Invariant Feature Selection for Battery State of Health Estimation in Heterogeneous Hybrid Electric Bus Fleets

Yuantao Fan[1,*], Mohammed Ghaith Altarabichi[1], Sepideh Pashami[1,2], Peyman Sheikholharam Mashhadi[1] and Sławomir Nowaczyk[1]

[1]Kristian IV:s väg 3, 301 18 Halmstad, Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, Sweden

[2]Isafjordsgatan 28 A, 164 40 Kista, Research Institutes of Sweden (RISE), Sweden

Abstract

Batteries are a safety-critical and the most expensive component for electric buses (EBs). Monitoring their condition, or the state of health (SoH), is crucial for ensuring the reliability of EB operation. However, EBs come in many models and variants, including different mechanical configurations, and deploy to operate under various conditions. Developing new degradation models for each combination of settings and faults quickly becomes challenging due to the unavailability of data for novel conditions and the low evidence for less popular vehicle populations. Therefore, building machine learning models that can generalize to new and unseen settings becomes a vital challenge for practical deployment. This study aims to develop and evaluate feature selection methods for robust machine learning models that allow estimating the SoH of batteries across various settings of EB configuration and usage. Building on our previous work, we propose two approaches, a genetic algorithm for domain invariant features (GADIF) and causal discovery for selecting invariant features (CDIF). Both aim to select features that are invariant across multiple domains. While GADIF utilizes a specific fitness function encompassing both task performance and domain shift, the CDIF identifies pairwise causal relations between features and selects the common causes of the target variable across domains. Experimental results confirm that selecting only invariant features leads to a better generalization of machine learning models to unseen domains. The contribution of this work comprises the two novel invariant feature selection methods, their evaluation on real-world EBs data, and a comparison against state-of-the-art invariant feature selection methods. Moreover, we analyze how the selected features vary under different settings.

Keywords

State of Health Estimation, Invariant Feature Selection, Transfer Learning, Casual Discovery, Genetic Algorithm

1. Introduction

The transition from fossil fuel to electromobility in the transportation ecosystem is well underway and accelerating as an increasing number of hybrid and full-electric vehicles are manufac-

HAII5.0: Embracing Human-Aware AI in Industry 5.0, at ECAI 2024, 19 October 2024, Santiago de Compostela, Spain.

*Corresponding author.

yuantao.fan@hh.se (Y. Fan); mohammed_ghaith.altarabichi@hh.se (M. G. Altarabichi); sepideh.pashami@hh.se (S. Pashami); peyman.mashhadi@hh.se (P. S. Mashhadi); slawomir.nowaczyk@hh.se (S. Nowaczyk)

0000-0002-3034-6630 (Y. Fan); 0000-0002-6040-2269 (M. G. Altarabichi); 0000-0003-3272-4145 (S. Pashami); 0000-0002-0051-0954 (P. S. Mashhadi); 0000-0002-7796-5201 (S. Nowaczyk)

tured and commissioned. As the market size of electric buses is growing worldwide, they have become a crucial part of a sustainable and environmentally friendly transportation solution [1]. The level of electrification in the buses depends mainly on the powertrain configuration, which includes Battery Electric, Fuel Cell Electric, Hybrid Electric, etc. Hybrid electrification is a stepping stone towards full electrification; so far, both are of similar market share in EU. The most common battery type that powers the electric drive-line is the lithium-ion battery [2] since the technology was matured for broad mass more than a decade ago [3]. As a result of major manufacturers, like Volvo, gradually electrifying their products, new hybrid electric buses are regularly launched for public transportation [4, 5]. These hybrid electric buses (HEBs) are mounted with prismatic or cylindrical lithium-ion cells using liquid-cooled packs.

However, the introduction of new electrical components means entering new, less well-known territories. For example, operating conditions, e.g., types of operation or temperature, can have a considerable impact on accelerated component wear in electric drive-line. This raises a question about the robustness of the system. Most conventional on-board diagnostic and monitoring methods [6, 7, 8, 9, 10, 11, 12] for batteries mainly utilize the internal parameters of the battery unit based on experiments in a laboratory environment. Notably, the degradation process of the battery heavily depends on the usage pattern of the buses it powers. Often, the batteries and their management unit are deployed to various bus fleets with different mechanical configurations (e.g., single vs. double deck) and operations (e.g., urban vs. inter-city transportation). Therefore, developing robust battery deterioration models that could generalize well from established to novel settings provides great value to the industry.

Drive batteries of HEBs are a good example of this challenge. Optimizing the deployment, usage, and maintenance of these components requires a better understanding of the component's health under diverse operational settings. Such batteries are expensive, and prolonging their effective life is crucial to achieving a sustainable transportation system, low-cost operation, and market satisfaction. This requires insights into what factors are relevant to battery deterioration; analyzing the importance of available onboard signals for SoH prediction is of great help. The reliability of this expensive component becomes more important as the electromobility is being implemented at scale. Continuous changes and improvements in the batteries' technologies challenge the experts and statistical models currently in use. For example, different generations of batteries have different definitions for health indicators, and modeling one generation might not transfer to the next one. Besides, the lack of historical data due to the relatively recent deployment of the first electric dive lines in the real world leads to high model uncertainty.

The State of Health (SoH) is defined as the ratio between the nominal capacity at time t and the initial capacity [8]. It is a commonly used health indicator for evaluating the aging and deterioration level of the battery. Accurate estimation of SoH can be considered as a proxy to estimate the remaining useful life (RUL) and predict the end-of-life (EOL) of batteries for maintenance planning. Conventionally, the SoH is computed by a model developed by the manufacturer of the battery, and they are usually based on the internal parameters of the battery measured online [6, 13], e.g., current, voltage, and temperature of the single cells [14]. Although the SoH estimated via battery management unit (BMU) units on-board was developed and specialized for the target energy system; however, in reality, it may not always function as intended, e.g., due to faults or mismatch of newly replaced sub-component versions. An alternative is to employ a data-driven approach, using machine learning models, e.g., [15, 16],

based on sensor data collected onboard.

In dynamic environments, the behavior of the systems is expected to change over time. In such cases, the conditions under which a machine learning model is trained are likely to differ from the conditions on which they are to be tested. Thus, the performance of machine learning models deteriorates unless domain transfer is explicitly addressed. Moreover, the change in conditions between domains is often associated with a change in the joint distribution of input and output between source and target domains. These problems are commonly known as forms of dataset shift [17] and were addressed by transfer learning or domain adaptation methods [18]. An example approach is to select only a subset of features that generalize well to a new domain [19]. More concretely, in our experiments with modeling Li-Ion drive batteries, the degradation behavior of hybrid buses varies considerably across different bus configurations and operating conditions. This means that the features that are important for model training might not be relevant for the different configurations on which it will be tested. Even worse, they could have detrimental effects, leading to negative transfer. Our objective is to select features useful not only for the task in the source domain(s) but also those that transfer well across different domains, including the unseen ones. Modern electric buses are complex machines equipped with many sub-systems controlled by electronic computing units and monitored by hundreds of sensors. However, not all sensor data collected are helpful in estimating the SoH of the batteries. Selecting a relevant subset of features that generalize to various domains, e.g., different EB models and operating conditions, is crucial for building an effective SoH prediction model.

In this paper, we propose two invariant feature selection methods, i.e., GADIF [20] and CDIF, and evaluate their performance for selecting a set of invariant features across different settings to estimate the SoH of batteries. We compare the two proposed approaches with a causal learning-based domain adaptation method that produces invariant models, presented by Rojas et al. in [21], and several classical feature selection methods under transfer learning settings, where the goal is to migrate to unseen domains. The experimental results show that using invariant features selected by the proposed approaches leads to better generality of resulting models.

2. Related Work

The vast majority of studies on data-driven methods for estimating the SoH of batteries (review available in [6, 7, 8, 9]) assumes the target population is the same as the population available at the training time. However, the deterioration characteristics of the battery might vary depending on the usage pattern of the equipment it powers. Vehicles may be deployed to a new area, i.e., an unseen domain, where the terrain and traffic situation are very different from its peers in the previous deployment. The same type of battery and the management unit may be installed in a vehicle with different physical configurations. Under such circumstances, differences between vehicle populations, i.e., domains, need to be considered for SoH prediction.

Recently, more works on SoH prediction were conducted from the perspective of transfer learning, where the source domain is different from the target domain. The most popular approach is to utilize temporal information and sequence modeling, e.g., the long short-term memory (LSTM) and the gated recurrent unit (GRU). For the SoH prediction task, this is done

under an inductive transfer learning setting, i.e., a small amount of labeled data from the target domain is available. In their work [16], Tan et al. presented an approach using model-based transfer using an LSTM network for SoH Prediction. The base model was trained for the generalization performance, and the last two fully connected layers were fine-tuned with a subset. Che et al., in their work [22], have adopted gate recurrent neural networks for SoH prediction under a transfer learning setting. The model was first trained using a relevant battery and fine-tuned with an early degradation cycle from the testing battery. To address the variability in the length of the cycle sequences, Zhou et al. presented an interpretable transfer learning method [23] that uses cycle synchronization for data alignment and explored a few variations of GRUs to estimate the SoH of batteries in the target domain.

Genetic algorithms (GA) are widely used for feature selection [24, 25, 26] and shown to be very promising in many applications [27, 28, 29, 30]. However, few works have considered the scenarios in which heterogeneity exists between the training and testing samples. One needs to account for useful features for SoH regression in the source domain being different from the ones in the target domain.

Causal feature learning [31, 32] is another promising approach for selecting informative features. It exploits the causal relationships learned from the observations since these relationships imply the underlying mechanism of the variables. In particular, causal relationships that do not vary over different source domains are of great interest and are likely to be robust and able to generalize to future domains. Rojas et al. have proposed, in their work [21], an approach based on causal transfer learning to produce invariant models (IMCTL) that deal with covariate shift in a subset of features. Similarly, Persello et al. presented a kernel-based approach [33] for selecting features that minimize the dataset shift between the training and testing data. Magliacane et al. proposed an approach [34] selecting a subset of features that lead to the best predictions in the source domains while satisfying the invariance condition, determined by a joint causal inference framework [35] by Mooij et al. Another perspective is to learn a common set of features for multiple tasks [36, 37], presented by Argyriou et al. Work [38] presented by Taghiyarrenani et al. focused on constructing a new feature space that aligns multiple domains with conditional shifts in the joint distribution, using a novel loss function considering pairwise similarities between samples from different domains. With their dedicated effort to studying capacity degradation in batteries [39], Sahoo et al. have proposed a two-layer artificial neural networks model, with the second layer being fine-tuned using a small portion of the target data and a novel feature that yields a low generalization error on a battery of different types and working conditions. Few works on causality-based feature selection have considered the settings where the source domains are different from target domains [32]. Neither works proposed by Magliacane et al. [34], and Rojas et al. [21] have considered domains that consist of time series data collected from multiple systems.

In this study, we emphasize the specific challenges of EBs setting, which motivate the need for proposed approaches, beyond state-of-the-art. The first one, inspiring GADIF, is employing a fitness function for GA that takes into account heterogeneity in the available sample population and extracting features that are invariant w.r.t. samples from different domains is of interest. The second one, inspiring CDIF, is that causal graphs need to be learned from each system individually, using data with time dependencies, and invariant features shall be selected based on the presence of causal relationships across multiple different domains.

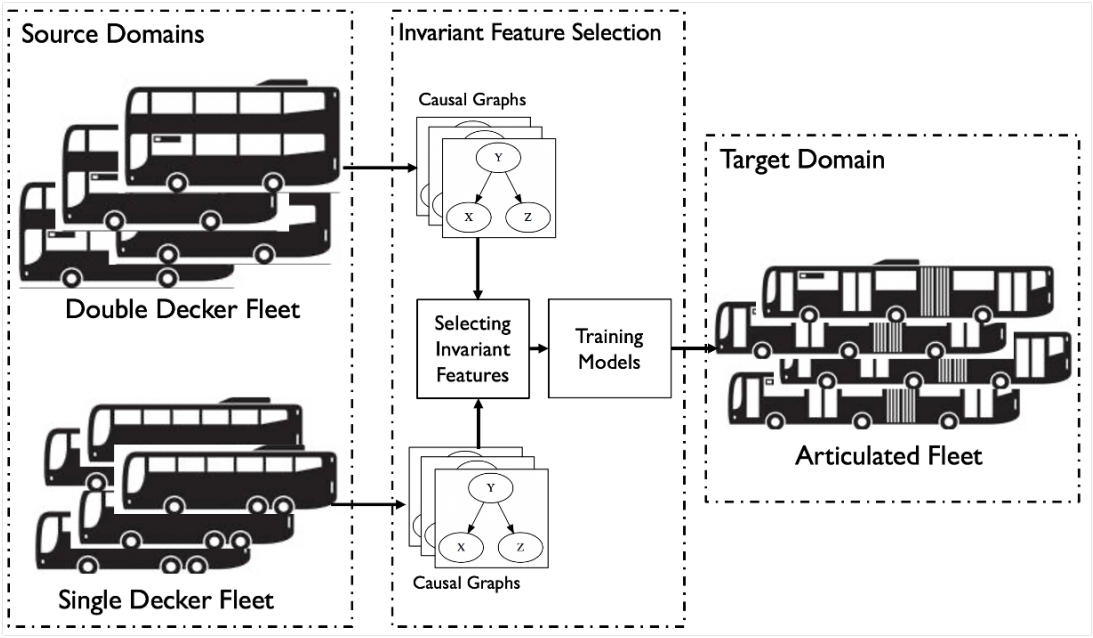


Figure 1: An illustration of the proposed CDIF method.

3. Method

This study targets an unsupervised domain adaptation problem, to be specific, a multi-source domain generalization problem, i.e., deploying machine learning models trained using labeled pairs of sample observations from one or more source domains $D_S = \{D_{S_1}, D_{S_2}, ..., D_{S_M}\}$ to an unseen target domain D_T, without any labels. Given a source domain D_{S_i}, and its entire vehicle population $V_{S_i} = \{v_1, v_2, ..., v_M\}$, the observations (labeled pairs) collected from vehicle v_i are denoted using $X_{v_i} = \{(\vec{x}_{(i,1)}, y_{(i,1)}), ..., (\vec{x}_{(i,n)}, y_{(i,n)})\}$. In this case, M is the number of vehicles, and X_{v_i} has dimensions of $(n \times k)$, where n is the number of samples and k is the number of features. A sample instance $\vec{x_{(.)}}$ is a k-dimensional real-valued vector $\vec{x}_{(.)} \in \mathbb{R}^k$. Data set of a source domain D_{S_i} is merely the union of samples from all vehicles V_{S_i} in that domain, i.e., $D_{S_i} = \bigcup_{i=1}^{M} X_{v_i}$. On the contrary, sample observations in the target domain D_T are not labeled, i.e., $D_T = \bigcup_{v_i \in V_T} \{\vec{x}_{(i,1)}, ..., \vec{x}_{(i,n)}\}$. Nevertheless, the feature space χ of sample instances in the source and the target domains are the same.

In this study, we propose two approaches, i.e., using a genetic algorithm for domain invariant features (GADIF) and causal discovery for selecting invariant features (CDIF), to select invariant features that are expected to generalize based on several source domains $\{D_{S_1}, D_{S_2}, ..., D_{S_m}\}$, where m is the number of source domains of which labeled data is available. Similar to [21], our work relies on the assumption that if the conditional distribution of the target label given a feature subset is invariant across several source domains, then the same is likely to hold in the target domain. In other words, there exists a subset of features that is invariant in predicting the SoH values across all domains. One way to verify how well this assumption holds is to evaluate whether the presence of causal relation between any features to the target variable is

consistent across various domains (an illustration is available in Figure 2).

3.1. Causal discovery for invariant features selection

We propose to use a causal discovery algorithm to learn causal relations in each vehicle and select a set of features that were the common causes for the target variable y, among vehicles across all source domains. More specifically, we consider the probabilistic causation between the two variables and compute the probability of the causal relations, e.g., that the feature u_i causes y, presented in each source population as a scoring criterion for selecting invariant features. An illustration of the CDIF method is presented in Figure 1; invariant features were extracted from two sets of causal graphs learned from individual vehicles in every source domain. For each vehicle v, a causal graph $G_v = (U_v, E_v)$ is generated using a causal discovery method with additive noise models, proposed by Hoyer et al. in their work [40], where U_v denotes a set of nodes, $U_v = \{u_1, u_2, ..., u_k\}$, corresponding to the k number of features in the training instances of vehicle v, and E_v denotes a set of edges e^v_{ij}, indicating causal relations direct from feature i to feature j if $e_{ij} = 1$, for vehicle v. Correspondingly, $e^v_{ij} = 0$ indicates no causal relation directly from feature u_i to u_j.

Given a source domain D_{s_m} and its corresponding vehicle population V_{s_m}, a $\mathbf{card}(V_{s_m})$, i.e. cardinally of set V_{s_m}, number of causal graphs were learned, and the number of occurrences of any causal relation presented in this vehicle population was computed as $\sum_{v=1}^{\mathbf{card}(V_{s_m})} e^v_{ij}$. The probability of any causal relation given a specific source domain D_{s_m}, with a vehicle population V_{s_m}, is then:

$$P(\, e^v_{ij} = 1 \mid v \in V_{s_m}) = \frac{1}{\mathbf{card}(V_{s_m})} \sum_{v \in V_{s_m}} e^v_{ij}, \tag{1}$$

where e^v_{ij} indicates the causal relation of feature u_i to u_j in vehicle v. In this work, we only consider the target variable y as u_j. Given a number of source domains $\{D_{S_1}, D_{S_2}, ..., D_{S_m}\}$ and its corresponding vehicle population $\{V_{S_1}, V_{S_2}, ..., V_{S_m}\}$, we propose two criteria for selecting invariant features:

$$U_\theta = \bigcap_{S \in \{V_{S_1}, V_{S_2}, ..., V_{S_m}\}} \{u_i \mid \frac{1}{\mathbf{card}(S)} \sum_{v \in S} e^v_{ij} \geq \theta,\ i \in \{1, 2, ..., k\}\} \tag{2}$$

$$U_\epsilon = \bigcup_{\substack{S_1, S_2 \in \{V_{S_1}, V_{S_2}, ..., V_{S_m}\}, \\ S_1 \neq S_2}} \{u_i \mid |\frac{1}{\mathbf{card}(S_1)} \sum_{v \in S_1} e^v_{ij} - \frac{1}{\mathbf{card}(S_2)} \sum_{v \in S_2} e^v_{ij}| \leq \epsilon,\ i \in \{1, 2, ..., k\}\} \tag{3}$$

where θ and ϵ are the thresholds for selecting features u_i in U_θ and U_ϵ. Invariant features included in U_θ are the ones of high presence, determined by θ, in the causal relation across all source domains; features included in U_ϵ are the ones of similar presence, determined by ϵ, in at least a pair of the source domains.

3.2. Genetic algorithm for domain invariant features

The proposed Genetic Algorithm (GA) [20] identifies good predictors of the target label in D_T, based on their generalization performance across a number of source domains

$\{D_{S_1}, D_{S_2}, ..., D_{S_m}\}$ in a leave-one-domain-out cross-validation setting. The fitness function of GADIF was designed to evaluate a feature subset U by testing its performance across a number of source domains $\{D_{S_1}, D_{S_2}, ..., D_{S_m}\}$.

Our algorithm GADIF is initiated with a population of individuals encoding feature subsets as chromosomes of binary strings, with 1 indicating the inclusion of the feature of the corresponding index. The algorithm proceeds from one generation to the next by applying crossover and mutation operators to the selected individuals in order to produce offspring. The process reassembles natural selection by choosing individuals with the best fitness as the most likely to propagate to the following generation. GADIF uses a variant of GA called CHC [41] to carry out the evolutionary search, which is essentially a wrapper method for feature selection, and can be computationally expensive. Our feature selection method evaluates feature subsets according to their performance across all available labeled source domains. The optimization search of GADIF maximizes a fitness function that extends the equation provided by [42]:

$$f(U) = \alpha P + (1 - \alpha)(1 - \frac{N_f}{N_t}) \tag{4}$$

where P is the leave-one-domain-out cross-validation performance (e.g. mean absolute error) of a machine learning model M trained in a wrapper setting using U, and α is a hyperparameter for the tradeoff between the model performance across several source domains P, and the reduction in a number of selected features N_f with respect to the total number of features N_t.

4. Experimental Results

The application scenario we focus on in this study is to build robust machine learning models using domain invariant features. That means features that generalize well from multiple source domains to an unseen target domain. For example, it can be a new population with novel operating conditions that were not available at training time. Therefore, we have chosen a dataset from heterogeneous fleets of commercial buses consisting of approximately 1500 hybrid energy vehicles. They have different physical configurations (e.g., Double-Decker vs. Single-Decker) and were deployed for different usage (e.g., long-haul vs. city operations) in various locations. All of them are equipped with lithium-ion batteries. Although both numerical and categorical features were present in the dataset, we only focused on a set of numerical features (10 in total) to model SoH degradation, as these features are of the best data quality and availability. The SoH values are collected from BMU, which is a component that does not necessarily last longer than batteries. The age of the energy storage systems (ESS) was post-processed (using repair records of the battery and BMU) and added to the feature set.

4.1. Experiment settings

Two sets of three experiments (six scenarios in total) were listed in Table 1 and were proposed to address the need for domain adaption in the presence of a discrepancy between the source and the target domain. The first set of experiments (i.e., 1 to 3) focuses on finding domain invariant features that could generalize over different operating conditions. The second set of experiments (i.e., 4 to 6) focuses on generalizing over different physical configurations. They

correspond to real-world scenarios in the applications with heterogeneous groups of vehicles, where a freshly deployed fleet is of varying configuration, or operates under different conditions; thus, it requires transfer learning.

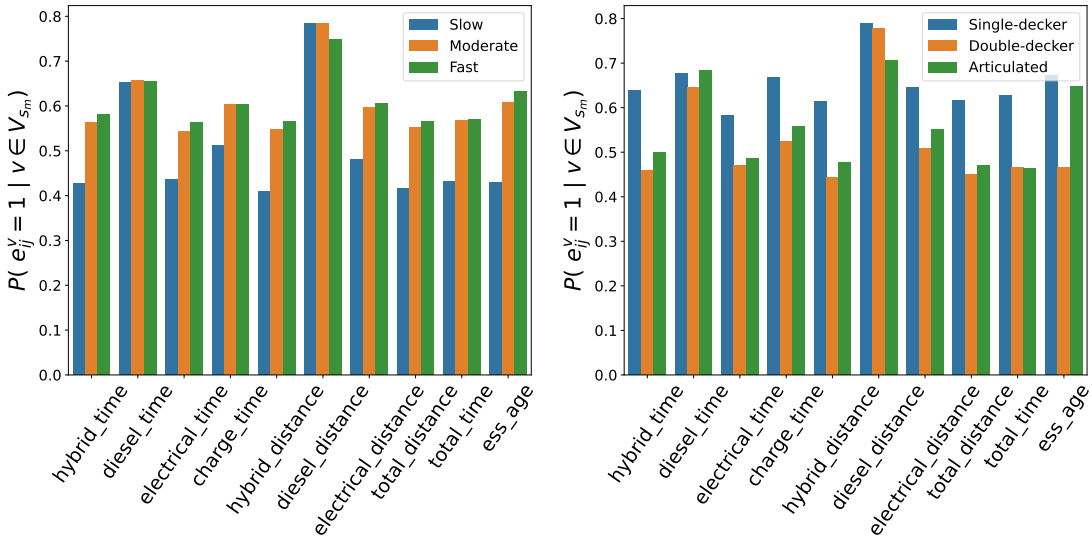
The first three scenarios focus on the transfer learning setting of fleets with different transport cycles (operating conditions), with respect to the average speed. There are three different vehicle populations/domains, i.e., the "slow", the "moderate", and the "fast". The first domain of vehicles belongs to the transport cycle "slow", which corresponds to buses operating in city centers. Due to heavier traffic, frequent stops, and speed limits, this group of vehicles obtains an average speed of under 12.5 kph. The second transport cycle domain "moderate" includes buses traveling at a higher speed range (12.5 to 17.5 kph) operating within the outer circle of cities. The group of "fast" vehicles travels at an average speed larger than 17.5 kph, which corresponds to buses operating between cities traveling long-hauls and highways. The second set of scenarios (i.e., 4 to

Table 1
Experiment scenarios

Scenario #	Source Domain D_S	Target Domain D_T
1	Moderate, Fast	Slow
2	Slow, Fast	Moderate
3	Slow, Moderate	Fast
4	Double-decker, Articulated	Single-decker
5	Single-decker, Articulated	Double-decker
6	Single-decker, Double-decker	Articulated

Table 2
Performance (MAE) of SoH predictions for different feature selection methods across all six experiment scenarios (note that GA* uses both the source and target data for feature selection).

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
GA*	1.50±0.01	1.59±0.01	1.88±0.01	2.05±0.00	1.53±0.01	1.48±0.01
Pearson	1.54±0.06	1.74±0.05	2.02±0.01	**2.16±0.12**	1.54±0.03	2.08±0.16
RF	1.62±0.07	1.72±0.07	1.99±0.04	2.19±0.14	1.54±0.03	1.78±0.10
LR	1.75±0.06	1.81±0.07	2.12±0.02	2.29±0.09	1.67±0.01	2.07±0.41
SFS	1.54±0.07	1.73±0.07	2.03±0.09	2.19±0.12	1.60±0.03	2.10±0.15
XGB	1.54±0.07	1.74±0.05	1.99±0.11	**2.16±0.12**	1.57±0.02	2.08±0.16
GADIF	1.54±0.02	**1.65±0.01**	**1.96±0.02**	2.27±0.03	**1.53±0.01**	**1.67±0.03**
IMCTL	2.08±0.68	1.97±0.43	5.55±4.98	2.70±0.71	3.47±0.99	4.33±3.56
IMCTL-G	10.75±5.11	4.01±4.02	6.80±4.85	11.99±0.66	10.70±1.01	10.25±1.75
CDIF(U_θ)	**1.53±0.07**	3.93±1.93	4.85±0.93	3.17±2.19	1.68±0.06	3.13±2.34
CDIF(U_ϵ)	1.53±0.08	1.68±0.13	2.10±0.14	2.22±0.23	1.65±0.03	2.96±1.86
All Features	**1.53±0.07**	1.72±0.09	2.06±0.14	2.19±0.21	1.68±0.05	2.07±0.18



(a) experiment scenario 1-3

(b) experiment scenario 4-6

Figure 2: Probability of causation of each feature to the target variable y in trained causal graphs

6) focuses on the transfer learning setting of vehicles with different physical configurations. The first domain in the second set was the single-decker buses; the second domain was the double-decker buses; the third domain was articulated buses, which are longer in length compared to vehicles in the other two domains. Figure 2 provided an overview of the presence of the causal relations on each feature u_i to the target variable y in different populations. It is shown in the figure that the presence of features in the "moderate" and the "fast" population are quite similar, as are the "double-decker" and the "articulated" vehicle population.

For each experiment scenario, several well-known filter- and wrapper-based feature selection methods were tested to compare with the proposed approaches, including Pearson index; feature importance measured by random forest (RF), extreme gradient boost (XGB), and coefficient from linear regression (LR) model; and sequential feature selection (SFS) [43] with linear regression. In addition, a recently proposed approach based on invariant models for causal transfer learning (IMCTL) [21] by Rojas et al., along with its variation utilizing greedy search (IMCTL-G) is tested and compared to the proposed approach based causal discovery, i.e., casual discovery for invariant features (CDIF U_θ and U_ϵ). Moreover, we compared the Genetic Algorithm (GA) using labeled data from the entire dataset (including all domains) with the proposed GA variation GAIF using a special fitness function. GA trained with data from all domains served as a reference and is expected to achieve the best performance. To evaluate the goodness of the features selected by different methods under a fair setting, a linear regression model was chosen for the SoH regression task, using Python implementation in scikit-learn [44]. Mean absolute error (MAE) was adopted as the evaluation metric for the SoH prediction task, and 4-fold cross-validation was performed on the vehicle level to measure the uncertainty.

The result section is organized as follows. We first present a comparison (in Table 2) of the

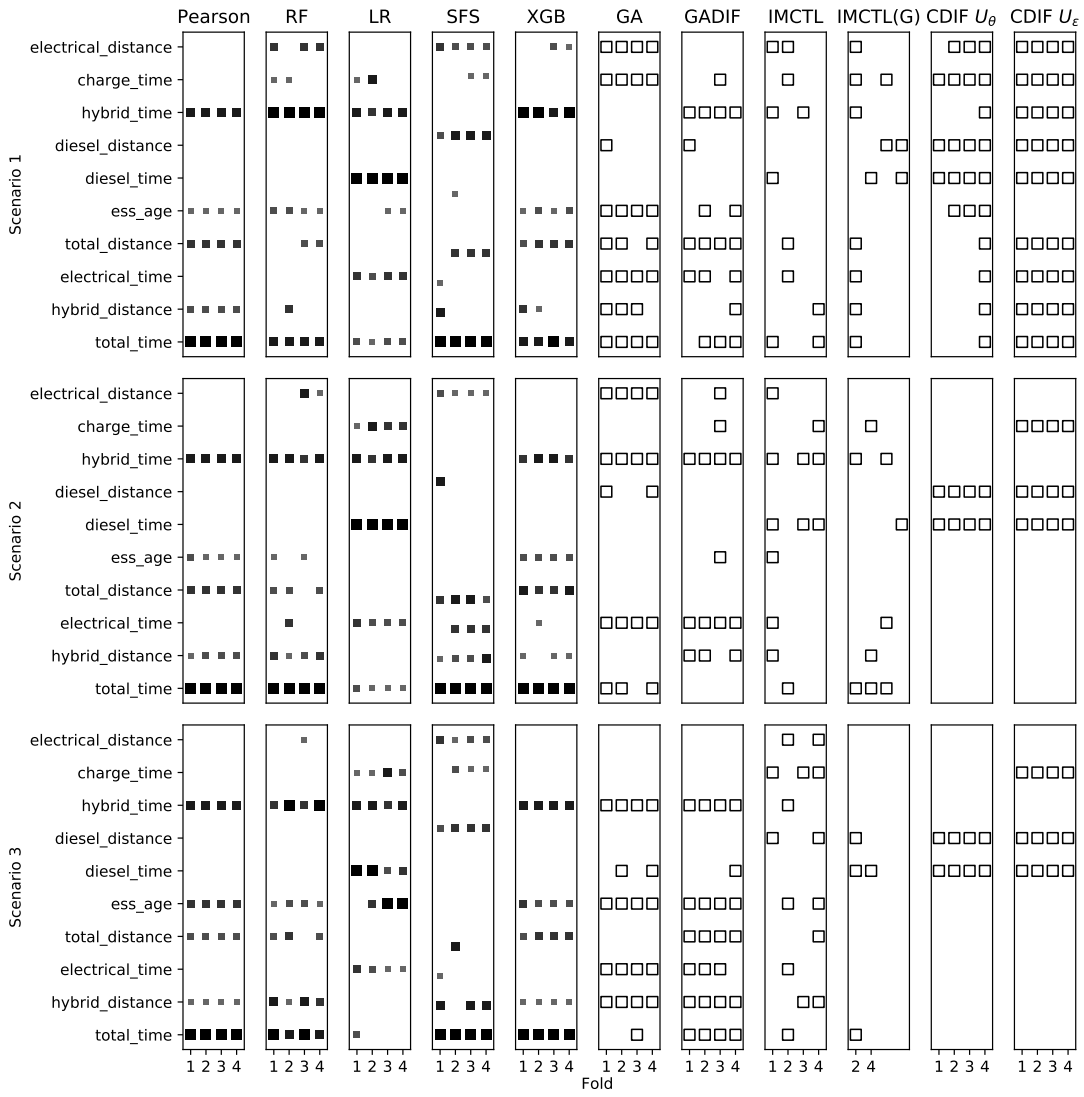


Figure 3: Features selected as invariant across different folds under experimental scenarios 1-3.

proposed approach against well-known methods across all six scenarios. Next, the uncertainty in the feature selection over different folds of the experiments is analyzed (Figures 3 and 4). We also study how the performance of different conventional feature selection methods depends on design parameter k, the number of features to be selected. We show the performance of CDIF across different thresholds and numbers of features selected in Figure 7.

4.2. Methods performance comparison

Table 2 presented the MAE of modeling the SoH using features set suggested by different methods under six scenarios. For methods based on the Pearson index, RF, LR, SFS, and XGB,

the best 5 features were selected. GA* corresponds to the approach using the genetic algorithm with the entire population from both the source and the target domain. GA* served as a reference and is expected to achieve the best performance, compared to other methods that only use the source population for feature selection. The two variations of CDIF, i.e., CDIF U_θ employed a threshold θ of 0.45, and CDIF U_ϵ employed a threshold of 0.18. The selection of the thresholds, i.e., θ and U_ϵ, is a trade-off between regression performance and number of selected features. The proposed approach (GADIF), with an α of 0.5, outperforms other well-known methods under scenarios 2, 3, 5, and 6 (except for GA*, which is expected since it utilizes data from the target domain while GADIF does not). CDIF U_θ outperformed other methods under scenario 1. For the rest of the scenarios, CDIF outperformed only IMCTL and IMCTL-G, which were also based on causal learning; however, it was not competitive against classical methods.

GADIF performs better under scenarios 5 and 6 compared to scenario 4. The source population under scenarios 5 and 6 contains vehicles of both simple and a more complex/heavier structure, i.e., single-decker and with one of the other two types. Intuitively, with the designed fitness function, GADIF can extract domain invariant features from a population with larger variation compared to scenario 4, in which only a vehicle with a complex/heavy structure is included while in the target only resides vehicle with a simple/lighter structure. In addition, the standard deviation in MAE values over 4-fold cross-validation of GA-based methods is rather small (around 0.01), which indicates that the difference in the prediction performance of GA-based methods between different scenarios is statistically significant.

4.3. Uncertainties in SoH modeling performance

The uncertainty (e.g., standard deviation computed over the 4 folds) in the MAE performance measure, presented in Table 2, was from the variations in the sample population. Therefore, features selected by different methods might vary as well. Figures 3 and 4 illustrated features selected by different methods: feature sets (top 5 features) suggested by Pearson index under scenarios 1-6 remain the same over all four folds; features sets suggested by CDIF U_ϵ remain the same across all four folds under most scenarios (except for the 5-th run); feature sets suggested by RF, LR, XGB, and CDIF U_θ are quite stable, with less than seven mismatches. The deeper and larger the filled squares are, the more important the corresponding feature is; unfilled squares correspond to features selected by GA, GADIF, IMCTL, IMCTL-G, and CDIF; these methods are binary and do not entail a degree of importance. Neither GA nor GADIF requires one to specify the number of features for the selection, and features selected from both methods vary similarly over the four folds; on the other hand, features selected by IMCTL, and IMCTL-G were quite different over the four folds for all scenarios.

The stability of feature selection methods was estimated by computing the distance between the selected features from experiments of different cross-validation folds over the six scenarios. A stable and consistent feature selection method would select identical feature sets from different scenarios, yielding a zero-sum of all pairwise distances, computed from, e.g., feature sets selected from different cross-validation folds. We computed the sum of all pairwise distances between feature sets selected from the 4-folds using the Jaccard index and compared different methods using a boxplot of summed distances from all six scenarios. Results are presented in Figure 5.

The result shows the stability of the proposed approaches GADIF and CDIF U_θ were not

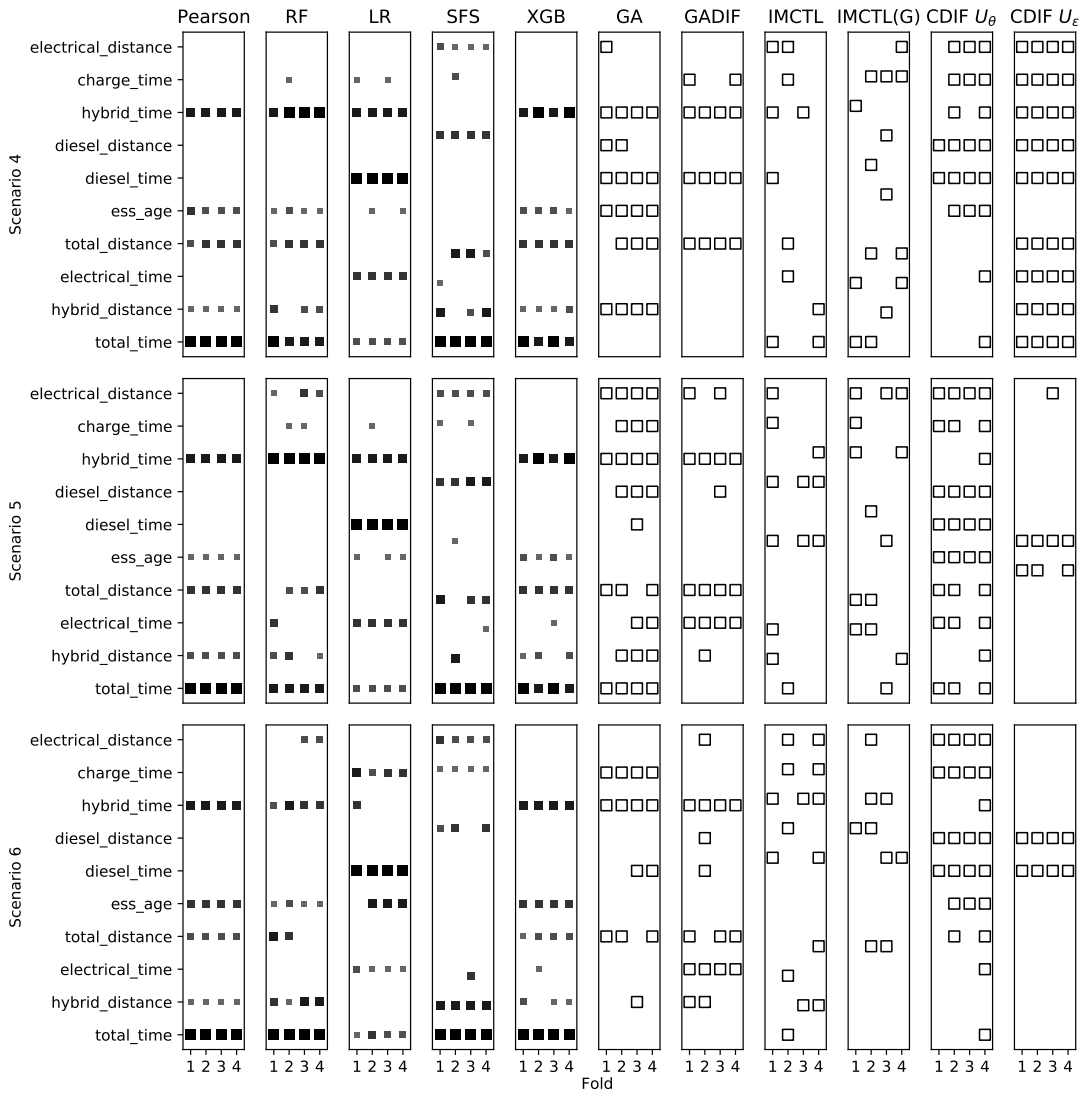


Figure 4: Features selected as invariant across different folds under experimental scenarios 4-6.

significantly different from the majority of the methods (i.e., RF, LR, SFS, and GA); feature sets selected by Pearson index, and the proposed approach CDIF U_ϵ were the most stable across the six scenarios, while XGB appears to slightly outperform most of the methods; feature sets selected using IMCTL and IMCTL-G were highly unstable.

4.4. Performance based on design parameter k

Figure 6 illustrates the performance of using different sets of features selected by different methods over the feature set size k, under experiment scenario 2. The gray horizontal line corresponds to the mean MAE of GADIF, and the margin corresponds to the standard deviation

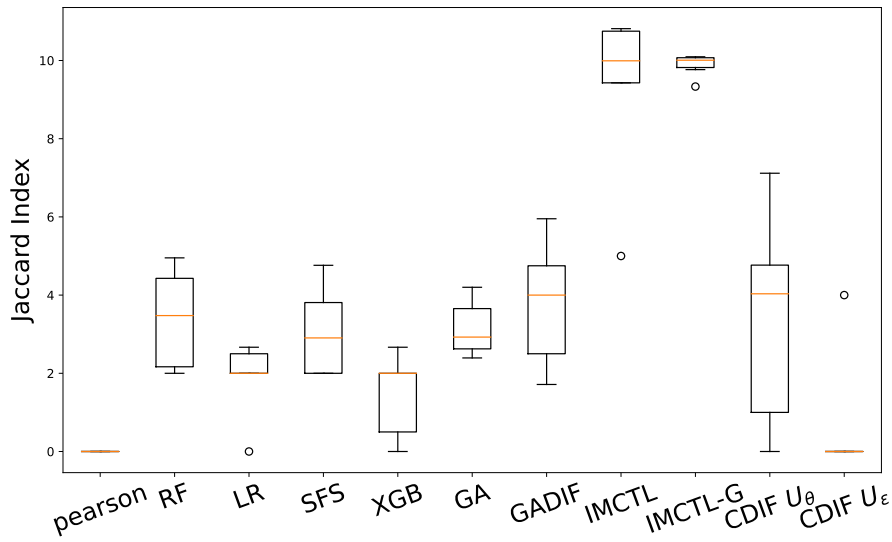


Figure 5: Stability measured with Jaccard Index of selected feature sets by different methods (a lower value indicates greater stability)

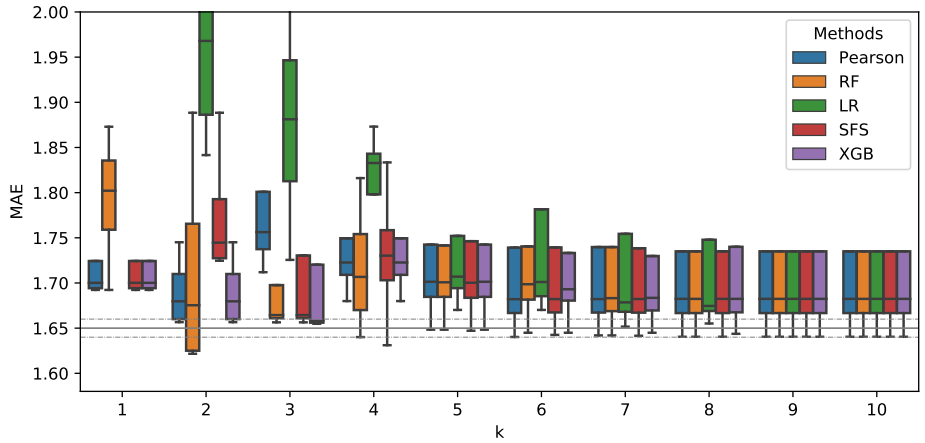
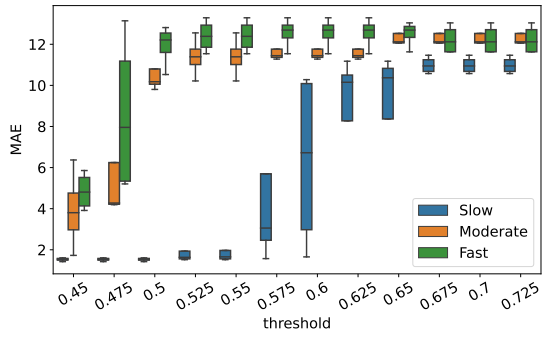


Figure 6: Performance (MAE) comparison of features selected by different methods for varying values of k, the feature set size, under experiment scenario 2.

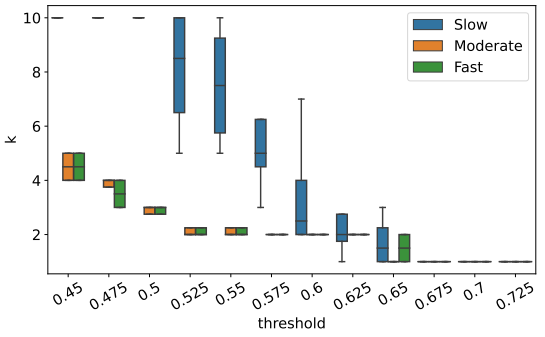
over 4-fold cross-validation. For most methods, it can be clearly seen that the performance starts to converge from a k of 5. GADIF does not need to specify the number of features to select and outperform other methods, regardless of the chosen k.

Subfigures in Figure 7(a), 7(c), 7(e), and 7(g) illustrate the performance of CDIF U_θ and CDIF U_ϵ for different values of the thresholds θ and ϵ. Correspondingly, Figure 7(b), 7(d), 7(f), and 7(h) showcases the number of features selected versus the threshold. For experiment scenario 1 with

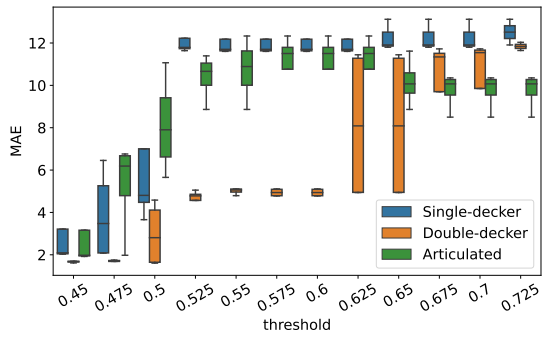
the feature set U_θ, the MAE for the "slow" target population was significantly lower compared to scenarios 2 and 3. This is due to the presence of many features that cause the target variable y. Therefore, more features were selected, which yielded better overall prediction power. In contrast, for scenarios 2 and 3, the presence of most features that cause the target variable y is quite low in the "slow" population compared to the other two populations. Thus, only a few features were selected for the regression task. For all six scenarios with the feature set U_θ and U_ϵ, MAE changes negatively correspond to the number of features selected, and hence, in general, the more features selected, the better the performance, until a saturation point that there is no significant improvement, as is shown in Figure 6.



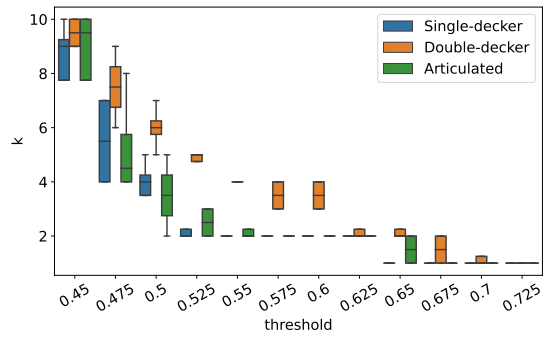
(a) MAE vs. Threshold of U_θ, Scenario 1-3



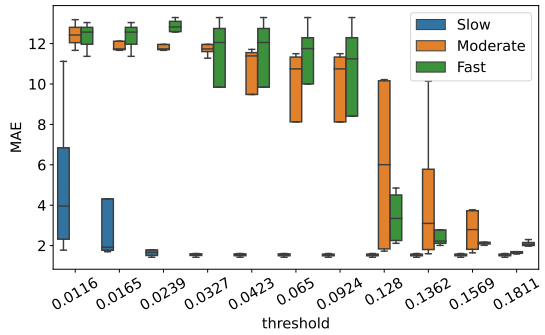
(b) k vs. Threshold of U_θ, Scenario 1-3



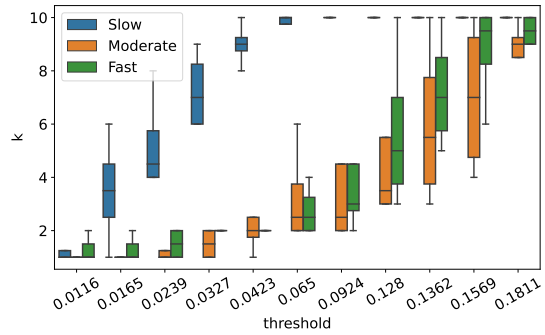
(c) MAE vs. Threshold of U_θ, Scenario 4-6



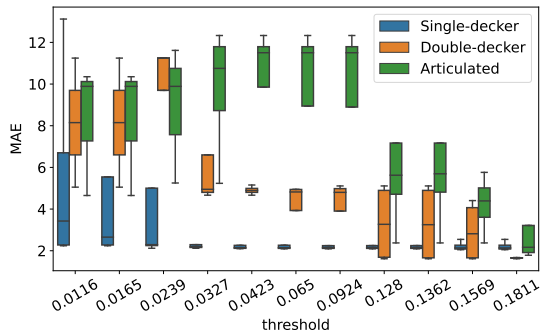
(d) k vs. Threshold of U_θ, Scenario 4-6



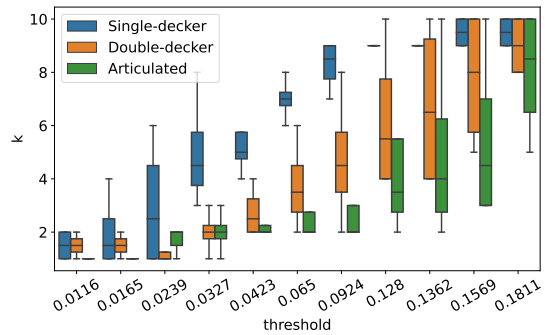
(e) MAE vs. Threshold of U_ϵ, Scenario 1-3



(f) k vs. Threshold of U_ϵ, Scenario 1-3



(g) MAE vs. Threshold of U_ϵ, Scenario 4-6



(h) k vs. Threshold of U_ϵ, Scenario 4-6

Figure 7: Performance (MAE) of CDIF, over the threshold and number of features k selected

5. Conclusions

This paper aims to develop and evaluate methods for selecting domain-invariant features. Such features are expected to lead to robust machine learning models and lead to a better generalization performance to unseen target populations. Two approaches, one based on causal discovery (CDIF) and the other based on an evolutionary algorithm (GADIF), were proposed to identify invariant features across several source domains in a multi-source domain generalization. The performance of the proposed approaches was compared with well-known feature selection methods in predicting the SoH of the batteries in real-world data from heterogeneous fleets of HEBs, running under different operating conditions and physical configurations. The main contribution of the paper is the comparison of well-known methods with the proposed approaches on their ability to choose invariant features for training SoH regression models that could generalize well to the unseen target domains. The experiment results show that GADIF outperforms other methods by selecting features that generalize better to unseen domains in 4 out of 6 testing scenarios, while CDIF outperforms other methods in one scenario. Future efforts for further development of CDIF include exploring the alternatives in generating the causal graph for each vehicle; criteria for suggesting threshold over different settings; extending the experiments with settings involving larger numbers of source domains; a more holistic evaluation metric with a balanced weighting on both performance accuracy and stability in the selected domain invariant features.

Acknowledgments

The work was carried out with support from the Knowledge Foundation and Vinnova (Sweden's innovation agency) through the Vehicle Strategic Research and Innovation Programme FFI.

References

[1] J.-Q. Li, Battery-electric transit bus developments and operations: A review, International Journal of Sustainable Transportation 10 (2016) 157–169.

[2] M. Mahmoud, R. Garnett, M. Ferguson, P. Kanaroglou, Electric buses: A review of alternative powertrains, Renewable and Sustainable Energy Reviews 62 (2016) 673–684.

[3] P. Pichler, M. Kapaun, Lithium ion batteries for hybrid busses and hybrid commercial vehicles-being ready for the broad mass.?; li-ion batterien fuer hybridbusse und hybridnutzfahrzeuge. startklar fuer die breite masse.? (2009).

[4] S. Borén, Electric buses' sustainability effects, noise, energy use, and costs, International Journal of Sustainable Transportation 14 (2020) 956–971.

[5] Volvo group invests in optibus to speed up on bus electrification and digitalization, https://www.sustainable-bus.com/its/optibus-volvo-investment/, ???? Accessed: 2021-11-12.

[6] M. H. Lipu, M. Hannan, A. Hussain, M. Hoque, P. J. Ker, M. H. M. Saad, A. Ayob, A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: Challenges and recommendations, Journal of Cleaner Production 205 (2018) 115–133.

[7] A. Fotouhi, D. J. Auger, K. Propp, S. Longo, M. Wild, A review on electric vehicle battery modelling: From lithium-ion toward lithium–sulphur, Renewable and Sustainable Energy Reviews 56 (2016) 1008–1021.
[8] A. Barré, B. Deguilhem, S. Grolleau, M. Gérard, F. Suard, D. Riu, A review on lithium-ion battery ageing mechanisms and estimations for automotive applications, Journal of Power Sources 241 (2013) 680–689.
[9] M. Berecibar, I. Gandiaga, I. Villarreal, N. Omar, J. V. Mierlo, P. V. den Bossche, Critical review of state of health estimation methods of li-ion batteries for real applications, Renewable and Sustainable Energy Reviews 56 (2016) 572–587.
[10] J. Du, X. Zhang, T. Wang, Z. Song, X. Yang, H. Wang, M. Ouyang, X. Wu, Battery degradation minimization oriented energy management strategy for plug-in hybrid electric bus with multi-energy storage system, Energy 165 (2018) 153–163.
[11] J.-H. Lee, I.-S. Lee, Estimation of online state of charge and state of health based on neural network model banks using lithium batteries, Sensors 22 (2022) 5536.
[12] N. Xu, Y. Xie, Q. Liu, F. Yue, D. Zhao, A data-driven approach to state of health estimation and prediction for a lithium-ion battery pack of electric buses based on real-world data, Sensors 22 (2022) 5762.
[13] B. Pattipati, K. Pattipati, J. P. Christopherson, S. M. Namburu, D. V. Prokhorov, L. Qiao, Automotive battery management systems, IEEE, 2008.
[14] S. Rothgang, M. Rogge, J. Becker, D. U. Sauer, Battery design for successful electrification in public transport, Energies 8 (2015) 6715–6737.
[15] D. Roman, S. Saxena, V. Robu, M. Pecht, D. Flynn, Machine learning pipeline for battery state-of-health estimation, Nature Machine Intelligence 3 (2021) 447–456.
[16] Y. Tan, G. Zhao, Transfer learning with long short-term memory network for state-of-health prediction of lithium-ion batteries, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS 67 (2020) 8723–8731.
[17] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, A. Schwaighofer, Dataset shift in machine learning, Mit Press, 2009.
[18] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.
[19] S. Uguroglu, J. Carbonell, Feature selection for transfer learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 430–442.
[20] M. G. Altarabichi, Y. Fan, S. Pashami, P. S. Mashhadi, S. Nowaczyk, Extracting invariant features for predicting state of health of batteries in hybrid energy buses, in: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2021, pp. 1–6.
[21] M. Rojas-Carulla, B. Schölkopf, R. Turner, J. Peters, Invariant models for causal transfer learning, The Journal of Machine Learning Research 19 (2018) 1309–1342.
[22] Y. Che, Z. Deng, X. Lin, L. Hu, X. Hu, Predictive battery health management with transfer learning and online model correction, IEEE Transactions on Vehicular Technology 70 (2021) 1269–1277.
[23] K. Q. Zhou, Y. Qin, C. Yuen, Transfer-learning-based state-of-health estimation for lithium-ion battery with cycle synchronization, IEEE/ASME Transactions on Mechatronics (2022).

[24] S. Sindhiya, S. Gunasundari, A survey on genetic algorithm based feature selection for disease diagnosis system, in: Proceedings of IEEE international conference on computer communication and systems ICCCS14, IEEE, 2014, pp. 164–169.
[25] M. Rostami, K. Berahmand, S. Forouzandeh, A novel community detection based genetic algorithm for feature selection, Journal of Big Data 8 (2021) 1–27.
[26] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Applied Intelligence (2021) 1–39.
[27] Z. Ullah, S. R. Naqvi, W. Farooq, H. Yang, S. Wang, D.-V. N. Vo, et al., A comparative study of machine learning methods for bio-oil yield prediction–a genetic algorithm-based features selection, Bioresource Technology 335 (2021) 125292.
[28] A. A. Ewees, M. A. Al-qaness, L. Abualigah, D. Oliva, Z. Y. Algamal, A. M. Anter, R. Ali Ibrahim, R. M. Ghoniem, M. Abd Elaziz, Boosting arithmetic optimization algorithm with genetic algorithm operators for feature selection: case study on cox proportional hazards model, Mathematics 9 (2021) 2321.
[29] N. Maleki, Y. Zeinali, S. T. A. Niaki, A k-nn method for lung cancer prognosis with the use of a genetic algorithm for feature selection, Expert Systems with Applications 164 (2021) 113981.
[30] P. Z. Lappas, A. N. Yannacopoulos, A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment, Applied Soft Computing 107 (2021) 107391.
[31] K. Chalupka, F. Eberhardt, P. Perona, Causal feature learning: an overview, Behaviormetrika 44 (2017) 137–164.
[32] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, X. Wu, Causality-based feature selection: Methods and evaluations, ACM Computing Surveys (CSUR) 53 (2020) 1–36.
[33] C. Persello, L. Bruzzone, Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning, IEEE transactions on geoscience and remote sensing 54 (2015) 2615–2626.
[34] S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, J. M. Mooij, Domain adaptation by using causal inference to predict invariant conditional distributions, Advances in neural information processing systems 31 (2018).
[35] J. M. Mooij, S. Magliacane, T. Claassen, Joint causal inference from multiple contexts (2020).
[36] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in 'advances in neural information processing systems 19', 2007.
[37] A. Argyriou, M. Pontil, Y. Ying, C. Micchelli, A spectral regularization framework for multi-task structure learning, Advances in neural information processing systems 20 (2007).
[38] Z. Taghiyarrenani, S. Nowaczyk, S. Pashami, M.-R. Bouguelia, Multi-domain adaptation for regression under conditional distribution shift, Available at SSRN 4197949 (2022).
[39] S. Sahoo, K. S. Hariharan, S. Agarwal, S. B. Swernath, R. Bharti, S. Han, S. Lee, Transfer learning based generalized framework for state of health estimation of li-ion cells, Scientific Reports 12 (2022) 1–12.
[40] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf, Nonlinear causal discovery with additive noise models, Advances in neural information processing systems 21 (2008).

[41] L. J. Eshelman, The chc adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination, in: Foundations of genetic algorithms, volume 1, Elsevier, 1991, pp. 265–283.

[42] S. M. Vieira, L. F. Mendonça, G. J. Farinha, J. M. Sousa, Modified binary pso for feature selection using svm applied to mortality prediction of septic patients, Applied Soft Computing 13 (2013) 3494–3504.

[43] S. Raschka, Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack, The Journal of Open Source Software 3 (2018). URL: http://joss.theoj.org/papers/10.21105/joss.00638. doi:10.21105/joss.00638.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.