ratio. The results show that using IGCI for causal graph learning in combination with weighted SHD achieves the highest AP and AUC. Overall, COSMO with causal graphs greatly outperforms Isolation Forest in detecting faulty batteries in this dataset.

## 6. Conclusion and Future Work

This study has presented an extension of the COSMO anomaly detection framework through the integration of causal discovery algorithms, with the specific aim of improving early fault detection in the energy storage systems of heavy-duty battery electric vehicles. By modeling each battery module as a causal graph derived from multivariate sensor data, we were able to obtain a more robust and interpretable representation of its behavior across diverse operational conditions. The proposed approach captures underlying dependencies among signals, thereby reducing the confounding impact of environmental and usage variability that typically complicates anomaly detection in real-world settings.

Our results demonstrate that causal graph representations enhance COSMO's ability to distinguish between healthy and potentially faulty modules. By employing a weighted structural Hamming distance, which prioritizes causal links most relevant to the ESS, the framework effectively identifies deviations that precede documented battery pack replacements. This not only validates the feasibility of incorporating causal structures into unsupervised fleet-based anomaly detection but also highlights their value for predictive maintenance applications where early intervention can mitigate costly failures and downtime.

Beyond performance gains, the integration of causal discovery offers an important step towards explainability in predictive maintenance. Causal graphs allow for tracing potential pathways of degradation, thereby supporting root-cause analysis and providing practitioners with interpretable evidence for maintenance decisions. Such transparency is essential for building trust in data-driven health monitoring systems, particularly in safety-critical applications like heavy-duty electromobility.

Table 1. Performance comparison between causal discovery algorithms, distance measures, and versus isolation forest

| **Method** | **Feature Type** | **Performance** | |
|---|---|---|---|
| | | **AP** | **AUROC** |
| ANM-SHD | Distance-to-Peers | 0.0191 | 0.5255 |
| | P-value | 0.0240 | 0.6611 |
| CDS-SHD | Distance-to-Peers | 0.0537 | 0.5639 |
| | P-value | 0.0536 | 0.7669 |
| IGCI-SHD | Distance-to-Peers | 0.0926 | 0.5585 |
| | P-value | 0.3230 | 0.7692 |
| ANM-wSHD | Distance-to-Peers | 0.0190 | 0.5069 |
| | P-value | 0.0165 | 0.5338 |
| CDS-wSHD | Distance-to-Peers | 0.0489 | 0.5663 |
| | P-value | 0.0573 | 0.7161 |
| IGCI-wSHD | Distance-to-Peers | 0.0728 | 0.5456 |
| | P-value | **0.3924** | **0.7822** |
| Isolation Forest | Anomaly Score | 0.0094 | 0.3346 |
| | Label | 0.0126 | 0.4538 |

Future work will focus on extending this approach in several directions. First, larger-scale evaluations across more diverse fleets will help establish generalizability and robustness. Second, the incorporation of temporal causal discovery methods could capture evolving dependencies that better reflect long-term battery aging processes. Finally, combining causal representations with complementary approaches such as physics-informed modeling or transfer learning could further strengthen the reliability of early fault detection. Taken together, these advances will contribute towards safer, more efficient, and sustainable deployment of heavy-duty battery electric vehicles.

## REFERENCES


Bhaskar, K., Kumar, A., Bunce, J., Pressman, J., Burkell, N., & Rahn, C. D. (2023). Data-driven thermal anomaly detection in large battery packs. *Batteries*, *9*(2), 70.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., & Schölkopf, B. (2012). Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*.

Dhanakshirur[1], M., Laumann, F., & Park, J. (2024). Distance to compare causal graphs. In *Causal inference: 6th pacific causal inference conference, pcic 2024, shanghai, china, july 5–6, 2024, revised selected papers* (Vol. 2200, p. 25).

Dong, G., & Lin, M. (2021). Model-based thermal anomaly detection for lithium-ion batteries using multiple-model residual generation. *Journal of Energy Storage*, *40*, 102740.

Fan, Y., Nowaczyk, S., Antonelo, E. A., et al. (2016). Predicting air compressor failures with echo state networks. In *Phm society european conference* (Vol. 3).

Fan, Y., Nowaczyk, S., & Rögnvaldsson, T. (2015). Evaluation of self-organized approach for predicting compressor faults in a city bus fleet. *Procedia Computer Science*, *53*, 447–456.

Fonollosa, J. A. (2016). Conditional distribution variability measures for causality detection. *arXiv preprint arXiv:1601.06680*.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, *21*.

Kalainathan, D., & Goudet, O. (2019). Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*.

Li, X., Wang, Q., Xu, C., Wu, Y., & Li, L. (2024). Survey of lithium-ion battery anomaly detection methods in electric vehicles. *IEEE Transactions on Transportation Electrification*.

Pashami, S., Nowaczyk, S., Fan, Y., Jakubowski, J., Paiva, N., Davari, N., ... others (2023). Explainable predictive maintenance. *arXiv preprint arXiv:2306.05120*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rögnvaldsson, T., Nowaczyk, S., Byttner, S., Prytz, R., & Svensson, M. (2018). Self-monitoring for maintenance of vehicle fleets. *Data mining and knowledge discovery*, *32*(2), 344–384.

Shin, H., Lee, J., & Kim, P. (2023). Causality-seq2seq model for battery anomaly detection. *International Journal of Aeronautical and Space Sciences*, *24*(1), 284–293.

Xiong, R., Sun, W., Yu, Q., & Sun, F. (2020). Research progress, challenges and prospects of fault diagnosis on battery system of electric vehicles. *Applied Energy*, *279*, 115855.