IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2023, accepted 30 August 2023, date of publication 4 September 2023, date of current version 12 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311895

RESEARCH ARTICLE

Data Driven Energy Economy Prediction for Electric City Buses Using Machine Learning

ROMAN MICHAEL SENNEFELDER[1], RUBÉN MARTÍN-CLEMENTE[2], (Member, IEEE), RAMÓN GONZÁLEZ-CARVAJAL[3], (Fellow, IEEE), AND DIMITAR TRIFONOV[1]

[1]EVO Engineering GmbH, 80807 Munich, Germany
[2]Departamento de Teoría de la Señal y Comunicaciones, Universidad de Sevilla, 41004 Seville, Spain
[3]Departamento de Ingeniería Electrónica, Universidad de Sevilla, 41004 Seville, Spain

Corresponding author: Ramón González-Carvajal (carvajal@us.es)

This work was supported by the Research Project Autonomous Low-Cost And Digital Monitoring of Electrical Generation and Distribution INfrastructure (ALADDIN) "Proyecto TED2021-131052B-C22 financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/Plan de Recuperación, Transformación y Resiliencia (PRTR)."

ABSTRACT Electrification of transportation systems is increasing, in particular city buses raise enormous potential. Deep understanding of real-world driving data is essential for vehicle design and fleet operation. Various technological aspects must be considered to run alternative powertrains efficiently. Uncertainty about energy demand results in conservative design which implies inefficiency and high costs. Both, industry, and academia miss analytical solutions to solve this problem due to complexity and interrelation of parameters. Precise energy demand prediction enables significant cost reduction by optimized operations. This paper aims at increased transparency of battery electric buses' (BEB) energy economy. We introduce novel sets of explanatory variables to characterize speed profiles, which we utilize in powerful machine learning methods. We develop and comprehensively assess 5 different algorithms regarding prediction accuracy, robustness, and overall applicability. Achieving a prediction accuracy of more than 94%, our models performed excellent in combination with the sophisticated selection of features. The presented methodology bears enormous potential for manufacturers, fleet operators and communities to transform mobility and thus pave the way for sustainable, public transportation.

INDEX TERMS Battery electric buses, energy demand prediction, feature extraction, machine learning, meta modeling.

I. INTRODUCTION

Traffic causes approximately 25% of greenhouse gas (GHG) emissions in Europe, and this percentage is increasing [1]. Therefore, widespread electrification of the mobility sector is one of the most positive actions that can be taken in relation to climate change and sustainability [2], [3]. It seems clear that electric buses, because of their low pollutant emissions, are set to play a key role in the public urban transportation of the future. Although the initial investment in electrification may be high - e.g. purchase costs of BEBs are up to twice as high as those of Diesel buses [4] - it is quickly amortized because the inherent efficiency of electric vehicles far exceeds that of internal combustion engine vehicles (up to 77% [5]) and thus operational respectively life cycle costs are significantly lower [6]. In addition, electrification of the powertrain brings many other advantages, such as a reduced noise level or pollution [7], [8], [9], [10]. On the downside, the battery charging time of an electric bus is significantly longer than the refueling time of a diesel bus, while the opposite is true for the range [11]. Ultimately, widespread electrification of the mobility sector is one of the most positive actions that can be taken in terms of climate change and sustainability, but more research is needed to ensure efficient operation, as it also poses significant challenges.

The starting point for this study was a problem proposed by Seville's public bus operator. In short, they wanted to replace their diesel fleet with all-electric vehicles, but first they had to size the vehicles' batteries and determine the best charging locations around the city. In practice, this means

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

using computers to predict consumption on each route [12]. Unfortunately, this can currently only be done with complex physical models that require long simulation times, or with data-driven models that are less computationally intensive once trained, but require numerous driving, mechanical, and road measurements as inputs (see Section I-A). This is where the present research comes in. In this paper we use the bus operator's database and a physics-based model of soon-to-be-deployed electric buses to develop data-driven models that predict the energy requirements of the vehicles. Amongst others, what distinguishes our contribution from previous data-driven approaches is the small number of physical variables involved: we show that, to accurately predict the consumption on a route using machine learning, we only need to know the instantaneous speed of the vehicle and the number of passengers on the bus. Specifically, our approach consists of three steps:

1) We calculate the energy consumed by the bus on each route using a physics-based model, validated by the vehicle manufacturer, that uses speed and mass as inputs, including the bus's own weight and the weight of its payload. Both variables are taken from the operator's database.
2) We extract a comprehensive set of time and frequency features from the speed signal.
3) We train machine learning regression models to predict the energy consumption from bus payload mass and the above set of features, and identify those with the best predictive value. Interestingly, the feature that turns out to be the most relevant, i.e., the spectral entropy of velocity, has so far gone unnoticed in this field of research.

Ultimately, our results are useful for planning the transition from a conventional to a green bus fleet, and even for adding new functionalities that will be useful to planners: for example, the algorithms may be run on the battery management systems to provide an alternative way of monitoring the current state of charge of the batteries.

The paper is structured as follows. First, we identify the challenges in this field and review the state of the art in section I. Secondly, our material, methodology and methods are explained in Section II. Experimental results are presented and discussed in section III. Finally, section IV concludes our paper and shows possible future developments.

A. STATE OF THE ART

The prediction of energy demand for battery electric vehicles (BEVs) in general, and battery electric buses (BEBs) in particular, have been thoroughly investigated. This is not surprising, as [13] shows that BEBs are a viable replacement for conventional vehicles and are also less sensitive to variations in mission profiles than diesel buses. It is important to note also that the duty cycle and driving conditions of a BEB are very different from those of other BEVs, shifting the focus from kinematic relationships to route, schedule, and passenger load.

The majority of previous studies utilize complex physics-based vehicle models, though they vary in focus and objective [14], [15], [16], [17], [18], [19], [20], [21]. In [14], for example, the authors examine the impact of powertrain efficiency, rolling resistance, and auxiliary power on the energy consumption of battery electric vehicles (BEVs). While drive train efficiency and rolling resistance are relevant to the physical movement of the vehicles, auxiliary power demand is especially important at the lower speeds ($<$ 40 km/h) where city buses typically operate, motivating the need for accurate knowledge of auxiliary power to predict overall energy consumption. The study of De Cauwer et al. [15] integrates a physical model of the vehicle and a data-driven methodology with the aim to detect and quantify correlations between the kinematic parameters and the vehicle's energy consumption. Commonly used kinematic parameters are complemented by additional factors such as the travel distance and time or the temperature. Wang et al. [17] studied the influence of rolling resistance, which depends on the road surface, as well as various weather conditions, on power demand. The prediction model in [18] consists of a longitudinal dynamics model complemented by additional dedicated measurements from a dynamometer, as well as coastdown tests, to reduce the model's uncertainty. Similarly, in [21] the authors introduce a novel and computationally efficient electro-mechanical model of a BEB to study the influence of factors such as payload mass, temperature and rolling resistance on consumption. All these approaches provide valuable insight on the interrelation of factors of influence; nevertheless, they involve intricate equations and require accurate modeling of the vehicles and their components to generate results. Like all physics-based models, they are of limited practical use due to the long simulation times. In addition, most previous research has focused primarily on light-duty vehicles, and scaling to the heavy-duty class is complex due to completely different driving profiles and dynamics.

Data-driven approaches, which use machine-learning or deep learning algorithms and real-world driving data, or even mixed data-driven and physics-based approaches, can be found in [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], and [35]. For example, Chen et al. [22] review state of the art energy-consumption estimation models (rule-based vs. data-driven) for electric vehicles and study the case of electric buses using logistic regression and neural networks on real-world data. Additionally, they identify the research gap for energy consumption models of heavy duty vehicles e.g. city buses, buttressing the motivation of our work. Pamula et al. [23] used both deep learning and classical neural networks to forecast the energy demand of electric buses. These prediction models utilized actual data obtained from various bus lines. The models are based on input variables that fleet operators can easily measure, but also operational information such as bus routes and stop locations, travel time between bus stops, schedules and peak hour information. Kontu and Miles [24] investigate factors

of influence such as the route and driver characteristics. Ericsson [25] studied the effects of different driving patterns collected in real traffic on consumption and emissions of internal combustion vehicles. Starting with 62 features, a factorial analysis allows them to reduce this number to only 16. This work demonstrates, on the one hand, the influence of common kinematic driving pattern parameters, such as speed, acceleration, and deceleration, on energy consumption and, on the other, the paper evaluated the usefulness of feature analysis and selection. Simonis and Sennefelder [26] accurately describe the behavior of drivers as a function of a set of selected characteristics, which can be used next to predict energy demand of BEVs. Interestingly, Abdelaty and Mohamed [27], [28] used a Simulink model to estimate the energy consumption of BEBs, where the inputs were carefully selected from a mix of operational, topological, vehicular and external variables using machine learning algorithms and statistical models. They found that the battery state of charge and the road gradient were the most significant factors, while the vehicle's drag coefficients appeared to have a relatively minimal effect. However, temperature and thus auxiliary power demand are not well covered, which is one of the most important factors as Ji et al. demonstrate in their paper [36], in which they investigate real world data from a fleet of 31 BEBs in Meihekou City, China. The ambient temperature expands from $-27°$ C to $35°$ C which lasts in up to 47% increased energy consumption compared to optimum working condition. Expanding on this important topic, in another recent study by Perugu et al. [37] in Lancaster, California, BEBs energy consumption and charging behavior are examined: the vehicles face significant daily and seasonally varying temperatures from $-9°$ C up to $46°$ C and thus the variability in energy consumption can be attributed to the use of heating, cooling, venting and air conditioning (HVAC). Their results show the existence of relevant operational costs for the operator, which can increase up to 18% during summer. Anyway, this cost analysis might be different in other situations (location, terrain, traffic etc.) as cost assessment of BEBs is generally a vast field as can be seen in [6] and [4], depending on a magnitude of factors (production numbers, development costs, public grants, energy price etc.). In [38], Goehlich et al. perform a technology assessment for BEBs in Berlin, Germany. They use an energy simulation model to forecast the consumption in daily service and finally analyze the system's economics in terms of total costs of ownership (TCO). Using a thermal model of the cabin, they find that heating by Positive Temperature Coefficient (PTC) elements is generally more critical than cooling, and discover a worst-case additional HVAC consumption of up to 1.1 kWh/km, which is almost a third of the overall energy consumption.

All these studies show an enormous variety of techniques for estimating the energy consumption of BEBs. Although they provide valuable insights, the following research gaps are identified and addressed in this work:

- Most approaches use data that standard vehicles are often not equipped to measure, such as the location of bus stops or road gradient. In addition, variables that are highly dependent on the particular conditions of the experiment are frequently taken into account, such as the length of the trip. The relationship of the latter with vehicle energy economy is obvious – e.g., the further you drive the more energy is consumed. However, it must be used with caution for prediction, as machine learning algorithms may focus on it and overlook other relevant factors. By contrast, our algorithms take as initial input only the mass (estimated from the curb weight plus number of passengers) and the vehicle speed, which can be easily obtained by the user. Furthermore, we characterize speed profiles by extracting 40 features at different levels of abstraction in the frequency and time domains. This way, we uncover hidden and valuable information that leads to higher prediction accuracy, improved generalization, and thus high application relevance. In addition, we implement an intelligent route segmentation algorithm that makes the prediction robust to data non-stationarity, making the final framework more transferable and even more applicable.
- Despite the abundance of machine-learning techniques, only a few of them are commonly used. In this work, we consider the full range, from non-learning statistical approaches to supervised learning and probabilistic methods. Consequently, this work presents and comprehensively compares the full potential of novel machine learning methods for predicting the energy consumption of EVs. Ultimately, we investigate the performance of various powerful machine learning models, from the very technical detail to the long-term application.
- Most studies use data from a single vehicle on a single route or use speed profiles from Standardized Driving Cycles (SDCs). Therefore, the variety and diversity within the data is comparatively low. However, a major challenge in this area is that the relevant factors are diverse and the interrelationships are complex. Thus, the larger the variety in the data, the better the machine learning predictions will be. In contrast, the underlying fleet data for this work is measured from an entire fleet of 30 vehicles, which operate various routes a day and drivers change frequently even during the day. This allows us to capture a wide variety of traffic situations and driving styles, containing much more valuable information.
- Auxiliary power demand, including HVAC, is rarely considered in detail and often replaced by a constant term. However, especially in extreme low and high temperature regions, heating and cooling have a significant impact on the energy consumption and thus the range of BEBs. We have considered complete energy profiles, including HVAC, recovery, etc., which allows this work

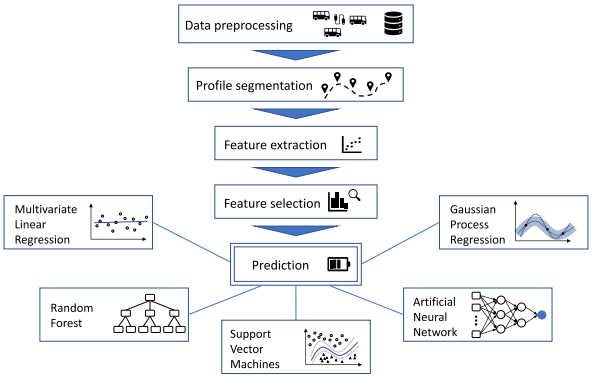

FIGURE 1. Graphical abstract of the papers methodology.

to address accurate total energy consumption at the trip level, which is relevant to transit operators.

II. MATERIAL AND METHODS

This section explains our work methodology. As a guideline, a schematic summary is provided in Figure 1.

A. DATA COLLECTION & PRE-PROCESSING

The operator of the public bus fleet in Seville, Spain, provided the data for this study. The data consists of measurements from 30 conventional diesel buses during regular operating service over 11 consecutive days in June 2019. Two different types of vehicles typically used in cities were equally tracked, namely 12 m regular buses and 18 m articulated buses. It is important to note that buses in Seville usually run several different routes each day, whereas in other cities buses run only one route all day. In addition, drivers change every few hours, adding much more diversity to the measurement. As a result, the data includes a wide variety of driving styles and traffic situations, which is what makes our database unique. The measurement setup did not require additional telematics system, as it was simply a selection of CAN protocol, and tracked up to 77 parameters of the vehicles, with speed profiles and load variations being the most relevant to this work. Since the energy consumption of conventional vehicles is rarely recorded, we used as a proxy the results of a previous study (see [12]) in which we extensively simulated the battery electric version and the energy consumption of these buses on the same routes. Therefore, the data of interest for this research consists of real world driving profiles (speed and mass) and calculated electric power and overall energy demand. This approach of simulating the energy demand instead of measuring it directly in an all-electric testing vehicle is quite common in the field (see [27], [39], [40]) because real BEB data are still scarce and, above all, because this option is more flexible and significantly cheaper and faster. Since the focus of this research is on data-based energy prediction and not on the comparison between measured and simulated data, we will only briefly explain the process below and refer the interested reader to the previous study [12]. The longitudinal vehicle simulation model used in that former study was a closed-loop Mathworks Simulink model, which followed the methodology in [41] and [42]. It was primarily designed to quantify the energy economy and performance of the vehicle, allowing different configurations and component variations to be tested. It explains the dynamic equations of motion of the bus (see e.g. [43]) and the energy flow, beginning at the battery and ending at the wheels (motoring) and *vice versa* (recuperation). The complete vehicle model with all its components, in particular the electric powertrain (e.g. energy storage, power electronics, e-machine, etc.), the mechanical vehicle itself (e.g. tires, body, wheelbase, etc.), as well as an estimate of the auxiliary power consumption, was initially designed by the manufacturer and validated with the real testing vehicle which they planned to put into operation in the target scenario. Furthermore, the model was plausibility-checked by an in-depth review of engineering experts in the field. Specifically, we performed the following validation steps. Firstly, vehicle dynamics (speed, acceleration, track deviation etc.), ambient inputs and consumption rates were analyzed to ensure that they were within physically plausible ranges. Secondly, simulations were performed using standardized on-road test cycles (Braunschweig City Driving Cycle, Manhattan Bus Cycle and European Transient Cycle) to compare the results with similar studies and publicly available tests. Finally, the results were extensively compared to the state-of-the art (see [38], [44], [45], [46]). In the end, the data can be considered as accurate as if they had been measured experimentally.

Finally, after collecting conventional fleet driving-data, an initial preprocessing step that consisted in discarding incomplete or error-ridden records, and the extensive simulation of the consumption of the BEBs, we had at our disposal 149 complete trips from twenty-four vehicles, totaling more than 2 832 hours of driving data (118 hours per bus on average, minimum value 39.75 h, maximum value 167.25 h, standard deviation 38.1 h). Note that the final data set of interest for this research consists of the vehicle speed and mass variation, and its total energy consumption.

B. SEGMENTATION INTO MICROTRIPS

Research in the field of BEB energy consumption is often based on the segmentation of the routes into *microtrips*. A microtrip is defined as the driving interval between two consecutive stops, and may or may not include periods of inactivity. Since vehicles of this class may be on the road for more than 16 hours a day, covering hundreds of kilometers, this segmentation is necessary to cope with the non-stationarity of driving conditions. During the trip there may be changes in the type of road (suburban, urban or highway) and in the maximum speed limits, which together with traffic conditions, traffic lights, intersections and so on, condition the driving style. This lack of stationarity is thus addressed by dividing the speed profile into segments that can be viewed as realizations of approximately stationary processes, i.e. the microtrips, so that the non-stationary speed profile is seen as a concatenation of stationary partitions.

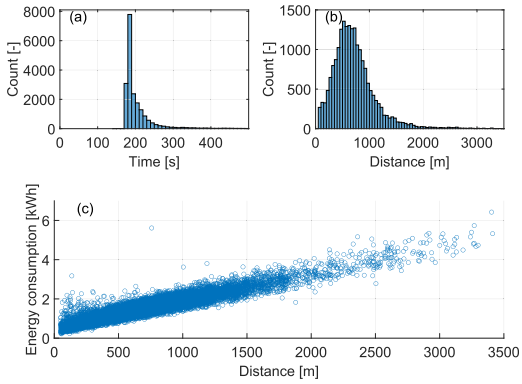

FIGURE 2. Distribution of travel time, distance traveled and energy consumption per microtrip for the entire database.

In the present case, we define microtrips by applying the following three criteria during segmentation: speed at the beginning and end of the microtrip must be zero, $v(t_{start}) = v(t_{end}) = 0$; the minimum trip length must equal 3 minutes, $t_{min} \geq 180$ s, and the minimum trip distance must be 50 m, $d_{min} \geq 50$ m. These values are consistent with those of typical standardized driving cycles for vehicles in this class, such as the Braunschweig City Driving Cycle (BCDC) or the Manhattan Bus Driving Cycle (MBDC) [47]. For illustration, Figure 2 shows in subplot (a) the distribution of duration and in subplot (b) the distribution of distance for all microtrips. Although the distance traveled varies considerably from one microtrip to another, duration is around 180 s in most cases. Scattering the energy consumed as a function of the covered distance, as in Figure 2 (c), reveals an obvious approximate linear relationship between these two variables (in fact, the linear regression explains 80% of the variation of the energy consumption). The conclusion (i.e., the more you drive, the more energy you require) is evident but this high level of determination encourages us to search for explanatory variables and models that allow a more robust regression and, therefore, an even more accurate prediction.

In the end, the segmentation algorithm provided a total number of 20 297 microtrips, with an average duration of 187 s (standard deviation = 92 s) and an average distance covered of 653 m (std = 472 m). The average energy consumption is 1.5 kWh (std = 0.7 kWh) per microtrip. Interestingly, both the Kolmogorov-Smirnov test and the Lilliefors test indicate that the data points follow Gaussian distributions in each microtrip at a significance level of 0.01.

C. FEATURE EXTRACTION

Although energy consumption can be calculated using physics-based methods [14], [15], [16], [17], [18], we choose to use a machine learning regression approach instead. Such models can be trained to learn the relationships between energy demand on the one hand and speed and payload mass as well as any other secondary set of variables on the other. Once the model is trained, making a prediction from new input data is computationally cheaper than solving the physical equations numerically. The process of finding meaningful input for machine learning models can be divided into two steps: feature extraction and feature selection. First, feature extraction can be interpreted as the definition or discretization of explanatory variables or *features*. For the present work, we start with an initial set of 40 features, shown in Table 1, which includes common measures of trend and variability, statistics and nonlinear descriptors. Observe the presence of the product of velocity and acceleration, which is a surrogate for inertial and drag power [48]. Note that one may has automatically obtained features from the data using techniques such as factor analysis [25], but we prefer these hand-crafted features since they are more intuitive to interpret than machine-built ones. The feature extraction step can be considered finished when all variables shown in Table 1 are calculated for each microtrip and stored in the matrix $\mathbf{F} \in \mathbb{R}^{n \times i}$, where f_{ni} is the n^{th} observation of feature i:

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1i} \\ f_{21} & f_{22} & \cdots & f_{2i} \\ \vdots & \vdots & \ddots & \\ f_{n1} & f_{n2} & \cdots & f_{ni} \end{pmatrix} \tag{1}$$

D. FEATURE SELECTION

Whichever approach is chosen, the features have to be ranked according to their predictive value. Identifying and discarding the less relevant ones, including those that are redundant, may prevent models from overfitting the training data set and, additionally, reducing the number of explanatory variables simplifies the model's complexity and reduces computational costs. Neighborhood component analysis feature selection (NCAFS) [49] and out-of-bag predictor importance estimate by permutation (oobImp) [50] are two representative techniques for selecting significant features. NCAFS minimizes the average loss of a neighborhood component analysis regression model while avoiding overfitting by means of a regularization term. The optimal weights resulting from the minimization indicate the importance of the features: when these weights are zero or close to zero, the associated features have no statistically significant impact on the predicted variable. On the other hand, oobImp is a powerful feature selection method which is intrinsically built into random forest regression models. The results of applying these techniques will be shown in the Experiments Section.

We finally emphasize, that to compare features with different units of measurement, they must first be either min-max-scaled between 0 and 1 or, as we did, normalized to zero-mean and unit variance.

E. PREDICTION

Let us now briefly describe the representative set of regression methods we tested. We will study their performance and provide implementation details in the 'Experiments' section.

1) MULTIVARIATE LINEAR REGRESSION (MLR)

Multivariate linear regression predicts the energy consumption as a weighted linear combination of the features [51].

TABLE 1. Overview of features used to characterize speed profiles.

Feature Name	Symbol	Unit
Maximum Speed	Max	km/h
Mean Speed	Mean	km/h
Median Speed	Median	km/h
25th Percentile	Q_{25}	km/h
75th Percentile	Q_{75}	km/h
Inter Quartile Range	IQR	km/h
Standstill	v_0	%
Percentage of time in which $5 < v \leq 15$ km/h	v_{5_15}	%
Percentage of time in which $15 < v \leq 30$ km/h	v_{15_30}	%
Percentage of time in which $30 < v \leq 40$ km/h	v_{30_40}	%
Percentage of time in which $v > 40$ km/h	v_{40}	%
Standard Deviation	StDev	km/h
Variance	$v_{variance}$	
Mean Absolute Deviation	MAD	km/h
Skewness	$v_{skewness}$	-
Kurtosis	$v_{kurtosis}$	-
Crest	v_{crest}	
Clearance	$v_{clearance}$	
Shape	v_{shape}	
Impulse	$v_{impulse}$	
Sqrt. Amplitude	$\sqrt{v_{amp}}$	
Abs. Amplitude	$\lvert v_{amp} \rvert$	
Spectral Entropy	$v_{entropy}$	-
Velocity Oscillation	v_{freq}	-
Number of Acceleration Shifts	nr_{acc_shifts}	-
Spectral Kurtosis	$Mean_{SpecKurt}$	-
Percentage Constant Drive	acc_{perc_const}	%
Mean acceleration constant drive	acc_{const_mean}	m/s^2
Percentage Accelerating	acc_{perc_acc}	%
Mean acceleration	acc_{pos_mean}	m/s^2
Percentage Decelerating	acc_{perc_dec}	%
Mean deceleration	acc_{neg_mean}	m/s^2
Relative Positive Acceleration	RPA	m/s^2
Relative Negative Acceleration	RNA	m/s^2
Average $v \cdot a$	MeanVA	m^2/s^3
Percentage of time $v \cdot a \leq 0$ m^2/s^3	va_0	%
Percentage of time when $0 < v \cdot a \leq 3$ m^2/s^3	va_{0_3}	%
Percentage of time when $3 < v \cdot a \leq 6$ m^2/s^3	va_{3_6}	%
Percentage of time when $v \cdot a > 6$ m^2/s^3	va_6	%
Total Mass (curb weight plus passengers)	m_{total}	kg

The model is mathematically expressed by:

$$\hat{y} = \beta_0 + \sum_i \beta_i f_i. \tag{2}$$

Although this approach seems simple, good results are often achieved in practice. Therefore, MLR is often the baseline against which other methods are compared.

2) RANDOM FOREST REGRESSION (RF)

Regression random forests fit a large number of classification and regression trees (CARTS) to different subsets of the data, which are generated by sampling with replacement from the original training dataset [52]. There are typically two options to do this: in 'bagging', all models are trained in parallel and independently of each other. Furthermore, any observation has the same probability to appear in a training subset. In 'boosting', on the contrary, models are trained sequentially. Then, we use the results of one model to decide which observations take part in the training of the next model. Finally, we average the predictions of all trees to improve accuracy and, additionally, control the overfitting.

Furthermore, the RF algorithm can be used to calculate how much each feature contributes to the global prediction

accuracy (comp. 'oobImp' in section II-C). This is interpreted as a measure of the relevance of the features and can be used as a criterion for discarding the less important ones.

3) SUPPORT VECTOR MACHINES (SVM)

In support vector regression, models are of the form [53]:

$$\hat{y}_i = \beta_0 + \sum_{n=1}^{N} \beta_n \kappa(\mathbf{x}_i, \mathbf{w}_n),$$

where $\mathbf{x}_i$ is i^{th} observation vector, $\mathbf{w}_1, \ldots, \mathbf{w}_N$ are the so-called 'support vectors', which are actually a subset of the training data, and $\kappa(\cdot)$, the 'kernel' function, measures the 'similarity' or 'distance' between $\mathbf{x}_i$ and $\mathbf{w}_n$. In this way, support vectors that are closer to $\mathbf{x}_i$ are weighted more heavily than those that are not. The model is adjusted to minimize the cost function:

$$\frac{1}{2}\sum_{n=1}^{N} \beta_n^2 + C \sum_i \left|y_i - \hat{y}_i\right|_\epsilon$$

where y_i is the actual value of the target variable, C is called 'box constraint', $\sum_{n=1}^{N} \beta_n^2$ is a regularizing term and $\left|y_i - \hat{y}_i\right|_\epsilon = \max(0, \left|y_i - \hat{y}_i\right| - \epsilon)$ is the Vapnik's ϵ-insensitive loss function [51].

4) REGRESSION USING ARTIFICIAL NEURAL NETWORKS (NN)

Linear regression assumes a linear relationship between the features and the target variable. However, this may a naive hypothesis in some cases. Artificial neural networks consist of several consecutive layers, each of which first linearly combines its input variables and then transforms the outcome nonlinearly. In this way, and according Cybenko's universal approximation theorem [53], they are expected to model any nonlinear relationship between the input features and the target variable.

5) GAUSSIAN PROCESS REGRESSION (GPR)

Following a completely different approach from that of the previous methods, Gaussian Process Regression determines a probability distribution over the set of all possible regression functions that fit the data [54]. Prediction is then usually carried out using the mean value of this distribution. Models are of the form:

$$\hat{y}_i = \sum_n \beta_n h_n(\mathbf{x}_i) + f(\mathbf{x}_i),$$

where $\{h_n(\cdot)\}$ is a set of functions that transform the feature vector, called 'basis functions', $\{\beta_n\}$ are weighting coefficients, and $f(\mathbf{x}_i)$ is drawn from a zero-mean Gaussian process whose covariance depends on the distance between the actual observation $\mathbf{x}_i$ and the training data (as measured by kernel functions). Due to its statistical nature, GPR not only predicts values but can also provide confidence interval estimates.

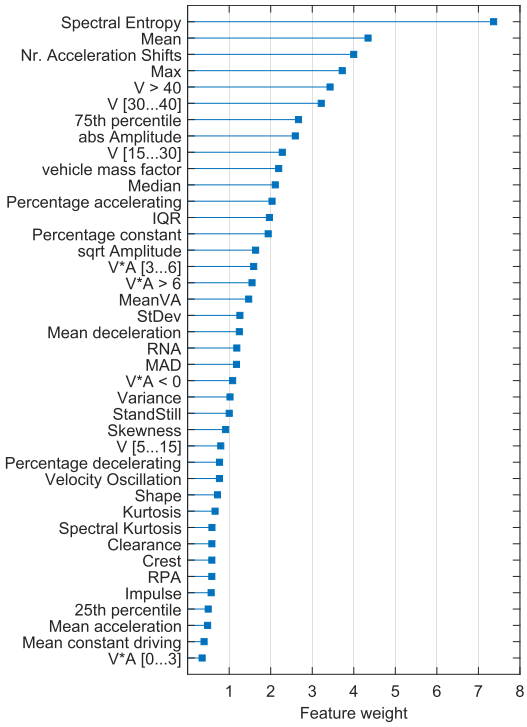

FIGURE 3. Feature importance derived by NCAFS. All features are ordered in descending order according to their relevance. The higher the weight of the feature, the more relevant it is for prediction. Top 27 features are used as inputs for regression models: MLR, SVM, NN and GPR.

III. SIMULATION RESULTS

Experimental results are shown in this Section. Unless otherwise stated, features are extracted from microtrips, not full trips. Accordingly, we predicted the energy consumption for microtrips, too. All simulations were performed using MATLAB R2021b using the functions `fitlm`, `fitrensemble`, `fitrsvm`, `fsrnca`, `fitrgp` and `fitrnet` to implement the different methods.

A. FEATURE SELECTION

First, we rank the features in Table 1 according to their relevance in predicting energy consumption. To this end, we will use two different approaches, i.e., neighborhood component analysis feature selection (NCAFS) and predictor importance estimation by permutation of out of bag predictor observation (oobImp). Additionally, we compare the results obtained with one approach and the other.

1) NCAFS

Figure 3 shows the features sorted according to their usefulness as judged by NCAFS. By far, the most relevant feature is the spectral entropy of the speed. Let us make some comments on this finding. The spectral entropy of a signal is a scalar value that summarizes the signal spectral power distribution. As time and frequency domains are alternative ways of representing the same information, the spectral entropy can be also considered as a measure of how the signal evolves with time. In a nutshell, we can say that the higher the spectral

entropy, the more the velocity resembles a white noise signal. On the other hand, the smaller it is, the more predictable the waveform is. In conclusion, as was expected, we see that the temporal structure of the speed bears enormous information and has a great impact on the vehicle's energy economy. Furthermore, the spectral entropy adequately describes this evolution. For the reader's reference, the spectral entropy is defined as:

$$H = -\sum_m P(m) \log_2 P(m) \tag{3}$$

where $P(m) = \frac{S(m)}{\sum_i S(i)}$ and $S(m)$ is the m$^{\text{th}}$ bin of the power spectral density, which equals the squared modulus of the signal's Discrete Fourier Transform. In other words, $P(m)$ is the percentage of energy of the m$^{\text{th}}$ frequency band and, if we consider it as a probability distribution, it follows that H is just its Shannon entropy.

A second group of relevant features is composed of average velocity, maximum speed, the percentages of time in which the speed is between 15 and 30, 30 and 40 or above 40 km/h and the 75$^{\text{th}}$ percentile of the speed. Regarding these intervals, a systematic trend is observed: the higher the speed, the more important the interval. The relevance of these features can be explained by several factors: firstly, all of them, multiplied by the duration of the microtrip, are a kind of surrogate for distance traveled, which is obviously related to consumption, as shown in Figure 2 (c). Secondly, the importance of high speed occurrences is quite obvious due to higher driving resistive forces. Interestingly, the interquartile range also has a comparably high impact probably due to its relationship to the 75$^{\text{th}}$ percentile.

Another highly relevant feature is the number of acceleration shifts. This is calculated as the sum of changes of the signum of the vehicle acceleration. We interpret this feature as the ability of the driver to anticipate varying traffic situations and its enormous impact is related to many aspects, such as the technical properties of the vehicle and e-drive systems. In addition, recuperation recovers only a part of the energy consumed, especially not at low speed, which is why mechanical braking events and the acceleration shifts are highly correlated to the energy economy. Another interesting finding is that the time spent accelerating and driving with constant speed is more relevant than the time spent decelerating, the mean deceleration or the relative negative acceleration (RNA). Interestingly, the inertial power surrogate $v \cdot a$ (speed by acceleration) is also a good explanatory variable with significant impact. Again, the higher the range of values for $v \cdot a$, the higher its impact becomes.

As expected, the total mass $m_{total} = m_{veh} + m_{pass}$, consisting of vehicle mass m_{veh} and passenger load m_{pass} is another characterizing and significant feature. In fact, the additional weight due to passengers in these vehicles can be up to 6 tons (typical empty weight: 13 t, maximum loaded weight: 19 t). In addition, auxiliary power consumption probably explains the impact of standstill. Especially for battery electric vehicles (BEVs) of heavy duty class, auxiliaries have a significant impact and make up to 50% of the overall consumption [45].

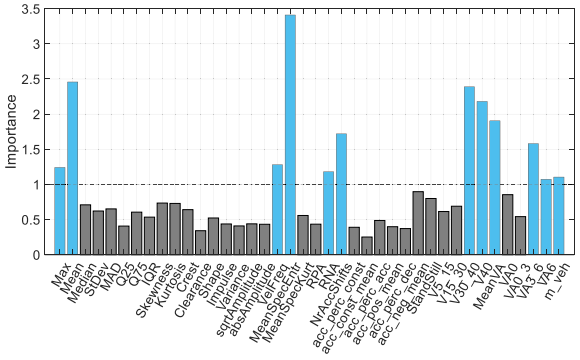

FIGURE 4. Predictor importance estimates according to oobImp for RF approach.

Finally, features like kurtosis, skewness, shape, clearance, crest factor or impulse seem to have lower impact. This is also reflected by their relatively low variance, which implies that they contain little information, and may go back to the fact that the distribution of bus speed profiles seems to be almost Gaussian, for which higher-order statistics actually vanish.

2) oobImp

The other feature selection algorithm (oobImp) is relevant for RF regression and explained in this paragraph. Again, all features in Table 1 were initially used as predictors, with the energy demand as the target variable. A bagged ensemble of 400 regression trees was trained to estimate feature importance values; where oobImp takes into account the accuracy of each regression tree on samples that have not been used for its training. Results are summarized in Figure 4.

Interestingly, features such as the spectral entropy, the mean and maximum speed, the number of acceleration shifts, the total mass, the percentages of time in each speed interval or the product of velocity and acceleration are again at the top of this ranking. As this result is essentially equivalent to that obtained by the NCAFS, the relevance of the features is therefore consistent in the sense that it is not sensitive to the algorithm used. Additionally, we calculate a predictive measure of association between pairs of features. Roughly speaking, the higher this value, the more reasonable it is to substitute one of the features of a pair for the other without appreciable loss of performance. Apart from the obvious relationships, for example, between standard deviation and variance, there does not seem to be any particular similarity between the proposed variables. This indicates that there is little redundancy in the feature set.

Summarizing the feature selection process, we reduce the initial feature space via NCAFS and oobImp by removing explanatory variables of low significance and high correlation. In the end, we derive two subsets of features:

1) Set 1: It considers the top 27 features found by NCAFS (see Figure 3). This set is used for regression models: MLR, SVM, NN and GPR.

2) Set 2: It contains the 12 most relevant features found by oobImp (Figure 4) for RF regression. This set is used for random forest regression only.

Some additional explanations will be given below.

B. PREDICTION

In the following, we describe the energy consumption prediction from the features defined above. As mentioned in section II-A, the input data from which the selected features are extracted is the real-time measurement of vehicle speed, while the predicted target value is the simulated energy demand. The quality of fit will be assessed by means of following indicators that complement each other:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

where y_i is the energy consumed in the i^{th} microtrip, $\hat{y}_i$ is the i^{th} predicted value, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and n is the total number of microtrips.

First of all, r^2, also called 'coefficient of determination', is a metric that can be interpreted as the proportion of the variance in the response explained by the model. Its value ranges from 0 to 1 (or from 0 to 100%): the higher the value, the better the model fits the data set. The Root Mean Squared Error (RMSE) is an intuitive and in statistics often used metric that describes the distance from predicted values to the observed data. The lower the RMSE value, the better a model fits the dataset. Another fundamental criterion for assessing prediction models is the Mean Absolute Deviation (MAE), which is the absolute arithmetic average deviation between predicted and actual values. Lastly, Mean Absolute Percentage Error (MAPE) is the mean relative deviation between actual and predicted values in percent.

Note, that for all experiments below, 80% of the data goes into the training set and the remaining data are used for testing. Experiments were independently repeated 20 times, where each time the splitting into train and test was performed at random, and the results were averaged. The accuracy of each method is shown in Table 2.

We see that there is apparently no great difference in the results obtained by the different algorithms. This is remarkable because it indicates that the features have a high predictive value regardless of which regression method is used. Having said that, notice that random forest algorithms have performed the regression using only 12 features (colored in cyan in Figure 4) while the other approaches use 27 (i.e. the top 27 features in Fig. 3). This is based on the experiments: best overall prediction results are obtained (GPR yields $r^2 = 94\%$) by using 27 features as regressors. On the other hand, random forest achieves comparable results with only 12 features ($r^2 \approx 90\%$). Furthermore, increasing the number of features to 27 does not improve random forest performance, whereas reducing the number of inputs to 12 significantly worsens the other algorithms. This provides us with flexibility and allows us to consider different use cases during planning. For example, to estimate in real time the remaining range of the vehicle, a trained random forest may be implemented in a digital processor embedded in the bus. The most complex algorithms, on the other hand, may run in the cloud and provide improved predictions and extra security.

Further details of how the experiments have been performed are given in the remainder of this Section.

TABLE 2. Results of regression models and their quality to predict energy demand on test data. Note that random forest approaches use 12 features as input variables while the prediction capabilities of the remainder algorithms is evaluated on 27 features.
Legend: MLR = Mutiple Linear Regression, RF = Random Forest, SVM = Support Vector Machine, NN = Neural Network, GPR = Gaussian Process Regression.

Method/Model	r^2 [%]	RMSE	MAE	MAPE [%]
MLR	82	0.27	0.18	14.3
RF bagged	91	0.19	0.13	10.6
RF boosted	89	0.21	0.14	10.8
SVM lin.	82	0.28	0.16	12.1
SVM gauss.	86	0.25	0.15	11.4
SVM pol.	90	0.20	0.14	11.1
NN simple	91	0.19	0.14	11.3
NN medium	91	0.19	0.14	11.7
NN complex	93	0.20	0.14	11.3
GPR quadr.	94	0.18	0.13	10.8
GPR sq. exp.	94	0.19	0.14	10.9

1) MULTIPLE LINEAR REGRESSION

MLR can be seen as the baseline against which other methods are usually compared. Here, the coefficient of determination r^2 for MLR shows that this simple approach explains 82% of the variability of the energy consumption. The model considers spectral entropy and constant acceleration most (highest coefficient weights). As can be seen in Figure 5, the linear approach works well for most microtrips ($> 97\%$), where consumption is between 0 and approximately 3 kWh. For microtrips with a consumption of more than 3 kWh, the deviation of predicted and test data - the spread of error - raises and the model fails to handle outliers. Especially above 4 kWh a systematical underprediction occurs.

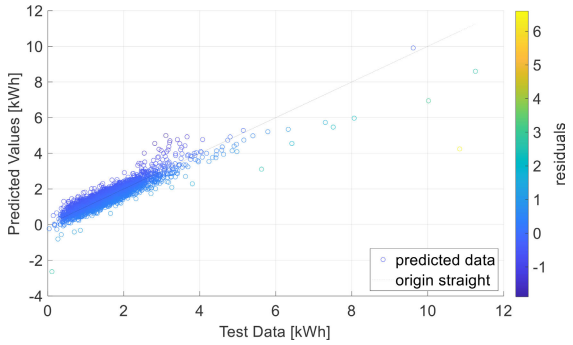

FIGURE 5. Prediction results by MLR.

2) RANDOM FOREST REGRESSION

Two different aggregation approaches are tested, namely, 'boosting' and 'bagging'. The hyper-parameters of the models (minimum leaf size, maximum number of splits, learning rate for shrinkage, and so on) are optimized via five-fold cross-validation. Specifically, boosting is implemented with 64 learners, a learning rate for shrinkage of ≈ 0.4, at least one observation per leaf and a minimum leaf size of 2. The bagging method, on the other hand, has 450 learners and a minimum leaf size of 2. Generally, the bagged ensemble tended to achieve slightly better results than the boosted model. It was found that, in average, $r^2_{train} \approx 92\%$ and $r^2_{test} \approx 90\%$, which suggests a very high model fit (low bias) and allows excellent prediction with low variance. The other scores are shown in Table 2. Recalling that the RF models require the fewest inputs, the accuracy and robustness is remarkable.

3) REGRESSION USING SUPPORT VECTOR MACHINES

Since SVM does not incorporate its own feature selection method, we have used the ranking elaborated by NCAFS. Specifically, as explained above, the top 27 features are used as regressors. We experiment with several kernel functions and their hyperparameters were optimized via five-fold cross-validation. In particular, we investigate 3 different kernel types: linear, Gaussian and polynomial kernel (2^{nd} Order). Generally, all SVM models were robust and the deviation between r^2_{train} and r^2_{test} was kept below 2%. It was also found that the polynomial kernel of second order achieves the best results, see Table 2. We conclude that support vector regression is an intuitive and powerful tool, which offers great prediction results. Attention has to be paid during training to avoid overfitting and minimize model complexity. We trained and optimized several SVM prediction models and their optimized setting is listed in Table 3.

4) REGRESSION USING ARTIFICIAL NEURAL NETWORKS (NNs)

The challenge of avoiding overfitting during training is important for NNs. Therefore, we use at most 3 hidden layers with a maximum of 35 neurons per layer. The maximum number of iterations during training is limited to 1000 and the stopping criterion is defined from the mean square error. The features used as inputs to the NNs are the same as in the case of SVR. Optimum parameters are given in table 4.

We have experimented with three different models. Starting with a simple narrow, feedforward network with only one hidden layer and 10 neurons, a good trade-off between training $r^2_{train} \approx 93\%$ accuracy and prediction $r^2_{test} \approx 91\%$ was found. Its reduced computing time, $t_{train} < 1$ min in our machine, and a prediction capability of $\approx 520\,000$ observations per second, makes this topology convenient. The next, more complex, model is composed of two hidden layers with 10 neurons each which, as the previous one, uses a rectified linear activation function (ReLU). The trade-off between training $r^2_{train} \approx 93\%$ and prediction $r^2_{test} \approx 92\%$ is the best of all tested neural network prediction models. The computational performance is comparable to that of the simple model. Finally, a complex model was trained with 3 hidden layers and, respectively, 10 and 35 neurons in each layer. The model used a sigmoid activation function and was sensibly more precise during training ($r^2_{train} \approx 96\%$) than during testing ($r^2_{test} \approx 93\%$), which suggests a small overfitting. Results of other performance indicators are summarized in Table 2.

5) GAUSSIAN PROCESS REGRESSION

While the basis function is usually constant for GPR, kernels can adopt many different forms as long as they are definite symmetric. We optimized the hyperparameters of the models via cross-validation and Table 5 shows the final configuration. As in previous experiments, the top 27 features in Figure 3 are used as regressors. A squared exponential

TABLE 3. Results of hyperparameter optimization for SVMs. We refer the reader to Section II-E3 for the meaning of hyper-parameters C and ϵ. Kernel scale is a scaling parameter for the data.

Kernel	Parameters		
	Box Constraint C	Kernel Scale σ	Epsilon ϵ
Linear	0.48	3.70	0.048
Gaussian	0.48	25.0	0.048
Polynomial (2^{nd} Order)	0.48	4.06	0.048

TABLE 4. Overview NN Model Properties.

Type	Layers		Activation
	Nr. of hidden layers	Size 1^{st}, 2^{nd}, 3^{rd} layer	
Simple NN	1	25	ReLU
Medium NN	2	10, 10	ReLU
Complex NN	3	35, 10, 10	Sigmoid

TABLE 5. Overview of GPR Properties. We refer the reader to Section II-E5 for the meaning of hyper-parameters β. σ^2 is an estimate of the variance of the prediction error.

Kernel	Basis Function	Beta β	Sigma σ
Quadratic	Constant	4.6860	0.1489
Squared Exponential	Constant	1.5403	0.1521

kernel was found to outperform all other models, resulting in $r^2_{train} \approx 96\%$ accuracy during training and $r^2_{test} \approx 94\%$ during testing. Comparing the overall results in Table 2, we conclude that GPR is a powerful methodology which obtains great prediction results in this context.

C. ANALYSIS OF RESIDUALS

Figure 6 provides a deeper insight into the behavior of the algorithms. It shows prediction and residuals of just the best model of each method: RF bagged, SVM polynomial kernel, NN complex, GPR squared exponential. First, it reveals the good fit of the models, as they all predict the test data with high accuracy. Furthermore, the concentration of residuals around zero is easily detectable. That said, we again observe the same peculiarity as in the linear regression (see again Fig.5): the prediction works well when trips have a consumption of less than about 5 kWh, but becomes inaccurate beyond this point. This will not happen very often, since the average energy consumption per microtrips is only 1.5 KWh (standard deviation = 0.7 KWh), and therefore values higher than 5 KWh have little relevance in the context of a full trip. However, it suggests the existence of a phenomenon that appears when consumption is high and that requires specific modeling. Further research is therefore needed.

Some additional comments are in order: the distribution of residuals are about the same range in RF, SVM and NN. MLR has the largest deviation overall while GPR shows both best fit and least outliers. Interestingly, SVM fits outliers quite efficiently, although the error is larger than with NN or GPR. Although NN has comparatively high accuracy, the residuals reveal that this method also has problems with extreme values. There might be a trend of underprediction in SVM and overprediction in RF, but it is still almost not detectable ($< 1\%$). Regarding the inter-quartile-range of the errors, we see that all are approximately in the same range except MLR which is slightly wider. Finally, GPR offers high accuracy over the entire displacement range and a small deviation for extreme values, which allows us to argue that it is the most robust and utmost accurate method. A closer examination of the standard deviation of the errors again shows that GPR is the best performer, followed by NN and SVM; MLR is more widely again.

D. FULL TRIP ENERGY CONSUMPTION PREDICTION

To evaluate the presented methods at a more applicable and intuitive level, we repeat the experiments and predict the energy demand of complete trips. To this end, we split the 149 available full trips into a training set (111 trips) and a test set (38 trips); a ratio of approximately three to one. Models were trained using the microtrips of those first 111 journeys. Therefore, we ensure that no training data is used during the validation. For each test route, predictions were made on consecutive microtrips and accumulated to estimate the consumption of the entire trip.

TABLE 6. Performance indicators of full trips for entire data base.

Method/Model	r^2 [%]	RMSE	MAE	MAPE [%]
MLR	90	9.52	7.39	3.44
RF	86	11.27	8.88	4.03
SVM	87	10.5	8.41	3.84
NN	89	9.95	7.88	3.60
GPR	89	9.93	7.94	3.63

Figure 7 shows the reduction of the residual energy left in the vehicles battery, which can be interpreted as state of charge, as a function of the distance traveled during a randomly-chosen test trip, as well as the values predicted by the models. As can be seen, during most of the trip (30-60 km) models tend to overestimate the energy consumption. After about 80 km there is a turning point and the actual consumption becomes higher than predicted. In any case, even with the above inaccuracies, the predictions of all models are very close to reality during the entire trip.

Figure 8 shows the energy demand in several full trips and the results predicted by the models. For better readability, only 12 out of the 38 test trips are plotted. Figures 7 and 8 confirm the overall satisfactory prediction performance.

To evaluate the approach in more detail, Table 6 lists the performance indicators now calculated for full trips (see Table 2 for comparison).

As can be seen, all methods can predict the energy consumption of full trips with excellent accuracy. Interestingly, all prediction methods achieve similar performance indicators, especially in case of r^2. The RMSE and MAE, which can be interpreted as the necessary safety buffer that the vehicle battery capacity should have in full trips, are similar and comparatively low. In addition, the MAPE for full trip prediction (on average $\approx$ 3.7%) is consistently better than for microtrips. Considering that MAPE is probably one of the most relevant factors for fleet operators, these results validate our approach and show that it is tailored to these use cases. Figure 8 also reveals a random distribution of both underestimation (such as in trips 71 and 80) and overprediction (e.g., see trips 41 or 52), which suggests that the models are not biased to one side or the other. Finally, Figure 9 shows the overall distribution of prediction errors, that occur on each segment (micro trip) during the complete trips. None of the models has a notable margin of error and the overall mean error is remarkable small. Interestingly, the standard

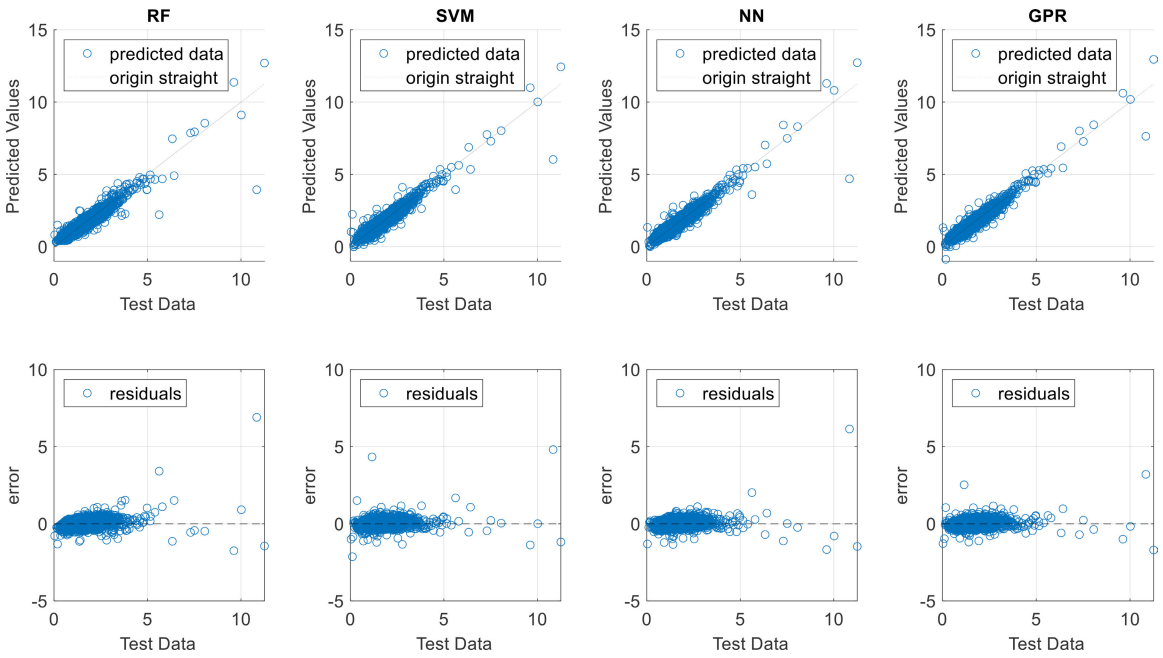

FIGURE 6. First row: model prediction results. Second row: residuals.

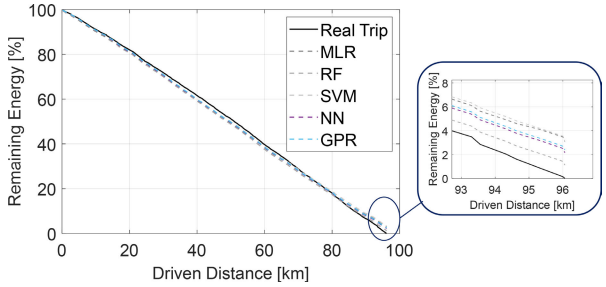

FIGURE 7. Prediction and actual consumption of randomly chosen full trip.

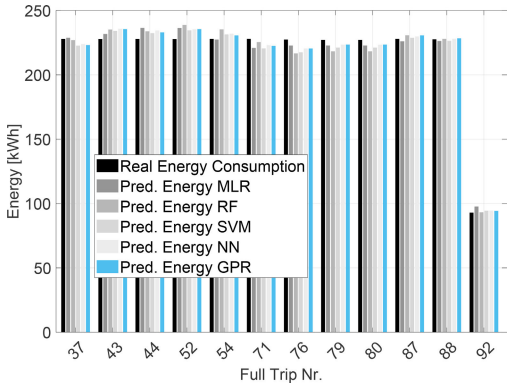

FIGURE 8. Prediction of full trip energy economy.

deviation and inter-quartal range (75th - 25th percentile) are approximately the same, indicating that outliers and extreme values are well covered.

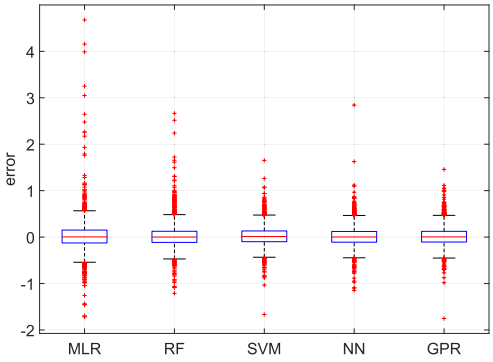

FIGURE 9. Distribution of prediction errors for all segments (microtrips) during a full trip summarized on the whole test journeys.

E. DISCUSSION

Our work continues the investigation where other authors have left off. As in Ericsson's study on internal combustion vehicles [25], we have used regression algorithms to predict the value of a target variable (in our case, the energy demand of battery electric buses) from a selection of features in the data. This data-driven approach has been chosen, rather than a well-tested physics-based model like in [14], [15], [17], [18], and [55], because data driven models have presumably a lower computational cost, a higher robustness and goodness of fit as they can uncover phenomena not contemplated in analytical physics-based models. Undoubtedly, technological progress will promote machine learning algorithms in almost any field of research especially in energy demand prediction for BEBs as e.g. Abdelaty et al. demonstrated in [27],

[28], and [29]. The ultimate goal in this area is to identify and understand the factors that most affect a BEB's energy consumption and to create a framework that is both accurate and robust in its predictions. At this point, the present study proves that it is possible to make an utmost accurate prediction of up to 94% and more from only two primary parameters namely passenger load and speed. All other secondary variables are derived from the latter, e.g. the spectral entropy. These secondary explanatory variables possibly implicitly encode relevant features found in familiar studies s.a. road condition or stop density. This finding significantly reduces the number of variables to be measured on the one hand and on the other it animates researchers in the field to explore the feature space and selection methods more deeply.

Like other authors, see [21], [23], [29], and [56], we have also compared and discussed our models regarding their overall applicability. Operators can clearly benefit from our most robust and accurate model based on NCAFS and GPR or they could choose the reduced bagged RF combined with oobImp, as it is less complex and could be easily implemented with a lazy learning algorithm on a vehicle platform with low computational capabilities. In this way, they can make predictions about the battery state of charge and the remaining range of the electric vehicle. Thus, the entire fleet may be centrally monitored, managed and operated, enabling an efficient deployment of the vehicles.

Since our main contributions to the field have been already detailed above, we turn to discuss some challenges. The first and most obvious, and also common to all data-driven approaches, is to provide us with sufficient and high quality real-world measurement data; unfortunately, the availability of such data is limited, especially in the case of BEBs. One possible solution to the limited availability is to artificially generate driving data to be used in simulations. We also want to emphasize at this point, although the initial set of features comes from experts in the field, it is still a hand-crafted selection and there may be other features of high relevance that have not yet been discovered. Furthermore, the feature selection process performs very well but there are other methods to evaluate the importance of features as e.g. the ''Relative Object Purity Ratio'' presented in [57] or the ''SHapley Additive exPlanations'' values presented in [58] to be investigated, which may lead to different results. In this context, the segmentation algorithm is robust, but may need to be adapted when applied to other use cases. However, since all data preprocessing and machine learning model training is already highly automated, this is not a major drawback.

IV. CONCLUSION

This paper offers a data-driven approach that uses both simulated and real-world data for planning problems and electrification of public transport. The results confirm that the energetic relevant features obtained by feature selection and regression analysis perfectly characterize the energy consumption of BEBs under different real driving conditions. It is a practical approach for fleet operators who want to retrofit or replace their conventional buses with electric vehicles and build the corresponding infrastructure. We emphasize in this context the so-called ''Vehicle Routing Problem'', e.g. mentioned by [59] and [60]. The energy demand on each route needs to be known *a priori* to correctly size the batteries, decide on the optimal bus operating modes (all-electric, hybrid electric, *et cetera*), and select the best charging strategies (i.e. opportunity vs. conventional charging). The worst-case scenario – the most energy-intensive route – is the limiting factor. Ultimately, this knowledge is essential for fleet operators to identify critical operational limits in advance, avoid potential showstoppers, and gain confidence in new technologies. Thus, to achieve reliable and affordable service on all routes in the end.

As our main contribution, the paper presents a novel selection of explanatory variables that combine time and frequency characteristics of the speed waveform. To extract these features, the route is divided into microtrips. This 'segment-based' prediction provides robustness against non-stationarity. Starting with an initial set of 40 features, we have found a minimum number of characteristics with high predictive value. The most relevant of these features, i.e., the spectral entropy of velocity profiles, has so far even gone unnoticed in this field. This result confirms our assumption that it is in the velocity waveform, whose temporal structure is well captured by the spectral entropy, where the most essential information actually resides.

In future research, we plan to extend this approach to other scenarios, as the challenge is to find out how this methodology performs under different circumstances. The proposed approach is of particular interest to companies in the transportation and logistics sector. In particular, it is of interest to fleet operators that rely on heavy-duty trucks and often struggle to electrify their fleets because they lack a solid framework for making the right choices for the right vehicles. It could even be applied to other classes of vehicles or transport systems, such as passenger vehicles or rail transport. On the other hand, meteorological characteristics, road type and operational features for instance could be investigated more deeply. This is why we plan to investigate seasonally and locally changing conditions and recommend careful feature selection according to each use case. Finally, predictive analytics of additional target variables, such as the peak power of the system or the electric current demands on the batteries are of high interest and could be investigated by the presented methodology.

ACKNOWLEDGMENT

The authors would like to thank the public bus transport operator in Seville—TUSSAM for providing them with data.

REFERENCES

[1] *EU Transport in Figures: Statistical Pocketbook 2019*, Directorate-Gen. Mobility Transp., Eur. Commission, Publications Office, Brussels, Belgium, 2019.

[2] P. Hertzke, N. Müller, S. Schenk, and T. Wu. (May 2018). *The Global Electric-Vehicle Market is Amped up and on the Rise*. McKinsey Analysis. [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-global-electric-vehicle-market-is-amped-up-and-on-the-rise

[3] G. Kalghatgi and B. Johansson, "Gasoline compression ignition approach to efficient, clean and affordable future engines," *Proc. Inst. Mech. Engineers, D, J. Automobile Eng.*, vol. 232, no. 1, pp. 118–138, Jan. 2018.

[4] C. Johnson, E. Nobler, L. Eudy, and M. Jeffers. (2020). *Financial Analysis of Battery Electrictransit Buses*. [Online]. Available: https://www.nrel.gov/docs/fy20osti/74832.pdf

[5] A. Braun and W. Rid, "Energy consumption of an electric and an internal combustion passenger car. A comparative case study from real world data on the Erfurt circuit in Germany," *Transp. Res. Proc.*, vol. 27, pp. 468–475, Sep. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146517309419

[6] A. Lajunen and T. Lipman, "Lifecycle cost assessment and carbon dioxide emissions of diesel, natural gas, hybrid electric, fuel cell hybrid and electric transit buses," *Energy*, vol. 106, pp. 329–342, Jul. 2016.

[7] B. Propfe, M. Redelbach, D. Santini, and H. Friedrich, "Cost analysis of plug-in hybrid electric vehicles including maintenance & repair costs and resale values," *World Electric Vehicle J.*, vol. 5, no. 4, pp. 886–895, Dec. 2012.

[8] S. Trommer, V. Kolarova, E. Fraedrich, L. Kröger, B. Kickhöfer, T. Kuhnimhof, B. Lenz, and P. Phleps. (Dec. 2016). *Autonomous Driving—The Impact of Vehicle Automation on Mobility Behaviour*. [Online]. Available: https://elib.dlr.de/110337/1/ifmo_2016_Autonomous_Driving_2035_en.pdf

[9] V. Keller, B. Lyseng, C. Wade, S. Scholtysik, M. Fowler, J. Donald, K. Palmer-Wilson, B. Robertson, P. Wild, and A. Rowe, "Electricity system and emission impact of direct and indirect electrification of heavy-duty transportation," *Energy*, vol. 172, pp. 740–751, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544219301768

[10] M. S. Koroma, D. Costa, M. Philippot, G. Cardellini, M. S. Hosen, T. Coosemans, and M. Messagie, "Life cycle assessment of battery electric vehicles: Implications of future electricity mix and different battery end-of-life management," *Sci. Total Environ.*, vol. 831, Jul. 2022, Art. no. 154859. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969722019520

[11] T. Perger and H. Auer, "Energy efficient route planning for electric vehicles with special consideration of the topography and battery lifetime," *Energy Efficiency*, vol. 13, no. 8, pp. 1705–1726, Dec. 2020.

[12] R. M. Sennefelder, P. Micek, R. Martín-Clemente, J. C. Risquez, R. Carvajal, and J. A. Carrillo-Castrillo, "Driving cycle synthesis, aiming for realness, by extending real-world driving databases," *IEEE Access*, vol. 10, pp. 54123–54135, 2022.

[13] A. Lajunen, "Energy consumption and cost-benefit analysis of hybrid and electric city buses," *Transp. Res. C, Emerg. Technol.*, vol. 38, pp. 1–15, Jan. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X13002234

[14] J. Asamer, A. Graser, B. Heilmann, and M. Ruthmair, "Sensitivity analysis for energy demand estimation of electric vehicles," *Transp. Res. D, Transp. Environ.*, vol. 46, pp. 182–199, Jul. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920915300250

[15] C. De Cauwer, J. Van Mierlo, and T. Coosemans, "Energy consumption prediction for electric vehicles based on real-world data," *Energies*, vol. 8, no. 8, pp. 8573–8593, Aug. 2015. [Online]. Available: https://www.mdpi.com/1996-1073/8/8/8573

[16] M. Gallet, T. Massier, and T. Hamacher, "Estimation of the energy demand of electric buses based on real-world data for large-scale public transport networks," *Appl. Energy*, vol. 230, pp. 344–356, Nov. 2018.

[17] J. Wang, I. Besselink, and H. Nijmeijer, "Battery electric vehicle energy consumption modelling for range estimation," *Int. J. Electric Hybrid Vehicles*, vol. 9, no. 2, pp. 79–102, 2017.

[18] C. Beckers, I. Besselink, J. Frints, and H. Nijmeijer, "Energy consumption prediction for electric city buses," in *Proc. 13th ITS Eur. Congr.*, Brainport, The Netherlands, 2019, pp. 3–6.

[19] O. A. Hjelkrem, K. Y. Lervåg, S. Babri, C. Lu, and C.-J. Södersten, "A battery electric bus energy consumption model for strategic purposes: Validation of a proposed model structure with data from bus fleets in China and Norway," *Transp. Res. D, Transp. Environ.*, vol. 94, May 2021, Art. no. 102804.

[20] L. Maybury, P. Corcoran, and L. Cipcigan, "Mathematical modelling of electric vehicle adoption: A systematic literature review," *Transp. Res. D, Transp. Environ.*, vol. 107, Jun. 2022, Art. no. 103278.

[21] J. Vepsäläinen, K. Otto, A. Lajunen, and K. Tammi, "Computationally efficient model for energy demand prediction of electric city bus in varying operating conditions," *Energy*, vol. 169, pp. 433–443, Feb. 2019.

[22] Y. Chen, G. Wu, R. Sun, A. Dubey, A. Laszka, and P. Pugliese, "A review and outlook on energy consumption estimation models for electric vehicles," *SAE Int. J. Sustain. Transp., Energy, Environ., Policy*, vol. 2, no. 1, p. 5, Mar. 2021. [Online]. Available: https://www.osti.gov/biblio/1824218

[23] T. Pamuła and D. Pamuła, "Prediction of electric buses energy consumption from trip parameters using deep learning," *Energies*, vol. 15, no. 5, p. 1747, Feb. 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/5/1747

[24] A. Kontou and J. Miles, "Electric buses: Lessons to be learnt from the Milton Keynes demonstration project," *Proc. Eng.*, vol. 118, pp. 1137–1144, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877705815021104

[25] E. Ericsson, "Independent driving pattern factors and their influence on fuel-use and exhaust emission factors," *Transp. Res. D, Transp. Environ.*, vol. 6, no. 5, pp. 325–345, Sep. 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920901000037

[26] C. Simonis and R. Sennefelder, "Route specific driver characterization for data-based range prediction of battery electric vehicles," in *Proc. 14th Int. Conf. Ecological Vehicles Renew. Energies (EVER)*, May 2019, pp. 1–6.

[27] H. Abdelaty and M. Mohamed, "A prediction model for battery electric bus energy consumption in transit," *Energies*, vol. 14, no. 10, p. 2824, May 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/10/2824

[28] H. Abdelaty, A. Al-Obaidi, M. Mohamed, and H. E. Z. Farag, "Machine learning prediction models for battery-electric bus energy consumption in transit," *Transp. Res. D, Transp. Environ.*, vol. 96, Jul. 2021, Art. no. 102868. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920921001693

[29] H. Abdelaty and M. Mohamed, "A framework for BEB energy prediction using low-resolution open-source data-driven model," *Transp. Res. D, Transp. Environ.*, vol. 103, Feb. 2022, Art. no. 103170.

[30] R. Basso, B. Kulcsár, and I. Sanchez-Diaz, "Electric vehicle routing problem with machine learning for energy prediction," *Transp. Res. B, Methodol.*, vol. 145, pp. 24–55, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261520304549

[31] C. Sun, F. Sun, and H. He, "Investigating adaptive-ECMS with velocity forecast ability for hybrid electric vehicles," *Appl. Energy*, vol. 185, pp. 1644–1653, Jan. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261916301490

[32] Y. Liu, J. Li, J. Gao, Z. Lei, Y. Zhang, and Z. Chen, "Prediction of vehicle driving conditions with incorporation of stochastic forecasting and machine learning and a case study in energy management of plug-in hybrid electric vehicles," *Mech. Syst. Signal Process.*, vol. 158, Sep. 2021, Art. no. 107765. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327021001606

[33] T. M. Aljohani, A. Ebrahim, and O. Mohammed, "Real-time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and Markov chain model," *Electric Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106962.

[34] P. Li, Y. Zhang, Y. Zhang, Y. Zhang, and K. Zhang, "Prediction of electric bus energy consumption with stochastic speed profile generation modelling and data driven method based on real-world big data," *Appl. Energy*, vol. 298, Sep. 2021, Art. no. 117204.

[35] Y. Chen, Y. Zhang, and R. Sun, "Data-driven estimation of energy consumption for electric bus under real-world driving conditions," *Transp. Res. D, Transp. Environ.*, vol. 98, Sep. 2021, Art. no. 102969.

[36] J. Ji, Y. Bie, Z. Zeng, and L. Wang, "Trip energy consumption estimation for electric buses," *Commun. Transp. Res.*, vol. 2, Dec. 2022, Art. no. 100069. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772424722000191

[37] H. Perugu, S. Collier, Y. Tan, S. Yoon, and J. Herner, "Characterization of battery electric transit bus energy consumption by temporal and speed variation," *Energy*, vol. 263, Jan. 2023, Art. no. 125914. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544222028006

[38] D. Göhlich, A. Kunith, and T.-A. Ly, "Technology assessment of an electric urban bus system for Berlin," *WIT Trans. Built Environ.*, vol. 138, pp. 137–149, May 2014, doi: 10.2495/UT140121.

[39] K. Kivekäs, J. Vepsäläinen, K. Tammi, and J. Anttila, "Influence of driving cycle uncertainty on electric city bus energy consumption," in *Proc. IEEE Vehicle Power Propuls. Conf. (VPPC)*, Dec. 2017, pp. 1–5.

[40] A. Bunzel and B. Baker, "Energy consumption of electric city buses : Determination as a part of a technological and economic evaluation of bus lines with regards to their electrifiability," in *Proc. IEEE Int. Conf. Electr. Syst. for Aircr., Railway, Ship Propuls. Road Vehicles Int. Transp. Electrific. Conf. (ESARS-ITEC)*, Nov. 2018, pp. 1–5.

[41] B. H. Hahn and D. T. Valentine, *Essential MATLAB for Engineers and Scientists*. Amsterdam, The Netherlands: Elsevier, 2017.

[42] T. Markel, A. Brooker, T. Hendricks, V. Johnson, K. Kelly, B. Kramer, M. O'Keefe, S. Sprik, and K. Wipke, "ADVISOR: A systems analysis tool for advanced vehicle modeling," *J. Power Sources*, vol. 110, no. 2, pp. 255–266, Aug. 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378775302001891

[43] R. Rajamani, *Longitudinal Vehicle Dynamics*. Berlin, Germany: Springer, 2006, pp. 95–122, doi: 10.1007/0-387-28823-6.

[44] K. Kivekäs, A. Lajunen, J. Vepsäläinen, and K. Tammi, "City bus powertrain comparison: Driving cycle variation and passenger load sensitivity analysis," *Energies*, vol. 11, no. 7, p. 1755, Jul. 2018. [Online]. Available: https://www.mdpi.com/1996-1073/11/7/1755

[45] D. Göhlich, T.-A. Fay, D. Jefferies, E. Lauth, A. Kunith, and X. Zhang, "Design of urban electric bus systems," *Design Sci.*, vol. 4, Aug. 2018, Art. no. e15, doi: 10.1017/dsj.2018.10.

[46] Z. Gao, Z. Lin, T. J. LaClair, C. Liu, J.-M. Li, A. K. Birky, and J. Ward, "Battery capacity and recharging needs for electric buses in city transit service," *Energy*, vol. 122, pp. 588–600, Mar. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544217301081

[47] T. Barlow, S. Latham, I. S. McCrae, and P. Boulter, "A reference book of driving cycles for use in the measurement of road vehicle emissions," TRL Published Project, Wokingham, U.K., Tech. Rep. PPR 354, 2009.

[48] I. Fomunung, S. Washington, and R. Guensler, "A statistical model for estimating oxides of nitrogen emissions from light duty motor vehicles," *Transp. Res. D, Transp. Environ.*, vol. 4, no. 5, pp. 333–352, Sep. 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920999000139

[49] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, no. 1, pp. 161–168, Jan. 2012.

[50] W.-Y. Loh, "Regression tress with unbiased variable selection and interaction detection," *Statist. Sinica*, vol. 12, pp. 361–386, Apr. 2002. [Online]. Available: http://www.jstor.org/stable/24306967

[51] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. New York, NY, USA: Academic Press.

[52] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[53] S. Haykin, *Neural Networks and Learning Machines*. London, U.K.: Pearson, 2008.

[54] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[55] X. Wu, D. Freese, A. Cabrera, and W. A. Kitch, "Electric vehicles' energy consumption measurement and estimation," *Transp. Res. D, Transp. Environ.*, vol. 34, pp. 52–67, Jan. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920914001485

[56] D. Huang, H. Xie, H. Ma, and Q. Sun, "Driving cycle prediction model based on bus route features," *Transp. Res. D, Transp. Environ.*, vol. 54, pp. 99–113, Jul. 2017.

[57] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," *Transp. Res. C, Emerg. Technol.*, vol. 55, pp. 444–459, Jul. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X15000947

[58] S. Sundrani and J. Lu, "Computing the hazard ratios associated with explanatory variables using machine learning models of survival data," *JCO Clin. Cancer Informat.*, vol. 5, no. 5, pp. 364–378, Dec. 2021, doi: 10.1200/CCI.20.00172.

[59] R. G. Conrad and M. A. Figliozzi, "The recharging vehicle routing problem," in *Proc. Ind. Eng. Res. Conf.*, vol. 8. Norcross, GA, USA: IISE, 2011, pp. 1–6.

[60] D. Goeke and M. Schneider, "Routing a mixed fleet of electric and conventional vehicles," *Eur. J. Oper. Res.*, vol. 245, no. 1, pp. 81–99, Aug. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221715000697

ROMAN MICHAEL SENNEFELDER was born in Munich, Germany. He received the B.Sc. and M.Sc. degrees in electrical engineering and information technology from the Technical University of Munich, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the university of Seville and collaboratively working at EVO Engineering GmbH.

During his studies and also in the time after, he worked in the automotive engineering industry with focus on electric drive systems, components, and alternative mobility in general. Concurrently, he was involved in several research projects and collaborations on various e-mobility topics. Among others, his research interests include data science, machine learning, energetic analysis, range prediction, driving characterization, and the technical and operational efficiency improvement of electric and hybrid vehicles.

RUBÉN MARTÍN-CLEMENTE (Member, IEEE) received the M.Eng. degree in telecommunications engineering and the Ph.D. degree (Hons.) in telecommunications engineering from Universidad de Sevilla, Seville, Spain, in 1996 and 2000, respectively. He is currently the Head of the Department of Signal Theory and Communications, Universidad de Sevilla. He was a Visiting Researcher with the University of Regensburg, Regensburg, Germany, in 2001 and 2009, and the University of Nice, France, in 2015, 2016, and 2018, respectively. Among other areas, his research interests include signal processing and machine learning with emphasis on independent component analysis and its application to biomedical problems. He has authored or coauthored numerous publications on these topics.

RAMÓN GONZÁLEZ-CARVAJAL (Fellow, IEEE) was born in Seville, Spain. He received the M.Sc. degree in electrical engineering and the Ph.D. degree (Hons.) from Universidad de Sevilla, Seville, in 1995 and 1999, respectively. Since 1996, he has been with the Department of Electronic Engineering, School of Engineering, Universidad de Sevilla, where he has been an Associate Professor, since 1996, and a Professor, since 2002. He was an Invited Researcher with the Klipsch School of Electrical Engineering, New Mexico State University (NMSU), Las Cruces, NM, USA, in 1999 and from 2001 to 2004, and also with the Department of Electrical Engineering, Texas A&M University, College Station, TX, USA, in 1997. He was an Adjunct Professor with the Klipsch School of Electrical Engineering, NMSU. He has authored over 150 papers in international journals and 200 in international conferences. His current research interests include embedded systems and the IoT for smart cities and automotive applications.

DIMITAR TRIFONOV was born in Varna, Bulgaria. He received the bachelor's and master's degrees in mechanical engineering from the Technical University of Munich (TUM), Germany, in 2018 and 2021, respectively. He is currently a Hardware Test Engineer with EVO Engineering GmbH, Munich, where his main focus is testing and validation of power electronic components of electric vehicles. His current research interests include energy management strategies for optimizing the battery lifetime and range of electric vehicles, on which topic he has coauthored several publications.