IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 11, 2021, accepted June 25, 2021, date of publication July 2, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094366

Robust Vehicle Classification Based on Deep Features Learning

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

ABSTRACT This paper aims to introduce a scientific Semi-Supervised Fuzzy C-Mean (SSFCM) clustering approach for passenger cars classification based on the feature learning technique. The proposed method is able to classify passenger vehicles in the micro, small, middle, upper middle, large and luxury classes. The performance of the algorithm is analyzed and compared with an unsupervised fuzzy C-means (FCM) clustering algorithm and Swiss expert classification dataset. Experiment results demonstrate that the classification of SSFCM algorithm has better correlation with expert classification than traditional unsupervised algorithm. These results exhibit that SSFCM can reduce the sensitivity of FCM to the initial cluster centroids with the help of labeled instances. Furthermore, SSFCM results in improved classification performance by using the resampling technique to deal with the multi-class imbalanced problem and eliminate the irrelevant and redundant features.

INDEX TERMS Vehicle classification, fuzzy C-means clustering, semi-supervised learning, feature learning.

I. INTRODUCTION

During the last decades there have been major changes in passenger vehicle sizes around the world. One of the main functionalities of these changes is global competition, which forces car manufacturers worldwide to quickly improve and introduce competing vehicles. The implication is significantly changes in the vehicle segments (increased share of SUVs), vehicle dimensions (increased size of the vehicles within each segment), and other design parameters like powertrain type and power. To date, numerous vision-based vehicle classification methods are available. However, illumination changes, shadows, partial detections, occlusion, and camera viewpoint changes have strong impact on these techniques [1]–[6]. This complexity further increases for vehicles that belong to different classes despite having similar dimensions or have visually similar appearance but dissimilar dimensions [7]. Furthermore, the review of related literatures [8]–[11] shows that the growth of car models over time poses an additional challenge to the accurate vehicle classification based on dimensions. In fact, the most popular names of car models have remained intact since they were developed, making it difficult to follow the evolution of sizes during vehicle classification (Table 1). In this setting, Fig. 1 illustrates the core challenge of vision-based vehicle classification which is high intra-class (within the cluster of classes) variation and relatively low inter-class (between classes of multi-class classification) variation [12]–[14].

This issue reveals the need of using automated and efficient classification techniques for different vehicle types for a variety of applications. For a vehicle classification system to be useful in real-world conditions, it must be robust to inter-class interferences in the data by considering new measurement method based on the discriminative features.

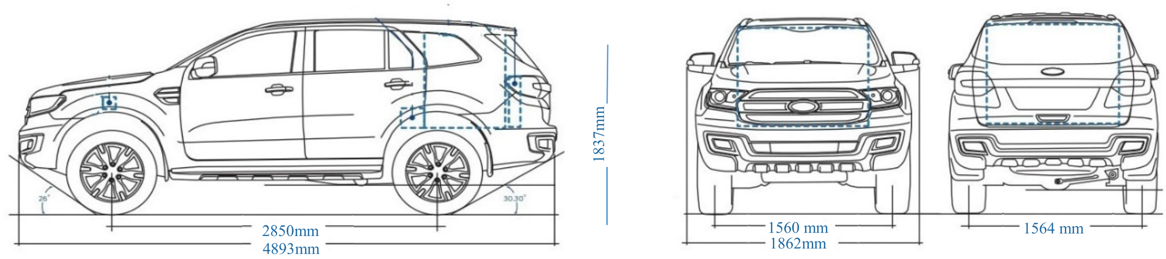
One approach that has received much attention in recent years is semi-supervised learning (SSL) techniques. SSL is a branch of machine learning that aims to classify large size of the unlabeled data using the minimal size of labeled information set to build better learning [15]–[19]. SSL are particularly relevant to scenarios where the input labeled set being able to utilize the classifier during the training process and then applied out-of-sample approaches to make predictions on unlabeled data. In recent years, SSL approaches attempt to improve the performance in sequential learning of supervised and unsupervised learning by utilizing information generally associated with the other and repeating the

The associate editor coordinating the review of this manuscript and approving it for publication was Shun-Feng Su.

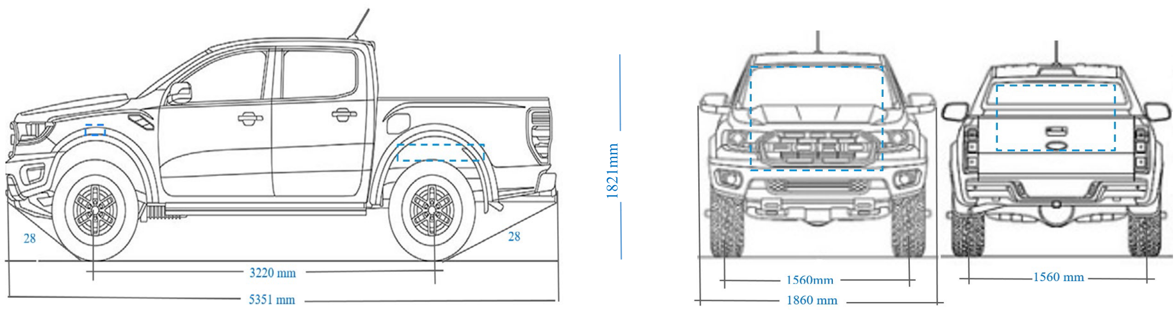
TABLE 1. Sample of passenger car growth.

	Car Models					
Percentage of growth	*VW Polo I* *VW Polo VI*	*VW Golf I* *VW Golf VIII*	*VW Passat B1* *VW Passat B8*	*BMW E21* *BMW G21*	*Mercedes-Benz W 124* *Mercedes-Benz W 213*	*Opel Corsa A* *Opel Corsa F*
Length (cm)	+12.50	+12.63	+11.16	+8.03	+8.37	+2.53
Width (cm)	+11.46	+10.49	+13.66	+13.66	+9.77	+14.29
Height (cm)	+8.96	+5.67	+7.09	+5.07	+4.90	+5.11

Source: https://de.wikipedia.org (Accessed: May 2021)



(a) Ford Everest 2019



(b) Ford Ranger 2019

FIGURE 1. Demonstration of two major challenges in vision-based vehicle classification: intra-class variation and inter-class similarity (Source: https://www.automobiledimension.com/ (Accessed: May 2021)).

process until convergence [20]–[22]. SSL algorithms can be classified into the Supervised Deep Learning using Pseudo Labels [23], Spatio-spectral LapSVM [24], Semi-Supervised Ladder Network [25], Temporal Ensembling with augmentation [26], Mean Teacher, weight-averaged consisting targets [27] and semi-supervised fuzzy rough convolution neural network (FRCNN) [28]. Most of these algorithms often consider a balanced training set which may increase the overlapping between majority and minority class instances.

This paper is an extension of previous work originally focused on the preprocessing data technique by feature extraction to classify data into the five different categories as part of a multi-class vehicle classification system based on both technical and geometric parameters using fuzzy and non-fuzzy clustering methods [7]. However, the classification performance metrics was found to be slightly affected with the imbalanced problem. The problem of imbalanced dataset appears in most traditional classification methods when the proportion of majority class instances across the known classes is biased or skewed, i.e. the number of instances in a majority class is much higher than the number of instances in a minority class [29]–[31].

In recent literatures many different types of algorithms and techniques are proposed to deal with the problem of class imbalance datasets, such as random sampling, learning algorithms, and feature selection [32]–[38]. Inspired by these methods, we develop a new semi supervised approach which utilizes both supervised and unsupervised data to alleviate the multi-class imbalanced problem, eliminating the redundant features and also exploiting hidden information during the clustering process.

In this paper, we propose a novel Semi-Supervised Fuzzy C-Mean (SSFCM) clustering approach based on the feature learning technique, which is a highly useful technique for representation learning with high dimensional datasets containing high-level of uncertainties. In our approach, Fuzzy

C-Mean (FCM) clustering use to predict labels for unsupervised data. For classification, we suggest that one can use unsupervised and supervised data along with the predicted labels to extract the discriminative information for classification. Following this, in order to enhance the prediction ability, we combine feature reduction techniques (feature extraction technique and feature selection technique using Random oversampling (ROS)) to handle the multi-class imbalanced problem and eliminate the irrelevant and redundant features for improving the classification accuracy. To evaluate the performance of these algorithms we use both technical and geometric parameters, and apply them to a vehicle fleet dataset that includes all new registered passenger cars in Switzerland.

This work focuses mainly on the performance of vehicle multi-class classification and vehicle intra-class classification. The aim of our investigation is to accurately analyze vehicle changes enabling automated vehicle classification of large databases. The experimental results demonstrate that the performance metrics of proposed SSFCM approach are on average better than those of other tested methods. Furthermore, the potential of proposed approach with ROS in improving the imbalance multi-class classification performance on vehicle dataset increases the quality of clustering results over FCM.

From the experimental validation, our model can achieve acceptable accuracy (∼86%) using only 10% of the labels rate for each class. All the experiments demonstrate that the proposed method is feasible and effective in categorizing vehicles based on their features.

The rest of the paper is structure as follows. Section 2 briefly summarizes the main results of related research works. In section 3 we present a detailed description concise details of the datasets, algorithms and experiments. Lastly, in section 4 we present the majors findings of this work and provide recommendations for further research.

II. RELATED WORK

A large number of works have applied machine learning techniques for the classification of vehicles, including unsupervised (clustering) and supervised (classification) methods [16], [23], [24], [26], [27], [39]–[67]. Among the unsupervised techniques, FCM clustering algorithm is the most popular. Javadi *et al.* [42] applied FCM with dimension and speed features to classify vehicles into ''private car'', ''light trailer'', ''lorry or bus'' and ''heavy trailer'', reaching an accuracy of 96.5% on a dataset composed of 400 vehicle images. Yao *et al.* [43] developed an axle-based vehicle detection and classification method using FCM to identify and segment vehicle axle pixels from camera images. With this vision-based approach, they reached a vehicle detection rate of 62.8%. Saraçoğlu and Nemati [44] used FCM for image segmentation based on the vehicle dimensions combined with Support Vector Machine classification to classify the vehicles as ''small vehicle'', ''big vehicle'' or ''others''. Researchers have also carried out multiple comparisons of FCM with other methods, which demonstrated that the FCM algorithm is capable to overcome some of the problems faced with noise sensitivity defect and non-linear data clustering [45]–[47]. Velmurugun and Santhanam [48] compared the clustering performance and effectiveness of the K-means (KM), K-Medoids, and FCM clustering algorithms using different shapes to cluster arbitrarily distributed data and finding mutual exclusion clusters. Moreover, Joyti and Kumar [49] compared the performance of KM and FCM algorithms in terms of computational time, while Gosh and Dubey [50] computed the performance and clustering accuracy of these two methods based on the efficiency of the clustering output as well as the computational time.

In contrast, other works have applied supervised methods for the classification of vehicles. Zhang *et al.* [51] developed a length-based method for vehicle detection and classification, reaching an accuracy of 97% for truck classification. Arunkumar *et al.* [52] used a neural network classifier based on geometrical features and appearance-based attributes to classify passenger vehicles into brands. Moussa [53] also used geometric-based and appearance-based features for multi-class (''small'', ''medium'', and ''large size'') and intra-class (''pickup'', ''sport utility vehicle'', and ''van'') vehicle classification using a support-vector network mod-el. Lastly, Cheung *et al.* [41] presented a vehicle classification method based on measurements with magnetic sensors to classify the vehicles into six types (''passenger vehicle'', ''SUV'', ''van'', ''bus'', ''mini-trucks'', and ''truck''). Their algorithm achieved an accuracy of 80% to 90% when vehicle length was used as a feature, compared to only 60% when the length was not considered.

These works highlight the potential of using machine learning techniques for feature-based vehicle classification. However, in most cases they separate vehicles in classes that differ greatly in appearance and size (e.g. mopeds, passenger cars, vans trucks, buses), while separating more alike subclasses within these categories (e.g. the passenger car subclasses) poses a greater challenge, especially for vision-based methods, and requires expert knowledge.

Lately, feature learning techniques have showed outstanding performance for addressing uncertainty problem for clustering and classification [24], [54]–[59]. The classification performance highly depends on the quality of features generated from the data as input to the classifier process. However, only a limited number of studies have been done on combine feature learning techniques to improve classification performance on the high dimensional and multi-class imbalanced datasets. The classification method proposed in this paper is a new semi-supervised clustering scheme SSFCM that incorporates semi-supervised information in FCM algorithm to considerably improve its effectiveness [22], [16], [23], [26], [27], [60]–[63]. More details about the feature learning techniques can be found in the article by Jiang *et al.* [64], in which they combined several feature extraction methods with a support vector machine classifier to group the vehicles in six categories, namely ''large bus'', ''passenger car'', ''motorcycle'', ''minibus'', ''truck'' and ''van''. This study

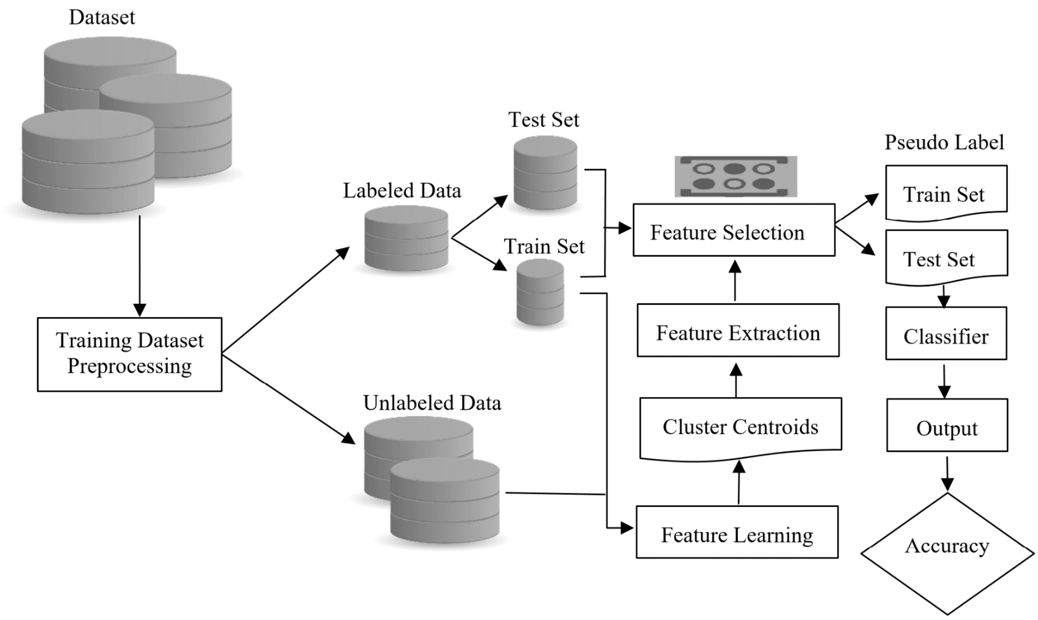


FIGURE 2. **The structure of the proposed semi-supervised deep learning process.**

achieved a classification accuracy of 97.4%. Balid *et al.* [65] implemented a deep learning-based classification model that uses the vehicle length as a key feature and classifies vehicles into ''passenger vehicles'', ''single unit trucks'', ''combination trucks'', and ''multi-trailer trucks'' with a classification accuracy of 97%. Maungmai and Nuthong [66] used a convolutional neural network method to classify the vehicle type as ''small'', ''medium'', ''large'', and ''unknown'', and vehicle color as ''black'', ''blue'', ''white'', ''green'', ''yellow'', ''red'', and ''unknown''. The convolutional neural network classifier achieved an accuracy of 84.65% for the vehicle type and 75.44% for vehicle color classification. The results comparison with utilized decision trees, random forest, and densely deep neural network classifier show that the classification of vehicle type accuracy increased by 1.8%, and vehicle color increased by 0.8%. Dong *et al.* [67] proposed a vehicle type classification method using a semi-supervised convolutional neural network from high-resolution vehicle frontal view images. The algorithm achieved 88.11% accuracy.

III. MATERIALS AND METHODS

A. SEMI-SUPERVISED CLUSTERING

Semi-supervised learning falls somewhere between unsupervised learning and supervised learning, which combines both labeled and unlabeled data. It solves classification problems, such as the imbalance of data categories, by using supervised learning algorithm that helps boost the classification accuracy. There are three main types of semi-supervised learning: semi-supervised classification [56], [68]–[71], semi-supervised dimension reduction [24], [72] and semi-supervised clustering [25], [73]–[75].

Semi-supervised classification focuses on minimizing the square loss of predicted errors in the labeled examples, while part of instances is unlabeled; semi-supervised dimensionality reduction attempts to reduce the variance of classifiers in high dimensional data or find the low-dimensional space; semi-supervised clustering aims to increase the accuracy of the defined clusters by identifying better clusters than the ones obtained from unlabeled data [19], [75]–[79]. Typically, semi-supervised clustering methods obtains worse representation results in the original feature space. To make the semi-supervised clustering more efficient, it is reasonable to do semi-supervised clustering with deep feature learning [62], [63], [16], [80]. The framework of the proposed clustering approach is depicted in Fig. 2.

Unlike the most widely used approaches, in SSFCM clustering based feature extraction technique, we consider the three types of information (diffusion label, extract core data, and ex-tract feature vectors) in order to improving accuracy of classification and decrease class imbalance and multi-class overlapping problems. This framework includes two main layers. The first layer contains the labeled (supervised) dataset which is split into train set to create a classifier based on the core dataset and test set to evaluate its output. The second layer contains the combination of recordings from train set not used for creating the classifier along with the unlabeled dataset and provides the input for feature learning process. The cluster centroids that are evaluated in feature learning process are used as an input for extracting feature vectors. Then feature selection process is implemented to eliminate irrelevant and redundant features. Lastly, a SSFCM model is constructed based on the selected feature vectors of the train set and is validated using predicted labels of test set.

B. SEMI-SUPERVISED FUZZY C-MEAN CLUSTERING

Fuzzy C-means (FCM), as an overlapping clustering algorithm, is one of the most popular fuzzy clustering meth-

Algorithm 1 Fuzzy C-Means Membership and Centroid

Input: Data X whose number of elements N, A, C, m, max. iteration number (T), error threshold (ε)

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.

ods [81]. This technique is a soft clustering algorithm. By this we mean that each data point has a probability of belonging to each cluster with partial membership values ranged from 0 to 1. However, due to the non-convexity of its objective function, it may fall into local optimal solution during optimization. To address this issue, we propose semi supervised fuzzy C-means clustering (SSFCM) algorithm that incorporates deep feature learning in FCM to further improve its effectiveness and eliminate redundant information [22], [28], [75].

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

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

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

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

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

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

where N is number of data elements, C is the number of clusters; X_k represents the data k of $X = \{X_1, X_2, X_3, \ldots, X_N\}$ 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; A is a positive and symmetric (n × n) weight matrix; U is the membership degree of data elements 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 algorithm comprises of the following steps:

After calculating deep FCM membership degrees and centroids using the algorithm 2, we select the features ($s \subset K$) using random oversampling (ROS) and Euclidean distance metric techniques. The purpose of ROS approach is to maintain a balance between the features subsets of labeled classes and unlabeled data elements [22], [38], [82]. The Euclidean distances is the most applied (dis)similarity or distance metric

Algorithm 2 The Training Strategies for Semi-Supervised Fuzzy C-Means

Input: N data elements $X = \{X_1, X_2, \ldots, X_N\}, C, K$, labeled dataset ($L$), unlabeled dataset ($UN$), membership degree ($U$), max. iteration number (T), error threshold (ε)

Output: u_{iL}, u_{UNL}, v_{iL}, v_{UNL}

Set t = 0

1. Initialize v_{iL}, v_{UL}
2. Update t = t + 1
3. Compute u_{iL}, u_{UNL}
4. Compute v_{iL}, v_{UNL}
5. If $t > T$ 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.

Algorithm 3 Semi Supervised Fuzzy C-Means Classifier

Input: N data elements $X = \{X_1, X_2, \ldots, X_N\}$ 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 $Q = \emptyset$

1. For $i \in \{1, \ldots, c\}$ do
2. For $j \in \{1, \ldots, N\}$ do
3. Employ V_{iL}^{s} to calculate *max* Sim_i
4. If maximum average of *max* $Sim_i \in i^{th}$ labeled class, then
5. Append X_j to i^{th} labeled class
6. Update Q if a labeled class is achieved
7. For all $V_{iL}^{s} \in V_{L}^{s}$ do
8. Return Q

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 the each selected feature of unlabeled data and labeled classes [17].

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

C. PERFORMANCE MEASURE

Two methods, micro-averaging and macro-averaging, are used to evaluate the accuracy and performance of classification [22], [83]–[85]. In selecting the best approach for evaluating the performance of a given classifier is important to consider the class-balance and expected outcomes. A macro-average tend to estimate the metric for each class separately and then compute the overall average. In contrast, the micro-average aggregates the contributions of all classes to estimate the average metric. One specific performance evaluation allows assessing a classifier based on the particular perspective and frequently fails to measure others.

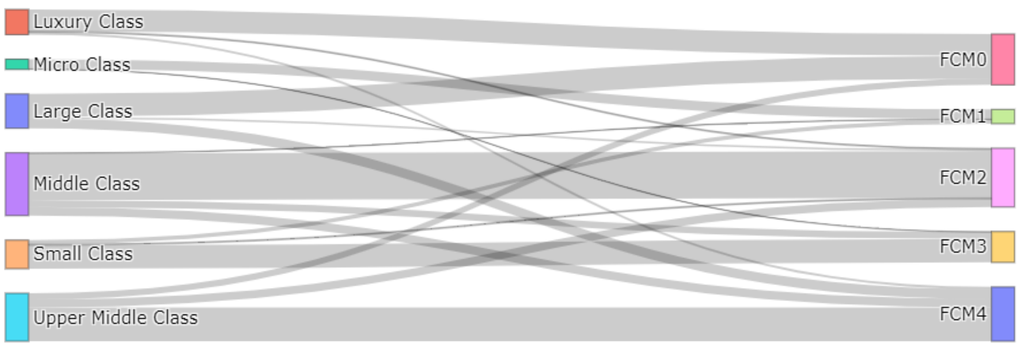


FIGURE 3. Sankey diagram of the multi-class imbalanced datasets. The colors indicate the vehicle classification by unsupervised Fuzzy C-means algorithm (right) and Swiss expert (left).

Therefore, in this study the macro-averaged accuracy (MAcc) and micro-averaged F-Measure (MFM) are used for the validation of our approach.

The MAcc can be defined as:

$$Pr = \frac{TP}{TP + FP} \tag{7}$$

$$MAcc = \frac{\sum_{i=1}^{n} (Pr)_i}{c} \quad (1 \leq i \leq c) \tag{8}$$

Precision and recall are defined as follows:

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (1 \leq i \leq c) \tag{9}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (1 \leq i \leq c) \tag{10}$$

The MFM is then computed as:

$$F-Measure = \frac{2P_iR_i}{P_i + R_i} \tag{11}$$

$$MFM = \frac{\sum_{i=1}^{c} (FM)_i}{c} \tag{12}$$

Here, Pr denotes precision in macro-average method; TP_i (True Positives) is the proportion of the data points classified correctly to class i; FP_i (False Positives) is the proportion of the data points that do not belong to class i but are classified to class i incorrectly; FN_i (False Negatives) is proportion of the data points that are not classified to class i but which actually belong to class i; n is the number of data points; and c is total number of classes.

IV. EXPERIMENTS

A. DATA PREPARATION

In experiments, the vehicle database containing annually new registered passenger cars, vehicle technical specifications and vehicle expert classification data obtained from Swiss Motor Vehicle Information System [86], Federal Office Technical Information [87] and Vehicles Expert Partner [88] respectively. In the preprocessing step we filter the database in order to extract vehicles that do not meet the definition of typical passenger cars, such as small pickup trucks, standard pickup trucks, vans, special purpose vehicles (SPVs), sports cars and multi-purpose vehicles (MPVs) [7]. By considering the goal of this paper, the dataset is separated into two parts, training part and testing part. Training dataset contains 275,601 passenger cars registered in Switzerland in 2018 along with 22 parameters including geometric parameters such as height, length, width, wheelbase, angles of the vehicle front and rear, etc., and technical parameters such as power, engine capacity, energy efficiency, drivetrain, etc.

In the first step of learning process the training dataset is considered to contain two types of patterns: unsupervised (unlabeled) and supervised (labeled) dataset. For this purpose, we used labeled dataset from the traditional unsupervised algorithm, where the FCM clustering algorithm is used in determining passenger cars multi-class classification based on its features [7]. Fig. 3 demonstrates the benchmark of Swiss vehicle datasets in 2018 that shows brief properties of multi-class imbalanced datasets. The total 366 unique samples is grouped into six classes: micro class containing 18 samples, small class containing 50 samples, middle class containing 110 samples, upper middle class containing 84 samples, large class containing 60 samples and luxury class containing 44 samples, where the average imbalance ratio is 3. The average accuracy rate of unsupervised FCM clustering algorithm in compare to the Swiss expert classification was approximately 79% where the optimal number of unsupervised FCM clusters is equal to five [7].

Due to the some limitations of the unsupervised FCM clustering algorithm, the accuracy of features based clustering is not very high. The main reason is that the feature of misclassified samples has a low degree of category membership. Therefore, in this study we used the labeled data with a membership degree greater than 0.95 as the core dataset to extract the accurate classification of misclassified samples and provide the basic for the later step of training. The labeled data is divided into train set and test set. In feature learning process, the unlabeled data and the previous labeled along with the labels of the core dataset are used as input. Prior to model development, new features are extracted to reduce the number of features. In feature extraction step, the cluster centroids learnt by using algorithm 2 and each patch is transformed to a feature vector. In feature selection step,

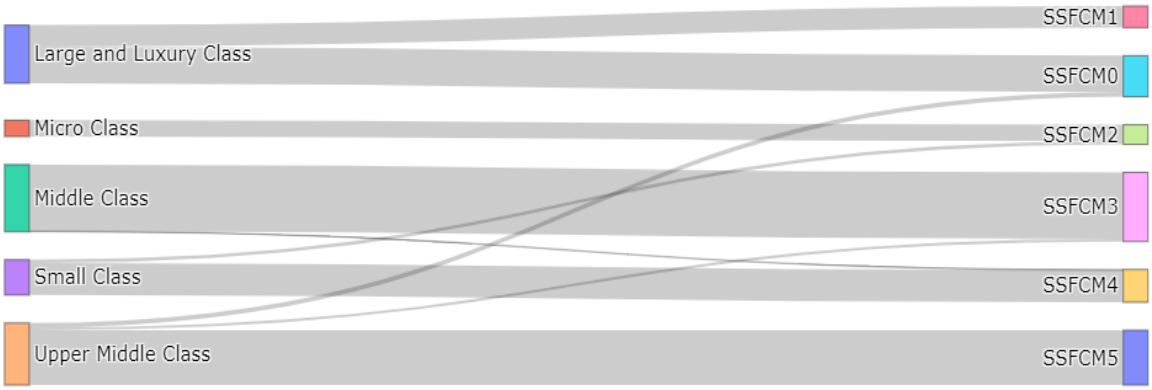


FIGURE 4. Sankey diagram shows the performance improvement with SSFCM approach (right) with labeled rate of 10% in comparison to unsupervised Fuzzy C-means algorithm (left).

TABLE 2. Comparison of the accuracy of the SSFCM method and the traditional FCM algorithm.

	FCM	SSFCM		
Labeled rate	-	10%	20%	30%
True Positive Prediction	277	311	318	333
Accuracy	0.79	0.86	0.87	0.91

Fuzzy C-Mean (FCM) and Semi-Supervised Fuzzy C-Mean (SSFCM).

resampling (ROS) technique is used in order to increase the number of extracted features from minority groups and make it equal to the number of features in majority groups. Finally, algorithm 3 is used to select the best features based on the Euclidean distance and remove redundancy from the feature vector. After we initialized all parts, pseudo labels (PL) of labeled data are assigned to unlabeled data. Following this, all training data with their pseudo labels is carried out to pre-train a SSFCM classification algorithm according to the extracted discriminative features. Finally, new SSFCM is constructed using only the labeled data with true labels.

To evaluate the performance of the proposed approach, we used 10, 20, and 30 percent of the labeled samples from each class to show the effect of the number of labeled data and the rest are used as unlabeled samples. As essential input arguments of unsupervised FCM, the maximum number of iterations was set to 100, the convergence value to $\varepsilon = 1\text{e-}6$ and the degree of fuzziness to $m = 2$. The classification performance of the SSFCM with different rates of labeled samples on the imbalanced multi-class benchmark dataset is presented in Table 2 and compared to the performance of the unsupervised FCM.

The experiment results show that the use of semi-supervised learning in fuzzy clustering algorithm can significantly enhance the classification accuracy. Since in the traditional FCM algorithm the initial cluster centers are selected randomly, the algorithm can be easily trapped in local minimum. In contrast, in our approach, the seed training set which contains some labeled data determines the initial centers. As a result, we get considerably better performances (Fig. 4). When the rate of labeled data for each category increases, the center in seed training set is more nearer to the center in dataset. Hence, the accuracy rate slightly improves along with the labeled rate, which remains at about 91%.

To further demonstrate the advantage of our approach, we compared SSFCM with other semi-supervised state-of-the-art methods k-nearest neighbors (kNN) [23], AdaBoost (Ada) [89], Ensemble selection for multi-class imbalanced datasets (DES-ML) [90], Random Forest Classifier (RF) [91], and Naive Bayes (NB) [92].

The classification accuracy results in Table 3 show that the proposed SSFCM method achieved in most cases the best performance on the passenger vehicle dataset for all labeled rates (only DES-ML with a labeled rate of 10% is slightly higher). The average accuracy over all labeled rates is highest for the SSFCM, showing that the proposed approach outperforms the other semi-supervised approaches.

B. PERFORMANCE MEASURE

To verify the underlying assumption that feature extraction leads to improve classification results compared to the initial classifier's predictions, we use two performance measures to investigate the classification performance of the traditional unsupervised FCM with the original features and SSFCM algorithm, namely MAcc and MFM.

Table 4 shows that, compared with the FCM-based classifier, the performance of the SSFCM with only 10% labeled rates is better for all vehicle classes. On average over all classes, SSFCM reaches a MAcc of 85.58% and a MFM of 0.85, which is higher than that of traditional unsupervised FCM. Furthermore, the visual comparison of the performance metrics resulting from the unsupervised FCM and the SSFCM algorithms reported in Fig. 5 (MAcc) and Fig. 6 (MFM), illustrates the superior performance of the SSFCM in confront to the traditional FCM.

The better results of SSFCM approach show the positive impact of taking the class imbalance problem into account for the higher classification accuracy. Therefore, the

TABLE 3. Accuracy of semi-supervised methods on dataset with labeled rate of 10%, 20%, 30% from each class.

Methods	Accuracy			Average
	10%	20%	30%	
kNN	0.71	0.74	0.78	0.74
Ada	0.48	0.43	0.44	0.45
DES-ML	0.87	0.81	0.88	0.85
RF	0.85	0.79	0.86	0.83
NB	0.42	0.40	0.43	0.42
SSFCM	0.86	0.87	0.91	**0.88**

k-nearest neighbors (kNN), AdaBoost (Ada), Ensemble selection for multi-class imbalanced datasets (DES-ML), Random Forest Classifier (RF), Naive Bayes (NB), and Semi-Supervised Fuzzy C-Mean (SSFCM).

TABLE 4. Clustering performance measured by MAcc and MFM for FCM and SSFCM with a 10% labeled rate.

Classification	FCM-MAcc	FCM-MFM	SSFCM-MAcc	SSFCM-MFM
Micro Class	68.00%	0.7907	**71.23%**	**0.8321**
Small Class	75.93%	0.7810	**89.24%**	**0.9128**
Middle Class	78.85%	0.7664	**83.72%**	**0.7897**
Upper Middle Class	62.11%	0.6592	**91.90%**	**0.8281**
Large & Luxury Class	87.64%	0.8083	-	-
Large Class	-	-	**94.79%**	**0.8385**
Luxury Class	-	-	**82.57%**	**0.8857**
Average	74.50%	0.7611	**85.58%**	**0.8478**

macro-averaged accuracy (MAcc), micro-averaged F-Measure (MFM), Fuzzy C-Mean (FCM), and Semi-Supervised Fuzzy C-Mean (SSFCM).

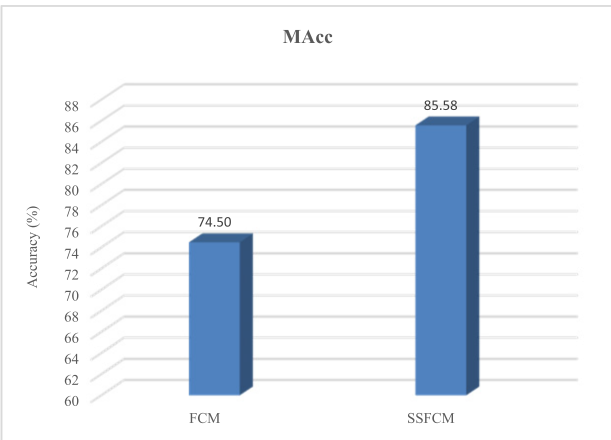


FIGURE 5. Comparison of MAcc of FCM and SSFCM with a 10% labeled rate.

13

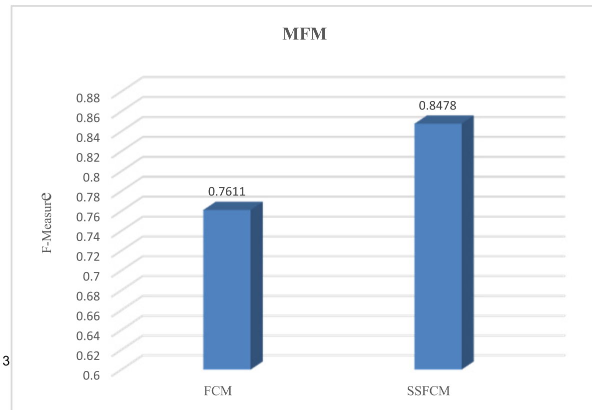


FIGURE 6. Comparison of MFM of FCM and SSFCM with a 10% labeled rate.

experimental results demonstrate that the SSFCM algorithm can extract richer information from vehicle dataset and obtain higher discriminative recognition rates than FCM-based classifier does. This might be due to the variety in geometry between the classes that can be trained effectively with the feature extractor. Our proposed approach can, not only effectively address the problem of multi-class imbalance data but also improve the classification performance.

Lastly, we carried out our experiment to assess the performance of our proposed vehicle classification approach for intra-class vehicle classifications, specifically for the separation of sport utility vehicle (SUV) from non-SUVs. The analysis of drivetrain technologies for each segment based on the expert classification demonstrates that more than two third of the SUVs are four-wheel drive (4WD), while most of the non-SUVs designs are front-wheel drive (FWD) and rear-wheel drive (RWD) [88]. Therefore, the drivetrain technologies taken into account as an assumption of labeled dataset for more accurate vehicle classifications, particularly for big data analysis. Out of the total 366 unique samples only 156 samples meet 4WD assumption. To evaluate the performance, the supervised labeled dataset was split into two subsets (SUV and non-SUVs passenger cars) along with their geometrical and technical features. According to the dataset evaluation SUV (*SUV1*) contains 96 samples and non-SUVs (*SUV0*) contains 270 samples, respectively. Fig. 7 demonstrates multi-class and intra-class vehicle classification based on the SSFCM approach. Vehicles are categorized into six main classes (micro, small, middle, upper middle, large and luxury classes) and two sub-classes (SUV and non SUVs). The performance of the SSFCM with only 10% labeled rates for the intra-class vehicle classification reaches a MAcc of 87% and a MFM of 0.91.

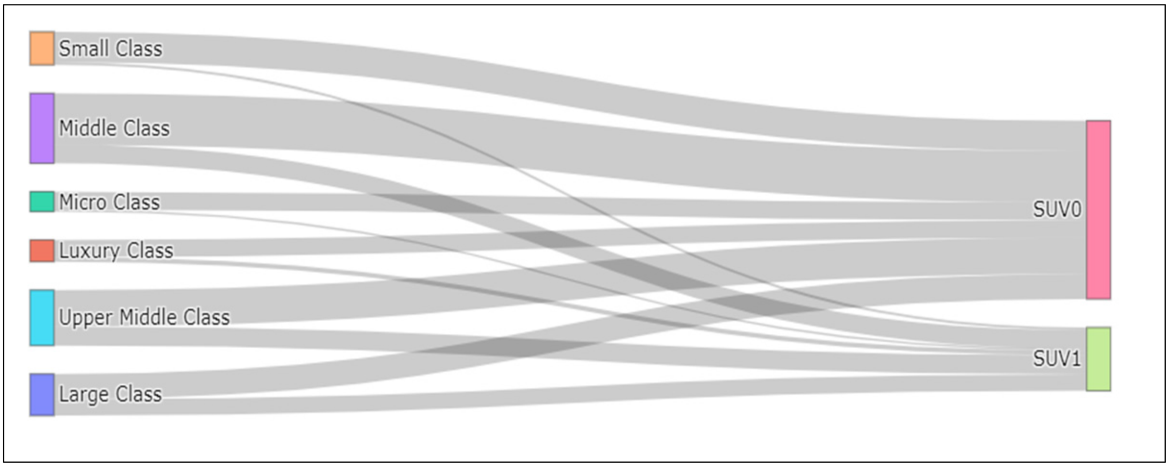


FIGURE 7. Sankey diagram of intra-class vehicle classifications of dataset. SSFCM classification with labeled rate of 10% (left) and SUV/non SUV classification (right).

V. CONCLUSION

To summarize, we designed a novel SSFCM algorithm for passenger vehicle classification, and compared an initial evaluation thereof with an expert segmentation to investigate higher levels of vehicle classification. To incorporate the unlabeled data in the training of SSFCM, we used their predicted labels generated by FCM clustering algorithm. In this way we generate more useful features for classification and it also helps to effectively solve the multi-class imbalance problem.

Our SSFCM algorithm first uses a feature extraction technique and then clusters dataset based on the maximum average of the maximum similarity between the selected features. In experiments with the new registered passenger car dataset, the SSFCM technique was significantly more accurate than other tested unsupervised and supervised approaches. As a result, the average of both evaluated performance measures (MAcc and MFM) are the highest for our proposed approach.

The proposed approach enables accurate automated vehicle classification of large databases, which in turn facilitates the analysis of fleet changes. Moreover, our automated approach has several important advantages over the expert-based segmentations, such as reduced classification costs and training times, and elimination of subjectivity factors that often hinder the comparison of vehicle classification databases across the world. A further area of potentially fruitful research would be to investigate better ways for generating supervised labels that are more consistent to the underlying semi-supervised labels, as the quality of the supervised labels has an effective impact on the quality of the features extraction and the final classification performance of semi supervised deep fuzzy approaches. The analyze time complexity of targeted resampling technique will further help in investigating the performances of the resampling techniques on multi-class problems.

ACKNOWLEDGMENT

The authors thank the Swiss Federal Roads Office (ASTRA) for providing the data from the Swiss Vehicle Information System (MOFIS) and the vehicle technical dataset, as well as the Vehicles Expert Partner (auto-i-dat) for providing the expert classification data.

REFERENCES

[1] J. Wu, H. Xu, Y. Zheng, Y. Zhang, B. Lv, and Z. Tian, "Automatic vehicle classification using roadside LiDAR data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 6, pp. 153–164, Jun. 2019.

[2] L. Zhuo, Z. Zhu, J. Li, L. Jiang, H. Zhang, and J. Zhang, "Feature extraction using lightweight convolutional network for vehicle classification," *J. Electron. Imag.*, vol. 27, no. 5, 2018, Art. no. 051222, doi: 10.1117/.1.JEI.27.5.051222.

[3] S. Kamkar and R. Safabakhsh, "Vehicle detection, counting and classification in various conditions," *IET Intell. Transp. Syst.*, vol. 10, no. 6, pp. 406–413, Aug. 2016.

[4] Z. Yao, H. Wei, Z. Li, T. Ma, H. Liu, and Y. J. Yang, "Developing operating mode distribution inputs for MOVES with a computer vision–based vehicle data collector," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2340, no. 1, pp. 49–58, Jan. 2013.

[5] J. Y. Ng and Y. H. Tay, "Image-based vehicle classification system," in *Proc. 11th Asia–Pacific ITS Forum Exhib.*, Kaoshiung, Taiwan, 2011, pp. 1–5.

[6] J.-W. Hsieh, S.-H. Yu, Y.-S. Chen, and W.-F. Hu, "Automatic traffic surveillance system for vehicle tracking and classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 175–187, Jun. 2006.

[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*, May 2021, Art. no. 036119812110107, doi: 10.1177/03611981211010795.

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

[9] 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.

[10] 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.

[11] 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.

[12] A. Ambardekar, M. Nicolescu, G. Bebis, and M. Nicolescu, "Vehicle classification framework: A comparative study," *EURASIP J. Image Video Process.*, vol. 2014, no. 1, p. 29, Dec. 2014, doi: 10.1186/1687-5281-2014-29.

[13] K. Valev, A. Schumann, L. W. Sommer, and J. Beyerer, "A systematic evaluation of recent deep learning architectures for fine-grained vehicle classification," *CoRR*, vol. 10649, pp. 1–11, Jan. 2018, doi: 10.1117/12.2305062.

[14] R. Kuma, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: An efficient baseline using triplet embedding," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–9, doi: 10.1109/IJCNN.2019.8852059.
[15] S. L. Shi, Q. Zhong, and J. M. Xu, "Robust algorithm of vehicle classification," in *Proc. Int. Conf. Softw. Eng., Artif. Intell., Netw., Parallel/Distrib.*, Jul. 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 and L. Y. Jiang, "Vehicle classification for large-scale traffic surveillance videos using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2016, pp. 2921–2929.
[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. Schälkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, U.K.: 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] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," Stanford InfoLab, Tech. Rep., 2002. [Online]. Available: http://ilpubs.stanford.edu:8090/528/
[26] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*. [Online]. Available: http://arxiv.org/abs/1610.02242
[27] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.
[28] 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.
[29] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.
[30] X. Ao, P. Luo, X. Ma, F. Zhuang, Q. He, Z. Shi, and Z. Shen, "Combining supervised and unsupervised models via unconstrained probabilistic embedding," *Inf. Sci.*, vol. 257, pp. 101–114, Feb. 2014.
[31] X. Tao, Q. Li, W. Guo, C. Ren, Q. He, R. Liu, and J. Zou, "Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering," *Inf. Sci.*, vol. 519, pp. 43–73, May 2020.
[32] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding," in *Proc. 8th Int. Conf. Signal Process.*, vol. 3, 2006, pp. 1–4.
[33] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, Apr. 2009.
[34] R. Longadge, S. Dongre, and L. Malik, "Class imbalance problem in data mining review," *Comput. Sci. Netw.*, vol. 2, no. 1, pp. 1–6, Feb. 2013.
[35] S. Vluymans, D. S. Tarragó, Y. Saeys, C. Cornelis, and F. Herrera, "Fuzzy rough classifiers for class imbalanced multi-instance data," *Pattern Recognit.*, vol. 53, pp. 36–45, May 2016.
[36] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940–7957, 2016, doi: 10.1109/ACCESS.2016.2619719.
[37] Z. Liu, D. Tang, Y. Cai, R. Wang, and F. Chen, "A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data," *Neurocomputing*, vol. 266, pp. 641–650, Nov. 2017.
[38] 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.
[39] D. Zhao, Y. Chen, and L. Lv, "Deep reinforcement learning with visual attention for vehicle classification," *IEEE Trans. Cognit. Develop. Syst.*, vol. 9, no. 4, pp. 356–367, Dec. 2017.
[40] Z. Cebeci and F. Yildiz, "Comparison of K-means and fuzzy C-means algorithms on different cluster structures," *J. Agricult. Informat.*, vol. 6, no. 3, pp. 13–23, Oct. 2015.
[41] S. Y. Cheung, S. Coleri, B. Dundar, S. Ganesh, C. W. Tan, and P. Varaiya, "Traffic measurement and vehicle classification with a single magnetic sensor," *Transp. Res. Rec.*, vol. 1917, no. 1, pp. 173–181, 2005.
[42] S. Javadi, M. Rameez, M. Dahl, and M. I. Pettersson, "Vehicle classification based on multiple fuzzy C-means clustering using dimensions and speed features," *Procedia Comput. Sci.*, vol. 126, pp. 1344–1350, Apr. 2018.
[43] Z. Yao, H. Wei, Z. Li, and J. Corey, "Fuzzy C-means image segmentation approach for axle-based vehicle classification," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2595, no. 1, pp. 68–77, Jan. 2016.
[44] R. Saraçoälu and N. Nemati, "Vehicle detection using fuzzy C-means clustering algorithm," *Int. J. Appl. Math. Electron. Comput.*, vol. 8, no. 3, pp. 85–91, Sep. 2020.
[45] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY, USA: Plenum, 1981, doi: 10.1007/978-1-4757-0450-1.
[46] A. Sheshasayee and P. Sharmila, "Comparative study of fuzzy C-means and K-means algorithm for requirements clustering," *Indian J. Sci. Technol.*, vol. 7, no. 6, pp. 853–885, 2014.
[47] A. Jain and M. Law, "Data clustering: A user's dilemma," in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.*, vol. 3776, 2005, pp. 1–10, doi: 10.1007/11590316_1.
[48] T. Velmurugan and T. Santhanam, "A survey of partition based clustering algorithms in data mining: An experimental approach," *Inf. Technol J.*, vol. 10, no. 3, pp. 478–484, 2011.
[49] B. D. Jyoti and G. A. Kumar, "A comparative study between fuzzy clustering algorithm and hard clustering algorithm," *Int. J. Comput. Trends Technol.*, vol. 10, no. 2, pp. 108–113, 2014.
[50] S. Ghosh and K. S. Dubey, "Comparative analysis of K-means and fuzzy C-means algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 4, pp. 35–39, 2013.
[51] G. Zhang, R. P. Avery, and Y. Wang, "Video-based vehicle detection and classification system for real-time traffic data collection using uncalibrated video cameras," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1993, no. 1, pp. 138–147, Jan. 2007.
[52] K. L. Arunkumar, A. Danti, and H. T. Manjunatha, "Classification of vehicle make based on geometric features and appearance-based attributes under complex background," in *Recent Trends in Image Processing and Pattern Recognition*, vol. 1035, K. C. Santosh and R. S. Hegadi, Eds. Singapore: Springer, 2018, pp. 41–48, doi: 10.1007/978-981-13-9181-1_4.
[53] G. S. Moussa, "Vehicle type classification with geometric and appearance attributes," *Int. J. Archit. Environ. Eng.*, vol. 8, no. 3, pp. 273–278, 2014.
[54] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 507–514.
[55] F. D. L. Torre and T. Kanade, "Discriminative cluster analysis," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 241–248.
[56] 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.
[57] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, vol. 15, 2011, pp. 215–223.
[58] 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.
[59] 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.
[60] 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.

[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. J. L. Jiang, 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.*, Singapore, Feb. 2019, pp. 294–298.

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

[68] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2005, pp. 57–64.

[69] Z. Xu, R. Jin, J. Zhu, I. King, M. R. Lyu, and Z. Yang, "Adaptive regularization for trans-duc-tive support vector machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2125–2133.

[70] K. Huang, Z. Xu, I. King, and M. R. Lyu, "Semi-supervised learning from general unlabeled data," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 273–282.

[71] Z. Xu, R. Jin, J. Zhu, I. King, and M. R. Lyu, "Efficient convex relaxation for transductive support vector machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1641–1648.

[72] Y. Yu, G. Yu, X. Chen, and Y. Ren, "Semi-supervised multi-label linear discriminant analysis," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 688–698.

[73] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained K-means clustering with background knowledge," in *Proc. Int. Conf. Mach. Learn.*, vol. 1, 2001, pp. 577–584.

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

[75] 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.

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

[77] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semisupervised clustering: A brief survey," *Rev. Mach. Learn. Techn. Process. Multimedia Content*, vol. 1, pp. 9–16, 2004.

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

[79] 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.

[80] G. Chen, "Deep transductive semi-supervised maximum margin clustering," 2015, *arXiv:1501.06237*. [Online]. Available: http://arxiv.org/abs/1501.06237

[81] 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.

[82] 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.

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

[84] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, 2001.

[85] B. Handaga and D. M. Mat, "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. Wan Mohd, and E. El-Qawasmeh, Eds. Berlin, Germany: Springer, 2011, pp. 575–589, doi: 10.1007/978-3-642-22191-0-50.

[86] (Mar. 2019). *Das Motorfahrzeuginformationssystem der Eidgen ssischen Fahrzeugkontrolle*. [Online]. Available: https://www.experience-online.ch/de/9-case-study/2023-mofis.

[87] ASTRA. (Mar. 2019). *Bundesamt Für Strassen,*. [Online]. Available: https://www.astra. amin.ch/astra/de/home.html

[88] (Mar. 2020). *Schweizer Partner Für Fahrzeugdaten*. [Online]. Available: https://www.auto-i-dat.ch

[89] 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.

[90] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets," *Inf. Sci.*, vols. 445–446, pp. 22–37, Jun. 2018.

[91] 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.

[92] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 534–545, Jan. 2010.

NAGHMEH NIROOMAND received the M.A. degree from Eastern Mediterranean University, Cyprus, and the Ph.D. and IAPM degrees from Queen's University, Canada, in 2016, and the Ph.D. degree from SSPH, Switzerland, in 2018. From 2018 to 2019, she worked as a Research Fellow with the Transport and Mobility Laboratory, EPFL Lausanne, and a Senior Scientist with Empa, from 2019 to 2021. She is currently a Techno-Energy Economist of the Automotive Powertrain Technologies Laboratory, Swiss Federal Laboratories of Material Science and Technology (Empa), Switzerland. Prior to joining EMPA, she was an Associate Research Economist with 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 University of Applied Sciences in Bern. He performed two internships at the Haagen-Smit Laboratory of the California Air Resources Board in El Monte (USA) to study zero and ultra-low emission technologies in the transport sector. He is currently a Lecturer with ETH Zurich. He is also the Head of the Automotive Powertrain Technologies Laboratory, Swiss Federal Laboratories of Material Science and Technology (Empa). He is 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 Zurich, Switzerland, in 2016. From 2016 to 2018, she worked as a Postdoctoral Researcher and a Senior Scientist with Swiss Federal Laboratories of Material Science and Technology (Empa), Switzerland, from 2019 to 2021. She currently 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.