IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 3, 2021, accepted December 13, 2021, date of publication December 14, 2021,
date of current version December 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135604

Segment-Based CO_2 Emission Evaluations From Passenger Cars Based on Deep Learning Techniques

NAGHMEH NIROOMAND, CHRISTIAN BACH, AND MIRIAM ELSER
Automotive Powertrain Technologies Laboratory, Swiss Federal Laboratories for Materials Science and Technology, CH-8600 Dübendorf, Switzerland
Corresponding author: Naghmeh Niroomand (naghmeh.niroomand@empa.ch)

ABSTRACT The overall level of emissions from the Swiss passenger cars is strongly dependent on the fleet composition. Despite technology improvements, the Swiss passenger cars fleet remains emissions intensive. To analyze the root of this problem and evaluate potential solutions, this paper applies deep learning techniques to evaluate the inter-class (namely micro, small, middle, upper middle, large and luxury class) and intra-class (namely sport utility vehicle and non-sport utility vehicle) differences in carbon dioxide (CO_2) emissions. This paper takes full use of novel semi-supervised fuzzy C-means (SSFCM), random forest and AdaBoost models as well as model fusion to successfully classify passenger vehicles and enable segment-based CO_2 emission evaluations.

INDEX TERMS CO_2 emissions, feature learning, semi-supervised deep learning, vehicle classification.

I. INTRODUCTION

More than 5 years after the adoption of the Paris agreement, which aims to limit global warming to below 2 °C (preferably 1.5 °C), global greenhouse gas emissions continue growing steadily [1]. According to the 2016 EU Reference Scenario, without an ambitious commitment towards decarbonization, transport related carbon dioxide (CO_2) emissions are expected to decrease only by 8% between 2010 and 2050 and will reach their largest share by the end of the projection period (2050) [2], [3]. Underlying this limited decrease are a significant increase in the number of passenger cars, the slow market penetration of electric cars and a limited shift towards alternative fuels.

Switzerland is responsible for less than 0.2% of global man-made fossil CO_2 emissions [4]. However, the transport sector represents the largest consumer of fossil fuels in Switzerland and caused about 32% of Switzerland's CO_2 emissions in 2019 (i.e., around 15 million tonnes CO_2 eq., excluding international aviation and shipping). Road transport was responsible for 98% of these emissions, with only small contributions from national shipping (0.8%), aviation (0.8%) and rail transport (0.2%). Among the different forms of road transport, passenger cars accounted for almost two thirds of the total emissions (73%), followed by freight transport (21%), buses (3%) and motorcycles (2%) [5]. Therefore, in order to meet the CO_2 reduction targets of Switzerland [6], fossil energy carriers in the mobility sector have to be substituted by ones based on renewable energies and the overall energy consumption has to be reduced. Although substituting fossil energy by renewable ones is essential to meet the CO_2 reduction targets, decisions about investments and new policies are not moving fast enough to decarbonize the economy in compliance with the Paris agreement.

On the other hand, during the last decades there have been large technical and dimensional changes in new passenger vehicles, mostly related to technology improvements and intra-class variations. Particularly relevant are the changes in the dimensions of the vehicle segments (i.e., increased size of the vehicles in most segments), within single vehicle segments (i.e., increased share of SUVs), and other design parameters like increased efficiency of the engine and engine displacement down-sizing. Understanding the impact of these changes on the fuel consumption and CO_2 emissions is crucial to develop successful strategies to decarbonize the road transport.

Since the division of vehicles into segments by experts is not standardized and therefore not always uniform, and

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

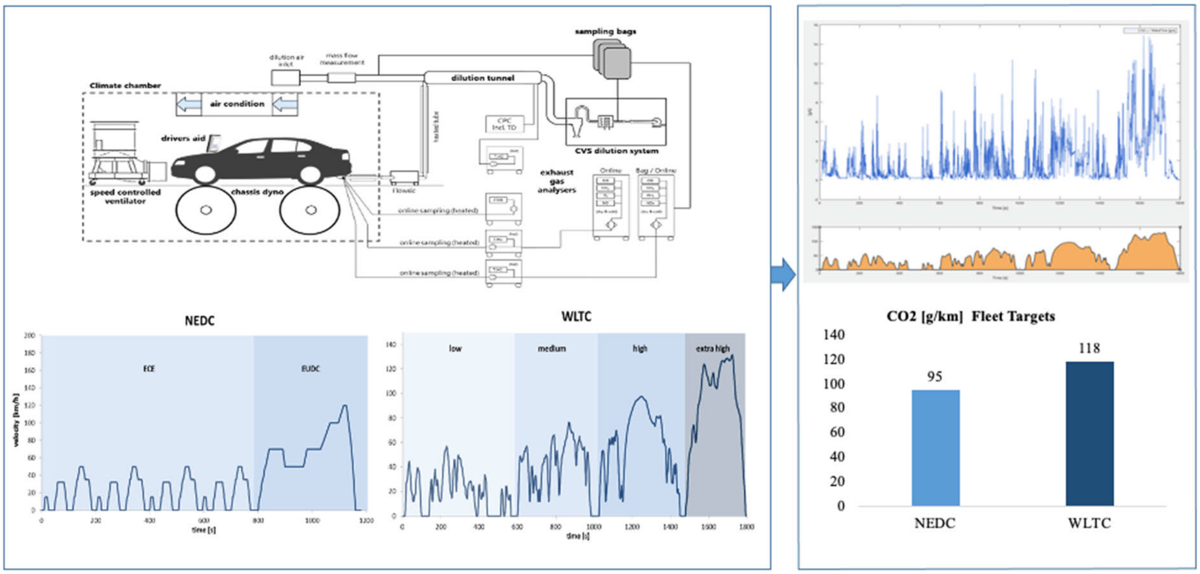


FIGURE 1. Type approval CO_2 emission measurements are conducted for new vehicles using regulatory cycles. Under the New European Driving Cycle (NEDC), the average level of CO_2 emissions from new passenger cars registered may not exceed 95 g CO_2 /km in 2021. This corresponds to a WLTP (world harmonized light-duty vehicles test procedure) target of 118 g CO_2 /km. The switch to the WLTP procedure results on average on 20 % higher emissions than with the NEDC [13].

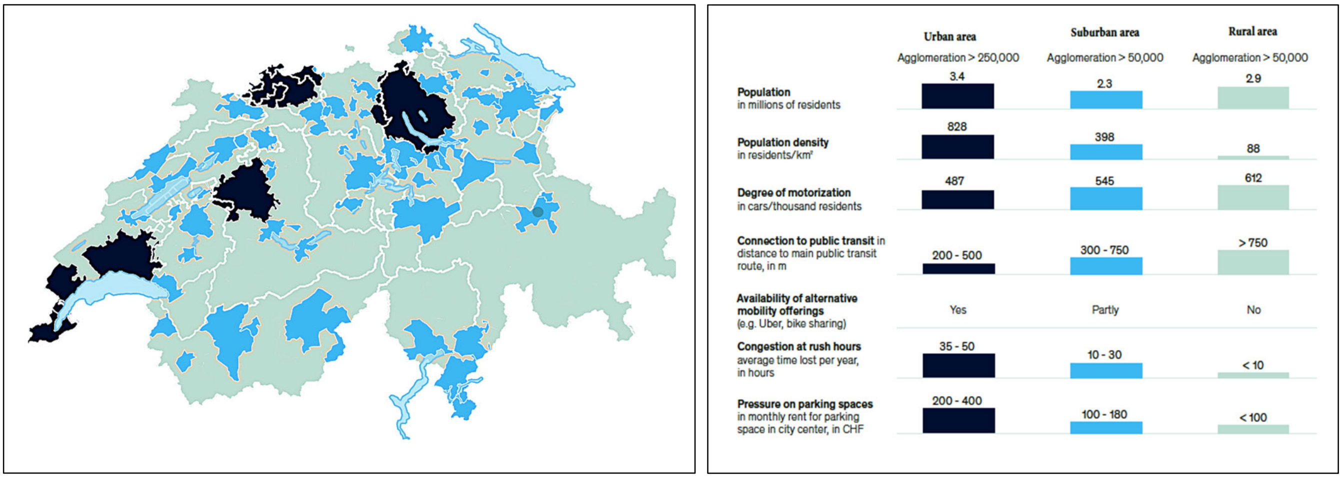
some vehicle models have recently positioned themselves as ‘‘crossovers’’ between established vehicle categories [7], [8], it has become increasingly difficult and inaccurate to segment the vehicle population using conventional classification methods. Using mathematical approaches, vehicles can be uniformly divided into segments based on similarity features. The development of a mathematical approach to accurately segment passenger vehicles is essential for determining the re-al CO_2 emissions from road traffic in the future. While road traffic has so far had its own energy system, which was comparatively easy to assess in terms of CO_2 emissions, increasing electrification of road traffic will difficult the distinction of energy consumption from road traffic and other stationary energy uses. Moreover, the estimated overall impact of the introduction of the world harmonized light-duty vehicles test procedure (WLTP) on average CO_2 emissions is in the order of 15-25%, which would lead to on average 18-30 g/km higher CO_2 emissions for the new passenger cars [9]–[12] (Fig. 1). Moreover, due to the limited informative value of the CO_2 type approval values on the real CO_2 emissions, the wide margin of uncertainty regarding vehicle classification and the type approval extension based on the new definitions in the test protocol [11], this segmentation is an important step on the way to a new CO_2 assessment of road traffic.

In this study, by segmenting the passenger vehicles based on technical and dimensional characteristics, we aim to better understand the impact of inter-class (between classes of a multi-class) and intra class (within each class) variations to the passenger vehicle fleet CO_2 footprint [14]. In our approach, several semi-supervised clustering algorithms are compared and used to predict labels from unsupervised clustering algorithms based on a feature learning technique, which is a highly useful method for representation learning with high-dimensional datasets containing high-level uncertainties [15]–[24]. This paper is an extension of a previous work originally focused on developing a machine learning based methodology for the mathematical inter-class and intra-class segmentation of passenger vehicles. Here we improved the classification performance of this method by adding emission and technical features as an input. Based on this novel approach, we can then predict accurate segment-based CO_2 emissions, which allows for detailed analyses of the main factors influencing the average fleet CO_2 emissions. Our results show that the proposed method is a viable and effective to categorize vehicles based on their technical, emission and dimensional features.

Section II briefly introduces the Swiss transportation system. Section III presents the related research. Section IV describes the methods. Section V provides concise details on the used datasets, the algorithms, the performed experiments and the discussion of the results and last, section VI provides the majors findings of our work and recommendations for further research.

II. SWISS TRANSPORT SECTOR AND CO_2 EMISSIONS

In terms of mobility, Switzerland can be divided in three main regions, namely urban, suburban, and rural areas. There are major differences in the sustainability challenges posed within these regions due to the urbanization. Fig. 2 illustrates that the growth of the number of cars has placed additional pressures on traffic congestion and parking spaces, particularly in higher density areas. This creates opportunities for offering alternatives to cover the existing transportation



	Urban area Agglomeration > 250,000	Suburban area Agglomeration > 50,000	Rural area Agglomeration > 50,000
Population in millions of residents	3.4	2.3	2.9
Population density in residents/km²	828	398	88
Degree of motorization in cars/thousand residents	487	545	612
Connection to public transit in distance to main public transit route, in m	200 - 500	300 - 750	> 750
Availability of alternative mobility offerings (e.g. Uber, bike sharing)	Yes	Partly	No
Congestion at rush hours average time lost per year, in hours	35 - 50	10 - 30	< 10
Pressure on parking spaces in monthly rent for parking space in city center, in CHF	200 - 400	100 - 180	< 100

FIGURE 2. Distribution of population and key parameters describing road mobility in urban, suburban and rural areas in Swizterland. Source: Statistik der Schweizer Städte [26], Global Traffic Scorecard (Inrix) [27].

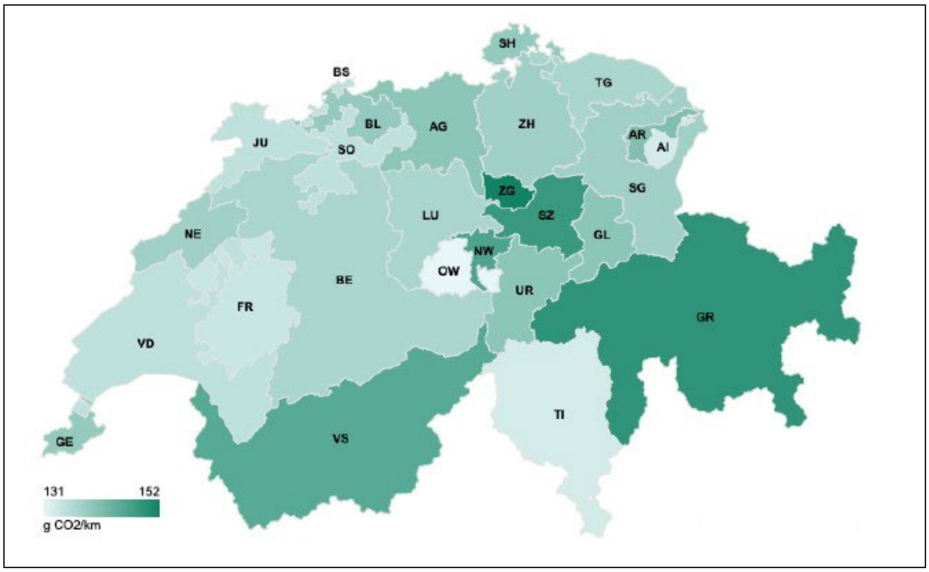


FIGURE 3. Average CO_2 emissions (in g CO_2 /km) of new passenger car registrations in Switzerland in 2018.

needs, including public transit network and shared mobility services. In contrast, in rural areas, which represent about one third of the total Swiss population, due to the lack of attractive and feasible transport alternatives, private automobile remains the most common form of transportation. In addition, despite the high rate of the population accepting public transport modes in Switzerland (59%), two thirds of the total passenger kilometers are still completed by car [25].

Over 6.5 million motor vehicles were registered in 2018 in Switzerland, more than 4.6 million of which were passenger cars. Around one third of all passenger cars were more than ten years old and over 1.6 million cars were completely outdated. Only 0.4% of all passenger cars had purely electric propulsion systems. Among the new registrations, petrol was the most common fuel (68%), while almost 30% of the sold vehicles had a diesel engine [28]. The mean type approval CO_2 emissions of newly registered cars continuously dropped from around 190 g CO_2/km in 2003 to around 134 g CO_2/km in 2016. After this steady decline, the mean CO_2 emissions of the new registrations rose again to 137.8 g CO_2/km in 2018. As a result, the specified target value of 130 g CO_2/km that came into force in 2012 was not entirely met. Fig. 3 shows that new cars in the Southeast of the country are generally less fuel efficient and produce more CO_2 emissions compared to the new cars in the Northwest.

It is forecasted that the share of electric vehicles in the Swiss passenger car fleet will increase from 1.5% in 2021 to 38-74% in 2050, depending on the considered scenario [29]. In addition to the considerable savings in terms of fossil fuel consumption, the increasing share of electric vehicles will drastically reduce the CO_2 emissions compared to vehicles powered by fossil fuels [30].

III. RELATED WORK

Over the last decades, as a result of new European CO_2 regulations, car manufacturers have used a whole range of technical and dimensional solutions to meet specific annual CO_2 emission targets. The review of related literature shows that the large changes in the passenger car models over time poses an additional challenge to the accurate vehicle classification [31]–[34].

In spite of the partial achievement of the targets based on the type approval CO2 emissions (laboratory tests), real-world CO2 emissions have decreased by only about 10% [35]. Subsequently, the gap between the calculated and real-world CO2 emissions has widened from 9% to 42%, resulting in 31 g CO_2/km of fake emissions savings [36], [37]. This gap varies considerably across countries due to the significant variation within vehicle classes. However, the lack of a standard classification method, hinders the comparison of these results between different countries [38], [39].

In order to close the gap between the CO2 emissions results estimated by two major techniques (top-down approaches focusing on fuel market interactions and bottom-up approaches focusing on technological details), researchers have developed multiple simulation programs, such as greenhouse gas emission models and vehicle energy calculation tools, for the compilation of emission inventories [40]–[43], [45]. From this point of view, the simulation is useful to compensate the limitations of the laboratory test methods. For example, Seo *et al.* [46] developed a vehicle type classification simulation method using a bottom-up approach to calculate national CO2 emissions. This study concluded that CO2 emissions of medium and heavy-duty vehicles (MHDV) represented 25.5% of the total on-road emissions, although only 4.2% of all vehicles were MHDV. Jimenez *et al.* [47] reviewed the influence of vehicle classification, vehicle characteristics, vehicle brand and registration year on the real-world CO2 emissions. They employed a database of 650 passenger cars. This study explained the impact of these factors on the gap between real-world and type-approval emission values. Ntziachristos *et al.* [48] reported that the deviations in fuel consumption are directly reflected in CO2 emissions. This study computed the observed 11 % gap for in-use petrol and 16% for in-use diesel with the type-approval procedure by controlling engine capacity, vehicle mass and power. They used a database of 924 passenger cars from Europe. The results indicated that the large vehicle class has the highest deviation in test score.

All these studies show that simulation techniques are capable to overcome some of the limitations faced with fuel-based approach in terms of estimating the CO2 emissions of each vehicle class. However, the simulation techniques cannot consider intra-class variations in CO_2 emissions, they are difficult to use when conducting a detailed analysis and they require expert knowledge.

Lately, feature learning techniques have shown an outstanding performance for addressing uncertainty problems for clustering and classification [19], [49]–[55]. The classification performance highly depends on a quality of features generated from the data as input to the classifier process. However, only a limited number of studies have combined feature learning techniques to improve the classification performance on a high dimensional dataset and predict the vehicle CO2 emissions. For more details about the feature learning techniques, refer to the article by He *et al.* [56], in which the authors implemented feature learning classification to analyze vehicular emissions. In particular, they applied decision tree, random forest, AdaBoost, and XgBoost models based on the fuel type and registration date. This study achieved a prediction accuracy of 70 % by artificially controlling the registration date for different users. Saleh *et al.* [57] used deep learning with a support vector machine (SVM) model to predict CO2 emissions by monitoring energy consumption. The low value of Root Mean Square Error of the model indicates the high accuracy of the prediction. Ghahramani *et al.* [58] proposed an unsupervised learning approach to estimate CO2 emissions from road transport with a focus on taxi trips. This study identified the most polluting trips and the vehicles associated with these trips in order to replace them with alternative alternatives powertrains, such as electric vehicles.

The classification method proposed in this paper is a new semi-supervised clustering scheme (SSFCM) that incorporates semi-supervised information in fuzzy C-means (FCM) algorithm to considerably improve its effectiveness [59]–[63]. In this field, Jiang *et al.* [64] combined several feature extraction methods with a support vector machine classifier to group the vehicles in six categories ‘‘large bus’’, ‘‘passenger car’’, ‘‘motorcycle’’, ‘‘minibus’’, ‘‘truck’’ and ‘‘van’’. This study achieved a classification accuracy of 97.4%. Balid *et al.* [65] implemented deep learning-based classification using the vehicle length as a key feature. Their method classifies vehicles into passenger vehicles, single unit trucks, combination trucks, and multi-trailer trucks and achieved a classification accuracy of 97%. Maungmai and Nuthong [66] used a convolutional neural network method to classify the vehicles type as ‘‘small’’, ‘‘medium’’, ‘‘large’’, and ‘‘unknown’’, and vehicle color as ‘‘black’’, ‘‘blue’’, ‘‘white’’, ‘‘green’’, ‘‘yellow’’, ‘‘red’’, and ‘‘unknown’’. The results comparison shows that, using decision trees, random forest, and densely deep neural network classifier, the classification accuracy of vehicle type and vehicle color increased by 1.8% and, 0.8%, respectively. Dong *et al.* [67] proposed a vehicle type classification method using a semi-supervised convolutional neural network using high-resolution vehicle frontal view images. The algorithm achieved 88.1% accuracy.

IV. MATERIALS AND METHODS

A. SEMI-SUPERVISED CLUSTERING

Semi-supervised clustering aims to boost the accuracy of the defined clusters by identifying better clusters than the ones obtained from the unsupervised learning algorithm [19], [68]–[74]. Typically, semi-supervised clustering methods

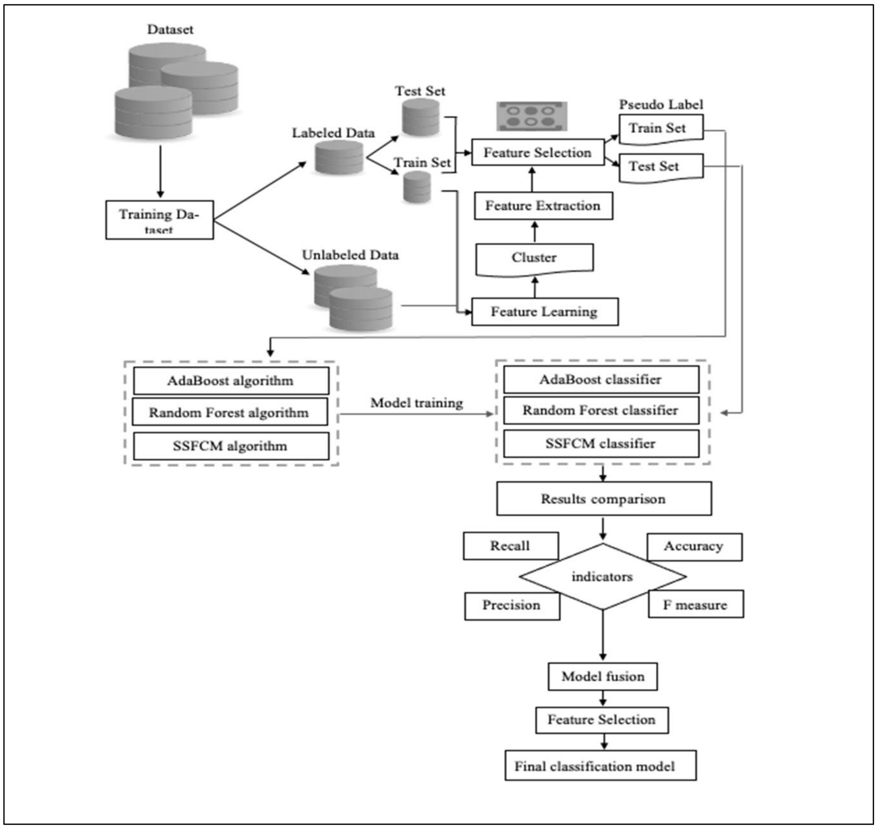


FIGURE 4. The structure of the proposed semi-supervised deep learning approach.

result in a worse representation of the results in the original feature space. To make the semi-supervised clustering more efficient, it is reasonable to combine semi-supervised clustering with deep feature learning [63], [75]–[77]. The framework of the proposed clustering approach is depicted in Fig. 4.

Unlike the most widely used approaches in semi-supervised clustering based on the feature extraction technique, we consider three types of information (diffusion labels, extracted core data, and extracted feature vectors) in order to improve the classification accuracy and mitigate the class imbalance and multi-class overlapping problems. This framework includes three main layers. First, labeled data is divided into train set and test set in order to build a classifier and evaluate its output, respectively. Then, recordings from the train set along with the unlabeled data are used as input to the feature learning process. The output of the feature learning step are the cluster centroids that are used to project data from train and test sets into a new learnt space and extract feature vectors in the feature extraction step. In the classification step, AdaBoost [78], Random Forest [79], and SSFCM models are built on the vectors of the train set and then used to predict the labels for the feature vectors of the test set. Finally, the performance parameters of the three single models are compared to the model fusion ones to evaluate their performances in terms of data classification and prediction.

B. SEMI-SUPERVISED FUZZY C-MEAN CLUSTERING

Fuzzy C-means (FCM), as an overlapping clustering algorithm, is one of the most popular fuzzy clustering methods [80]. FCM is a soft clustering algorithm, meaning that each data point has a probability of belonging to each cluster with partial membership values ranged from 0 to 1. How-ever, due to the non-convexity of its objective function, it may fall into a local optimal solution during optimization. To address this issue, we propose a semi supervised fuzzy C-means clustering (SSFCM) that incorporates deep feature learning in FCM to further improve its effectiveness and eliminate redundant information [81]–[83].

This method aims to minimize the objective function (J) as follows:

$$Min\ J\left(X;U,V\right)=\sum\nolimits_{k=1}^{N}\sum\nolimits_{i=1}^{C}u_{ki}^{m}D_{kiA}^{2}\ \left(1\leq m<\infty\right) \tag{1}$$

$$s.t.\ \sum\nolimits_{i=1}^{c}u_{ki}=1\quad\left(0\leq u_{ki}\leq 1\right) \tag{2}$$

$$v_i=\frac{\sum\nolimits_{k=1}^{N}u_{ki}^{m}X_k}{\sum\nolimits_{k=1}^{N}u_{ki}^{m}} \tag{3}$$

$$u_{ki}=\frac{1}{\sum\nolimits_{j=1}^{C}\left(\frac{D_{kiA}}{D_{kjA}}\right)^{2/(m-1)}} \tag{4}$$

$$D_{kiA}^{2}=\|X_k-v_i\|_A^2=\left(X_k-v_i\right)^T A\left(X_k-v_i\right) \tag{5}$$

where N is number of data elements, C is the number of clusters; X_k represents the data k of X = {X_1,X_2,X_3,…,X_N} in the i^{th} cluster; u_{ki} represents the weighted squared errors function known as membership function; m is a weighting exponent that determines the degree of fuzziness and that was set to 2 in order to ensure high membership values for each data point to its closest cluster; A is a positive and symmetric (n × n) weight matrix; U is the fuzzy partition matrix of the dataset X into c cluster; v_i is vectors of center in i^{th} cluster; K denotes the features, and $\|x_k - v_i\|_A^2$ denotes to the Euclidean distance function and it is computed in the A norm between j^{th} data and i^{th} cluster center.

The SSFCM methodology is composed of the following four steps. First, with algorithm 1 we find the FCM memberships and centroids:

Algorithm 1 Fuzzy C-means Membership and Centroid

```
Input: N data elements X = {X1, X2,…,XN}, weight matrix (A), number of clusters (C), degree of fuzziness (m = 2), max iteration number (T), error threshold ($\varepsilon$)
Output: $u_{ki}$, $v_i$
    Set t = 0
1. Initialize $v_i$
2. Update t = t + 1
3. Compute $u_{ki}$
4. Compute $v_i$
5. If t > T or $||u_t - u_{t-1}|| < \varepsilon$ then stop; otherwise
6. Repeat from step 3.
```

Next, algorithm 2 is used to calculate deep FCM memberships and centroids:

Algorithm 2 The Training Strategies for Deep Fuzzy C-Means

```
Input: $N$ data elements X = {$X_1$, $X_2$,…,$X_N$}, number of clusters ($C$), clusters feature ($K$), labeled dataset ($L$), unlabeled dataset ($UN$), membership degree ($U$), max iteration number ($T$), error threshold ($\varepsilon$)
    Set t = 0
1. Initialize $v_i^k$ (random for labeled data)
2. Update t = t + 1
3. Compute $u_{iL}$, $u_{iUNL}$
4. Compute $v_{iL}^{k+1}$, $v_{i\mathrm{UNL}}^{k+1}$
5. If a stopping criterion, t > T or $||J_t - J_{t-1}|| < \varepsilon$, is fulfilled for all labeled and unlabeled objective functions separately then stop; otherwise
6. Repeat from step 3.
```

Then, using algorithm 3 we select the features (s ⊂ K) by using the random oversampling (ROS) technique. The purpose of the ROS approach is to maintain a balance between the feature subsets of labeled classes and unlabeled data elements [84], [85].

Algorithm 3 Feature Extraction of Deep Fuzzy C-Means

```
Input: $N$ data elements X = {$X_1$, $X_2$,…,$X_N$}, clusters feature ($K$), labeled dataset ($L$), unlabeled dataset ($UN$), $\mu$ ($D$) mean of the elements of D, set of the centroids ( $v_{iL}^k$, $v_{i\mathrm{UNL}}^k$)
Output: Set of extract features of labeled and unlabeled dataset
    Set Q = Ø
1. Compute $D_{Lk} = \|x_{iL} - v_{iL}^k\|$
2. Compute $D_{UNLk} = \|x_{iUNL} - v_{iUNL}^k\|$
3. Compute means $D_{Lk}\&D_{UNLk}$ of elements
   $\mu_i\left(D_{iL}\right), \mu_i\left(D_{iUNL}\right)$
4. feature extraction ($f_k(x) = \max(0, \mu\,(D) - D_k)$
5. for all L and UNL features do
6. Return Q
```

Next, we apply the Euclidean distance technique, which is the most applied (dis)similarity or distance metric to measure the similarity between the labeled and unlabeled feature vectors. The outcome is the maximum average of the maximum relevant and minimum redundant features between each selected feature of unlabeled data and labeled classes [86]:

$$\max \mathrm{Sim}_i(X_j, V_L^s) = \min d_{jiL} = \min \left|X_j - V_{iL}^s\right| \qquad (1 \leq i \leq c), \quad X_j \epsilon X_{UNL} \tag{6}$$

Last, in algorithm 4 the maximum average of the maximum similarity between the selected features are estimated and used in the classifiers.

Algorithm 4 Semi Supervised Fuzzy C-Means Classifier

```
Input: N data elements X = {$X_1$,$X_2$,…,$X_N$} with minimum features in any subset ($s$), set of the centroid ($V_{iL}^s$, $V_{UNL}^s$) of selected features
Output: Predicted labeled data ($Q = \{q_{L+1}, q_{L+2}, \ldots, q_{L+N}\}$)
    Set Q = Ø
1. For i ϵ {1, …, c} do
2. For j ϵ {1, …, N} do
3. Employ $V_{iL}^s$ to calculate $max\ Sim_i$
4. If maximum average of $max\ Sim_i \epsilon i^{th}$ labeled class, then
5. Append $X_j$ to $i^{th}$ labeled class
6. Update Q if a labeled class is achieved
7. For all $V_{iL}^s \epsilon V_L^s$ do
8. Return Q
```

C. STATE-OF-THE-ART METHODS

Two ensemble learning methods, Random Forest and AdaBoost, are used to enhance the accuracy and performance of the classification [87], [88]. The Random Forest model is a parallel learning process that uses a bagging technique for the data training [89]. This data sampling technique aims to reduce the variance and bias in the model by generating multisets (multiple decision trees) for training from the original data. In the parallel process none of these decision trees is dependent on other trees.

Algorithm 5 Random Forests Classifier

```
Input: Training set ($S$), decision tree in forest ($B$), the subsample size ( $\mu$), max iteration number (T)
Output: Set Q = Ø
1. for t ϵ {1, …, T} do
2. for b ϵ {1, …, B} do
3. Sample $\mu$ instances from S with replacement $S_t$
4. Build classifier $Q_t$ using B on $S_t$, then
5. Return T
```

On the other hand, the AdaBoost model is a sequential learning process that uses the training data to make subsequent decision stumps [90]. In the sequential process the decision stump is dependent on the previous decision stump. In fact, the error made in the first decision stump through mis-classification of few datasets influences the next decision stump by assigning higher weights for those training data.

Algorithm 6 AdaBoost Classifier

```
Input: Data X whose number of elements $N$, training set ($S$), decision tree in forest ($B$), subsample size ( $\mu$), max iteration number (T)
Output: Set Q = Ø
1. for t ϵ {1, …, T} do
2. Initialize data weights {$D_n$} to 1/N
3. find best weak classifier $y_m(x)$ by minimizing weighted error function $J_m$:

   $J_m = \sum_{n=1}^{N} D_n^{(m)} 1[y_m(x_n) \neq t_n]$

4. Compute $err_m = \sum_{n=1}^{N} D_n^{(m)} 1[y_m(x_n) \neq t_n] / \sum_{n=1}^{N} D_n^{(m)}$
5. assign weight $\alpha_m = \log(\frac{1-err_m}{\varepsilon err_m})$ to classifier $y_m(x)$
6. update the data weights: $D_n^{(m+1)} = D_n^{(m)} \exp\{\alpha_m 1[y_m(x_n) \neq t_n]\}$
7. Normalize $D_n^{(m+1)}$ to be proper distribution
8. Make prediction using the final model: $Y_M(x) = sign(\sum_{m=1}^{M} \alpha_m y_m(x))$
```

D. PERFORMANCE MEASURE

To assess the performance of the different algorithms, we compute the confusion matrix and use it to determine the precision (P_i), recall (R_i), F-Measure and adjusted rand index (ARI) as given in the following:

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (1 \leq i \leq c) \tag{7}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (1 \leq i \leq c) \tag{8}$$

$$F - Measure = \frac{2P_iR_i}{P_i + R_i} \tag{9}$$

$$RI = \frac{TP + TN}{TP + FN + TN + FP} \quad (0 \leq RI \leq 1) \tag{10}$$

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (-1 \leq ARI \leq 1) \tag{11}$$

Here, TP_i (True Positives) is the proportion of data points classified correctly to each class i; FN_i (False Negatives) is the proportion of data points that are not classified to class i but actually belong to class i; TN_i (True Negatives) is the proportion of data points that are correctly not assigned to class i; FP_i (False Positives) is the proportion of the data points that are incorrectly assigned to class i.

E. MODEL FUSION

The Model fusion method is a deep learning process, by which different classification predictive modeling algorithms associated with individual weights are trained and combined in order to enhance the final estimation. This method turns out to be a stronger meta-classifier as it combines different classification models using a majority voting classifier estimator, partially overcoming the weaknesses of single classifiers and achieving higher classification accuracy. The commonly used voting classifiers include the hard voting classifier and soft voting classifier. The hard voting classifier takes the majority vote applying equal weights to each classifier (mode of all the predicted la-bels is taken) while the soft voting classifier takes the majority vote based on applying different weights to each classifier (probability of all the predicted labels is taken) [56, 93]. The voting classifier predictions can be defined as:

$$\begin{aligned} &H_{vote}(x) \\ &= \max\left\{\sum\nolimits_j lab(x, j, 1), \sum\nolimits_j lab(x, j, 2)\right. \\ &\quad \left., \ldots, \sum\nolimits_j lab(x, j, c)\right\} (1 \leq j \leq T)(1 \leq c \leq K) \end{aligned} \tag{12}$$

$$\begin{aligned} &S_{vote}(x) \\ &= \max\left\{\frac{\sum_i p(x, j, 1)}{n_T}, \frac{\sum_i p(x, j, 2)}{n_T}, \ldots, \frac{\sum_i p(x, j, c)}{n_T}\right\} \end{aligned} \tag{13}$$

where $H_{vote}(x)$ denotes the vote result of hard voting, lab (x, j, c) is an indicator function that shows if x belongs to label c calculated by j^{th} classifier, $S_{vot}e(x)$ is the vote result of hard voting, p (x, j, c) is the probability for classifier of exceeding some threshold values, n_T refers to the total number of classifiers and k is the number of labels.

V. EXPERIMENTS

A. DATA PREPARATION

The core dataset of this work is the Swiss Motor Vehicle Information System (MOFIS) [92], which contains over 6.5 million passenger vehicles along with their type approval numbers, geometric and weight properties, ownership details, technical information and date of registration. In addition, we also use vehicle technical specifications and vehicle expert segmentation data from the Technical Type Approval

TABLE 1. Performance of tested models on dataset with labeled rate of 10% from each class.

Method	Accuracy rate	Precision rate	Recall	F1
SSFCM	0.954	0.953	0.881	0.916
AdaBoost	0.891	0.871	0.823	0.846
Random Forest	0.902	0.89	0.86	0.875
Hard Voting	0.921	0.935	0.871	0.902
Soft Voting	0.942	0.956	0.878	0.915

Information from the Federal Roads Office (ASTRA) [28] and a Vehicles Expert Partner [93], respectively.

The data-mining framework consists of three main components: filtering of raw data, extraction of the vehicle sample by registration year, and identification of suitable clustering attributes. In a first step we filter the dataset by removing the vehicles that do not meet the definitions of typical passenger cars, such as small pickup trucks, standard pickup trucks, vans, special purpose vehicles (SPVs), sports cars and multi-purpose vehicles (MPVs) [18]. By considering the goal of this paper, the dataset is then separated into two parts, a training part and a testing part. The training dataset contains 308,824 new registered passenger cars in 2018 along with 30 features including emissions (carbon dioxide (CO_2), carbon monoxide (CO), nitrogen oxides (NOx), particulate matter (PM2.5), etc.), weight properties, dimensional features (length, height, width, axle, etc.) and vehicle technical specifications (power, engine capacity, drive, torque, etc.) for each car. It is important to note that we used two different values of CO_2 emissions, namely the average type approval values provided in the ASTRA database (measured CO_2) and the vehicle specific type approval values that also consider the vehicle weight and the gearbox and are reported in the MOFIS database (calculated CO_2).

B. EXPERIMENTAL SETUP AND RESULTS

In the first step of the learning process the training dataset is considered to contain two types of patterns: unlabeled and labeled data. The labeled dataset results from applying the unsupervised FCM clustering algorithm to the total 366 unique new registered passenger cars (based on make, model and manufacturer code) based on the dimensional features: micro class containing 18 samples, small class containing 50 samples, middle class containing 110 samples, upper middle class containing 84 samples, large class and luxury class containing 104 samples. The average accuracy rate and adjusted rand index of the FCM clustering algorithm in comparison to the Swiss expert classification was approximately 79% and 75%, respectively [18]. Due to some limitations of the unsupervised FCM clustering algorithm, only the labeled data with true labels (in vehicle class, measured CO_2 and calculated CO_2) with a membership degree greater than 0.95 were used as the core dataset to extract the accurate classification of misclassified samples and provide the base for the later step of training. Following this, we selected 10% of the data from each class as training labeled samples.

The preliminary statistical analysis on the correlation between the emissions, vehicle segments, sub-segments and preselected influencing factors demonstrated a high correlation between the features. In the feature learning process, the unlabeled data and the previous labeled data along with the labels of the core dataset are used as input. Each group of labeled and unlabeled data has a set of features in common. In order to eliminate multicollinearity, principal component analysis (PCA) was performed on the data. Prior to model development, new features are extracted to reduce the number of features. In the feature extraction step, the cluster centroids are defined using algorithm 2 and each patch is transformed to a feature vector.

In the feature selection step (algorithm 3), the resampling (ROS) technique is used in order to in-crease the number of extracted features from minority groups until it equals the number of features in the majority group. Then, algorithm 4 (based on the Euclidean distance) is used to select the best features and remove redundancy from the feature vector. After we initialized all parts, pseudo labels of labeled data are assigned to the unlabeled data in the training data. Following, this unlabeled data with pseudo labels is used to pre-train the SSFCM, random forest (algorithm 5) and AdaBoost (algorithm 6) classification algorithms by extracting discriminative features. Finally, model fusion is applied using only the labeled data with true labels.

The experimental results show that the single clustering models using SSFCM, random forest and AdaBoost algorithms and the fusion model all enhance the classification accuracy in comparison to the traditional FCM algorithm (overall accuracy of 79%). Among them, the soft voting fusion model and the SSFCM provide the most accurate results, 94.2 and 95.4% respectively. The F-measure value (F1), which represents the model performance, is 91.6% for the SSFCM clustering algorithm and 91.5% for the fusion model with soft voting (Table 1).

From the results of the model fusion, we extract the final features reported in Fig. 5 which we use to re-run the single algorithms and select the final classification model.

The underlying assumption of feature extraction is that it leads to improved classification results in comparison to the initial classifier's predictions with the original features. To verify that this assumption holds for our task, we use the prediction accuracy and other verification measures to check the classification performance of traditional FCM with the original features and the SSFCM, random forest and AdaBoost algorithms with feature extraction. It can be seen in

TABLE 2. Prediction accuracy and verification clustering results.

Method	Training accuracy	Test accuracy
SSFCM	0.952	0.904
AdaBoost	0.781	0.715
Random Forest	0.903	0.837

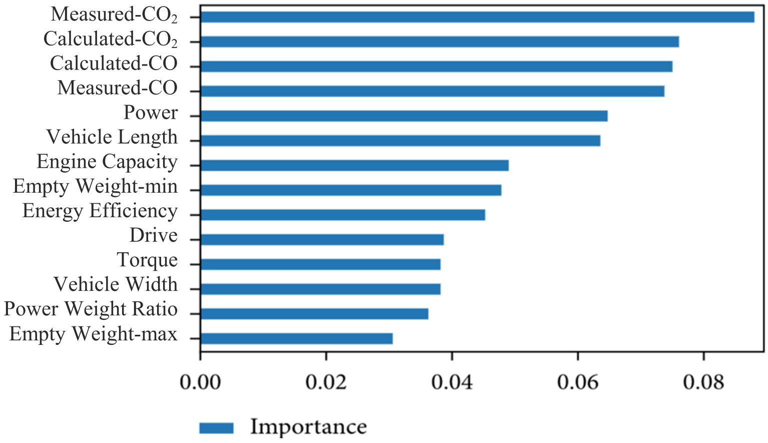


FIGURE 5. Importance of features of the proposed semi-supervised clustering algorithms.

Table 2 that, compared to the FCM-based classifier, the use of feature extraction techniques increases the classification performance. Among all tested approaches, SSFCM provides the best results in terms of both prediction accuracy (95.2%) and verification measures (90.4%) and is therefore selected as our final classification model.

The experimental results demonstrate that the SSFCM algorithm can extract richer information from the vehicle dataset and obtain more discriminative recognition rates than other classifiers do. Therefore, the proposed approach can not only effectively address the problem of multi-class im-balanced data but also improve the prediction performance.

C. DISCUSSIONS

Using the SSFCM model, we estimate the average CO_2 emissions of all new passenger vehicles registered in Switzerland in 2018 to be 138.9 g CO_2/km, which only deviates by 1.1% from the official estimate of the Swiss Federal Office of Energy (SFOE) of 137.8 g CO_2/km [94]. Moreover, for all 26 Swiss Cantons, we find that the correlation between our estimates and those from the SFOE are very high (R2 $>$ 0.95). Thus, although slightly different approaches were used to estimate the CO_2 emissions in both cases, the results are highly correlated.

The overall level of emissions from the Swiss passenger cars is strongly affected by the fleet composition, which is shifting in time between classes (from the upper-middle class to the large and luxury classes) and within each class (from non-SUV to SUV).

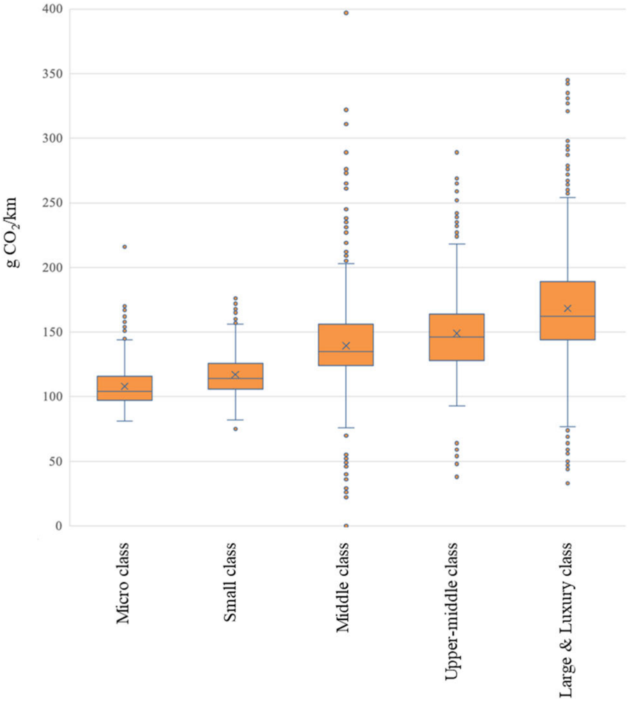


FIGURE 6. Distribution of the CO_2 emissions among the different vehicle classes.

Fig. 6 shows the distribution of the CO_2 emissions among the different vehicle classes. Both, the median CO_2 emissions and the spread around the median, show a clear upwards trend with the size of the vehicle class.

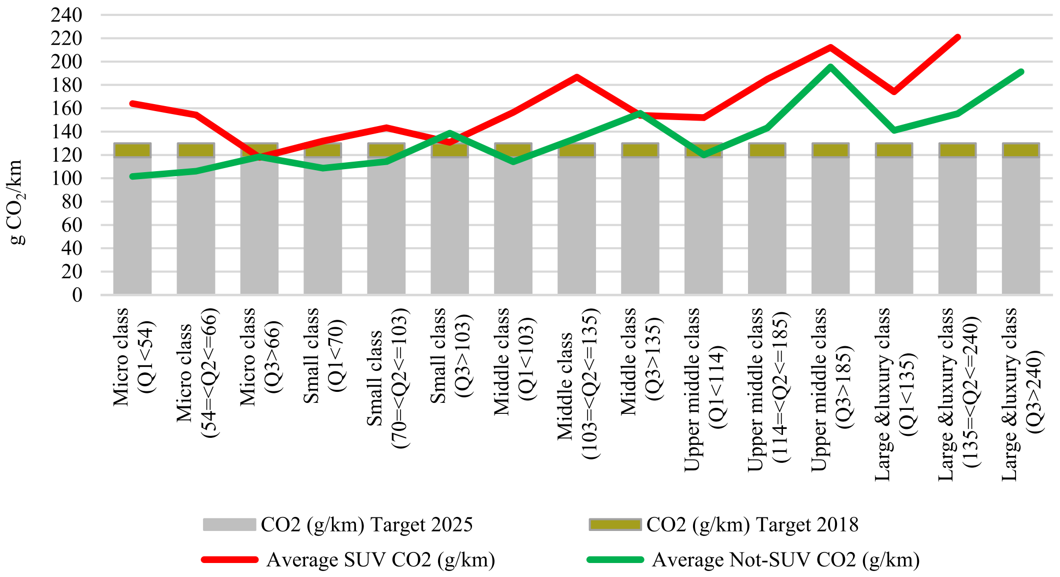


FIGURE 7. New registered passenger cars average CO_2 emissions intensity based on the interquartile range power range (Q) by both vehicle inter-class and intra-class classification in 2018.

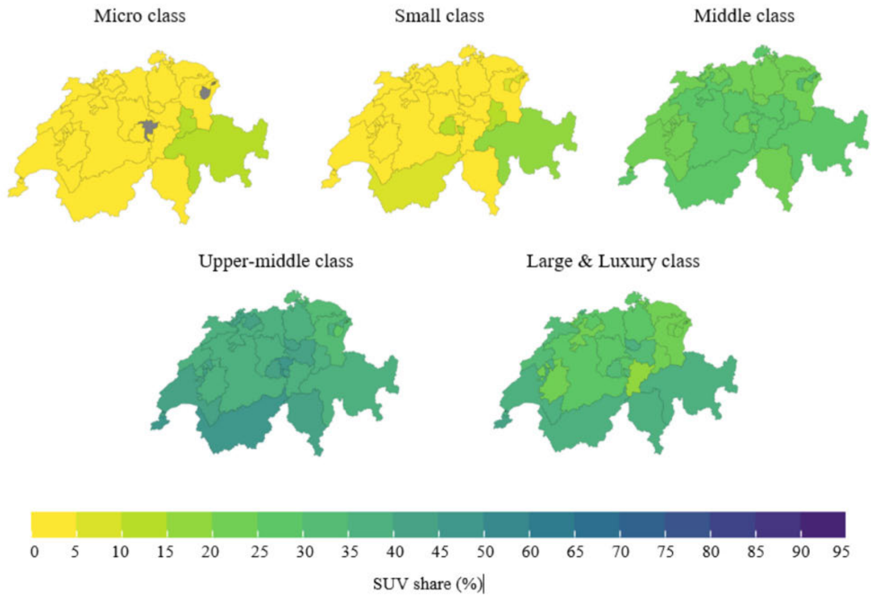


FIGURE 8. Spatial distribution of the share of SUV vehicles (in %) among different vehicle classes.

Fig. 7 shows the average CO_2 emissions (in g CO_2/km) calculated for each vehicle segment resulting from our inter- and intra-class classification. For each vehicle class we report the results based on the interquartile range distributions of the engine power. Overall, we see a significant variation of the CO_2 emissions between vehicle classes, sub-classes and power ranges. Comparing the different vehicle segments, we see that the CO_2 emissions increase with the vehicle size. Moreover, SUV vehicles tend to have significantly higher emissions than non-SUVs. In terms of engine power, it can be seen that within each class, an increase in the engine power generally leads to significantly higher CO_2 emissions.

Fig. 8 shows the spatial distribution of the share of SUVs within the different vehicle segments. It can be seen that the share of SUVs is the lowest for the micro and small classes (between 0 and 20%), followed by the middle class (between 20 and 35%) and the large and luxury class (between 20 and 45%), and is the highest for the upper-middle class (between

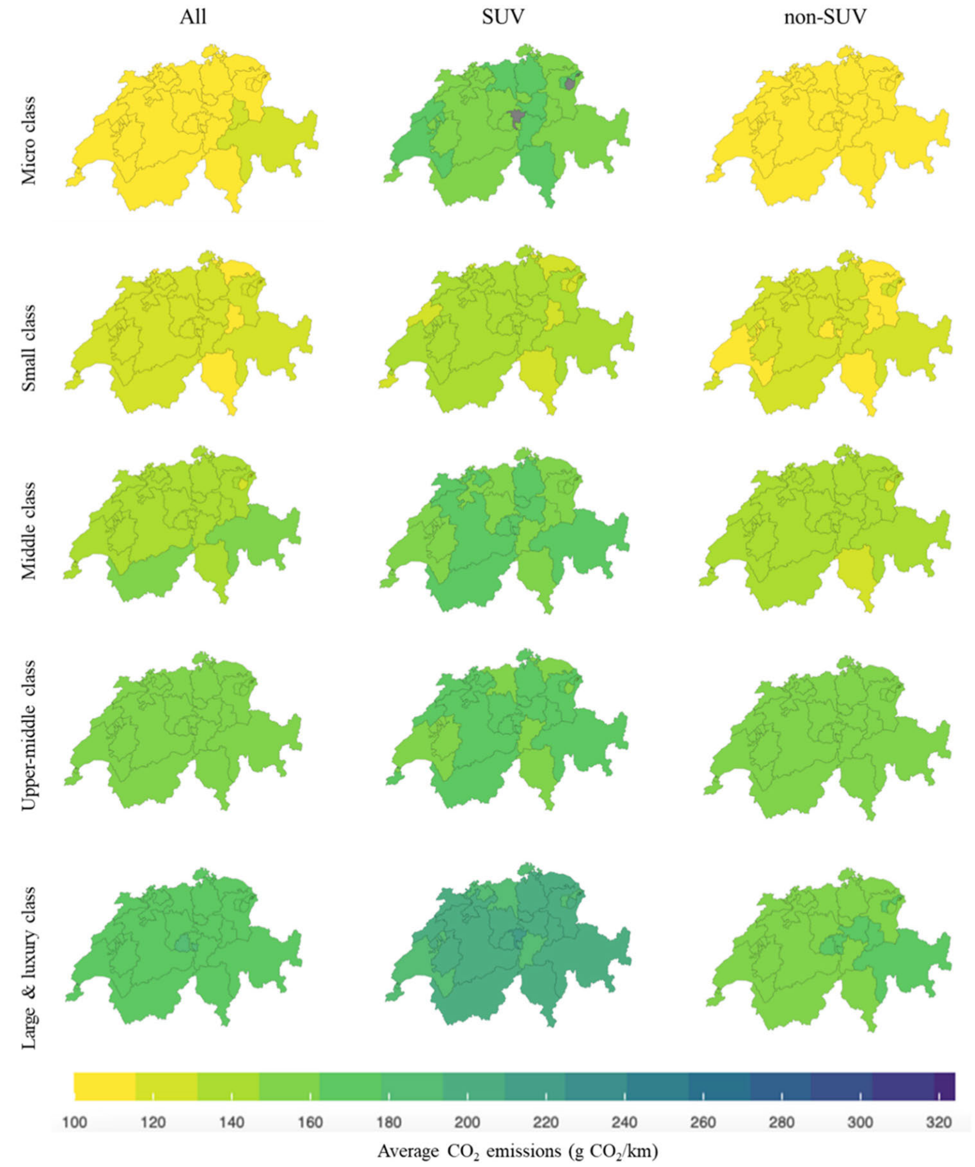


FIGURE 9. Spatial distribution of the average CO_2 emissions (in g CO_2 /km) among different vehicle classes and sub-classes.

30 and 50%). Moreover, we observe that in general the share of SUVs within each class is higher in the southern and center cantons.

Fig. 9 reports the spatial distribution of the average CO_2 emissions for each vehicle class and sub-class. As previously observed, the results demonstrate a high variability in the CO_2 emissions between the different inter and intra-classes. Among the different segments, the CO_2 emissions increase with the size of the vehicle, from around 100 g CO_2/km for the micro class to around 200 g CO_2/km for the large & luxury class. Moreover, SUV vehicles exhibit generally higher CO_2 emissions than non-SUVs. This difference is extremely remarkable in the case of micro class vehicles, although, as shown in Fig. 7, the share of SUVs is very small in this case. However, this indicates that shifting from a middle class non-SUV vehicle to a micro class SUV could lead to an increase in the CO_2 emissions. In general, the spatial distribution of the CO_2 emissions for each vehicle class and sub-class is quite homogeneous and we do not see any significant trends between different regions of the country.

VI. CONCLUSION

To This paper develops a novel approach to mathematically segment new registered passenger cars and assess the segment-based spatial distributions of the CO_2 emissions. A variety of semi-supervised clustering algorithms are adopted to classify a dataset of new registered passenger cars based on multiple technical, dimensional and emission features. Among all tested classifiers, the SSFCM technique was the most accurate, providing a classification accuracy of about 90.4% for verification measures.

The proposed approach enables accurate automated vehicle classification of large databases, which in turn facilitates the analysis of fleet changes. Another important advantage of the clustering based mathematical segmentation is that it removes the subjectivity factors affecting expert-based segmentations, reducing classification errors and making databases from across the world comparable. Finally, the automatized clustering approach also reduces classification costs and training time.

Despite technology improvements, the Swiss passenger car fleet remains emission intensive. Our results indicate large variabilities in the average CO_2 emissions of different vehicle classes. While a shift of the fleet towards smaller vehicles is likely to diminish CO_2 emissions, the emissions intensity could be more effectively reduced by shifting the vehicles proportion within each class (e.g., switching from SUV to non-SUV or to lower power vehicles in the same vehicle class). Therefore, the combination of the inter-class and intra-class classification provides crucial insights for developing fleet transformation strategies to decarbonize the passenger vehicle fleet. A further area of potentially fruitful research would be to use CO_2 estimates from real world measurements instead of type approval values for a more precise evaluation of the fleet CO_2 emissions.

ACKNOWLEDGMENT

The authors thank the Federal Roads Office (FEDRO) for providing the Swiss Vehicle Information System (MOFIS) data and the vehicle technical dataset and the Vehicle Expert Partner (Au-to-i-dat) for providing the expert segmentation data. In addition, the authors would like to express their gratitude to Philippe Zimmermann from the Automotive Powertrain Technologies Laboratory (Empa) for his support in CO_2 emissions estimation.

REFERENCES

[1] *The Paris Agreement | UNFCCC*. Accessed: Sep. 2021. [Online]. Available: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

[2] (2016). *EU Reference Scenario 2016, Energy, Transport and GHG Emissions. Trends to 2050, Euro-pean Commission*. [Online]. Available: https://ec.europa.eu/ener-gy/sites/ener/files/documents/ref2016_report_final-web.pdf

[3] P. Capros, D. Vita, A. Tasios, N. Siskos, P. Kannavou, M. Petropoulos, A. Evangelopoulou, and S. Zampara, "EU reference scenario 2016," Energy, Transp. GHG Emissions Trends 2050, Eur. Commission Directorate, General Energy, Directorate, General Climate Action Directorate, General Mobility Transp., EU Reference Scenario, Luxembourg, Europe, Tech. Rep. 20160712, 2016.

[4] *World Energy Outlook 2018*, Int. Energy Agency, Paris, France, 2018.

[5] (2021). *Federal Office for the Environment (FOEN)*. Accessed: Dec. 2021. [Online]. Available: https://www.bafu.admin.ch/bafu/en/home/topics/climate/state/data/greenhouse-gas-inventory/transport.html

[6] Federal Energy Research Commission CORE. (2020). *Energy Research Masterplan of the Federal Government 2021-2024*. [Online]. Available: https://pubdb.bfe.admin.ch/en/publication/download/10329

[7] N. Niroomand, C. Bach, and M. Elser, "Vehicle dimensions based passenger car classification using fuzzy and non-fuzzy clustering methods," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2675, no. 10, pp. 184–194, Oct. 2021, doi: 10.1177/03611981211010795.

[8] N. Niroomand, C. Bach, and M. Elser, "Robust vehicle classification based on deep features learning," *IEEE Access*, vol. 9, pp. 95675–95685, 2021, doi: 10.1109/ACCESS.2021.3094366.

[9] J. Krause, C. Thiel, D. Tsokolis, Z. Samaras, C. Rota, A. Ward, P. Prenninger, T. Coosemans, S. Neugebauer, and W. Verhoeve, "EU road vehicle energy consumption and CO_2 emissions by 2050—Expert-based scenarios," *Energy Policy*, vol. 138, Mar. 2020, Art. no. 111224.

[10] N. Zacharof, G. Fontaras, B. Ciuffo, S. Tsiakmakis, and K. Anagnostopoulos, "Review of in use factors affecting the fuel consumption and CO_2 emissions of passenger cars Euro commission," Publications Office Eur. Union, Luxembourg City, Luxembourg, Tech. Rep. EUR 27819 EN, 2016.

[11] D. Tsokolis, S. Tsiakmakis, A. Dimaratos, G. Fontaras, P. Pistikopoulos, B. Ciuffo, and Z. Samaras, "Fuel consumption and CO2 emissions of passenger cars over the new worldwide harmonized test protocol," *Appl. Energy*, vol. 179, pp. 1152–1165, Oct. 2016.

[12] J. Pavlovic, A. Marotta, B. Ciuffo, S. Serra, G. Fontaras, K. Anagnostopoulos, S. Tsiakmakis, V. Arcidiacono, S. Hausberger, and G. Silberholz, "Correction of test cycle tolerances: Evaluating the impact on CO_2 results," *Transp. Res. Proc.*, vol. 14, pp. 3099–3108, Jan. 2016.

[13] A. Dimaratos, D. Tsokolis, G. Fontaras, S. Tsiakmakis, B. Ciuffo, and Z. Samaras, "Comparative evaluation of the effect of various technologies on light-duty vehicle CO_2 emissions over NEDC and WLTP," *Transp. Res. Proc.*, vol. 14, pp. 3169–3178, Jan. 2016.

[14] *EC, A European Strategy for Low-Emission Mobility. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 2016, COM (2016) 501 Final*, document 52016DC0501, 2016.

[15] S. L. Shi, Q. Zhong, and J. M. Xu, "Robust algorithm of vehicle classification," in *Proc. Softw. Eng., Artif. Intell., Netw., Parallel/Distrib. Comput.*, 2007, pp. 269–272.

[16] W. Shi, Y. Gong, C. Ding, Z. Ma, X. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11209. Cham, Switzerland: Springer, 2015, pp. 299–315.

[17] X. Zhu. (2008). *Semi-Supervised Learning Literature Survey*. [Online]. Available: http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html

[18] L. Zhuo, L. Y. Jiang, Z. Zhu, J. Li, J. Zhang, and H. Long, "Vehicle classification for large-scale traffic surveillance videos using convolutional neural networks," *Mach. Vis. Appl.*, vol. 28, no. 7, pp. 793–802, 2016.

[19] G. Forestier and C. Wemmert, "Semi-supervised learning using multiple clusterings with limited labeled data," *Inf. Sci.*, vols. 361–362, pp. 48–65, Sep. 2016.

[20] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2006.

[21] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *J. Mach. Learn. Res.*, vol. 12, pp. 1149–1184, Jul. 2011.

[22] A. Arshad, S. Riaz, and L. Jiao, "Semi-supervised deep fuzzy c-mean clustering for imbalanced multi-class classification," *IEEE Access*, vol. 7, pp. 28100–28112, 2019, doi: 10.1109/ACCESS.2019.2901860.

[23] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

[24] Y.-Z. Ren, G.-J. Zhang, and G.-X. Yu, "Random subspace based semi-supervised feature selection," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Jul. 2011, pp. 113–118.

[25] *Swiss Federal Office of Energy (SFOE), CO_2 Emission Regulations for New Cars and Light Commercial Vehicles*. Accessed: Sep. 2021. [Online]. Available: https://www.bfe.admin.ch/bfe/en/home/efficiency/mobility/co2-emission-regulations-for-new-cars-and-light-commercial-vehicles.html

[26] *Statistik der Schweizer Städte 2020, Bundesamt für Statistik*. Accessed: Sep. 2021. [Online]. Available: https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/publikationen.assetdetail.12767482.html

[27] *Inrix Global Traffic Scorecard 2019*. Accessed: Sep. 2021. [Online]. Available: https://inrix.com/scorecard

[28] *Bundesamt für Strassen, ASTRA*. Accessed: Mar. 2019. [Online]. Available: https://www.astra amin.ch/astra/de/home.html

[29] (Nov. 2021). *Federal Office for Spatial Development (ARE): Transport Outlook 2050: Final Report*. Accessed: Dec. 2021. [Online]. Available: https://www.are.admin.ch/are/en/home/mobility/data/transport-outlook.html

[30] *Federal Department of the Environment 2020, Transport, Energy and Communication, Effects of the CO_2 Emission Regulations for New Passenger Cars 2012-2018*. Accessed: Oct. 2021. [Online]. Available: http://pubdb.bfe.admin.ch

[31] L. Opland, "Size classification of passenger's cars: Pre-study on how to size classify passengers' cars by inventorying the existing classification models," M.S. thesis, Chalmers Univ. Technol., Gothenburg, Sweden, 2007. [Online]. Available: https://hdl.handle.net/20.500.12380/44868

[32] K. Yousaf, A. Iftikhar, and A. Javed, "Comparative analysis of automatic vehicle classification techniques: A survey," *Int. J. Image, Graph. Signal Process.*, vol. 4, no. 9, pp. 52–59, Sep. 2012.

[33] H.-J. Cho and M.-T. Tseng, "A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation," *Math. Comput. Model.*, vol. 58, nos. 1–2, pp. 438–448, Jul. 2013.

[34] Y. Chen and G. Qin, "Video-based vehicle detection and classification in challenging scenarios," *Int. J. Smart Sens. Intell. Syst.*, vol. 7, no. 3, pp. 1077–1094, 2014.

[35] *Stanford Earth Matters Magazine, COVID Lockdown Causes Record Drop in Carbon Emissions for 2020*, Stanford University. Accessed: Oct. 2021. [Online]. Available: https://earth.stanford.edu/news

[36] F. Grelier, "CO_2 emissions form cars: The facts," Eur. Fed. Transp. Environ. AISBL, Brussels, Belgium, Tech. Rep., Apr. 2018.

[37] G. Fontaras, N.-G. Zacharof, and B. Ciuffo, "Fuel consumption and CO_2 emissions from passenger cars in Europe–laboratory versus real-world emissions," *Prog. Energy Combustion Sci.*, vol. 60, pp. 97–131, May 2017.

[38] Eionet (2019). *Eionet Central Data Repository*. [Online]. Available: https://cdr.eionet.europa.eu/it/eu/art17/envxuwp6g, Accessed: May 2021.

[39] R. Suarez-Bertoa, V. Valverde, M. Clairotte, J. Pavlovic, B. Giechaskiel, V. Franco, Z. Kregar, and C. Astorga, "On-road emissions of passenger cars beyond the boundary conditions of the real-driving emissions test," *Environ. Res.*, vol. 176, Sep. 2019, Art. no. 108572.

[40] J. Pavlovic, K. Anagnostopoulos, M. Clairotte, V. Arcidiacono, G. Fontaras, I. P. Rujas, V. V. Morales, and B. Ciuffo, "Dealing with the gap between type-approval and in-use light duty vehicles fuel consumption and CO_2 emissions: Present situation and future perspective," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2672, no. 2, pp. 23–32, Dec. 2018.

[41] H. Dai, P. Mischke, X. Xie, Y. Xie, and T. Masui, "Closing the gap? Top-down versus bottom-up projections of China's regional energy use and CO_2 emissions," *Appl. Energy*, vol. 162, pp. 1355–1373, Jan. 2016.

[42] S. D. Tuladhar, M. Yuan, P. Bernstein, W. D. Montgomery, and A. Smith, "A top–down bottom–up modeling approach to climate change policy analysis," *Energy Econ.*, vol. 31, pp. S223–S234, Dec. 2009.

[43] D. P. van Vuuren, M. Hoogwijk, T. Barker, K. Riahi, S. Boeters, J. Chateau, S. Scrieciu, J. van Vliet, T. Masui, K. Blok, E. Blomen, and T. Kram, "Comparison of top-down and bottom-up estimates of sectoral and regional greenhouse gas emission reduction potentials," *Energy Policy*, vol. 37, no. 12, pp. 5125–5139, Dec. 2009.

[44] N. Karali, T. Xu, and J. Sathaye, "Reducing energy consumption and CO_2 emissions by energy efficiency measures and international trading: A bottom-up modeling for the U.S. Iron and steel sector," *Appl. Energy*, vol. 120, pp. 133–146, May 2014.

[45] P. Thunis, B. Degraeuwe, K. Cuvelier, M. Guevara, L. Tarrason, and A. Clappier, "A novel approach to screen and compare emission inventories," *Air Qual., Atmos. Health*, vol. 9, no. 4, pp. 325–333, May 2016.

[46] J. Seo, J. Park, Y. Oh, and S. Park, "Estimation of total transport CO_2 emissions generated by Medium- and heavy-duty vehicles (MHDVs) in a sector of Korea," *Energies*, vol. 9, no. 8, p. 638, Aug. 2016.

[47] J. L. Jiménez, J. Valido, and N. Molden, "The drivers behind differences between official and actual vehicle efficiency and CO_2 emissions," *Transp. Res. D, Transp. Environ.*, vol. 67, pp. 628–641, Feb. 2019.

[48] L. Ntziachristos, G. Mellios, D. Tsokolis, M. Keller, S. Hausberger, N. E. Ligterink, and P. Dilara, "In-use vs. Type-approval fuel consumption of current passenger cars in Europe," *Energy Policy*, vol. 67, pp. 403–411, Apr. 2014.

[49] F. D. L. Torre and T. Kanade, "Discriminative cluster analysis," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 241–248.

[50] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1033–1047, Jul. 2010.

[51] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 15, 2011, pp. 215–223.

[52] Y.-Z. Ren, G.-J. Zhang, and G.-X. Yu, "Random subspace based semi-supervised feature selection," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Jul. 2011, pp. 113–118.

[53] Y. Ren, G. Zhang, G. Yu, and X. Li, "Local and global structure preserving based feature selection," *Neurocomputing*, vol. 89, pp. 147–157, Jul. 2012.

[54] L. Li, J. M. Garibaldi, D. He, and M. Wang, "Semi-supervised fuzzy clustering with feature discrimination," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0131160.

[55] Q. Li, F. Qiao, and L. Yu, "A machine learning approach for light-duty vehicle idling emission estimation based on real driving and environmental information," *Environ. Pollut. Climate Change*, vol. 1, no. 106, p. 2, 2016, doi: 10.4172/2573-458X.1000106.

[56] Z. He, G. Ye, H. Jiang, and Y. Fu, "Vehicle emission detection in data-driven methods," *Math. Problems Eng.*, vol. 2020, pp. 1–13, Oct. 2020, doi: 10.1155/2020/4875310.

[57] C. Saleh, N. R. Dzakiyullah, and J. B. Nugroho, "Carbon dioxide emission prediction using support vector machine," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 114, Feb. 2016, Art. no. 012148, doi: 10.1088/1757-899X/114/1/012148.

[58] M. Ghahramani and F. Pilla, "Analysis of carbon dioxide emissions from road transport using taxi trips," *IEEE Access*, vol. 9, pp. 98573–98580, 2021, doi: 10.1109/ACCESS.2021.3096279.

[59] G. Padmapriya and K. Duraiswamy, "Association of deep learning algorithm with fuzzy logic for multi document text summarization," *J. Theor. Appl. Inf. Technol.*, vol. 62, no. 1, pp. 166–173, 2014.

[60] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

[61] A. Arshad, S. Riaz, L. Jiao, and A. Murthy, "A semi-supervised deep fuzzy C-mean clustering for two classes classification," in *Proc. IEEE 3rd Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Oct. 2017, pp. 365–370.

[62] A. Arshad, S. Riaz, L. Jiao, and A. Murthy, "Semi-supervised deep fuzzy c-mean clustering for software fault prediction," *IEEE Access*, vol. 6, pp. 25675–25685, 2018.

[63] A. Arshad, S. Riaz, L. Jiao, and A. Murthy, "The empirical study of semi-supervised deep fuzzy C-mean clustering for software fault prediction," *IEEE Access*, vol. 6, pp. 47047–47061, 2018.

[64] L. Jiang, J. Li, L. Zhuo, and Z. Zhu, "Robust vehicle classification based on the combination of deep features and handcrafted features," in *Proc. IEEE Trustcom/BigDataSE/ICESS*, Aug. 2017, pp. 859–865.

[65] W. Balid, H. Tafish, and H. H. Refai, "Intelligent vehicle counting and classification sensor for real-time traffic surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 6, pp. 1784–1794, Jun. 2018.

[66] W. Maungmai and C. Nuthong, "Vehicle classification with deep learning," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Singapore, Feb. 2019, pp. 294–298.

[67] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2247–2256, Mar. 2015.

[68] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. 21t Int. Conf. Mach. Learn. (CML)*, 2004, pp. 81–88.

[69] Y. Ren, X. Hu, K. Shi, G. Yu, D. Yao, and Z. Xu, "Semi-supervised den peak clustering with pairwise constraints," in *Proc. 15th Pacific Rim Int. Conf. Artif. Intell.*, 2018, pp. 837–850.

[70] N. Grira, M. Crucianu, and N. Boujemaa, "Active semi-supervised fuzzy clustering," *Pattern Recognit.*, vol. 41, no. 5, pp. 1834–1844, May 2008.

[71] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: A brief survey," in *Proc. Rev. Mach. Learn. Techn. Process. Multimedia Content*, 2004, pp. 9–16.

[72] W. Qiu, "Based on similarity metric learning for semi-supervised clustering," *Sensors Transducers J.*, vol. 177, no. 8, pp. 238–245, 2014.

[73] Y. Qin, S. Ding, L. Wang, and Y. Wang, "Research progress on semi-supervised clustering," *Cognit. Comput.*, vol. 11, no. 5, pp. 599–612, Oct. 2019.

[74] Y. Re, X. Hu, K. Shi, G. Yu, D. Yao, and Z. Xu, "Semi-supervised denpeak clustering with pairwise constraints," in *Proc. 15th Pacific Rim Int. Conf. Artif. Intell.*, 2019, pp. 837–850.

[75] A. Arshad, S. Riaz, L. Jiao, and A. Murthy, "Semi-supervised deep fuzzy C-mean clustering for software fault prediction," *IEEE Access*, vol. 6, pp. 25675–25685, 2018.

[76] W. Shi, Y. Gong, C. Ding, Z. M. X. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 299–315.

[77] G. Chen, "Deep transductive semi-supervised maximum margin clustering," 2015, *arXiv:1501.06237*.

[78] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1119–1130, Aug. 2012.

[79] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, Feb. 2011.

[80] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, Jan. 1973.

[81] A. Arshad, S. Riaz, and L. Jiao, "Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification," *IEEE Access*, vol. 7, pp. 28100–28112, 2019, doi: 10.1109/ACCESS.2019.2901860.

[82] S. Riaz, A. Arshad, and L. Jiao, "A semi-supervised CNN with fuzzy rough C-mean for image classification," *IEEE Access*, vol. 7, pp. 49641–49652, 2019, doi: 10.1109/ACCESS.2019.2910406.

[83] Y. Ren, X. Hu, K. Shi, G. Yu, D. Yao, and Z. Xu, "Semi-supervised denpeak clustering with pairwise constraints," in *Proc. 15th Pacific Rim Int. Conf. Artif. Intell.*, 2018, pp. 837–850.

[84] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.

[85] T. M. Khoshgoftaar, C. Seiffert, J. V. Hulse, A. Napolitano, and A. Folleco, "Learning with limited minority class data," in *Proc. 6th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2007, pp. 348–353.

[86] X. Zhu. (2008). *Semi-Supervised Learning Literature Survey*. [Online]. Available: http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html

[87] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, "Investigating random undersampling and feature selection on bioinformatics big data," in *Proc. IEEE 5th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Apr. 2019, pp. 346–356, doi: 10.1109/BigDataService.2019.00063.

[88] B. Handaga and M. M. Deris, "Similarity approach on fuzzy soft set based numerical data classification," in *Software Engineering and Computer Systems* (Communications in Computer and Information Science), vol. 180, J. M. Zain, W. M. W. Mohd, E. E. Qawasmeh, Eds. Berlin, Germany: Springer, 2011, doi: 10.1007/978-3-642-22191-0_50.

[89] S. W. Kwok and C. Carter, "Multiple decision trees," *Uncertainty Artif. Intell.*, vol. 9, pp. 327–335, Jan. 1990.

[90] R. Kumar and R. Verma, "Classification algorithms for data mining: A survey," *Int. J. Innov. Eng. Technol.*, vol. 1, no. 2, pp. 7–14, 2012.

[91] C. Macdonald and I. Ounis, "Voting for candidates: Adapting data fusion techniques for an expert search task," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Arlington, VA, USA, Nov. 2006, pp. 387–396.

[92] *Das Motorfahrzeuginformationssystem der Eidgenössischen Fahrzeugkontrolle, MOFIS*. Accessed: Mar. 2019. [Online]. Available: https://www.experience-online.ch/de/9-case-study/2023-mofis

[93] *Schweizer Partner für Fahrzeugdaten*. Accessed: Mar. 2020. [Online]. Available: https://www.auto-i-dat.ch

[94] E. Limarzo, S. Dickenmann, and C. Schreyer. *Auswirkungen der CO_2-Emissionsvorschriften für Neue Personenwagen 2012-2018*. Accessed: Aug. 2021. [Online]. Available: https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-78134.html

NAGHMEH NIROOMAND received the M.B.A. degree from Eastern Mediterranean University, Cyprus, in 2010, the I.A.P.M. degree from Queen's University, Canada, in 2014, the first Ph.D. degree from Eastern Mediterranean University, in 2016, and the second Ph.D. degree from SSPH+, Switzerland, in 2018. She worked as a Visiting Fellow with the Transport and Mobility Laboratory, EPFL, Lausanne, from 2018 to 2019; and a Senior Scientist at the Swiss Federal Laboratories of Material Science and Technology (Empa), Switzerland, from 2019 to 2021. She is currently a Techno-Energy Economist with the Automotive Powertrain Technologies Laboratory, Empa. Prior to joining Empa, she was an Associate Research Economist at Cambridge Resources International, USA. Her current research interests include vehicle fleet and operational analysis, retro-perspective analyze vehicle specific changes in function of spatial technology and economic frame conditions, and economies of synthetic energy carriers.

CHRISTIAN BACH received the B.Sc. degree in automotive engineering from the Bern University of Applied Sciences. He performed two internships at the Haagen-Smit Laboratory, California Air Resources Board, El Monte, USA, to study zero and ultra low emission technologies in the transport sector. He is currently the Head of the Automotive Powertrain Technologies Laboratory, Swiss Federal Laboratories of Material Science and Technology (Empa). He is also a Lecturer at ETH Zürich, and a member of several expert groups in Switzerland.

MIRIAM ELSER received the B.Sc. and M.Sc. degrees in physics from the University of Milan, Italy, in 2010 and 2012, respectively, and the Ph.D. degree in sciences from ETH Zürich, Switzerland, in 2016. She has worked as a Postdoctoral Researcher, from 2016 to 2018, and a Senior Scientist, from 2019 to 2021, at the Swiss Federal Laboratories of Material Science and Technology (Empa), Switzerland. She leads the Vehicle Systems Group, Automotive Powertrain Technologies Laboratory, Empa. Her current research interests include vehicle fleet and operational analysis, measurement, and modeling of vehicular emissions, and real-world testing of sensors for automated vehicles.