Reliable Energy Consumption Modeling for an Electric Vehicle Fleet

Millend Roy
Microsoft Research
Bengaluru, India

Akshay Nambi*
Microsoft Research
Bengaluru, India

Anupam Sobti
Microsoft Research
Bengaluru, India

Tanuja Ganu
Microsoft Research
Bengaluru, India

Shivkumar Kalyanaraman
Microsoft
Bengaluru, India

Shankar Akella
Ashok Leyland
Chennai, India

Jaya Subha Devi
Ashok Leyland
Chennai, India

S A Sundaresan
Ashok Leyland
Chennai, India

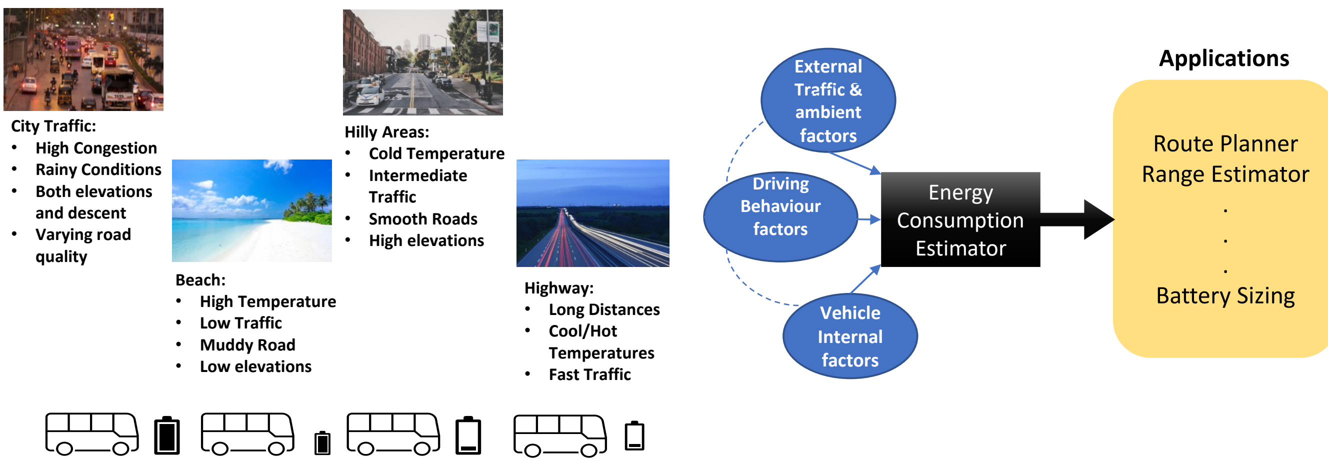

Figure 1: A bus company has to choose amongst various electric buses with different battery capacities and state of charge for a variety of trip/route conditions. The conditions might vary in different parameters (e.g., road type, traffic, driving behavior, environmental factors, etc.). To optimize for route planning, battery sizing, range estimation, one has to accurately predict the Total Energy Consumption for each of these trip conditions. In this paper, we use a data-driven approach to energy consumption modeling.

ABSTRACT

Accurately predicting the energy consumption of an electric vehicle (EV) under real-world circumstances (such as varying road, traffic, weather conditions, etc.) is critical for a number of decisions like range estimation and route planning. A major concern for electric vehicle owners is the uncertain nature of the battery consumption. This results in the "range anxiety" and reluctance from users for mass adoption of EVs, since they are concerned about untimely drainage of battery. Even at the organizational level, a company running a fleet of electric vehicles must understand the battery consumption profiles accurately for tasks such as route and driver planning, battery sizing, maintenance planning, etc.

In this paper, firstly, we highlight the challenges in modelling energy consumption and demonstrate the nature of data which is required to understand the energy consumption of electric vehicles under real-world conditions. Then, through a large and diverse dataset collected over 23,500 hours spanning ≈460,000 km with 27 vehicles, we demonstrate our two-stage approach to predict the energy consumption of an EV before the start of the trip. In our energy consumption modelling approach, apart from the primary features recorded directly before the trip, we also construct and predict secondary features through an extensive feature engineering process, both of which are then used to predict the energy consumption. We show that our approach outperforms Deep Learning based modelling for EV energy consumption prediction, and also provides explainable and interpretable models for domain experts. This novel method results in energy consumption modelling with $<5\%$ of Mean Absolute Percentage Error (MAPE) on our dataset and significantly outperforms state-of-the-art results in EV energy consumption modeling.

*Corresponding Author: akshay.nambi@microsoft.com

COMPASS '22, June 29-July 1, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9347-8/22/06…$15.00
https://doi.org/10.1145/3530190.3534803

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Transportation*; *Transportation*; • **Hardware** → **Batteries**.

KEYWORDS

electric vehicles, energy consumption estimation, sustainability

ACM Reference Format:
Millend Roy, Akshay Nambi, Anupam Sobti, Tanuja Ganu, Shivkumar Kalyanaraman, Shankar Akella, Jaya Subha Devi, and S A Sundaresan. 2022. Reliable Energy Consumption Modeling for an Electric Vehicle Fleet. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '22), June 29-July 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3530190.3534803

1 INTRODUCTION

Electric vehicles are a certain part of our future. With a focus on sustainability, more and more governments and organizations are adopting electric vehicles [26]. When running on batteries, accurate estimation of the energy usage is critical. From the user's perspective, uncertain estimates of energy consumption lead to a "range anxiety" [19, 20]. The user is not sure how long the vehicle is going to last and ends up charging the vehicle more frequently than required. Studies [14] show that vehicles typically use only 75-80% of battery capacity while driving. This uncertain nature of energy consumption also proves to be a challenge for organizations running fleets of electric vehicles. Tasks like route planning, fleet management, determining accurate pricing, battery maintenance costs, etc. are greatly affected by the ability to estimate the energy required for operations [15, 17]. Other infrastructure like charging station locations and schedules also depend heavily on accurate estimates of energy consumption by the vehicles [27].

Accurate energy consumption is a difficult problem due to its dependence on a variety of factors – (1) Vehicle Dynamics (2) Auxiliary Devices (3) Road Conditions (4) Weather Conditions (5) Traffic Conditions (6) Driver Behavior (7) Battery Characteristics, and, (8) Manufacturing Variability. Several models have been proposed for modeling the energy consumption in electric vehicles [2, 6, 23]. These types of models can be classified as analytical, computational and statistical models [18]. The *analytical* approach refers to modeling every part of the energy consumption, namely, the motor efficiency, power-train efficiency, and, the energy restored using regenerative braking in an analytical physics-based manner. This allows a much deeper understanding of the operating condition assumptions and the range of the vehicle is estimated based on the initial state of charge of the battery and the energy consumption based on the vehicle model. The *statistical* approach, goes backwards from system level data and uses machine learning algorithms to estimate the typical energy usage for a certain set of driving conditions. However, the statistical methods restrict themselves to using data from the vehicle operation itself. On the other hand, *computational* methods use higher level data like the GPS coordinates, driving behaviour estimation, weather conditions, etc. in their modeling and therefore refrain from a strictly analytical approach.

In this work, we use a computational modeling approach to model the Total Energy Consumption (TEC) – energy consumption in kWh for a specific trip – for use in an organizational setting to make decisions such as route optimization, range estimation, facility location and fleet management. The distinction of application is important since we model TEC on a trip-level before the trip starts. Therefore, we cannot use the data generated during the trip for energy consumption estimation. *Thus, the goal of this work is to accurately predict the energy consumption of the trip before the start, using information obtained before the start of the trip.* To this end, we develop a novel machine learning pipeline to accurately estimate the Total Energy Consumption (TEC) per trip for a wide variety of scenarios. Our machine learning pipeline provides < 5% Mean Absolute Percentage Error (MAPE) in energy estimation. We achieve this using a two-stage energy-estimation approach. We use a primary set of features readily available before the start, such as trip distance, trip duration, etc., to predict a set of secondary features (which are constructed through extensive feature engineering process), like braking frequency, number of stops, etc. Then, both the primary and secondary features are used in conjunction to regress the energy consumption for the entire trip. The key novelty of this work is: (1) construction of features that provide deep understanding of energy usage, (ii) predicting these features ahead of trip without any trip data, and (iii) showcasing the applicability of these features to model energy consumption accurately on our dataset and also a public dataset. We also compare this approach with the deep learning paradigm of learning features directly from data. Apart from getting better accuracy with our approach, our pipeline is also explainable and fully debuggable.

In this paper, we mostly focus on electric buses (eBus), however, the proposed approach can also be applied to other EVs such as cars (as we show in Section 6.5). The current public datasets to evaluate the proposed approach lack in the following ways: (1) The amount of data available for EVs, is fairly limited. (2) The attributes recorded might not be enough for accurate estimation of energy consumption. (3) There are no datasets available for electric buses operating in a wide variety of passenger loads. To overcome these issues, we collect our own dataset from a fleet of electric buses serving the population in a major city. We collected data from 27 buses running over 23,500 hours with a cumulative distance of 459,326 km. We collect 128 features concerning various aspects of the trip including weather conditions, auxiliary devices, traffic density, geographic information, etc. Specifically, we also collect driver behaviour attributes like harsh braking, harsh acceleration, etc. during our data collection. Apart from detailed evaluation of the proposed model on our dataset, we also show the applicability of the model and features learnt to another public dataset.

Contributions. Our key contributions are the following:

1. We provide insights on the features to be recorded for the effective and reliable energy consumption model of a large electric vehicle fleet. We demonstrate this with a dataset, collected from 27 eBuses.
2. We employ an extensive feature engineering process leveraging domain knowledge to highlight the important factors for a computational approach and present a novel two-stage machine learning pipeline for accurate energy estimation.

(3) We compare our method with end-to-end deep learning baselines and show that even deep learning methods are not able to construct the intermediate features reliably that we are able to derive.
(4) We extensively evaluate and validate the proposed model on our dataset with case studies of different decisions including longevity and generalization to new routes and vehicles. Our proposed models are fully explainable and debuggable.

2 RELATED WORK

We focus our review of related work on computational models of energy consumption. There are two primary aspects of learning computational models of energy consumption - the data and the models. While some works use data generated using simulations of a particular vehicle's dynamics [19], we only consider methods that record real data in our review.

Data Review

Early works in energy consumption estimation for EVs established the complexity of the problem. Li et al. [16] describe the major factors that contribute to the energy consumption of the electric vehicles. The authors found that the topography and use of HVAC systems were the primary contributors to the variation in energy consumption. Their experiments, however, considered the case that the other auxiliary devices such as headlights were always turned off which would also be a primary source of energy consumption. The data was limited to a Nissan Leaf 2011 model on 5 km routes within Sydney. Alvarez et al. [1] used smartphones to record data inside the car and predict the range. The authors use speed, acceleration and jerks recorded through the smartphone to estimate driver behaviour in addition to the GPS data for energy consumption estimation. The data is collected in a Mitsubishi Electric car with 10 users on the same 6 km stretch in Madrid.

The recently proposed public Vehicle Energy Dataset (VED) [21] consists of attributes that capture various aspects of a drive: *trip characteristics* - distance, speed, GPS coordinates; *environmental conditions* - Outside Air temperature; *Auxiliary Devices* - Heater Power, AC; *Power and Battery Info* - State of Charge (SOC), Voltage and Current. The dataset is recorded over an year consisting of data from 374,000 miles (600,000 km) of drives which includes hybrid and pure-electric drives with a total of ≈50,000 km of data. Another publicly available dataset is the SpritMonitor dataset [2, 9] which consists of weather, tire, heater information in addition to the distance, speed and time attributes. A total distance of 4630 km with Mitsubishi i-MiEV and Volkswagen e-golf vehicles were collected. Other datasets used in recent works [23] include the 2011 Denmark data [14]. The data consists of 1.4 million km of rides with 164 electric vehicles. The recorded data has been augmented with weather information, wind speed, etc. with the help of GPS coordinates and weather prediction. In contrast, our dataset is recorded for a fleet of 27 buses in a major city traveling ≈460,000 km over a period of 23,500 hours. A total of 128 attributes concerning trip length, distance, speed, driver behaviour, environmental conditions, auxiliary devices, etc. is sampled every 6 seconds. We discuss our dataset in detail in Section 3.

The various relevant datasets are summarized in Table 1. Among all the datasets discussed, VED is a public dataset with good distance coverage. In Section 6.5 we show the applicability of our approach to this dataset.

Model Review

Various models have been used to learn energy consumption from large-scale vehicle data. The input data for the problems is of the order of 20-30 attributes. With this size of the input data, most methods have stuck to classical machine learning methods like- Decision Tree, SVR, Random Forest,Multiple Linear Regressor (MLR) [6] or Artificial Neural Networks (Multi-layer perceptrons) over trip-level features [1, 16]. A hybrid Convolution Neural Network (CNN) and Bagged Decision Tree (BDT) model is proposed by [19]. More recently, Petcevicius et al. [23] suggest a method that breaks down the route of interest into segments. The model then predicts the speed profile for this segment. Based on the speed profile and other conditions like environmental conditions, road type, etc., the energy consumption (and the variance) is predicted for every segment and finally accumulated to get the Total Energy Consumption. The authors test sate-of-the-art (SOTA) deep learning methods like Long-Short Term Memory and Deep Neural Networks with dense connections.

Distinctions from our work

Our work differs from existing work in the following ways: (1) We collect a large dataset of a single type of electric vehicle - buses in our fleet. The task of energy consumption prediction is quite challenging as the passenger load in buses are quite dynamic as compared to passenger vehicles. (2) Our model uses a two-stage prediction approach. A primary set of features is used to predict feature engineered secondary set of features. Finally, both the primary and secondary set of features are used to predict the energy consumption for the trip. We test this approach on a variety of models and scenarios on real-world data. (3) We systematically identify the features that the energy consumption depends on and then model it via our machine learning pipeline. (4) We test our models extensively for longevity and generalization to unseen routes and vehicles. (5) We also show the applicability of the model and the engineered features, not only on our dataset, but also on a public dataset.

3 DATASET ANALYSIS

Table 1 discussed various datasets that have been proposed in the literature for the learning of energy consumption for electric vehicles. As highlighted earlier, the existing datasets have following limitations:

- *Limited data for electric vehicles*: The publicly available VED dataset also has only ∼7600 km of electric vehicle data.
- *Limited passenger load variability*: Cars have a limited variability in passenger load due to a lower capacity. In this work, we consider data from a fleet of buses. Therefore, the load on the vehicle varies significantly.
- *Limited features Collected*: In related work, while an effort has been made to record attributes concerning all aspects of the drive, in many cases, critical information such as motor

Dataset	Features	Public	Size (km)	Vehicles
VED [21]	Vehicle ID, GPS, Altitude, Vehicle Speed, Engine Signals, Ambient Temperature, Battery Usage, Auxiliary Power Signals, Battery V-I, Battery SOC	Yes	600,000 7598 (EV)	Hybrid, Battery, Fuel
Nissan [16]	Time, Distance, Speed, HVAC	No	5km $\times N_1$	Nissan Leaf 2011
Smartphone [1]	Time, Distance, Speed, Displacement, Acceleration	No	6km $\times N_2$	Mitsubishi Electric
SpritMonitor [2]	Distance, Time, Speed, Driving Style, Tires, Park Heating, A/C	Yes	4,630	Nissan iMiEV, Volkswagen E-Golf
Denmark [14]	Time, Distance, GPS, Air Temperature, Travel Time, Wind Speed, Altitude diff, Weekend/Non-weekend, Road Condition, Road Type	No	1,400,000	33 Citroen C-Zero, 56 Mitsubishi iMiEV 75 Peugeot iOn
Ours	Time, Speed, Distance, SOC, Ambient Temp, Saloon Temp, DC-DC Energy, Motor Energy, Auxiliary Energy	No	460,000	Electric Buses

Table 1: A comparison of EV datasets and the features used for energy consumption. N_1 represents 25 runs along the same 5km route. N_2 signifies 10 drivers driving through the 6km route where the number of trial runs were not disclosed.

	Trip Time (mins)	Trip Dist. (kms)	Speed (kmph)	Ambient Temp. (°C)	Saloon Temp. (°C)	CSOC (%)	DCEC (kWh)	Auxiliary (kWh)	Motor (kWh)	TEC (kWh)
mean	97.46	31.59	21.17	26.49	24.86	12.97	1.35	2.27	29.03	38.27
std	67.14	18.26	3.34	4.57	2.74	8.05	1.4	1.69	17.09	22.18
min	3	2.12	3.86	11.42	11.16	-2	-9	0.08	-96.69	1
25%	61.8	20.38	19.06	23.5	23.94	8	1	1.33	18.79	26
50%	79.8	28.62	21.05	26.91	25.75	11	1	1.86	26.24	35
75%	109.7	37.38	23.16	29.96	26.38	17	2	2.79	35.13	45
max	573	130.25	55.69	37.52	38.24	99	12	20.04	129.8	176

Table 2: Descriptive Statistics for our dataset with 27 eBuses.

	Trip Time (mins)	Trip Dist. (kms)	Speed (kmph)	Ambient Temp. (°C)	CSOC (%)	AC Energy (kWh)	Heater Energy (kWh)	TEC (kWh)
mean	9.51	4.58	39.63	12.34	4.29	0.06	0.05	0.76
std	8.94	3.57	12.02	11.47	4.15	0.1	0.15	0.71
min	0.81	0.1	3.49	-14.45	-0.12	0	0	0.01
25%	3.62	1.97	31.69	3.07	1.34	0	0	0.28
50%	7.15	3.6	39.02	11.5	3.05	0.03	0	0.57
75%	11.76	5.86	46.61	22.14	5.98	0.08	0.04	1
max	78.72	21.31	89.43	35.53	22.93	1.36	1.42	5.77

Table 3: Descriptive Statistics for VED dataset.

energy, auxiliary energy, temperature inside the car, etc. has not been recorded. In this dataset, we record these values to get an accurate picture of the vehicle conditions and the other factors that influence energy consumption of the vehicle.

Energy consumption modeling depends heavily on the choice of input features. In this work, we model the energy consumption on a trip-level. Therefore, the features are accumulated at the trip level before the energy consumption is estimated. Our dataset consists

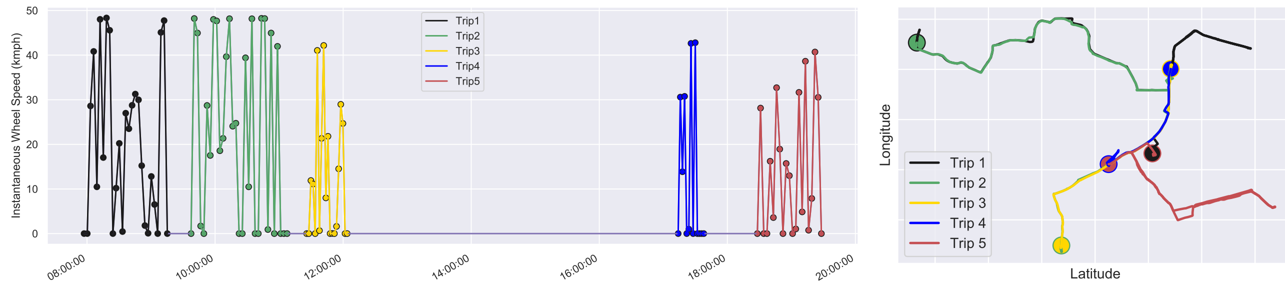

Figure 2: Trip segmentation based on 15 mins stoppage duration in our dataset. Each color represents a segmented trip.

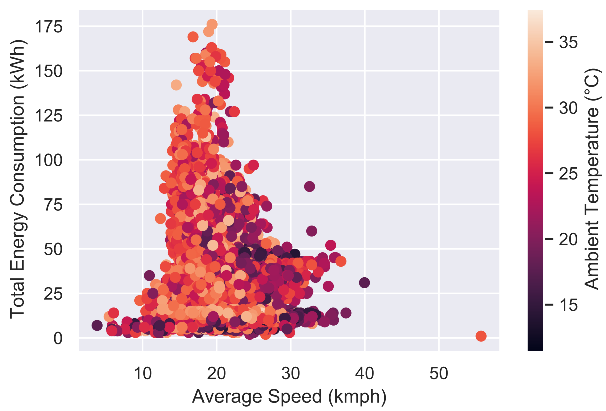

Figure 3: Total Energy Consumption for different Average Speed and Ambient Temperature values across trips in our dataset.

of 14413 trips as compared to 503 in the VED dataset and has over ~60× km of vehicle data collected.

A trip is defined as a period of vehicle data where the ignition is on and it has less than 15 minutes of stops, i.e., we instantiate a new trip if the vehicle stops for more than 15 minutes (continuously) while traveling on a route.

We define our trip in a way that active periods of the vehicle's motion are modeled, i.e., we remove periods where the vehicle is stopping for a long period of time for breaks, etc. This allows our features to represent the vehicle in motion only. Table 2 shows a roster of the trip-level features recorded and their values. A bus trip, as in our case, is 97 mins on average. In comparison, we show the statistics from the VED dataset in Table 3. An average trip lasts just 9 mins. The speed variation is much higher in the VED dataset since the data is for cars as opposed to buses in our dataset. The temperature variation in our data is from 11-37 degrees Celsius. The average SOC consumed is at 12%.

Figure 2 shows the example trip segments within a route (each marked with a different color) segmented based on the above definition. As one can observe, the trip represents sustained activity in the bus for varying amounts of time. Figure 2 also shows the geographical variation (masked lat, long) within the trip.

Our dataset consists of time series parameters reported every 6 seconds from the vehicle. The data is sent to a centralized server in real-time over a 4G network. The higher sampling frequency of the time series data helps in accurate and precise data collection of primary and secondary features- e.g. to compute the trip distance accurately based on latitude, longitude recorded, or to determine harsh braking, acceleration incidents within a trip. This further helps to provide accurate insight into the energy consumption. Given the goal of estimating energy consumption for a trip, we use only trip-level features, i.e., aggregating data for an entire trip.

Challenges in Energy Modeling

Accurately predicting total energy consumption of an EV before the trip starts is challenging since it is dependent on several directly observable/ non-observable factors. Figure 3 shows the distribution of total energy consumption (TEC) for different average speed and ambient temperature values. First, we show that TEC varies significantly across different vehicle speeds and ambient temperature. Even for trips with fixed average speed and ambient temperature, the TEC values have high variance. Thus, modeling a subset of features such as speed or temperature, by itself is not a good indicator of energy consumption levels. This necessitates a deep dive into understanding the factors affecting energy consumption holistically- including external factors like weather and traffic conditions, vehicle internal factors like saloon temperature, load etc., as well as driver behaviour parameters like harsh accelerations and braking.

4 PROPOSED METHOD FOR EV ENERGY CONSUMPTION MODELLING

Our goal is to predict the energy consumption of a eBus trip "before the start of the trip". Our approach to this goal has been to develop a model that can generalize to any route given the basic features that the energy consumption is dependent on. Therefore, our learning strategy does not take route specific energy consumption into account. Rather, we rely on the features described in this section for learning the energy consumption model. To this end, we propose various set of features, which aid our analysis towards energy consumption modelling. Apart from the features, we also discuss the ML pipeline and the models used along with the evaluation metrics for our task.

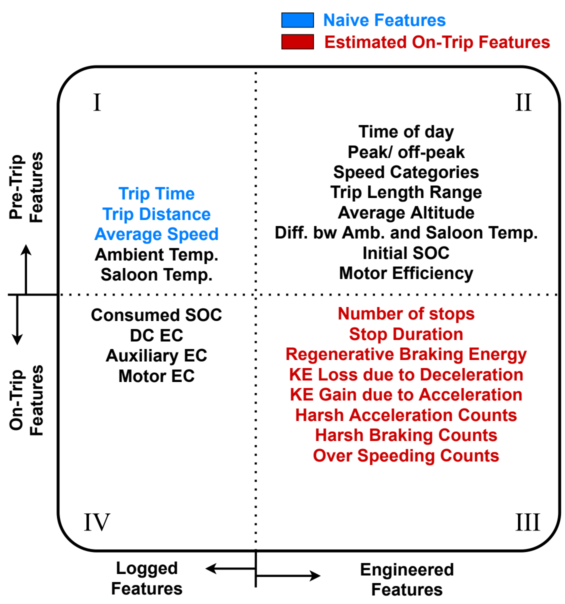

Figure 4: A comprehensive set of features proposed and constructed in this work.

4.1 Features

Figure 4 shows all the features constructed in this paper. Features in Quadrant I and IV are readily-available and logged from the data collected, and the features in Quadrant II and III are the key features constructed (estimated) based on extensive feature engineering, which is a core contribution of this work. Quadrant I and II features are available before the start of the trip (called as Pre-trip features). Quadrant III and IV features are derived during the trip (called as On-trip features). In this section, we present different feature sets used for energy consumption modelling.

4.1.1 NF: Naive Features. We start our analysis with a baseline feature set for energy consumption estimation. Here, we use just the three basic features readily available from pre-trip features - trip distance, time and average speed during the trip (shown in Blue color Quadrant I, Figure 4). In this work, we consider the trip distance, speed and time as pre-trip features since these are roughly known a priori for the bus route based on time of the day. Even approximate values of these features provide a good baseline for the energy consumption estimation. To further validate this assumption, we do a sensitivity analysis of these features in Section 5.

- **Trip Time:** The trip time is derived as the difference in timestamps at the beginning and the end of the trip.
- **Trip Distance:** Using the recorded GPS attributes, we calculate the distance of the trip as, $D = \sum_{i=1}^{N} H(P_1, P_2)$, where N is the number of samples (note that each sample is recorded after 6 seconds in our dataset) within the trip and H is the Haversine [5] distance between the geospatial points $P_1(lat_1, long_1)$ and $P_2(lat_2, long_2)$.
- **Average Speed:** Average speed is derived from the mean wheel based speed (tachometer) for the vehicle during the trip.

4.1.2 LF: Logged Features. In the Logged Features set, we include all the features logged by the vehicle during the trip. The intent here is to find the features which correlate the most with the output variable (TEC, in this case) and then use only these features for regressing the target. The following features are logged in the vehicle alongside the Naive Features:

- **Ambient Temperature:** The mean ambient temperature (in °C) for the trip.
- **Saloon Temperature:** The temperature inside the bus cabin is referred as the saloon temperature (in °C). This indicates the usage of the Heating, Ventilation and Air Conditioning (HVAC) system of the bus.
- **Consumed state of charge (CSOC):** We record the charge depletion of the high voltage battery during the trip. The consumed state of charge (in %) is.

$$CSOC = SOC_s - SOC_e \tag{1}$$

where SOC_e denotes the battery SOC at the end of the trip, SOC_s is the SOC at the start of the trip (as % of the battery capacity).

- **DC-DC Energy Consumption (DCEC):** The motor drive within a bus runs on High Voltage DC. Thus, a DC-DC converter is used to convert the battery voltage to HVDC. The energy lost in the switching activity is used as the DC-DC energy consumption feature.

$$DCEC = DCEC_e - DCEC_s \tag{2}$$

where $DCEC_e$ denotes the DC-DC Energy Consumption at the end of the trip, $DCEC_s$ is the DC-DC Energy Consumption at the start of the trip (in kWh).

- **Auxiliary Energy Consumption (AEC):** Auxiliary energy in the vehicle is consumed by devices such as lights, fans, power steering, windscreen wiper motor, head-lights and rear-lights and even the HVAC system. The total energy consumed by the auxiliaries directly effect the energy consumption from the battery. The auxiliary energy consumption (in kWh) is defined as,

$$AEC = \sum_{i=1}^{N} V_b(i) I_a(i)(t(i) - t(i-1)) \tag{3}$$

where N is the number of samples within the trip, V_b is the battery pack voltage (in volts), I_a is the total auxiliary input current (in Ampere), t represents the timestamp (in seconds).

- **Motor Energy Consumption (MEC):** The instantaneous power values of the motor voltage and current are used to accumulate the energy used during the trip time. The total Motor Energy Consumption (in kWh) is defined as,

$$MEC = \sum_{i=1}^{N} V_m(i) I_m(i)(t(i) - t(i-1)) \tag{4}$$

where N is the number of samples within the trip, V_m is the motor input voltage (in volts), I_m is the motor input current (in Amperes), t represents the timestamp (in seconds).

4.1.3 *Engineered Features (EF).*

Apart from the naive and logged features, we define a set of features based on domain knowledge and real-world experiments, that is critical to model TEC. The proposed features broadly belongs to traffic conditions, driving behavior, geography and environmental conditions, and travel specifications. A subset of these features can be computed before the start of the trip (Quadrant II) and the remaining are estimated based on the available data (Quadrant III) as described in Section 4.3. We enumerate the specific features below:

- **Traffic Conditions:** The acquisition of real-time traffic information is a challenge and the prediction of its behaviour before starting on a journey is all the more difficult. But, from the perspective of TEC estimation, traffic conditions directly correlate with the driving behaviour which influences the energy consumption. Hence, it becomes extremely important to engineer traffic behaviour. Following engineered features help capture traffic conditions apriori to the trip:
  - **Time of Day (TOD):** A whole day has been divided into different parts - Morning (from 6:00hrs to 12:00hrs), Afternoon (from 12:00hrs to 16:00hrs), Evening (from 16:00hrs to 21:00hrs) and Night (from 21:00hrs to 6:00hrs). These different time frames capture the traffic behaviour accordingly and hence provide the model with the required traffic information.
  - **Peak/ Off-Peak Hours:** Peak and off-peak hours are a direct indication of the traffic along a route in a city. The busiest hours or peak hours are selected from 6:00 to 8:00 hrs, 11:00 to 16:00 hrs and 20:00 to 23:00 hrs and the left out time periods as the off-peak hours. Selection of the time-frames were rightly acknowledged by the traffic supervisors in the area.
  - **Number of stops:** While in motion, a vehicle can stop at traffic signals and for passenger entry/get down. Traffic density levels is another factor which can also define the number of stops and is a variable quantity. Hence, more the number of stops, higher is the traffic density and vice versa. Every time, the wheel based vehicle speed parameter drops to zero for few seconds (e.g., >10 seconds), it gets counted towards the number of stops.
  - **Stop Duration (stopdur):** This feature represents the total stop time (in minutes) within a trip that corresponds to the stops because of high traffic density. This is computed by summing the time period blocks when the wheel based vehicle speed is zero within a trip.
- **Driving Behavior:** Driving behaviour significantly contributes to the TEC estimation. Although driver specific information is difficult to capture, one of the main applications of TEC estimation traces back to rewarding drivers based on the energy consumption for the trip and encourages efficient driving skills. This helps in reducing the battery degradation rate and preserving battery life cycles. Following are the features that indirectly help engineer driving behaviour:
  - **Speed Categories:** This feature is engineered directly from the average speed as one of the categorical variables. From the data distribution as well as the domain expertise, speed ranging from 15kmph to 25kmph is considered as normal. Less than 15kmph is termed as slow, possibly because of higher traffic densities. More than 25kmph is fast driving indicating the vehicle is being driven on a highway or possibly, where traffic density is less.
  - **Regenerative Braking Energy (RBE):** The regenerative action in electric vehicle occurs during braking which helps in accumulating energy being fed back to the battery for charging while still on motion. Miri et al. [18] demonstrate that the regenerative action proves to be inefficient at fairly low speeds and hence is set to zero. Thus, we use a speed threshold s_{th} over which the regenerative action is considered. The regeneration only happens while decelerating, therefore regenerative action is enabled if the deceleration is less than d_{th}. Above this limit, it indicates a higher braking torque demand resulting in dynamic braking where friction brakes are utilized. Further, the regenerative action is disabled at higher battery levels to prevent overcharging. We use $s_{th} = 15kmph$ and $d_{th} = 0.7ms^{-2}$, derived empirically. Using the above three conditions, the regenerative braking energy (in kWh/kg) is derived as,

$$RBE = \sum_{\substack{i=1 \\ v(i) \geqslant s_{th} \\ 0 < d(i) < d_{th} \\ soc(i) < 95}}^{N} (v(i)^2 - s_{th}^2) \tag{5}$$

    where N is the number of samples within the trip, v represents the vehicle speed (in kmph), d represents deceleration (in m/s^2), and, soc represents the state of charge of the battery (in %).
  - **Kinetic Energy Loss due to deceleration:** This feature captures the lost kinetic energy in the trip at the time of decelerating. It is computed by:

$$LKED = \sum_{\substack{i=2 \\ v(i) \leqslant v(i-1)}}^{N} (v(i-1)^2 - v(i)^2) \tag{6}$$

    where N is the number of samples within the trip, v represents the vehicle speed (in kmph) and LKED is Lost Kinetic Energy due to deceleration (in kWh/kg).
  - **Kinetic Energy Gain due to acceleration:** This feature captures the gained kinetic energy in the trip at the time of accelerating. It is computed by:

$$GKEA = \sum_{\substack{i=2 \\ v(i) \geq v(i-1)}}^{N} (v(i)^2 - v(i-1)^2) \tag{7}$$

    where N is the number of samples within the trip, v represents the vehicle speed (in kmph) and GKEA is Gained Kinetic Energy due to acceleration (in kWh/kg).
  - **Harsh Acceleration (HA) Counts:** When the acceleration of the vehicle rises above a threshold of $1.5m/s^2$, it gets added to the harsh acceleration counts.
  - **Harsh Braking (HB) Counts:** If the deceleration of the vehicle drops below a threshold of $1.5m/s^2$, it is considered as harsh braking.

  - **Over Speeding (OVS) Counts:** Whenever the vehicle speed exceeds a set threshold (say, 50 kmph), it is classified as a over-speeding category.
- **Geography and environmental conditions:** The geographical as well as the environmental conditions have a significant effect on the estimation of TEC. For instance, **average altitude (in metres)** of a trip represents the geographical conditions. Similarly, ambient temperature and saloon temperature moulds the environmental conditions to some extent. However, since we cannot measure HVAC condition or consumption, we monitor the **difference between ambient and saloon temperature** to help in capturing when the heater and air conditioner is switched on respectively (as shown in Figure 5). A positive difference value indicates that A/C is ON whereas a negative value proves that the temperature inside the vehicle is higher referring to the working of the heater.
- **Travel Specifications:** This set of features directly relates to the actual trip type and the specification of the vehicle used in the trip. These features add value to the TEC estimation as discussed below:
  - **Trip Length Range** This is a categorical feature which is directly computed from the trip time and the distance. A trip time within 45mins to 2hrs is considered an average time length of trip. Trips with greater than 2hrs trip time may indicate going to a destination far off or having high traffic density. Less than 45mins trip time means either the destination is very close or the traffic is less dense.
  - **State of Charge prior to the trip (initial SOC):** The initial state of charge of the battery indirectly determines the rate of discharge of battery while on a trip. Hence it is essential to have it as one of the features that can help the driver get an idea of the energy he would consume by the end of the trip.
  - **Motor Efficiency (η) :** These features help model the motor characteristics and the losses that may occur within the motor. Mathematically, it is computed by:

$$MOE = \sum_{i=1}^{N} T_m(i)\omega_m(i)(t(i) - t(i-1)) \tag{8}$$

$$\eta = \frac{MOE}{MEC} \tag{9}$$

where, T_m is torque output from the motor (in Nm), ω_m is angular speed of the motor (in rpm), MOE is motor output energy (in kWh), and MEC is motor energy consumption discussed under LF feature set.

As noted earlier, a subset of these engineered features can be readily *computed* pre-trip and a subset needs to be *estimated* for a given trip. To this end, we develop tiny regressors to predict these features ahead of the trip as discussed in Section 4.3.

Given that we have defined all the features above, we will now define combination of feature sets to study the importance of engineered and logged features. Specially, we define the following feature sets to perform ablation study and show the importance of the engineered features.

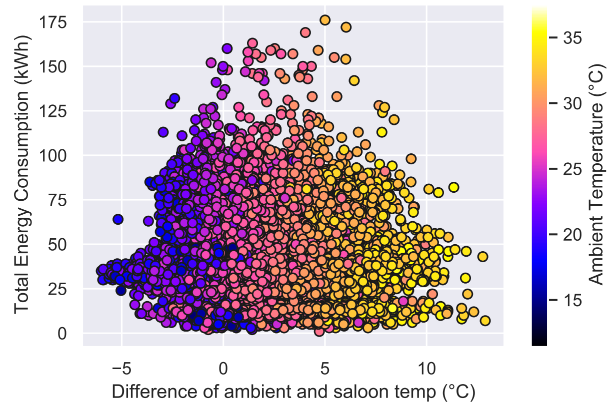

Figure 5: Identifying operation of A/C and heater from the difference between ambient and saloon temperature.

(1) **Naive Features:** (Quadrant I) These are the three basic features such as trip time, duration and avg. speed that are available pre-trip and marked in blue in Quadrant I.
(2) **Logged Features:** (Quadrant I & IV) These are all features that are recorded by the vehicle before and during the trip.
(3) **PF(L) Pre-Trip Features - Logged:** (Quadrant I) This feature set depicts the features that are available and logged before the start of the trip. *This represents the primary set of features in our work to model the energy consumption.*
(4) **PF(LEF) Pre-Trip Features - Logged and Engineered** (Quadrant I & II) This feature set depicts the features that are available/logged or can be computed before the start of the trip. This includes a subset of features from both logged and engineered features.
(5) **PF(ON) Pre-Trip and estimated On-Trip Features** (Quadrant I, II & III) This feature set includes combination of pre-trip logged and all proposed engineered features (those that are readily available pre-trip and those that are estimated using the predictors). This feature set is the proposed set that captures all the relevant features, which is used by our two-stage pipeline towards estimating energy consumption accurately prior to the trip.
(6) **ALL features:** (Quadrant I, II, III, & IV) This feature set includes all of pre and on trip features This is a comprehensive feature set with the combination of all pre-trip and on-trip features including both logged and engineered features. This acts as the best case feature set if the on-trip data is also available, which is infeasible when computing energy consumption apriori.

4.2 Ground-Truth Label Computations

The ground truth (GT) is the target variable which we try to predict using the machine learning models. We calculate the GT Total Energy Consumption (TEC) using the input features, i.e.,

$$TEC = BatteryOut_e - BatteryOut_s \tag{10}$$

where, $BatteryOut_e$ and $BatteryOut_s$ represents the outgoing energy from the battery (in kWh) at the end and start of the trip, respectively.

4.3 Two-Stage Pipeline to Model EV Energy Consumption

In this Section, we discuss our two-stage modelling approach to accurately predict Total Energy Consumption (TEC) per trip. Figure 6 presents the two-stage ML pipeline designed for per trip TEC estimation.

Stage 1: In this stage, we determine the potential pre-trip features that are readily available for energy consumption modeling. This includes both naive features like trip time, distance, speed and also the engineered features like Time of day, peak/off-peak, etc., in Quadrant II.

Stage 2: In stage 2, we develop tiny ML predictors to estimate the engineered features in Quadrant III, that are not readily-available pre-trip. The input to these predictors are pre-trip features from stage I and output corresponds to the specific engineered feature. We use various ML models- like LightGBM, AdaBoost, RandomeForest- to train the individual feature predictor for the engineered features such as harsh acceleration, braking, over speed, stop duration, number of stops, etc. We describe the specific models used, their hyper-parameters, and their performance in Section 5.1.5, Table 6.

Finally, our solution employs a regression model that takes both pre-trip and predicted on-trip features from stage 1 and 2 to accurately estimate the energy consumption of the trip. In this work, we employ and compare several traditional models such as multiple linear regressor (MLR) [13], kernel support vector regressor (KernelSVR) [8], decision tree (DT) [24], random forest (RF) [3], adaptive boosting (AdaBoost) [25] and light gradient boosting (LightGBM) [12] regressors along with deep multilayer perceptron (Deep MLP) networks for prediction of the Total Energy Consumption (in kWh).

Traditional models: Figure 7 shows the linear correlation of TEC with various features. We can see that TEC correlates well with trip distance, trip time and KE Gain due to acceleration. Thus, indicating multiple linear regression (MLR) model may perform well on our dataset. Apart from MLR, we also consider Support Vector Regressors using linear and radial basis function (RBF) kernels, which provide an edge in some cases because of its nature of easy adaptability. Decision Trees learn simple rules from the training data features that help in prediction using the iterative dichotomiser algorithm [11] working at its baseline. Our paper uses DTs with a max depth of 7. DTs suffer from low bias, high variance problems which are corrected using ensemble algorithms. Random Forest is one of the bagging based ensemble techniques that uses bootstrap aggregation of the DT based base learners to predict an output. In our paper, RF uses 200 base estimators for prediction to achieve satisfactory performances with a max depth of 7 for the base learners. Adaptive Boosting and Light gradient Boosting [12] are two of the boosting based ensemble methods which also use DTs as their base models. The computation time of LightGBM is lesser than most of the ensemble algorithms since the algorithm is accelerated by the use of leaf wise tree growth, without any noticeable deterioration of the model performances. LightGBM also uses 200 base estimators each with 100 maximum leaves and max depth of 8. AdaBoost uses 800 base estimators for the Total Energy Consumption prediction. All the above mentioned hyper parameter values were obtained using Grid Search Cross-Validation (GridSearchCV) method of the scikit-learn library [22].

Deep Learning (DL) Models: The deep learning paradigm has demonstrated a way to learn the required features from the raw data automatically without feature engineering in several application. Thus, to validate if the extensive feature engineering is essential in this particular use-case or if deep learning can learn intermediate representations directly, we use the Deep MLP to predict the TEC directly using the pre-trip PF(L) feature set, which includes only the logged pre-trip features and NOT the engineered features. We then compare the results from traditional models, i.e., PF(ON), which includes engineered features with the described DL models. The results are shown in Section 5.2. The architecture for the Deep MLP Network used is a 3 hidden layers network with 25, 20 and 9 hidden neurons respectively. This model was obtained after a series of experimentation using the scikit learn wrapper class offered by Keras [4] along with GridSearchCV which helped in hyperparameter tuning of the number of hidden layers and the number of hidden neurons accordingly. The model uses He Normal[10] initialization with Rectified Linear Unit (ReLU) as the Activation Function for the hidden neurons. The output layer activation function is set as linear, since it is a regression problem. Adaptive Moment Estimation (Adam) is used as the optimizer, with mean absolute percentage error as the loss function for the training process. Early stopping monitoring over the validation loss with a patience of 10 iterations is used to ensure the model is free from over-fitting issues. Finally, the model is trained over 200 epochs using a batch size of 32 to produce the best possible performance scores.

4.4 Performance Metrics

The estimated Total Energy Consumption (kWh) results are evaluated against a number of performance metrics.

- **R^2 score:** It is commonly known as the coefficient of determination. The closer the R-squared value is to 1, the better is the model. Mathematically,

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{11}$$

where m is the number of test samples, $\hat{y}_i$ is the predicted value of y_i and $\bar{y}$ is the mean value of y

- **MAPE:** Mean Absolute Percentage Error is sensitive to relative errors. The closer it is to 0, the better are our models. It can be denoted by:

$$MAPE = (\frac{1}{m}\sum_{i=1}^{m}\frac{\mid y_i - \hat{y}_i \mid}{y_i}) \times 100 \tag{12}$$

where m is the number of test samples and $\hat{y}_i$ is the predicted value of y_i

- **MedAPE:** Median Absolute Percentage Error is not sensitive to outliers. Hence it helps give a more accurate insight of how the models are performing. Similar to MAPE, values closer to 0 indicates better models.

$$MedAPE = median(\frac{\mid y_i - \hat{y}_i \mid}{y_i}) \times 100 \tag{13}$$

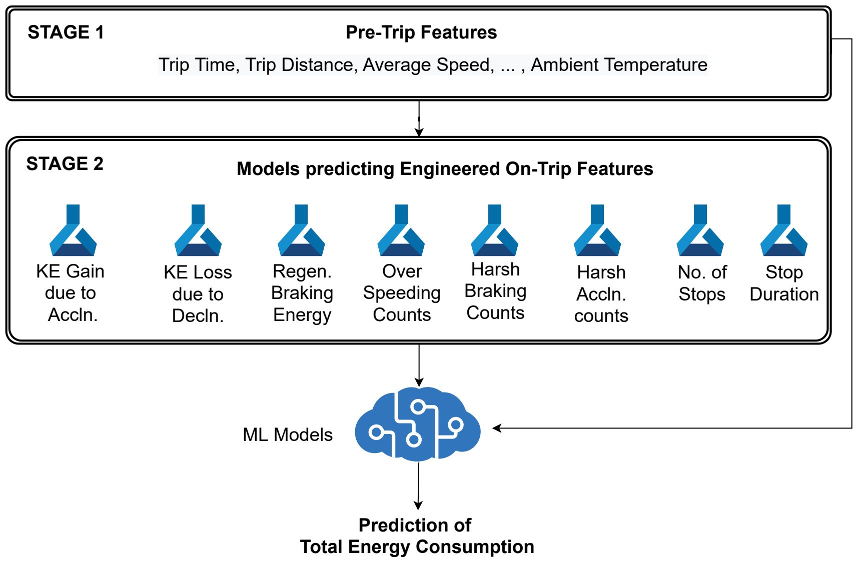

Figure 6: Two-stage pipeline to model EV energy consumption

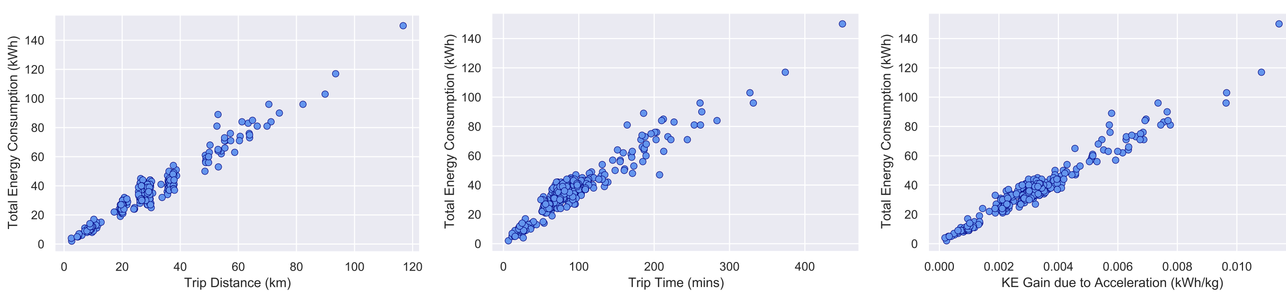

Figure 7: Positive correlation of TEC with Trip Distance, Trip Time, Kinetic Energy Gain due to Acceleration (top correlations).

- **MAE:** Mean Absolute Error is relative to the errors. It is computed by:

$$MAE = \frac{1}{m}\sum_{i=1}^{m} \mid y_i - \hat{y}_i \mid \tag{14}$$

- **RMSE:** Root Mean Squared Error is denoted by:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{15}$$

We use R^2 score, MAPE and MedAPE for evaluation in most experiments. Section 6.5 uses MAE and RMSE to compare model performances on VED dataset[23].

Model	Metric		
	R^2	MAPE	MedAPE
MLR	0.952	10.659	8.635
KernelSVR	0.952	9.974	8.433
Decision Tree	0.956	9.372	7.631
Random Forest	0.957	9.260	7.583
MLP	0.953	9.596	8.367
LightGBM	**0.957**	**9.262**	**7.473**
AdaBoost	**0.955**	**9.350**	**7.494**

Table 4: Traditional models performance on the naive feature set (NF)

5 EXPERIMENTS

We now show the model performance using different feature sets described in Section 4.1 on our dataset. Our dataset contains 14413 trips from 27 buses and the data is split randomly into 70/10/20(%), i.e. 10089, 1441, 2883 trips for training, validation and testing.

#	Feature	Model	Metric		
			R^2	MAPE	MedAPE
1	NF	LGBM	0.957	9.262	7.473
		AdaBoost	0.955	9.35	7.494
2	LF	LGBM	0.989	4.369	3.396
		AdaBoost	0.988	4.572	3.584
3	PF(L)	LGBM	0.965	8.072	6.485
		AdaBoost	0.963	8.154	6.769
4	PF(LEF)	LGBM	0.968	7.623	6.112
		AdaBoost	0.967	7.768	6.227
5	PF(ON)	LGBM	0.99	4.294	3.064
		AdaBoost	**0.993**	**3.027**	**2.213**
6	ALL	LGBM	0.994	3.745	2.867
		AdaBoost	**0.998**	**1.516**	**1.118**

Table 5: Performance of LGBM and AdaBoost models on various proposed feature sets.

5.1 Model performance on different feature sets

Effect of model performance on the user : Increasing the model performance results in improvements in all subsequent applications, like battery sizing, range estimation, etc. For example, for an average eBus with 300 kWh battery capacity, a 250 km (10 trips) range is expected. With an MAPE of 10% (NF Set), the *safe range* would be ~225 km. If the model has an MAPE of 3% (PF(ON)), the *safe range* would come to 242.5 km. Thus, the improvement in MAPE from 10 to 3% would result in extension of the safe range by ~17.5km. The model performance achieved through different feature sets is discussed in the following subsections.

5.1.1 Naive Feature Set: NF. Table 4 shows the results for classical and deep learning based machine learning models for regressing the Total Energy Consumption (TEC) of the trip using only the most intuitive features described as the Naive Feature set. The naive features from a trip establish the baseline for the error. As we can see, most of the models have an MAPE of 9-10%. In the remainder of the paper, we show the results only for LightGBM and Adaboost models for different feature sets, since these models also have a reasonable median error (MedAPE) compared to others.

5.1.2 Logged Features : LF. With Logged features, we use all the variables recorded directly from the dataset as input to the regression models. This gives us a clear target that can be achieved for the regression problem and help estimate the relative importance of the input features for regressing the output. Specifically, this set includes all the energy consumption features that are directly available during the trip with respect to DC-DC, and motor energy. In a typical setting these features are not available apriori. Table 5 (row 2) shows that the Logged features results in 0.98 R^2 score with MAPE of 4.3-4.5% (50% decrease in MAPE as compared to NF). This is mainly due to the inclusion of features which are available during the actual trip, e.g., motor energy, auxiliary energy, etc. This

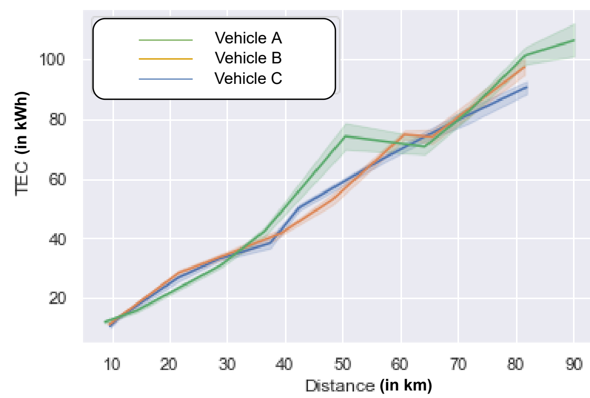

Figure 8: Variation of total enegy consumption (TEC) along with the bounds when the distance parameter is perturbed by ±10%.

Estimated On-Trip Features	Models	Metric Hyperparameters	R^2
HA	LightGBM	nEstimators = 200, max depth=8	0.71
HB	AdaBoost	nEstimators = 800, max depth =7	0.84
OVS	LightGBM	nEstimators = 200, max depth = 8	0.52
stopdur	LightGBM	nEstimators = 200, max depth=8	0.97
no. of stops	RF	nEstimators = 800	0.85
RBE	RF	nEstimators = 800	0.86
LKED	AdaBoost	nEstimators = 800, max depth =7	0.97
GKEA	AdaBoost	nEstimators = 800, max depth =7	0.97

Table 6: Performance of feature predictors to accurately estimate engineered on-trip features.

is considered as the best case baseline and our goal is to achieve this error without using any of the logged features (i.e. without using any actual on-trip features) and using the proposed engineered features and the two-stage pipeline.

5.1.3 Pre-Trip Features (Logged): PF(L). PF(L) includes only the pre-trip features described in Figure 4-Quadrant I . Table 5 (row 3) shows that PF(L) with only a limited number of features performs poorly with 8% MAPE error for estimating TEC. However, this is slightly better than the NF set which just uses the three pre-trip features.

5.1.4 Pre-Trip Features: PF(LEF). PF(LEF) includes both logged and engineered features that are available prior to the trip (Figure 4 -Quadrant I and II). Using this feature, the models are slightly able to improve upon PF(L) with an MAPE of 7.6% (Table 5 (row 4)). This indicates that pre-trip engineered features are contributing to reducing the energy prediction errors. We also inspected the importance of saloon and ambient temperature related features explicitly since these were typically not captured in other datasets. By excluding these features, the MAPE increases to 8.1%. Thus, it would always be helpful to capture these features, however, regardless, our model uses the rest of the features well to provide a comparable accuracy.

5.1.5 Pre-Trip and estimated On-Trip Features: PF(ON). PF(ON) feature set includes all the pre-trip and estimated on-trip features. As noted earlier, our two-stage ML pipeline operates on this feature set, where pre-trip features are used to estimate the on-trip features in Figure 4- Quadrant III. Table 6 shows the performance of feature predictors across all on-trip engineered features. We can see that most of the predictors have high R^2 score indicating a good performance of the individual predictors.

Our two-stage pipeline uses the combination of pre-trip and estimated features, to model the energy consumption. Table 5 (row 5) shows that with PF(ON) on our two-stage pipeline we can achieve MAPE error of 3-4%, which is 67% decrease in error compared to Naive features (row 1) and 33% decrease as compared to Logged features (row 2), which are derived by collecting the data during both pre and on trip. Thus, showing the impact of engineered features on total energy consumption estimation of EV vehicles.

We also demonstrate the sensitivity of this model to the variation in trip distance in Figure 8. We found that the change in TEC is < 5.5% for a 10% change in the trip distance. For a small distance error of 1 km, the change in TEC is <1.17%. Similar analysis has been done for speed and time and the resulting TEC variation was minimal.

5.1.6 All features: ALL. In this experiment, we use all the features presented to derive the least possible error for TEC estimation. Note that, since this uses on-trip logged features (Quadrant IV), this feature set is not feasible to use in real scenarios as it requires monitoring during the trip. Nonetheless, it can be used to compare the MAPE errors obtained from our two-stage pipeline. Table 5 (row 6) shows that MAPE using ALL features is the least across different sets. Our proposed two-stage pipeline with PF(ON) is very close to the MAPE error derived using ALL features, showcasing the benefit of the engineered features.

We test the generalization of these results thoroughly in Section 6 using the same PF(ON) feature set on our two-stage pipeline.

5.2 Feature engineering vs Automatic feature learning : Deep Learning Baseline

In the previous Section, we showed the MAPE errors obtained by our two-stage pipeline with PF(ON) feature set (See Table 5 (row 5)). We will now show the MAPE errors when we use a end-to-end neural network to automatically learn the intermediate features from input data (i.e. PF(L) features) and estimate TEC. Table 7 compares the result from deep MLP model with the LightGBM and AdaBoost models using our two-stage pipeline. Deep MLP fails to learn the automatic relationships from the raw data and we can see that MAPE of deep MLP model is 65% higher than the MAPE result derived using our two-stage pipeline with pre-trip and proposed engineered features. Thus, without feature engineering leveraging the domain knowledge, it is not feasible to apply a end-to-end network to model energy consumption in EVs, especially due to variations across several factors such as traffic, driver, vehicle, environment, etc.

5.3 Model Explanation

A major advantage of feature engineering process and usage of traditional models is the explainability of the results. To this end, we now show how engineered features and feature correlation help our models to estimate energy consumption accurately.

#	Feature	Model	Metric		
			R^2	MAPE	MedAPE
1	PF(L)	MLP	0.958	8.639	7.465
2	PF(ON)	LGBM	0.99	4.294	3.064
		AdaBoost	**0.993**	**3.027**	**2.213**

Table 7: Comparison of Deep learning MLP model (automatic feature learning) with our Two-stage pipeline (with feature engineering).

5.3.1 Validation of feature variation. To show feature explainability, let us take two trips from two different vehicles in our dataset. Trip 1005 and Trip 2220, both having same total energy consumption of 30kWh for the trip with different distances travelled, i.e. 22.25 km for the former and 35km for the latter. Both these trips were at the same time of day and we also noticed the auxiliary energy from these vehicles had similar energy consumption i.e. 1.5kWh and 1.4kWh. Thus, it is non-trivial to understand the cause of such high energy consumption in case of trip 1005 with shorter distance compared to trip 2220.

Through the proposed feature engineering process presented in Section 4.1, we can now determine what is the cause for such high variation in energy consumption. From our analysis using the engineered features, it turns out that the driver in trip 1005 was driving rashly with high harsh accelerations and braking as opposed to driver in trip 2220. Thus, if driver behavior is not modelled, it is non-trivial to provide reasoning for this variation. Further, with the help of domain knowledge and feature engineering, we cover most of the aspects that impact energy consumption.

5.3.2 Correlation Validation. To understand if the ML models are learning the correct correlation with respect to the features, we check the feature importance values of the models.

We validate this with the Pearson's Correlation Coefficient [7]. As per the Pearson's Correlation Coefficient of TEC and the respective feature, the top and worse correlations are plotted in Figure 7 and 9. Figure 7 shows the features that correlate the most with the model output variable, i.e. Total Energy Consumption. We found that trip distance, trip time, kinetic energy gain and lost during acceleration and deceleration respectively correlate the most with the energy consumption. The least correlated attributes were speed, initial soc and ambient temperature as shown in Figure 9. We observe that the feature importance for most correlated features are in line with the top correlations. Similarly, the bad correlation features have lower feature importance as shown in Table 8.

5.4 Discussions

We showed that our proposed methodology can reduce the MAPE errors close to 3% on our dataset and we now present few limitations and opportunities to further reduce the error in energy consumption estimation.

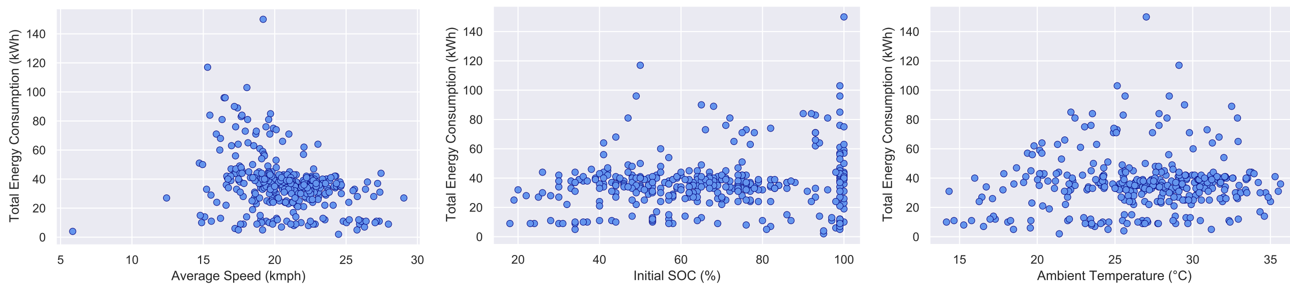

Figure 9: Correlation of TEC with Average Speed, Initial SOC, Ambient Temperature (bad correlations).

Features	Models LGBM	AdaBoost
trip time	**0.11143**	0.02861
trip distance	0.09633	**0.79084**
avg. speed	0.05133	0.00199
ambient temp.	0.06800	0.00202
diff. of amb. and sal. temp.	0.05100	0.00887
TOD	0.02000	0.00066
peak/ off-peak	0.00433	0.00014
speed categories	0.00033	0.00004
trip length range	0.00066	0.00015
avg. altitude	0.00423	0.00197
init. soc	0.04367	0.00179
motor eff.	0.06033	0.00458
no. of stops	0.03900	0.00326
stopdur	0.03733	0.00197
RBE	0.08933	0.00603
LKED	0.07566	0.09048
GKEA	0.10600	0.04638
HA	0.05100	0.00192
HB	0.08567	0.00816
OVS	0.00433	0.00009

Table 8: Feature Importance values across different models.

(1) Due to cost and privacy implications, there is a limited observability on passenger loads. In large vehicles such as buses, this can be a significant factor towards energy consumption modelling. One way to further improve MAPE errors is by indirectly monitoring the passenger load through vehicle torque etc.
(2) Due to limited memory and instrumentation overheads, we could only record the battery out and state of charge at integer precision. Therefore, our results only reflect the accuracy to the said precision.
(3) Battery state of health is another parameter that impacts energy consumption. The remaining battery life cycles determine the rate of discharge of the battery and hence contribute to the total energy consumption. Since we consider different vehicles, each vehicle may have batteries of varying remaining life, which if considered in the modelling beside the proposed features can significantly reduce the errors. We plan to include this feature as part of our future work.

6 GENERALIZATION STUDY

Till now, we used 70% of data from all vehicles to train the model and remaining was used for validation (10%) and testing (20%). We now present detailed analysis when our two-stage pipeline with PF(ON) feature set is applied to different amounts of training data, past data, number of vehicles and routes.

6.1 Training with lesser data

Table 9 shows the performance of our two-stage pipeline with PF(ON) features with the reduction in training data from 80 to 50 to 20%. The number of trips used for training are 11530, 7206 and 2882 subsequently for the 80-50-20(%) cases. The MAPE results show a slight deterioration as we reduce the training data, but is still within the acceptable ranges. For instance, MAPE increases by just 1% with reduction of training data from 80% to 50%. This demonstrates the full potential of the models trained on the proposed feature set: PF(ON) which generalizes quite well even with a smaller training set.

6.2 Longevity Study

We now show how the proposed model can generalize when trained on trips from one time period and tested on trips from another time period. We used 2 months of data with a total of 6266 trips as training data and another 2 months data with 8147 trips as test data (both having non-overlap time periods). The AdaBoost model performance for this case resulted in MAPE of 6.44% which shows that the model trained on data from one time period (past data) can generalize well to other time periods. Extensive longevity studies were not possible as the data collection is still ongoing and hence this analysis was restricted with two months of training data and two months of test data. We also performed other splits such as one month training data and three months test data and so on, and observed similar performance.

6.3 Vehicle Generalization

Until now, we showed the performance of our two-stage approach when the training data included data from all vehicles together. In this section, we use data from a subset of the vehicles for training and test on the other subset of vehicles (non-overlapping vehicles). We filter the vehicles used for training on the basis of domain knowledge and feature distribution, i.e., we first select the trips (and associated vehicles) which has diverse range of feature values in

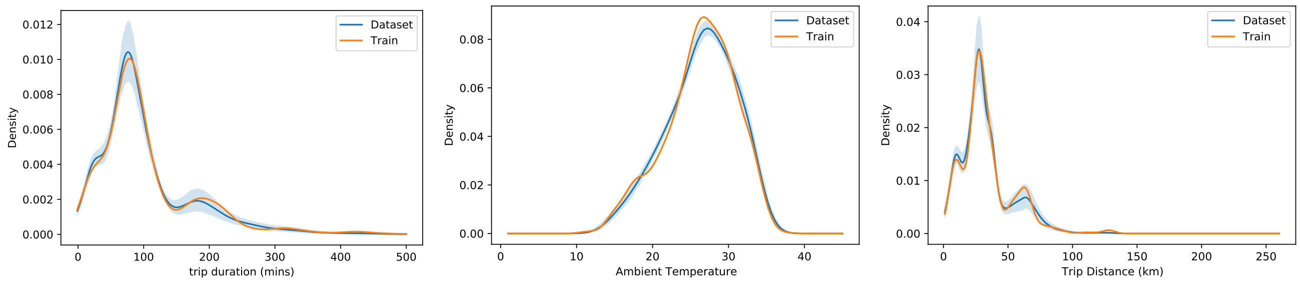

Figure 10: Comparison between train distribution vehicles and all vehicles in our dataset.

Training Data Ratio (%)	Number of Train Trips	R2	MAPE	MedAPE
		Metric		
80	**11530**	**0.993**	**3.027**	**2.213**
50	7206	0.988	4.134	2.703
20	2882	0.98	5.943	4.167

Table 9: Two-stage model performance showing the effect of varying training data size.

Selection Policy	Training Size Vehicles	Trips	R^2	Metric MAPE	MedAPE
Distribution-based	15	1455	0.97	7.69	6.38
	15	**8122**	**0.97**	**6.16**	**5.26**
Random	**22 (80%)**	**11461**	**0.97**	**6.98**	**5.16**
	14 (50%)	6690	0.97	6.97	5.42
	5 (20%)	2256	0.96	8.38	6.45

Table 10: Two-stage model performance for generalization over vehicles using distribution-based and random selection.

Figure 11: Two routes selected from our dataset for route generalization.

PF(ON) feature set. This ensures that the training has a good overlap across the dataset. Through this we identified that only 1455 trips across 15 vehicles were sufficient to cover the diverse distribution across all the features. We call this selection policy of vehicles as

Models	VED MAE	RMSE
LGBM	**1.22**	**1.792**
AdaBoost	1.382	1.986
LSTM [23]	1.47	7.78
DNN [23]	2.33	9.91

Table 11: Comparison of two-stage model results with Deep learning models to predict CSOC on VED dataset.

"distribution-based". Figure 10 shows the distribution-based selected vehicles' training distribution having a good overlap against the total dataset with all the vehicles together. The confidence band across the dataset distribution shows the variance in data for each of the features across the different vehicles.

Table 10 compares the distribution-based training vehicle selection procedure with randomly selected vehicles (80-50-20%) using AdaBoost model. Vehicle selection using distribution-based approach results in 15 vehicles, in one case, we use only the necessary trips from this 15 vehicles as training data (the ones which has highest distribution coverage) and in another case we take all the trips from these 15 vehicles. As we can see, with just 1455 trips ((10.1% of overall data), our model is able to achieve a decent MAPE of 7% and if we use all the trips from these 15 selected vehicles, i.e., 8122 trips (56.5% of overall data) we can reduce the MAPE error to 6%. Thus, we show models trained using data from one set of vehicles can be applied to new vehicles with unseen trip data.

We also performed random selection of vehicles with 80%, 50%, and 20% of vehicles selected for training, i.e., 22, 14,and 5 vehicles, respectively. Table 10 shows that as the % of vehicles considered in training reduces, the MAPE error increases significantly. Furthermore, MAPE results of distribution-based approach outperforms random selection policy with much lesser vehicles and trips used for training. Thus, the proposed model is still generalizing better for the carefully selected vehicles even with a lower amount of training data.

6.4 Route Generalization

We now show the efficacy of our model with PF(ON) feature set for generalization across routes. The idea is to use trip data from one particular route for training and test the model on data from a different route. Figure 11 shows two specific routes A and B in our dataset with length of 22.4km and 20.96km, respectively. We trained our two-stage ML pipeline with data from Route A consisting of 2047 trips with a total of 49196 km. The test data belongs to Route B consisting of 595 trips. Our two-stage pipeline to predict energy consumption for Route B has 7.3% MAPE. This error is within acceptable limits and is slightly higher when our model is trained with data from both the routes with MAPE error of 6.0%. This shows that our engineered features are able to capture diverse characteristics in the data and can be used to generalize across different routes.

6.5 Applicability to other EV datasets

We now show the applicability of our proposed two-stage pipeline with PF(ON) feature set on a public dataset, *viz.,* VED dataset [21]. VED dataset includes data from passenger EV vehicles, see Table 1 for more information. To this end, the state-of-the-art results on VED dataset is proposed in [23], which uses Deep learning models such as LSTM and DNN to estimate the consumed state of charge per trip on the VED dataset. To ensure fair comparison, we use the proposed two-stage approach with the engineered features to estimate consumed state of charge (CSOC) in VED dataset. Table 11 shows the performance of the proposed model against that presented in [23] for estimating CSOC on VED dataset. In [23], only MAE and RMSE values are presented for CSOC prediction and is shown in the last two rows of Table 11. We can see that the proposed two-stage model performance (first two rows) has much lower MAE errors (100% reduction) and RMSE errors (400% reduction) when compared to the approach presented in [23]. Thus, we showcase the applicability of the proposed approach and feature engineering to other EV datasets while outperforming state-of-the-art prediction results.

7 CONCLUSION

As the adoption of electric vehicles increases, accurate energy estimation is very important for large scale deployment of electric vehicles. In this work, we discussed challenges in estimating real-world energy consumption of electric vehicles and presented a data collection approach to address this. We modeled the energy consumption for a fleet of buses using features which describe various aspects of a trip including external factors (like weather and traffic conditions), vehicle internal factors, driver behaviour etc. Then, we showed that with a two-stage prediction approach where some intermediate features are predicted using machine learning methods along with directly observed features, we are able to achieve an MAPE of < 5% on our dataset. Further, we also show that our models generalize well for a new set of vehicles, routes and time periods on our dataset. The proposed approach also generalizes well to other EV datasets. Our dataset is the largest EV dataset ever discussed and the first dataset for a fleet of buses. We also show that our methodology with feature engineering is superior compared to the end-to-end deep learning baselines. As a future direction, we would like to explore the estimation of state of health (SOH) of a battery based on its historical usage and remaining life, and incorporate this information for TEC prediction.

With our analysis, we are hopeful that applications such as routing, fleet management, battery sizing, etc., would be able to benefit and make electric vehicle fleets an integral part to achieve a sustainable future.

REFERENCES

[1] Alberto Diaz Alvarez, Francisco Serradilla Garcia, José Eugenio Naranjo, Jose Javier Anaya, and Felipe Jimenez. 2014. Modeling the driving behavior of electric vehicles using smartphones and neural networks. *IEEE Intelligent Transportation Systems Magazine* 6, 3 (2014), 44–53.

[2] Abdollah Amirkhani, Arman Haghanifar, and Mohammad R Mosavi. 2019. Electric Vehicles Driving Range and Energy Consumption Investigation: A Comparative Study of Machine Learning Techniques. In *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE, 1–6.

[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[4] François Chollet et al. 2015. Keras. https://keras.io.

[5] P Dauni, M D Firdaus, R Asfariani, M I N Saputra, A A Hidayat, and W B Zulfikar. 2019. Implementation of Haversine formula for school location tracking. 1402 (dec 2019), 077028. https://doi.org/10.1088/1742-6596/1402/7/077028

[6] Cedric De Cauwer, Wouter Verbeke, Thierry Coosemans, Saphir Faid, and Joeri Van Mierlo. 2017. A data-driven method for energy consumption prediction and energy-efficient routing of electric vehicles in real-world conditions. *Energies* 10, 5 (2017), 608.

[7] David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* (2007).

[8] D. Gutiérrez-González, E. Parrado-Hernández, and A. Navia-Vázquez. 2005. Multi-kernel Growing Support Vector Regressor. In *Computational Intelligence and Bioinspired Systems*, Joan Cabestany, Alberto Prieto, and Francisco Sandoval (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 357–365.

[9] Arman Haghanifar. 2018. SpritMonitor Crawler. https://github.com/armiro/crawlers/tree/master/SpritMonitor-Crawler.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[11] Chen Jin, Luo De-lin, and Mu Fen-xiang. 2009. An improved ID3 decision tree algorithm. In *2009 4th International Conference on Computer Science Education*. 127–130. https://doi.org/10.1109/ICCSE.2009.5228509

[12] Guolin Ke, Qi Meng, Thomas Finely, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30 (NIP 2017)*. https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

[13] Kyi Lai Lai Khine and Thi Thi Soe Nyunt. 2019. Predictive Big Data Analytics Using Multiple Linear Regression Model. In *Big Data Analysis and Deep Learning Applications*, Thi Thi Zin and Jerry Chun-Wei Lin (Eds.). Springer Singapore, Singapore, 9–19.

[14] Benjamin Krogh, Ove Andersen, and Kristian Torp. 2015. Analyzing electric vehicle energy consumption using very large data sets. In *International Conference on Database Systems for Advanced Applications*. Springer, 471–487.

[15] Jin Li, Feng Wang, and Yu He. 2020. Electric vehicle routing problem with battery swapping considering energy consumption and carbon emissions. *Sustainability* 12, 24 (2020), 10537.

[16] Wen Li, Patrick Stanula, Patricia Egede, Sami Kara, and Christoph Herrmann. 2016. Determining the main factors influencing the energy consumption of electric vehicles in the usage phase. *Procedia Cirp* 48 (2016), 352–357.

[17] Leandro do C Martins, Rafael D Tordecilla, Juliana Castaneda, Angel A Juan, and Javier Faulin. 2021. Electric Vehicle Routing, Arc Routing, and Team Orienteering Problems in Sustainable Transportation. *Energies* 14, 16 (2021), 5131.

[18] Ilyès Miri, Abbas Fotouhi, and Nathan Ewin. 2021. Electric vehicle energy consumption modelling and estimation—A case study. *International Journal of Energy Research* 45, 1 (2021), 501–520.

[19] Shatrughan Modi, Jhilik Bhattacharya, and Prasenjit Basak. 2021. Convolutional neural network–bagged decision tree: a hybrid approach to reduce electric vehicle's driver's range anxiety by estimating energy consumption in real-time. *Soft Computing* 25, 3 (2021), 2399–2416.

[20] Lance Noel, Gerardo Zarazua de Rubens, Benjamin K Sovacool, and Johannes Kester. 2019. Fear and loathing of electric vehicles: The reactionary rhetoric of range anxiety. *Energy research & social science* 48 (2019), 96–107.

[21] Geunseob Oh, David J Leblanc, and Huei Peng. 2020. Vehicle energy dataset (ved), a large-scale dataset for vehicle energy consumption research. *IEEE Transactions on Intelligent Transportation Systems* (2020).

[22] 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, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[23] Linas Petkevicius, Simonas Saltenis, Alminas Civilis, and Kristian Torp. 2021. Probabilistic Deep Learning for Electric-Vehicle Energy-Use Prediction. In *17th International Symposium on Spatial and Temporal Databases*. 85–95.

[24] Lior Rokach and Oded Maimon. 2005. *Decision Trees*. Springer US, Boston, MA, 165–192. https://doi.org/10.1007/0-387-25465-X_9

[25] Joaquín Torres-Sospedra, Carlos Hernández-Espinosa, and Mercedes Fernández-Redondo. 2006. Adaptive Boosting: Dividing the Learning Set to Increase the Diversity and Performance of the Ensemble. In *Neural Information Processing*, Irwin King, Jun Wang, Lai-Wan Chan, and DeLiang Wang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 688–697.

[26] Jim Witham. 2021. Creating The Sustainable Electric Vehicle Revolution. https://www.forbes.com/sites/forbestechcouncil/2021/10/22/creating-the-sustainable-electric-vehicle-revolution/.

[27] Yu Zheng, Yue Song, David J Hill, and Ke Meng. 2018. Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems. *IEEE Transactions on Industrial Informatics* 15, 2 (2018), 638–649.