IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 9, 2019, accepted May 28, 2019, date of publication June 3, 2019, date of current version June 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920489

Impact of Driver Behavior on Fuel Consumption: Classification, Evaluation and Prediction Using Machine Learning

PENG PING[1], WENHU QIN[1], YANG XU[1], CHIYOMI MIYAJIMA[2], (Member, IEEE), AND KAZUYA TAKEDA[3], (Senior Member, IEEE)

[1]School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Informatics, Daido University, Nagoya 457-8530, Japan
[3]Graduate School of Informatics, Nagoya University, Nagoya 464-0814, Japan

Corresponding author: Wenhu Qin (qinwenhu@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61300101, and in part by the Key Research Plan of Jiangsu Province under Grant BE2017035.

ABSTRACT Driving behavior has a large impact on vehicle fuel consumption. Dedicated study on the relationship between the driving behavior and fuel consumption can contribute to decreasing the energy cost of transportation and the development of the behavior assessment technology for the ADAS system. Therefore, it is vital to evaluate this relationship in order to develop more ecological driving assistance systems and improve the vehicle fuel economy. However, modeling driving behavior under the dynamic driving conditions is complex, making a quantitative analysis of the relationship between the driving behavior and the fuel consumption difficult. In this paper, we introduce two kinds of machine learning methods for evaluating the fuel efficiency of driving behavior using the naturalistic driving data. In the first stage, we use an unsupervised spectral clustering algorithm to study the macroscopic relationship between driving behavior and fuel consumption, using the data collected during the natural driving process. In the second stage, the dynamic information from the driving environment and natural driving data is integrated to generate a model of the relationship between various driving behaviors and the corresponding fuel consumption features. The dynamic environment factors are coded into a processable, digital form using a deep learning-based object detection method so that the environmental data can be linked with the vehicle's operating signal data to provide the training data for the deep learning network. The training data are labeled according to its fuel consumption feature distribution, which is obtained from the road segment data and historical driving data. This deep learning-based model can then be used as a predictor of the fuel consumption associated with different driving behaviors. Our results show that the proposed method can effectively identify the relationship between the driving behavior and the fuel consumption on both macro and micro levels, allowing for end-to-end fuel consumption feature prediction, which can then be applied in the advanced driving assistance systems.

INDEX TERMS Driving behavior modeling, data mining, deep learning, vehicle fuel economy.

I. INTRODUCTION

A combination of emissions from coal combustion and urban vehicle use has become the primary source of air pollution in most of the world's major cities [1], [2]. According to the World Health Organization, transportation emissions are a significant and growing contributor to particulate air pollution, which makes up 30% of particulate matter emissions (PM) in European cities and 50% of PM emissions in OECD countries [3]. One study estimated that approximately 1.03 million deaths were associated with ambient PM 2.5 air pollution in the 74 largest cities of China in 2013, which accounted for 32% of all reported deaths [4]. As a result, much research has been focused on reducing vehicle emissions. As has been demonstrated in various studies [5]–[7], driving behavior, such as speed control,

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

preferred rate of acceleration, and vehicle control stability, have a major effect on fuel consumption, regardless of the type of vehicle being driven. By accurately identifying relationships between driving behavior and fuel consumption, Advanced Driving Assistant Systems (ADAS) can be designed to give more accurate and intelligent eco-driving advice [8], [9]. By studying the driving behavior's impact on the fuel consumption, we can know how some drivers cost more energy than others so as to help high energy cost drivers to achieve fuel-effect driving style. Besides, as the fundamental technology of the ADAS systems or eco-driving coaching system, the effective driving behavior-energy consumption model can be applied to decrease the commercial vehicle's fuel cost [10], optimal the charging station location [11], decrease the transportation's greenhouse gas emission [12] and so on. Thus, discovering the precise relationship between driving behavior and fuel consumption, in order to reduce vehicle emissions and increase fuel efficiency, has become an important studying area and the motivation of our study. However, effective analysis model for driving behavior's impact on fuel consumption is rarely studied. In this paper, we aimed to design a machine learning based method which can analyze and predict a reasonable relationship between the driving behavior and fuel consumption. The eco-driving system or ADAS system can obtain driving state from the proposed model so as to give more reasonable advice to the driver to keep fuel-efficient driving.

Quantitative analysis of the relationship between driving behavior and fuel consumption is a natural and direct approach. However, traditional fuel consumption models such as the Vehicle Specific Power (VSP) model [13], the Comprehensive Modal Emission Model (CMEM) [14] and the International Vehicle Emissions model (IVE) [15] are specifically designed to evaluate the fuel economy performance of engines, and the process of calibrating these models is very complex [16]. In contrast, most driving behavior modeling studies have focused on specific driving scenarios, such as lane changes [17], [18], arterial corridors [19], signalized intersections [20], and so on. These models focus on identifying safe or comfortable driving, which are difficult to link to fuel consumption. As a result, the integration of driving behavior parameters or models with traditional fuel consumption models is a problem which remains to be resolved. Many researchers have proposed two-stage methods, where statistical or machine learning methods are used to identify a driver's driving style, and then the features of that driving style are compared with the related fuel consumption features. J. E. Meseguer *et al.* used a three-layered neural network to classify drivers into quiet, normal and aggressive groups [21]. They then analyzed the fuel consumption features for each group. E. Gilman *et al.* used 17 driving behavior factors to identify fuel-efficient driving behavior for a driver coaching system [22]. The driving behavior factors were evaluated according to their distributions, calculated from a historical driving trip. R. Trigui *et al.* analyzed the impact of various driving behaviors on fuel efficiency using mathematical modeling [23]. The study first divided driving behavior into two levels; maneuvering level and control level behavior. Then, by identifying the various parameters of their model, the authors simulated three different behaviors; aggressive driving, eco-driving and normal driving. Their results showed that their proposed model could accurately match measured fuel consumption and real driving behavior. C. Lv *et al.* proposed an unsupervised machine learning method using Gaussian mixture models to recognize three typical driving styles, and then provided the optimal control strategy for each driving style in order to improve energy efficiency [24]. All of the studies cited here succeeded in identifying fuel-efficient driving behavior, however their lack of detailed consideration of the impact of various traffic condition limits the usefulness of their results as driving behaviors are also influenced by various static or dynamic environmental factors [25], [26].

Therefore, some researchers have also examined driving environment features, which can be deduced or directly obtained from naturalistic driving data, in their analyses of driver fuel consumption, resulting in more nuanced assessments. M. Ehsani *et al.* discussed in detail the effects of external environmental factors on vehicle fuel consumption [27], but did not carefully examine the effect of driving behavior, only mentioning that speed and acceleration are the two most important parameters. J. Rios-Torres *et al.* classified driving styles into three categories by analyzing real-world data, and then examined the effect of each driving style on fuel consumption [28]. The results of this study show that vehicle fuel consumption can vary widely compared with standard US Environmental Protection Agency (EPA) driving cycles, depending on the driver's driving style and the driving scenario.

The studies mentioned above investigating the relationship between driving behavior and fuel consumption have achieved good results, but unanswered questions remain. Most of these studies have employed statistical or rule-based methods to analyze the relationship between driving behavior and fuel consumption, so these methods require huge amounts of long-term driving data as well as prior knowledge of the data's statistical feature. The ordinary methods usually need lots of expert skills to extracted prior knowledge from the raw data set. And the results have limited universality because the experiments have mostly been conducted on a limited variety of traffic conditions. Although the machine learning method also need considerable amount of data, the learning-based method can learn the inner feature or the knowledge from the raw data automatically.

Thus, in this paper we propose an approach which employs two machine learning methods, in order to push the research of the fuel-efficient driving behavior one step further. In the first stage, we use an unsupervised machine learning method to analyze the fuel efficiency of driver behavior macroscopically, as shown in the upper section of Fig. 1 (circled in red). Inspired by some previous studies [29]–[31] in which machine learning was used for driving behavior analysis,

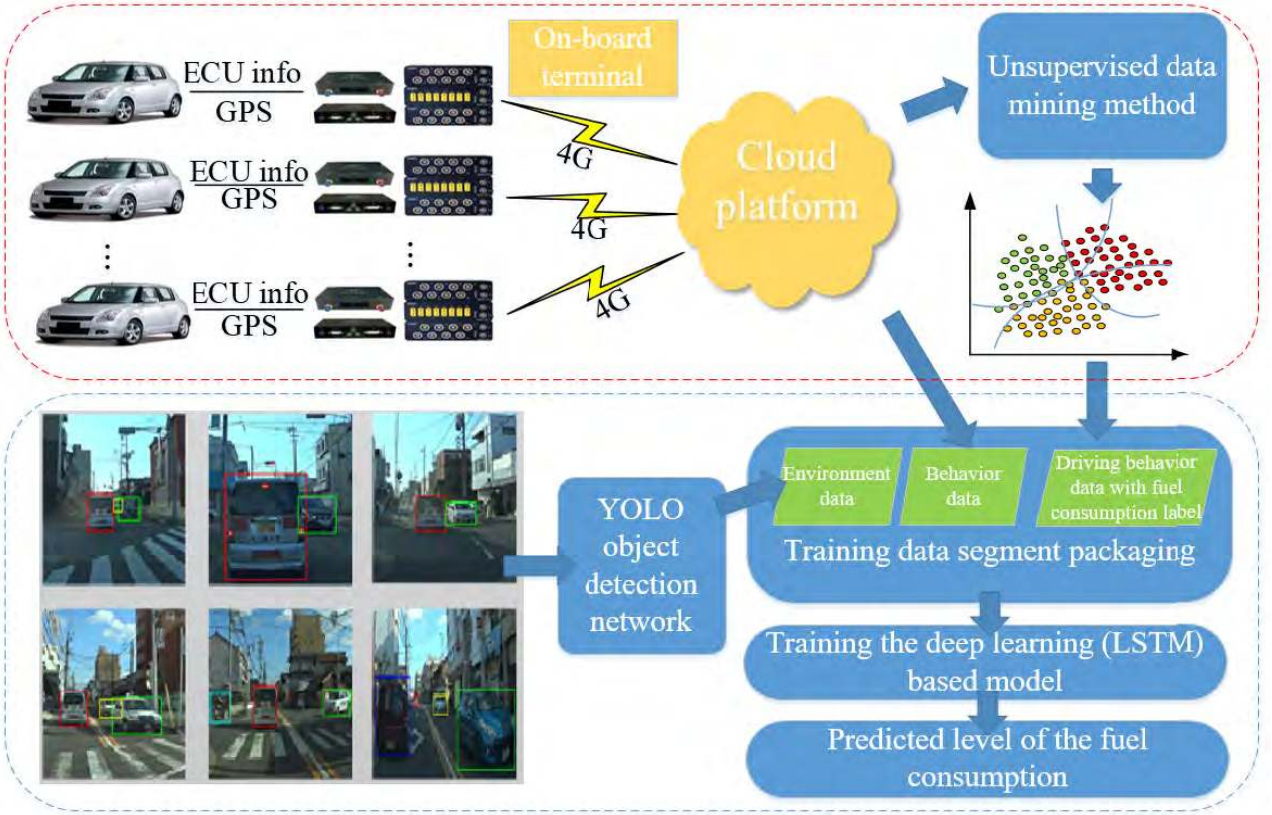

FIGURE 1. Two-stage architecture of the proposed driving behavior modeling method. In the first stage (outlined in red) unsupervised machine learning is used to obtain the macro-level fuel consumption features of driver behavior. In the second stage (outlined in blue) an LSTM is used to analyze short-term driving behavior and driving environment data to predict real-time fuel consumption.

in this study we employ a parallel spectral clustering algorithm [32] to classify the driving signal dataset collected from multiple drivers. Drivers are divided into three groups based on similarities in their driving styles. We then analyze the data to extract the data points which lie in the same fuel consumption zone. Due to the properties of spectral clustering, prior knowledge about the data is not required, so this clustering method is suitable for dealing with unique sets of driving data. A parallel calculating structure is also used to improve the efficiency of the clustering process.

The other machine-learning method used in this study is Long Short-Term Memory (LSTM), which is a powerful method for modeling behavior [33]. In contrast to previous studies which using LSTM to analyze the fuel consumption model [34], [35], in this paper we include more features of the dynamic traffic environment, in form of video frames, in our learning model, as shown in the lower part of Fig. 1 (circled in blue), so as to make the network more robust and general to a wider variety of traffic conditions. In addition to analyzing the fuel efficiency of a driver's historic or long-term driving behavior, our learning-based method is designed to also examine short-term driving data, making the prediction results adaptive to dynamic traffic conditions. The input end of the model uses video frame, GPS and ECU information, while the output is a real-time prediction of the level of fuel consumption. This structure allows end-to-end evaluation of the fuel-efficiency of driving behavior.

This paper is organized as follows: The paper's objectives and related research are described in the Introduction. Section II provides details about the spectral clustering algorithm we employed and describes the collection of driving behavior data using data mining. Section III describes our use of an LSTM to predict short-term fuel consumption features and describes the model's performance using representative fuel consumption feature prediction results. Finally, in Section IV we discuss our findings and conclusions.

II. DATA COLLECTION AND UNSUPERVISED EXTRACTION OF FEATURES OF FUEL-EFFICIENT DRIVING BEHAVIOR

A. DATA COLLECTION PROCESS

1) EXPERIMENT DESIGN

Research by Ericsson [26] suggests that driving behavior is affected by various factors such as street design, traffic management methods, traffic conditions, weather conditions and the driver's mental and physical condition. In order to evaluate the effect of the driver's condition on vehicle fuel consumption and simplify the verification process, in this study we fixed the vehicle type, trip route and weather conditions used in our experiment. The only variable factors are the drivers (i.e., their driving behavior) and the traffic conditions.

FIGURE 2. Two types of roads used for data collection. Left: Expressway loop with two lanes. Right: Ordinary road with one lane in each direction.

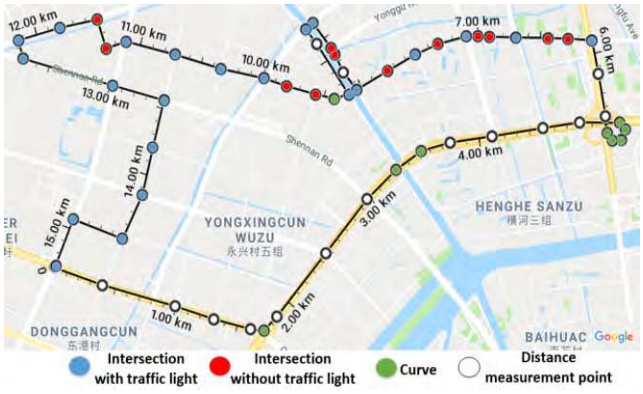

FIGURE 3. Overall map of the roads used for data collection. Yellow line is the expressway and the white line is the ordinary road.

If more than one route were used in the experiment, it would be difficult to determine which factors were primarily responsible for variation in fuel consumption. Therefore, all of the data for our experiment was collected using a fixed route which included some variation in road types. Examples of the two types of roads used in our study are shown in Fig. 2. The total distance of all of the road segments was about 15.2 km, which consisted of a 5.3 km expressway loop with two lanes in each direction and 9.9 km of ordinary road with one lane in each direction. The detail route map and road information are shown in Fig. 3.

Our data was collected using 30 normal passenger cars with a 1.2T (85kw) gasoline engine and a six-speed automatic transmission (6AT). Fuel consumption increases by 0.38±0.079% each time the air temperature decreases by 1°C [36]. Therefore, in order to avoid the possibility of variations in air temperature obscuring the relationship between driving behavior and fuel consumption, the data collection was conducted in the autumn from September to November. 202 drivers are selected to join the experiment, the information of the drivers is shown in Fig. 4. As the supervised and unsupervised learning method need lots of samples, so we try out best to find the experiment participants as much as possible. We choose these 202 drivers from our university's students and the cooperator's staffs. All the participants drove in the experimental route for 10 circuits a day and the whole experiment of single drivers last a week. When processing our experiment, we did not give time limitation or some special

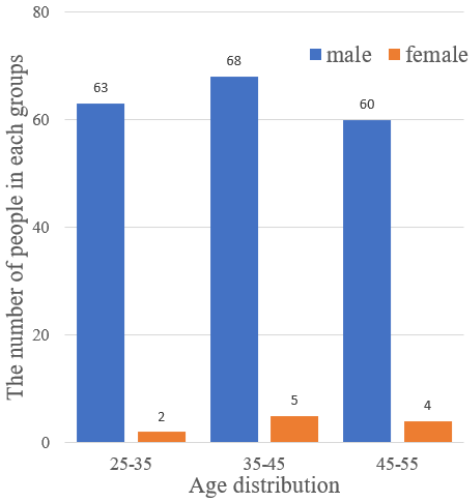

FIGURE 4. Age and sex distribution of all the experiment participants.

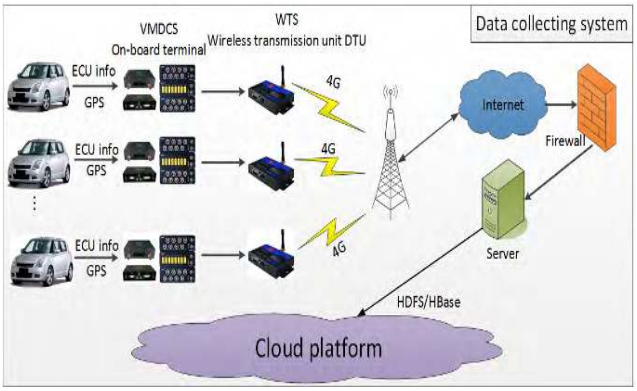

FIGURE 5. Data collection system (for driving data and GPS information).

driving tasks to the participants in order to avoid extra mental pressure. We just tell them the research goal, experimental route and drive as they usually do. Most of the experiment participants are in normal emotion and will be paid after the experiment.

2) DATA COLLECTION AND REDUNDANT DATA PRUNING

The data collection system (DCS) in Fig. 5 is divided into three parts: a vehicle-mounted data collection system (VMDCS), a wireless transmission system (WTS) and a data center (DC). The VMDCS uses On-Board Diagnostics (OBD) to obtain the vehicle's operating information from the ECU, and uses GPS to track the vehicle's position. The WTS uses a wireless transmission unit (WTU) installed on the vehicle which communicates with the base station via 4G broadband to upload the collected data. Messages from the WTS include a receiving module IP address so that the data can be transmitted to the DC via the internet. The DC server shows the vehicle's position and real-time vehicle information on the Web. The collected data is stored in an SQL database.

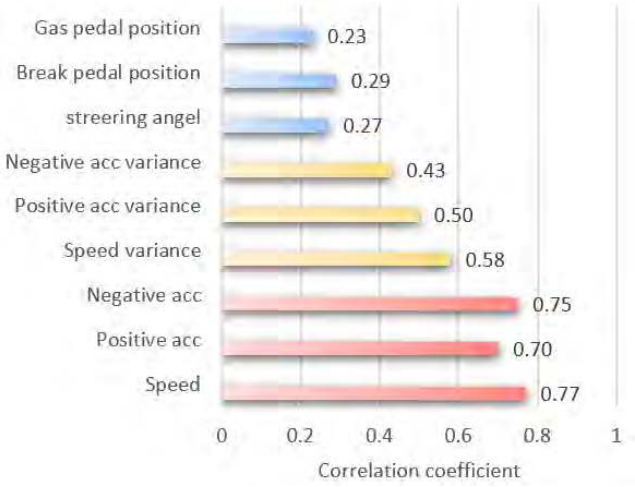

FIGURE 6. Correlation coefficients of various vehicle operation parameters with fuel consumption. Red bar: Strong correlation, Yellow bar: Moderate correlation, Blue bar: Weak correlation.

In order to improve calculation efficiency, we selected vehicle operation data with a strong relationship to driving behavior, and used the Pearson correlation coefficient (PCC) [37] to determine the relevance of each parameter to vehicle fuel consumption. We treated positive and negative acceleration as different parameters because their effects on fuel economy differ. For example, when calculating fuel cost, if negative acceleration is less than zero, instantaneous fuel consumption is zero. The calculated correlation coefficients for various features are listed in Fig. 6, where PCC value ρ is represented by different color bars according to the following standard guidelines; when $|\rho| > 0.5$ = strong correlation, when $0.5 > |\rho| > 0.3$ = moderate correlation, when $|\rho| < 0.3$ = weak correlation [38]. In Fig. 6, 'Negative acc' and 'Negative acc variance' have a negative correlation with fuel consumption, so in fact, the PCC of these two parameters are negative values. Then, before using an unsupervised clustering method to abstract the data distribution features, we first pruned the weakly correlated data parameters.

3) FUEL CONSUMPTION CALCULATION

To calculate fuel consumption, we integrated instant fuel consumption information from the ECU to obtain accumulated fuel consumption data. In order to verify the results of our calculations, we compared our calculated results with the results from a fuel consumption analyzer under various traffic conditions. The differences between these two fuel consumption measurement approaches are shown in Table 1.

From the data in Table 1, we can conclude that the difference between our calculation method and actual fuel consumption is less than 6%. As the route used in our experiment is only 15 km in length and the goal of the study is to evaluate the effect of driving behavior on fuel consumption, this difference can be ignored.

TABLE 1. Difference between calculated fuel consumption and fuel consumption analyzer results.

Road type / Vehicle load	Urban road	Expressway	Rural road
No-load	4.85%	1.28%	2.18%
Full-load	5.94%	0.81%	3.65%

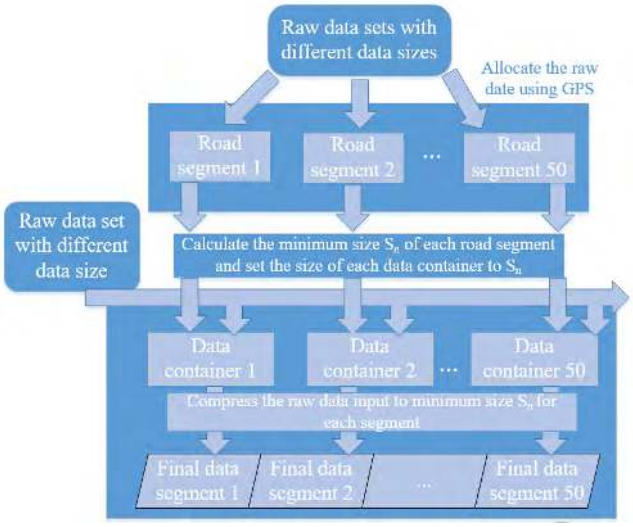

FIGURE 7. Data compression process based on road segment.

B. DATA SEGMENT CONSTRUCTION

As our research goal is to analyze and predict the impact of driving behavior on fuel consumption within a limited time frame (25 to 35 minutes), in this section we describe the spectral clustering method we used to compare inner similarity within the data set, so as to cluster data with similar features into the same cluster. Our spectral clustering method can only handle data sets of the same size. The data collection rate was 10Hz and we collected 15,000-21,000 data points per circuit of the driving route (we treated each circuit of the driving route as an independent data set). Since the amount of data collected in each data set varied, we needed to compress each data set to the same size.

As shown in Fig. 7, we firstly partitioned the raw data set into several subsets. The driving route was divided into 50 road segments according to their location distribution. And then the whole data will be divided according to their belonging road segment (each data points contain the GPS position). As each road segment contains a different number of data points, we needed to calculate each segment's minimum data size S_n. For example, S_1 is the minimum data size of the first road segment (calculated from the entire data set associated with the first road segment). Each data set allocated to road segment 1 is then compressed to size S_1. After data compression, each data set will have the same data size S_{all}, as shown in equation (1):

$$S_{all} = \sum_{i=1}^{50} S_i \tag{1}$$

In contrast to using maximum information entropy to select the size limit of the data, as in our previous study [39], the data compression method adopted in this paper allows us to retain most of the data points.

C. UNSUPERVISED DATA FEATURE EXTRACTION

1) SPECTRAL CLUSTERING ALGORITHM

Unsupervised machine learning is usually used for data distribution analysis or data set inner feature abstraction. In this paper, we adopt spectral clustering to study the features of our self-collected dataset. As described previously, we collected driving data sets of the same size from multiple drivers during natural driving along a fixed route. Spectral clustering performs data clustering as a graph partitioning problem without making any assumptions about the form of the data clusters. Due to this characteristic, we do not need to have prior knowledge of the driving behavior data. This is very important for our research because the data sets which are obtained from the data collection platform vary from driver to driver. Spectral clustering is a suitable method for working with these kinds of 'random' data sets. In additions, spectral clustering is reasonably fast, especially for sparse data sets of up to several thousands of points. Furthermore, spectral clustering is not dependent on the dimensions of the data sets. The first step of the spectral clustering process is to construct driving data layout graph G, which is an undirected similarity graph for the parameters of the data points, all of which are scalar. We use X to represent the entire raw driving data set:

$$\mathrm{X} = \{x_1, x_2, \ldots, x_N\}, \quad x_i \in R^{l \times S_{all}} \tag{2}$$

Each x_i contains the six selected fuel-efficiency linked parameters which were chosen as described above, so $\mathrm{l} = 6$ in this case. N is the total number of data samples. Graph G is weighted using the distances between each pair of vertices x_i and x_j, which are represented by non-negative weight $w_{i,j}$. Because there has been no definitive determination of how the designs of similarity graphs influence spectral clustering results [29], here we use a full-connection to construct similarity matrix W, and use a Gaussian function to calculate $w_{i,j}$ as follows:

$$w_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\delta^2}\right), \quad \delta = 10 \tag{3}$$

Similarity matrix $\mathrm{W} \in R^{N \times N}$ is constructed using the terms of $w_{i,j}$. Obviously, matrix W is a symmetric matrix for G, which is an undirected similarity graph. We then build degree matrix D, which is a diagonal matrix with degree $(d_1, \ldots, d_n)$ as the diagonal. The degree of vertex x_i is defined as:

$$d_i = \sum_{j=1}^{N} w_{i,j} \tag{4}$$

Two other parameters are defined, the volume of a cluster, Vol(C), and the border between two clusters, Cut (C_1, C_2), which are calculated as follows:

$$\mathrm{Vol(C)} = \sum_{i \in C} d_i \tag{5}$$

$$\mathrm{Cut}\,(C_1, C_2) = \sum_{i \in C_1} \sum_{j \in C_2} w_{i,j} \tag{6}$$

Next, similarity graph G is partitioned into disjointed sets. There are different graph cutting methods, such as MinCut [40], RatioCut [41] and NCut [42]. MinCut is simple and effective, but it often fails to satisfactorily solve the problem due to possible singularity problems. RatioCut and NCut take into consideration the vertices and edge weights to make the clusters more balanced, but RatioCut is relatively slow, so in this study we chose Ncut, which is an NP-hard problem [40], as our border determination method. In order to obtain optimal clustering results, we used the object function shown in (7), where $(A_1, \ldots, A_k)$ are the final clustering groups. This object function is used again in (10). $\overline{A}_i$ is the complementary set of A_i:

$$\begin{aligned} \mathrm{minNcut}(A_1, \ldots, A_k) &= \min\left(\frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i, \overline{A}_i)}{\mathrm{Vol(A_i)}}\right) \\ &= \min\left(\sum_{i=1}^{k}\frac{\mathrm{Cut}(A_i, \overline{A}_i)}{\mathrm{Vol(A_i)}}\right) \end{aligned} \tag{7}$$

A group of indicator vectors $h_j = \left(h_{1,j}, \ldots, h_{n,j}\right)^T$ are then defined as follows:

$$h_{i,j} = \begin{cases} \dfrac{1}{\sqrt{\mathrm{Vol}\left(A_j\right)}}, & x_i \in A_j \\ 0, & x_i \notin A_j \end{cases} \tag{8}$$

Matrix $\mathrm{H} \in R^{N \times k}$ which contains the k indicator vectors $h_{i,j}$ as columns, is then constructed. Normalized graph Laplacians [44] are then introduced as:

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{9}$$

Due to the following given properties:

$$\begin{cases} H'H = I \\ h_i' D h_i = 1 \\ h_i' L h_i = \mathrm{Cut}\left(A_i, \overline{A}_i\right) / \mathrm{Vol}\left(A_i\right) \end{cases} \tag{10}$$

the Ncut problem is then reformulated as:

$$\mathrm{argmin}_{A_1, \ldots A_k}\ \mathrm{Tr}\left(H'LH\right) \text{ subject to } H'DH = I \tag{11}$$

By substituting $\mathrm{T} = D^{-\frac{1}{2}} H$, we can change the Ncut problem into a simpler form:

$$\mathrm{argmin}_{T \in R^{N \times k}}\ \mathrm{Tr}\left(T'^{D^{-\frac{1}{2}}} L D^{-\frac{1}{2}} T\right) \text{ subject to } T'T = I \tag{12}$$

Then, according to the Rayleigh-Ritz theorem [32, 45], this standard trace minimization problem can be solved using

matrix U, which contains k eigenvectors as columns, corresponding to the first k eigenvalues (in increasing order) of L_{sym}. Finally, by taking each row of matrix U as new data sets, we then cluster them into k groups using a K-means clustering algorithm. If the unit in row i of matrix U belongs to group C_j, the original data x_i in the raw data set X also belongs to group C_j.

2) PARALLEL SPECTRAL CLUSTERING ALGORITHM

The time complexity of a spectral clustering algorithm is $O(n^3)$, where n represents the amount of data. If n is greater than 5,000, the time cost of spectral clustering using conventional calculation methods will be excessively high, therefore we introduce a method of parallel spectral clustering which employs cloud computing. The cloud computing platform Spark [46] is suitable for parallel calculations involving big data. By analyzing the inner calculation mechanism of our spectral clustering method, we see that three processes are responsible for most of the calculation time cost: construction of the similarity matrix, calculation of the eigenvalues of the graph Laplacians and the final K-means clustering.

The process of parallel spectral clustering using the Spark platform can be described as follows:

Step 1: Calculating the similarity matrix in a parallel manner.

First, we store the entire raw data set in a Hadoop distributed file system (HDFS), since data sets in HDFS can be accessed by the whole calculating cluster. We then use the Spark resilient distributed dataset (RDD) map method (shown in Fig. 8) to assign the spilt data set to several parallel calculating tasks. Because the similarity graph is fully connected, the similarity matrix is symmetrical. As a result, we just need to calculate $w_{i,j}, \forall 1 \leq i \leq j \leq N$. The detailed method for dividing the data to construct sub-sets is shown below:

$$\begin{aligned}
&\text{Raw data set: } X = x_1, x_2, \ldots, x_N \\
&\text{Fragment set:} \\
&X_1 = \left\{x_1, X_1'\right\}, X - X_1' = \emptyset \\
&X_1 = \left\{x_2, X_1'\right\}, X - X_1' = \{x_1\} \\
&\qquad\qquad \vdots \\
&X_N = \left\{x_N, X_N'\right\}, X - X_N' = \{x_1, x_2, \ldots, x_{N-1}\} \quad (13)
\end{aligned}$$

Fragment set X_1 will be assigned to Task 1, as shown in Fig. 8. The job of the Task 1 model is to calculate $\left(w_{1,1}, \ldots, w_{1,n}\right)$. Expanding to arbitrary Task i, the fragment set X_i will be offered to Task i to calculate $\left(w_{i,1}, \ldots, w_{i,n}\right)$. The final step is to integrate the results of all of the tasks in order to construct the similarity matrix. An overview of the method of calculating the similarity matrix in a parallel manner is shown in Fig. 8.

Step 2: Simplifying the calculation of the eigenvalues of the graph Laplacians.

Lanczos algorithm [47] is the method used to calculate the eigenvalues, and the calculation process is shown in Fig. 9.

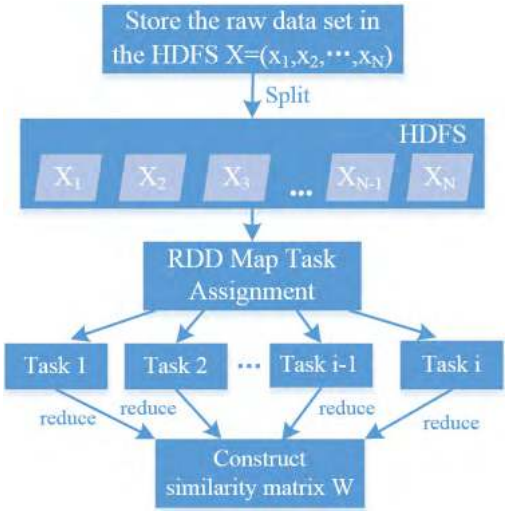

FIGURE 8. Method of calculating the similarity matrix in a parallel manner.

input: graph Laplacians L
output: diagonal matrix T

unit vector : $v_1 = 1$;
zero vector : $v_0 = 0$;
zero vector : $b_1 = 0$;
FOR $j = 1 : n$ *do*
$w_j = L \times v_j - b_j \times v_{j-1}$;
$a_j = \left(w_j, v_j\right); w_j = w_j - a_j \times v_j$;
$b_{j+1} = \left\|w_j\right\|; v_{j+1} = w_j / b_{j+1}$;
End For

FIGURE 9. Method of calculating the eigenvalues of the graph Laplacians.

Based on the process shown in Fig. 9, the following relationships can be derived:

$$V'LV = T, \quad V = \{v_1, v_2, \ldots, v_n\} \quad (14)$$

$$T = \text{tridiag(B, A, B)}, \quad B = \{b_1, \ldots, b_n\}, \quad A = \{a_1, \ldots, a_n\} \quad (15)$$

By observing the Lanczos algorithm calculation process, we find that most of calculation time cost is due to the process of $L \times v_j$, so we split L into n rows and multiply each row by v_j. We then merge the results to get the final value of $L \times v_j$. An overview of the calculation process is shown in Fig. 10. The parallel calculation process increases memory cost, but the inner memory assignment mechanism limits this problem to a tolerable level.

Step 3: K-means is an iteration process, so we split the data into several smaller data sets.

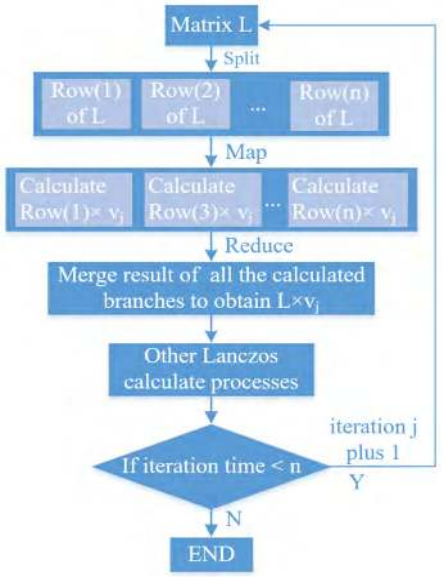

FIGURE 10. Parallel calculation of the eigenvalues of the graph Laplacians.

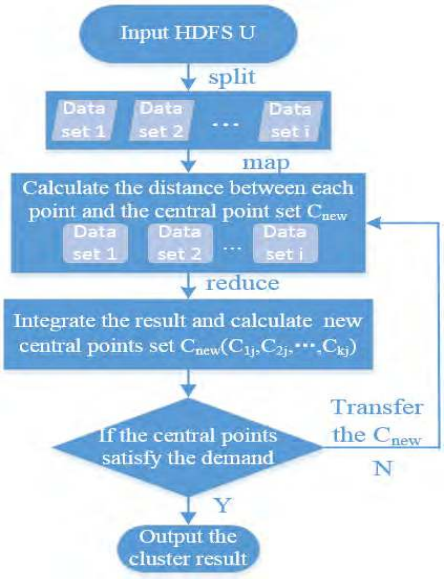

FIGURE 11. Method of calculating K-means in a parallel manner.

We first choose random center points for the whole data set and assign the center points to each data subset. The subset data will be used to calculate the distance between the subset data and the randomly chosen center points. Next, the subset data results are sent to a task which integrates all of the results of the data subsets, in order to obtain new center points for the whole data set. This process will continue until the center points satisfy the demands of the overall data set. Compared to the traditional K-means process, parallel K-means calculation converts global calculation into regional calculation, which simplifies the calculation object in order to reduce the time cost. The parallel K-means calculation process is shown in Fig. 11.

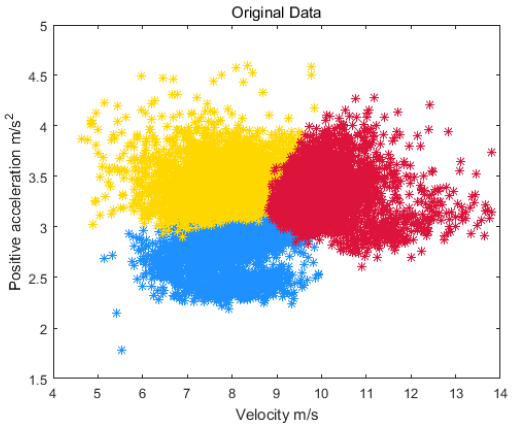

FIGURE 12. Driving data clustering results.

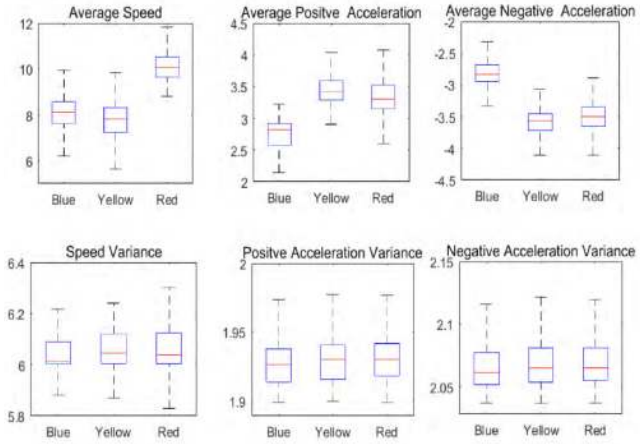

FIGURE 13. Driving data clustering results for all of the selected parameters (Blue, Yellow and Red refer to the data clusters shown in Fig. 12).

3) FEATURE EXTRACTION RESULTS

A total of 8,984 natural driving data samples (i.e., the number of completed trips) were selected during the data collection process described in Subsection A above. Using the parallel spectral clustering algorithm described above, the data samples were then clustered into three groups, with each group containing drivers with similar driving styles or behavior, as shown in Fig. 12. The X and Y axes of Fig. 12 represent velocity and positive acceleration, respectively. The points in the blue cluster represent the drivers who drove at low velocity with low positive acceleration. The points in the yellow cluster represent the drivers who preferred to drive at low velocity but who used high rates of acceleration. Points in the red cluster represent the drivers who preferred to drive at a high velocity and whose acceleration ranged from high to low. We break the clustering results down statistically using our six selected fuel consumption-related parameters in Fig. 13. In Fig. 14, the data points of each of the clusters are plotted on 2-D and 3-D graphs according to fuel-consumption and their serial number within the data set. Average fuel consumption for drivers on the blue line was 3.68 L/100 km, on the yellow line 5.14 L/100 km and on the red line 7.44 L/100 km.

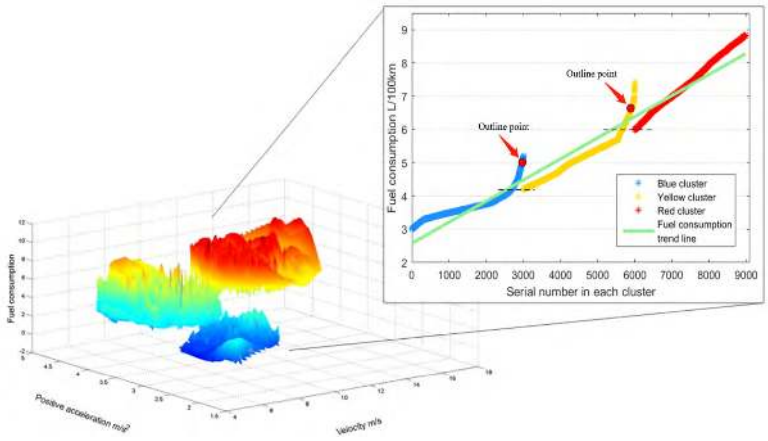

FIGURE 14. Fuel consumption distribution of the three-cluster group.

There are several phenomena illustrated in Fig. 13 which are worth noting. First, we find that fuel-consumption within each cluster differs and that fuel-consumption increases steadily from the blue to the yellow to the red cluster, i.e., there is a surprising amount of variation within each group, but this variation is constrained by a clear trend. Second, some outlier points exist, which represent drivers whose fuel consumption was actually higher than that of some of the drivers in the next cluster. A numerical analysis of these outlier points is shown in Table 2. We can see clearly in the Fig. 14 that the height of each cluster, which represents increasing fuel consumption, differs. We can also see that the three clusters have overlapping areas, which can be observed in the areas of the 3D graph containing blended colors. These overlapping areas represent the outlier points. Because the spectral clustering process is based on a data graph partition algorithm, the points on the periphery of each cluster group will tend towards randomness, which means the points on the boundaries will join the clusters randomly. Additionally, the six chosen parameters represent only the major factors affecting fuel-consumption, but not all of the factors related to vehicle operation. As a result, some data points which have high fuel-consumption attributes may also share other attributes with data points in the lower fuel consumption clusters. What's more, long-term fuel consumption is deduced by observing instantaneous fuel consumption, as shown in Table 1, so the calculated fuel consumption values could have an error rate of 0.8%-5.9%, which could also affect the final clustering results. Finally, the overall proportion of outlier data points is about 20.69%.

From the above results, we can conclude that drivers who operate their vehicles with relatively low fuel consumption are those who change their driving speed moderately and drive their vehicles at a relatively low average speed. The proposed parallel spectral clustering algorithm was able to accurately cluster the drivers according to their fuel-consumption using vehicle operation data, with an approximate clustering accuracy rate of 79.31%.

TABLE 2. Numerical analysis of outlier points.

Group	# of outlier points	Total # of points	Proportion of outlier points
Low fuel consumption	276	3000	9.20%
Medium fuel consumption	345	3001	11.49%

In order to verify the performance of the clustering method used in this study, we compared our clustering results with those of the kernel fuzzy C-means (KFCM) [30] and K-means clustering methods [48]. Performance of the three clustering methods are compared in Table 3.

From this comparison we can see that the proportion of outlier points when using KFCM is 4% higher than when using the proposed parallel spectral clustering method. When using the K-means method, the data points are less tightly clustered compared with the other two clustering methods, and the proportion of outlier points is the highest of all the clustering methods. Therefore, the proposed parallel spectral clustering method achieved the best clustering performance of the three methods.

We then compared the calculation efficiency of the proposed parallel calculating structure with normal spectral clustering. Different sample sizes were chosen to verify the proposed method's superior performance. The results are shown in Fig. 15. When the amount of data being calculated is greater than 10,000 data points, the time cost of normal spectral clustering using Matlab is almost 18 times higher than when using the parallel spectral clustering method. Furthermore, as the amount of data increases, the time cost of normal spectral clustering increases sharply.

In this section we described the clustering method used to obtain the macroscopic relationship between driving behavior and fuel consumption. In the next section, an LSTM-based

TABLE 3. Comparison of different clustering methods.

Clustering algorithm	Fuel consumption group	# of outlier points	Total # of points	Outlier point proportion	Total proportion of outlier points
Spectral cluster	Low	276	3000	9.20%	20.69%
	Medium	345	3001	11.49%	
KFCM	Low	417	3117	13.38%	24.53%
	Medium	334	2995	11.15%	
K-means	Low	535	2889	18.52%	28.71%
	Medium	327	3210	10.19%	

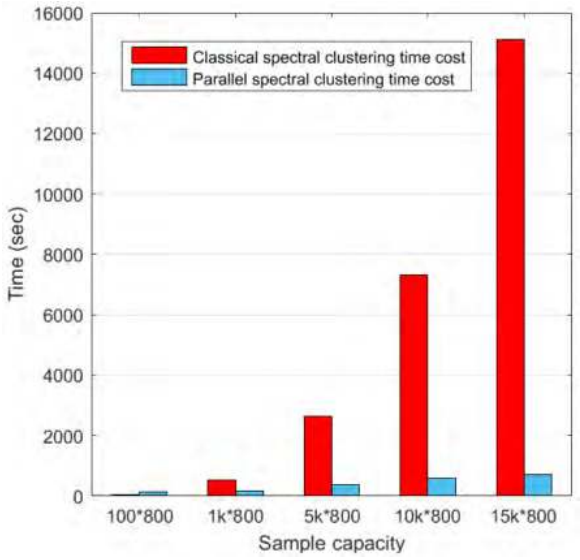

FIGURE 15. Comparison of calculation efficiency of classical and parallel spectral clustering methods.

method is proposed to analyze this relationship in a more detailed or microscopic way.

III. PREDICTION OF SHORT-TERM FUEL CONSUMPTION USING LSTM

The clustering-based method proposed in Section II above can only provide relatively long-term (25 to 35 minutes) assessment of the impact of a driver's behavior on fuel consumption. When attempting to perform relatively short-term prediction (30 seconds to 5 minutes), the clustering-based method does not work well for classifying driving behavior according to fuel efficiency. Besides, our clustering method is, in fact, a kind of classifier, so it has no prediction ability. Therefore, in this section we propose the use of a time series learning method (an LSTM network) to model the relationship between driving behavior and fuel consumption, allowing us to predict the short-term fuel consumption state of a driver's behavior. As a driving behavior pattern represents the driver's interaction with a dynamic driving environment, and fuel consumption can be treated as the cost result of this process, in this section we add dynamic driving environment information to our learning data. In the series data construction process described in this section, we first explain how we coded driving environment factors into a digital form. Then the environmental feature data and the behavior data are integrated into time-series data using a sliding window. Fuel consumption state will be the label for the constructed time-series data set. The LSTM-based model is then trained using the time-series data. The model's classification performance and prediction accuracy will be discussed at the end of this section.

A. TIME-SERIES DATA CONSTRUCTION

1) CODING OF ENVIRONMENTAL FACTORS

As explained in our previous study [49], we divided the environmental factors into two categories, dynamic environmental features (other vehicles, brake lights of leading vehicles, pedestrians, etc.) and static environmental features (features which remain invariable for relatively long periods of time, including road structures such as intersections and curves). The driving environment factors used for training our model are shown in Table 4. Some of the dynamic features are captured by a camera mounted on the vehicle. As shown in Figs. 2 and 16, two types of roads were used in this study. In Fig. 16, the gray car is the experimental vehicle, the red vehicle is the leading vehicle or leading vehicle in the right lane, the blue vehicle is a parked vehicle, the green vehicle is the first on-coming vehicle in the opposite lane and the yellow vehicle is the second on-coming vehicle in the opposite lane. In ordinary-road scenes (one lane in each direction), the motorcycle or motorbike and the pedestrian are also considered to be environmental factors which can affect the driver's behavior. Thanks to the development of object detection technology, we can easily extract these traffic environment factors. In this study we used YOLOv3 [50], a deep learning-based, real-time object detection method, to obtain the relative positions of these traffic factors. Using this position information, we can code the traffic factors into a digital form.

Examples of the raw output of the YOLO network are shown in the two images on the left of Fig. 17. Environmental factors beside the road which will not affect driving behavior are also detected by YOLO. As the camera position is fixed,

TABLE 4. Driving environment factors considered in the training data.

Dynamic features	On-road traffic factors	Leading vehicle's brake lights
		Position of the vehicle in the right lane
		Position of the vehicle in the left lane
		Positions of parked vehicles
		Position of merging vehicle
		Positions of pedestrian & bicycles
Static features	Road structure	Curves
		Uncontrolled intersections
		Controlled intersections

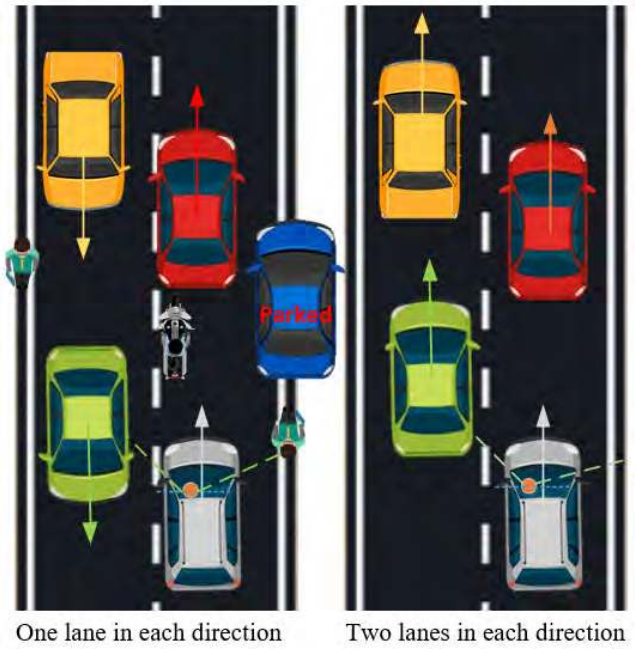

FIGURE 16. Road types and dynamic traffic factors considered in this study.

a lane detection program can be used to determine lane position. Using the lane boundary indicator (blue dotted line shown in upper left image of Fig. 17), we can remove the detected environmental factors which are not located within the range of the road lane. The other noise in YOLO's output is the multi-bounding box. We first identify the unneeded multi-bounding boxes by comparing the center points of

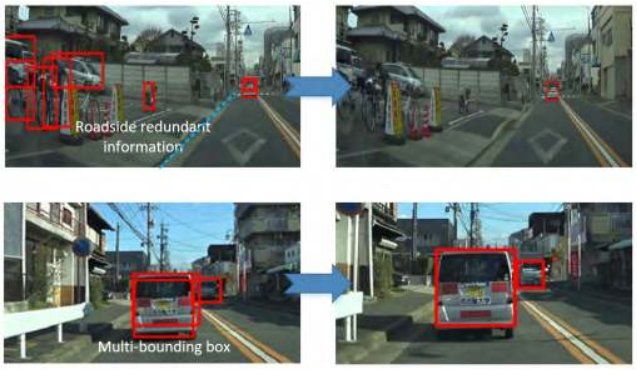

FIGURE 17. Correction of raw YOLO output.

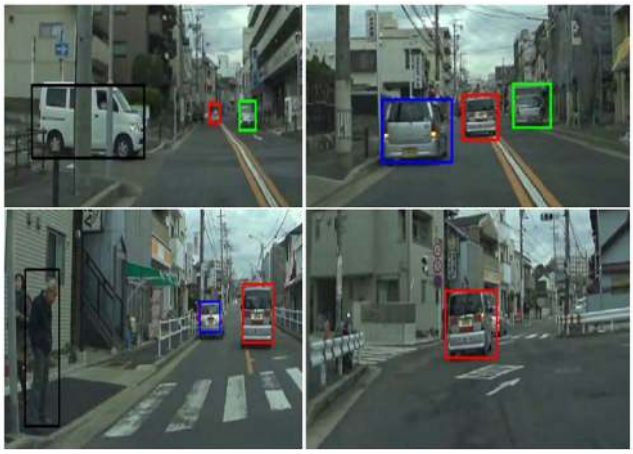

FIGURE 18. The classified traffic factors label for the object detected by YOLO.

each box, and then remove the box with the lower confidence rating.

After removing the redundant roadside data and the unneeded bounding boxes, we classify the environmental factors, using the feature categories listed in Table 4, according to their positions in the camera image, as shown in Fig. 18.

In our previous study [46], we discovered that providing the positions of the detected environmental factors helps the LSTM learn driving behavior more effectively. So, in this study, we use the same method to change the continuous positions of traffic objects into discrete locations using a mapping grid. As shown in Fig. 19, the positions of traffic factors, such as the vehicles in the photo, are labeled as the belonging to an area or zone, in this case areas A2 and B1. The size of each object is labeled according to the length of the yellow line under the object.

2) FUEL CONSUMPTION FEATURE LABELING AND TIME SERIES DATA CONSTRUCTION

In (16), B_T represents the driving behavior data set from one trip along the fixed driving route, while S represents the size of the data (the number of behavior data points) collected during the time period it took to complete the route. S is calculated by applying the method shown in Fig. 7 (compression of all of the data sets into the same size).

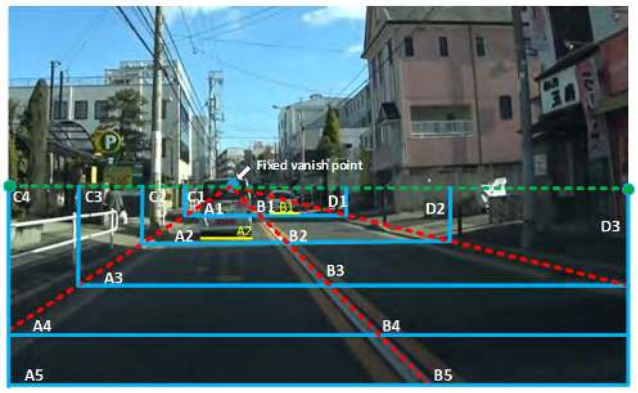

FIGURE 19. Position zones for identifying the locations of traffic factors [49].

The only difference in compressing process used in this section is that here, we divide the experimental road into 150 segments instead of 50 in order to obtain much more detailed data features. N in (16) represents the driving behavior categories strongly and moderately correlated with fuel consumption (N = 6, which are listed in Fig. 6).

$$B_T = \begin{pmatrix} b_{1,1} & \cdots & b_{1,S} \\ \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & b_{N,S} \end{pmatrix}_{N\times S} \tag{16}$$

In (17), E_T represents the environmental data set from one trip along the driving route. S is the size (number of environmental data points) of the environmental data collected during the period of time it takes to complete one circuit of the driving route. M represents the environmental factor number from the list in Table 3 (M = 13).

$$E_T = \begin{pmatrix} e_{1,1} & \cdots & e_{1,S} \\ \vdots & \ddots & \vdots \\ e_{M,1} & \cdots & e_{M,S} \end{pmatrix}_{M\times S} \tag{17}$$

When collecting the driving data, in addition to the camera frames we also collect the driving behavior data associated with each frame simultaneously, so that each set of behavior data corresponds to one camera frame. This allows us to integrate driving behavior data set B_T and environmental data set E_T into a single dataset X_T:

$$X_T = \begin{pmatrix} b_{1,1} & \cdots & b_{1,S} \\ \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & b_{N,S} \\ e_{1,1} & \cdots & e_{1,S} \\ \vdots & \ddots & \vdots \\ e_{M,1} & \cdots & e_{M,S} \end{pmatrix}_{(M+N)\times S} \tag{18}$$

Fuel consumption F can then be calculated as follows:

$$\mathrm{F}(X_T) = \{F_1, F_2, \ldots, F_i, \ldots, F_I\} \tag{19}$$

Function f(x) in (20) and (21) represents the hypothetical equation which describes the nonlinear relationship between driving behavior, driving environment and fuel consumption features. To deduce function $f(x)$ would be relatively difficult, so here we treat $f(x)$ as a 'black box', so our LSTM-based method is applied to simulate the computations of this 'black box'. As the input for the LSTM should be data in a time-series format, the raw training data must first be converted into time-series data. As shown in Fig. 20, we use a sliding window to construct each set of time-series data, and the data label is each data segment's fuel consumption F_i. The window size is 50 data points and the size of the sliding step is 15 data points, so in (21), step = 15 and j = 50. F_i is mapped into the data segment's distribution to obtain its ranking level. For example, in Fig. 20, F_1 belongs to the low fuel consumption level (marked with dotted points), so the "time-series data 1" will be labeled as "low fuel consumption". The green, yellow, and red labels represent the low, medium, and high fuel consumption group respectively. The fuel consumption group is judged by the other driver's historical records.

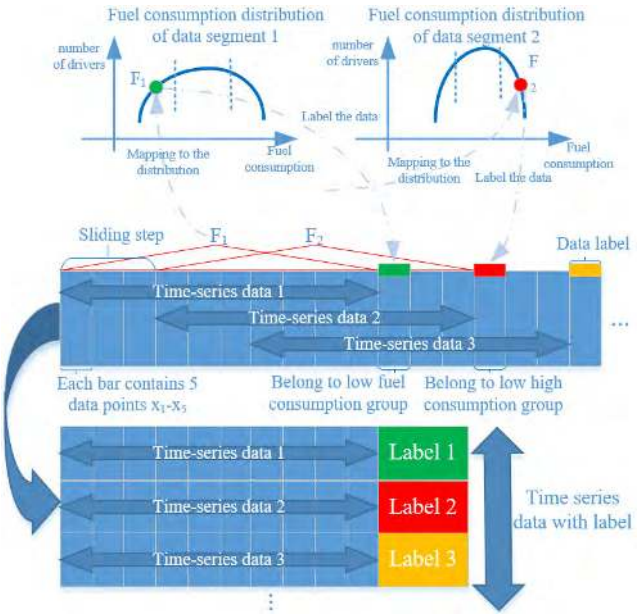

FIGURE 20. Time series data composition using sliding window.

$$F_1 = f\begin{pmatrix} b_{1,1} & \cdots & b_{1,j} \\ \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & b_{N,j} \\ e_{1,1} & \cdots & e_{1,j} \\ \vdots & \ddots & \vdots \\ e_{M,1} & \cdots & e_{M,j} \end{pmatrix} \tag{20}$$

$$F_i = f\begin{pmatrix} b_{1,1+(i-1)\times step} & \cdots & b_{1,(i-1)\times step+j} \\ \vdots & \ddots & \vdots \\ b_{N,1+(i-1)\times step} & \cdots & b_{N,(i-1)\times step+j} \\ e_{1,1+(i-1)\times step} & \cdots & e_{1,(i-1)\times step+j} \\ \vdots & \ddots & \vdots \\ e_{M,1+(i-1)\times step} & \cdots & e_{M,(i-1)\times step+j} \end{pmatrix} \tag{21}$$

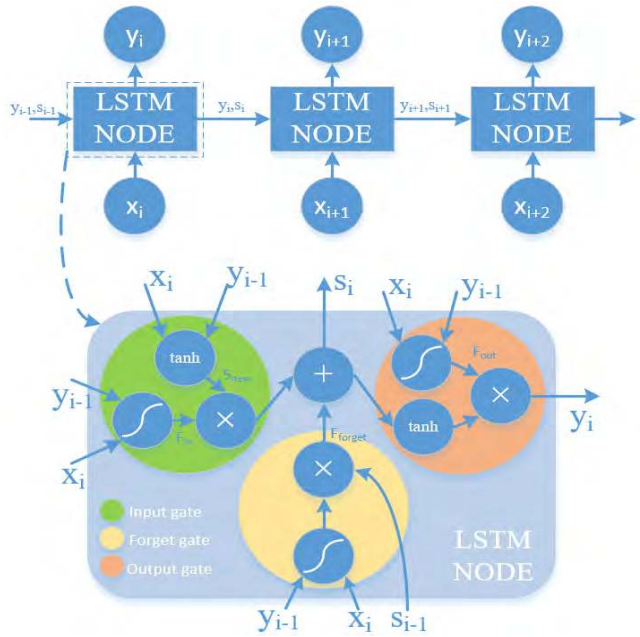

FIGURE 21. The unfolded structure of the LSTM network and the inner composition of an LSTM node.

The boundaries of the fuel consumption levels are defined by the trisection lines, and the equation for calculating the boundaries is shown in (22). Results (0,0,1), (0,1,0) and (1,0,0) represent low, moderate and high fuel consumption, respectively. F_i is the current data segment's fuel consumption and $F_{avg,i}$ is the expected value of the remaining driving process data in that data segment.

$$I_i = \begin{cases} (0,0,1), & F_i < 0.6F_{avg,i} \\ (0,1,0), & 0.6F_{avg,i} < F_i < 1.2F_{avg,i} \\ (1,0,0), & F_i > 1.2F_{avg,i} \end{cases} \tag{22}$$

After completing the labeling process, we can obtain our training data with fuel consumption feature labels. The labels are not only obtained by calculating detailed fuel consumption, but also obtained by comparing the fuel consumption distribution with all of the other drivers' fuel consumption distributions.

B. FUEL CONSUMPTION PREDICTION MODELING BASED ON LSTM

1) LSTM COMPONENTS AND THEIR MATHEMATICAL EXPRESSIONS

As the state of the art in information processing and behavior modeling, LSTM is widely used in machine translation [51], speech recognition [52], driving behavior analysis [53], and other applications. LSTM is in fact a kind of Recurrent Neural Network (RNN) [33, 54]. Standard RNNs usually suffer from the vanishing gradient problem, but LSTMs include a 'forget gate', which can prevent backpropagation errors from vanishing or exploding. The structure of the LSTM used in this study is shown in Fig. 21.

An LSTM is a recurrent network which produces a state as its output, and the state of current network is passed on to the next step in the network for further calculation. As shown in Fig. 21, each node of the LSTM network is composed of three main components, a 'forget gate', an 'input gate' and an 'output gate'. The 'forget gate' determines the effect of the information from the previous step on the calculations of the current network, which is the key feature of the LSTM, allowing it to avoid the problems of gradient vanishing or exploding. The function of the 'forget gate' can be expressed mathematically as follows:

$$F_{\text{forget}} = \sigma\left(W_f \cdot [y_{i-1}, x_i] + b_f\right) \tag{23}$$

$$\sigma(\mathrm{x}) = \frac{1}{1+e^{-x}} \tag{24}$$

As $\sigma(\mathrm{x})$ is a sigmiod function, F_{forget} is always smaller than 1. Furthermore, F_{forget} will be multiplied by previous network state S_{i-1} to form part of the new state S_i, so F_{forget} determines how much state S_{i-1} will affect current network state S_i.

The second part of the LSTM is the 'input gate', which mainly decides what should be newly added to the current network state. First, we should find which part of the previous state should be updated, so we use the following equations to define the update procedure:

$$F_{in} = \sigma\left(W_i \cdot [y_{i-1}, x_i] + b_i\right) \tag{25}$$

And then the updated value can be determined as follows:

$$S_{new} = \tanh\left(W_{new} \cdot [y_{i-1}, x_i] + b_{new}\right) \tag{26}$$

Current network state S_i can be obtained from the updated state value and the remaining previous network state:

$$S_i = F_{in} \times S_{new} + F_{\text{forget}} \tag{27}$$

The third part of the LSTM is the 'output gate', which uses current network state S_i to generate the final output. Using current inner state S_i, we decide which data we can output, then the data is multiplied by F_{out} (which ranges from 0 to 1) to determine which data can be output. The calculation process is shown in the following equation:

$$y_i = \tanh\left(S_i\right) \times \sigma\left(W_o \cdot [y_{i-1}, x_i] + b_o\right) \tag{28}$$

In this paper, input $x_i = X_T$ in (18), and the size of X_T, which is defined by the sliding window in Fig. 20, is 50.

2) LSTM NETWORK TRAINING PROCESS

First, we need to pre-process the training data. All of the time-series data is normalized into a range of 0 to 1. We code each data set's label into a one-hot form: high fuel consumption is $(0, 0, 1)^T$, medium fuel consumption is $(0, 1, 0)^T$ and low fuel consumption is $(1, 0, 0)^T$.

To build the LSTM network, we used TensorFlow [55], which is an end-to-end open source software platform for machine learning. The LSTM block is based on the LSTM node unit ''tf.nn.rnn_cell.LSTMCell'' [56] which is provided

TABLE 5. Hyper-parameters and the training strategy of the LSTM network.

	Parameter name	Parameter value
Hyper-parameters	Number of units in the LSTM's hidden layers	125, 150, 200
	Number of hidden layers in the LSTM	2
	Batch size	64
	Initial forget bias	1
	Initial learning rate	0.005
Training strategy	Optimizer	Adam Optimizer [57]
	Loss function	Sparse Softmax cross entropy

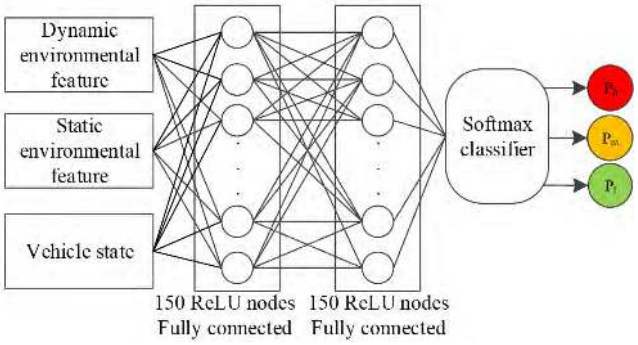

FIGURE 22. Structure of the neural network with two hidden layers, each of which contains 150 ReLU nodes.

by the TensorFlow API. The hyper-parameters and the training strategy of the LSTM network are shown in Table 5.

The output of the LSTM is put into a Softmax classifier, which calculates its probability of belonging to each class. The Softmax function can convert the output of the LSTM into a range from 0 to 1. The mathematical expression of the Softmax function is as follows:

$$C_i = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{29}$$

where C_i is the output confidence rate, i.e., the dataset's probability of belonging to a certain fuel consumption group.

C. RESULTS OF FUEL CONSUMPTION PREDICTION USING LSTM

1) TRAINING DATA

The entire data set is divided into six groups randomly, with each group containing 5,000 data points of time-series data. The six groups of data are divided as follows: four groups

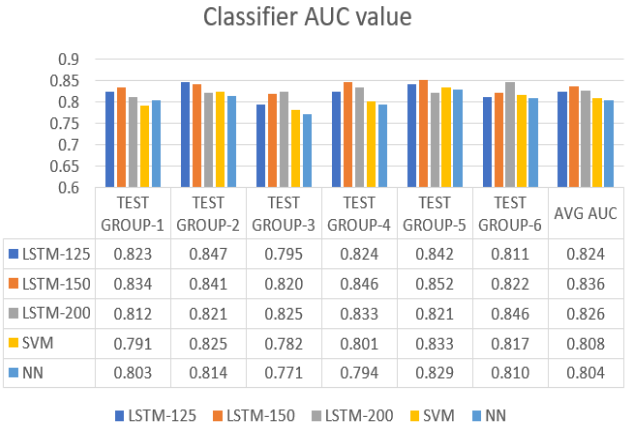

FIGURE 23. AUC values for each modeling method and each test group.

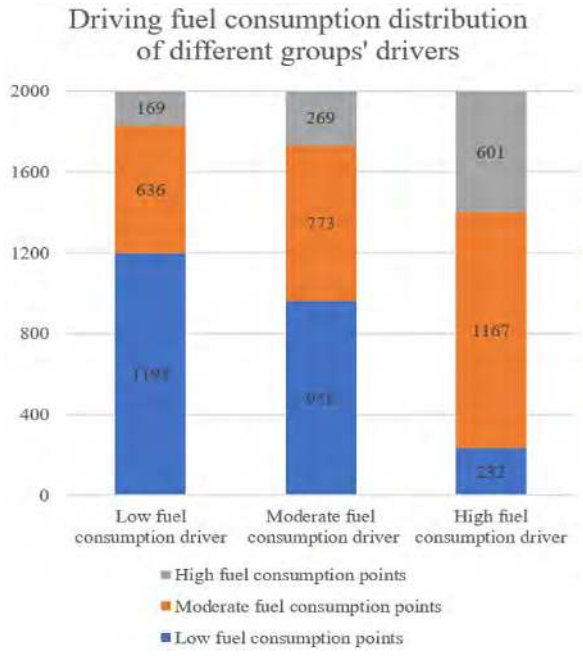

FIGURE 24. Fuel consumption data distributions of three representative drivers.

are used for training, one group is used for validation and one group is used for testing. Because the training process involves cross-validation, each group will be treated as a training data group, a validation group or a testing group.

2) COMPARISON OF LSTM PREDICTION RESULTS WITH THOSE OF OTHER MACHINE LEARNING METHODS

We compared the performance of two other machine learning methods with the performance of the proposed LSTM-based method. One of those methods was kernel-based Support Vector Machine (SVM) [58], and the other was a multi-layer neural network. In addition, LSTM networks with different number of nodes were also evaluated.

SVM is a very powerful machine learning method which maps the objects to be sorted into high-dimensional feature

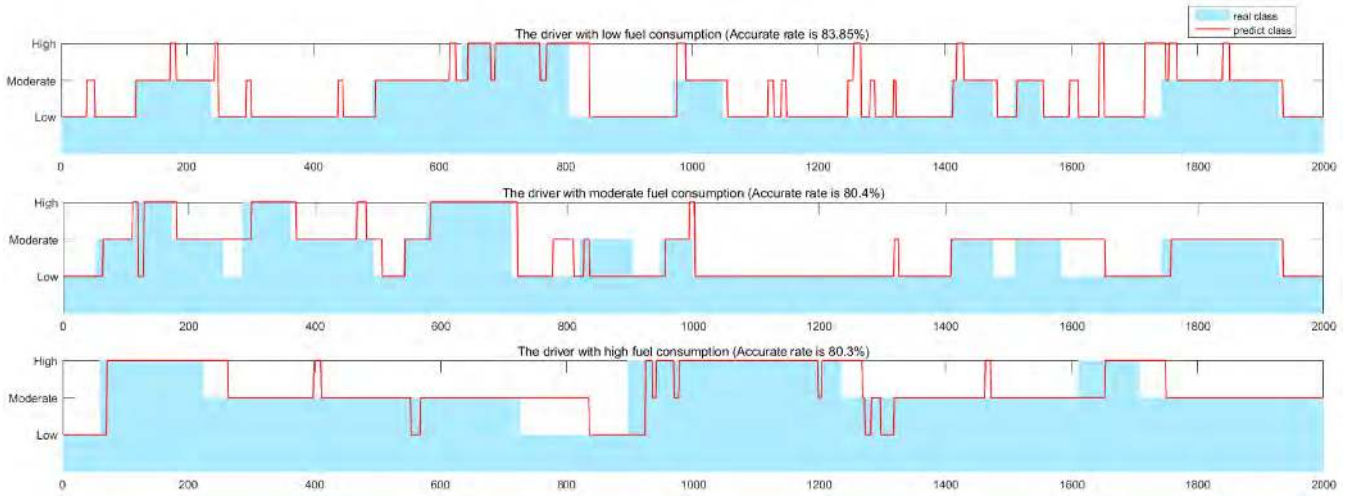

FIGURE 25. Short-term fuel consumption prediction performance using LSTM-based classifier and fuel consumption features for three representative drivers.

spaces. It is widely used for semantic parsing [59], image segmentation [60], facial recognition [61] and other applications. MATLAB's Statistics and Machine Learning Toolbox [62] was used to construct our SVM-based classifier.

The multi-layer neural network we used had two hidden layers, and each layer contained 150 rectified linear units (ReLU), as shown in Fig. 22. The output of the network is passed into a Softmax layer, and the probabilities of the data belonging to each of the three fuel consumption categories are calculated.

Two criteria were considered in our evaluation, the classifier accuracy rate and the area under the curve (AUC) of receiver operating characteristics (ROC) [63]. The classifier accuracy rate is a direct index which can be used to judge the performance of the prediction model, however it cannot evaluate the classification performance of the model. AUC is a probability value, which is the general standard for evaluating classifier performance. In Fig. 23 we show each classifier's performance for each of the six testing groups. We can see in Fig. 24 that the LSTM with 150 nodes achieved the best overall performance.

Next, we experimentally evaluated the short-term fuel consumption estimation performance of our proposed LSTM-based prediction method. Three representative drivers who belonged to different fuel consumption groups were selected to test the performance of our deep learning-based predictor. The fuel consumption data distributions for these three drivers are shown in Fig. 24.

The LSTM-based classifier's prediction accuracies for these three drivers are illustrated in Fig. 25.

The red lines represent the predicted fuel consumption category based on the driver's fuel consumption features over time, while the light blue bars represent the actual distribution of the fuel consumption features corresponding to the driver's behavior. The average prediction accuracy for the three selected drivers was 81%.

IV. DISCUSSION AND CONCLUSION

In this paper, we first used the unsupervised machine learning method of spectral clustering to classify drivers into three groups using six driving behavior-based fuel consumption features. We then analyzed the macro-behavior of each group, focusing on power demand (speed and acceleration) and control stability (variation in speed and acceleration). Our results showed that the proposed spectral clustering-based method could accurately identify drivers with different fuel consumption profiles, and clearly modeled the relationship between the real-world driving data and the corresponding fuel consumption features.

In addition to the estimation of fuel consumption using vehicle operation data, we also performed a qualitative analyses of driving behavior, as shown in Fig. 13. Speed and acceleration information reveal the amount of power demanded by a driver, while variance in speed and acceleration represent the range of dynamic control exercised by drivers [25], [26]. The results of our analysis showed that high fuel consumption drivers (those in the red cluster) tend to maintain a relatively steady, high demand for power, while their dynamic control of the vehicle is less stable. Their acceleration rates are higher and their pedal control behavior is less stable compared to drivers in the low fuel consumption cluster. Drivers in the median yellow cluster showed the lowest speed distribution, but their gas and brake pedal operation characteristics were similar to those of the low efficiency drivers in the red cluster. Drivers in the blue cluster had the lowest fuel consumption, since they tended to maintain a consistent speed, and their dynamic control of the vehicle was the most stable among the three groups. We also compared the spectral cluster method with other state of art clustering method such as k-means and KFCM. As show in Table.3, spectral cluster method can achieve the best clustering performance of the three methods.

However, there were drawbacks to our proposed method, in that the spectral clustering-based method requires

relatively long-term data to produce accurate classification results. So, for real-time and short-term fuel consumption feature prediction, this unsupervised method is not appropriate. Furthermore, the results of data mining can only show the impact of a driver's behavior on fuel consumption on a macro-level.

Therefore, in the second stage of our study we attempted to use a supervised machine learning-based LSTM method to build a link between short-term driving data and the fuel consumption features. The proposed LSTM-based model was able to accurately predict driver behavior, achieving a maximum AUC of 0.836, which is considered to be good human behavior prediction performance [64]. As shown in Fig. 23, the LSTM-based method achieved better classification performance than the SVM or NN-based methods. LSTM networks with different numbers of hidden nodes were also evaluated in this study, revealing that the LSTM with 150 hidden nodes achieved the best average AUC, compared to LSTMs with 125 or 200 hidden nodes. Three representative drivers were then selected for a more detailed evaluation of the model's performance. As shown in Fig. 25, the short-term fuel consumption performance of the drivers could be accurately predicted using the proposed method, although some prediction error did occur. However, an average overall prediction accuracy of more than 80% was achieved.The whole prediction process is end-to-end, as the input of the model is the driving behavior and dynamic traffic condition data. After the raw data is reformatted and then processed by the model, the output is a prediction of which fuel consumption group a particular driver belongs to.

In conclusion, we made three contributions in this paper; firstly, we propose a clustering-based data-mining method which can analyze the behavior and its fuel consumption result in a macro view. The method can serve as a group behavior assessment mechanism for the public transportation department or the commercial transportation company to evaluate the energy cost distribution. Secondly, we also propose a micro fuel consumption evaluation model by learning the driving behavior. The model shows good prediction ability which can be integrated into the ADAS system or the eco-driving coach system to evaluate and obtain the fuel-cost behavior of the single drivers. The predicted state can make the ADAS or eco-driving system give more reasonable and adaptive fuel-efficient driving strategy or detail manipulation. Thirdly, we widen the deep learning method's application area, to our knowledge, it is the first time that the deep learning method is used for learning the driving behavior's impact on fuel consumption feature.

There are some limitations in our study and in our proposed method. First, the shortcomings of the collected data will mainly affect the deep-learning based method. As the collected data are collected from two kinds of road and the traffic environment factors are not all coded into the time-series data, so the LSTM can just learn the limited feature from the fixed traffic condition and the environment it ever meet. When facing different road types, for example the road with four lanes, it will suffer prediction performance decreasing. Second, the prediction accuracy of the proposed LSTM-based method was not extremely high. We suspect this is mainly because the model input information included a limited number of traffic conditions, and because the form of this input information was relatively basic. As a result, the LSTM could not accurately predict fuel consumption in very complex or unknown situations. And our deep-learning based method is the model can just predict the fuel consumption level of the driving process so it is hard to give more detail fuel cost information. What's more, compared with other state of the art behavior prediction method, LSTM or deep learning network need lots of training data and training time. If other new behavior factors which affect the fuel consumption need to be added into the network, the model need to revised the original parameter and training process should be reprocessed. This will limit the generality of the model. Third, we only used one type of experimental vehicle, so we need to do further research to determine whether the proposed LSTM-based model can be adapted to other types of vehicles. At last, the driver's personal feature such as age, sex, driving experiences and so on, are not further studied in this study.

So, in our future work, firstly we aim to use larger scale naturalistic driving data to make our prediction model with more robustness. Then the other factors' effect, such as group personality feature or vehicle type, on the fuel consumption analysis should also be studied in order to make the fuel consumption prediction model more general.

ACKNOWLEDGMENT

The authors would like to thank the Vehicle Engineering Development Division of Mitsubishi Motors and Functional Safety Department of UISEE Technologies Beijing Company, Ltd for their valuable research assistance.

REFERENCES

[1] C. Sun, Y. Luo, and J. Li, "Urban traffic infrastructure investment and air pollution: Evidence from the 83 cities in China," *J. Cleaner Prod.*, vol. 172, pp. 488–496, Jan. 2018.
[2] H. Zhang, S. Wang, J. Hao, X. Wang, S. Wang, F. Chai, and M. Li, "Air pollution and control action in Beijing," *J. Cleaner Prod.*, vol. 112, pp. 1519–1527, Jan. 2016.
[3] *Health in the Green Economy: Health Co-Benefits of Climate Change Mitigation*, World Health Org., Geneva, Switzerland, 2011.
[4] D. Fang, Q. Wang, H. Li, Y. Yu, Y. Lu, and X. Qian, "Mortality effects assessment of ambient $PM_{2.5}$ pollution in the 74 leading cities of China," *Sci. Total Environ.*, vols. 569–570, pp. 1545–1552, Nov. 2016.
[5] H. Liimatainen, "Utilization of fuel consumption data in an ecodriving incentive system for heavy-duty vehicle drivers," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1087–1095, Dec. 2011.
[6] J. E. Meseguer, C. T. Calafate, J. C. Cano, and P. Manzoni, "Assessing the impact of driving behavior on instantaneous fuel consumption," in *Proc. 12th Annu. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2015, pp. 443–448.
[7] C. D'Agostino, A. Saidi, G. Scouarnec, and L. Chen, "Rational truck driving and its correlated driving features in extra-urban areas," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 1199–1204.
[8] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 4, pp. 6–22, Oct. 2014.
[9] J. N. Barkenbus, "Eco-driving: An overlooked climate change initiative," *Energy Policy*, vol. 38, no. 2, pp. 762–769, 2010.

[10] M. J. M. Sullman, L. Dorn, and P. Niemi, "Eco-driving training of professional bus drivers—Does it work?" *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 749–759, Sep. 2015.

[11] C.-H. Lee and C.-H. Wu, "An incremental learning technique for detecting driving behaviors using collected EV big data," in *Proc. ASE BigData SocialInform.*, 2015, p. 10.

[12] J. Négre and P. Delhomme, "Drivers' self-perceptions about being an eco-driver according to their concern for the environment, beliefs on eco-driving, and driving behavior," *Transp. Res. A, Policy Pract.*, vol. 105, pp. 95–105, Nov. 2017.

[13] J. L. Jimnez-Palacios, "Understanding and quantifying motor vehicle emissions with vehicle specific power and TILDAS remote sensing," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1999.

[14] M. Barth, F. An, T. Younglove, G. Scora, C. Levine, M. Ross, and T. Wenzel, "NCHRP PROJECT 25-11: Development of a comprehensive modal emissions model," in *Proc. 7th CRC Road Vehicle Emissions Workshop*, 2000, pp. 1–6.

[15] N. Davis, J. Lents, M. Osses, N. Nikkila, and M. Barth, "Development and application of an international vehicle emissions model," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1939, no. 1, pp. 156–165, 2005.

[16] Z. Xu, T. Wei, S. Easa, X. Zhao, and X. Qu, "Modeling relationship between truck fuel consumption and driving behavior using data from Internet of vehicles," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 3, pp. 209–219, 2018.

[17] G. Xu, L. Liu, Y. Ou, and Z. Song, "Dynamic modeling of Driver control strategy of lane-change behavior and trajectory planning for collision prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1138–1155, Sep. 2012.

[18] Z. Zheng, "Recent developments and research needs in modeling lane changing," *Transp. Res. B, Methodol.*, vol. 60, pp. 16–32, Feb. 2014.

[19] H. Xia, K. Boriboonsomsin, and M. Barth, "Dynamic ECO-driving for signalized arterial corridors and its indirect network-wide energy/emissions benefits," *J. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 31–41, 2013.

[20] X. Xiang, K. Zhou, W. B. Zhang, W. Qin, and Q. Mao, "A closed-loop speed advisory model with driver's behavior adaptability for eco-driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3313–3324, Dec. 2015.

[21] J. E. Meseguer, C. K. Toh, C. T. Calafate, J. C. Cano, and P. Manzoni, "Drivingstyles: A mobile platform for driving styles and fuel consumption characterization," *J. Commun. Netw.*, vol. 19, no. 2, pp. 162–168, 2017.

[22] E. Gilman, A. Keskinarkaus, S. Tamminen, S. Pirttikangas, J. Röning, and J. Riekki, "Personalised assistance for fuel-efficient driving," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 681–705, Sep. 2015.

[23] R. Trigui, S. Javanmardi, E. N. Bourles, H. Tattegrain, E. Bideaux, and J. F. Trégouët, "Driving style modelling for eco-driving applications," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 13866–13871, 2017.

[24] C. Lv, X. Hu, A. Sangiovanni-Vincentelli, Y. Li, C. M. Martinez, and D. Cao, "Driving-style-based codesign optimization of an automated electric vehicle: A cyber-physical system approach," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 2965–2975, Apr. 2019.

[25] A. E. Wåhlberg, "Long-term effects of training in economical driving: Fuel consumption, accidents, driver acceleration behavior and technical feedback," *Int. J. Ind. Ergonom.*, vol. 37, no. 4, pp. 333–343, 2007.

[26] E. Ericsson, "Variability in urban driving patterns," *Transp. Res. D, Transp. Environ.*, vol. 5, no. 5, pp. 337–354, 2000.

[27] M. Ehsani, A. Ahmadi, and D. Fadai, "Modeling of vehicle fuel consumption and carbon dioxide emission in road transport," *Renew. Sustain. Energy Rev.*, vol. 53, pp. 1638–1648, Jan. 2016.

[28] J. Rios-Torres, J. Liu, and A. Khattak, "Fuel consumption for various driving styles in conventional and hybrid electric vehicles: Integrating driving cycle predictions with fuel consumption optimization," *Int. J. Sustain. Transp.*, vol. 13, no. 2, pp. 123–137, 2018.

[29] C.-H. Lee and C.-H. Wu, "A novel big data modeling method for improving driving range estimation of EVs," *IEEE Access*, vol. 3, pp. 1980–1993, 2015.

[30] J. Wu, Y. Du, G. Qi, and M. Xu, "Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis," *IET Intell. Transp. Syst.*, vol. 9, no. 8, pp. 792–801, 2015.

[31] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving style analysis using data mining techniques," *Int. J. Comput. Commun. Control*, vol. 5, no. 5, pp. 654–663, 2010.

[32] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] L. Zeng, R. Wang, Q. Han, C. Chen, L. Ye, and X. He, "Driving behavior modeling and evaluation for bus enter and leave stop process," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, USA, 2017, pp. 1–6.

[35] S. Kanarachos, J. Mathew, and M. E. Fitzpatrick, "Instantaneous vehicle fuel consumption estimation using smartphones and recurrent neural networks," *Expert Syst. Appl.*, vol. 120, pp. 436–447, Apr. 2019.

[36] B. Beusen, B. Degraeuwe, C. Beckx, T. Denys, L. Govaerts, S. Broekx, M. Gijsbers, K. Scheepers, L. I. Panis, and R. Torfs, "Using on-board logging devices to study the longer-term impact of an eco-driving course," *Transp. Res. D, Transp. Environ.*, vol. 14, no. 7, pp. 514–520, 2009.

[37] *SPSS Tutorials: Pearson Correlation*. Accessed: Jun. 19, 2019. [Online]. Available: https://libguides.library.kent.edu/SPSS/PearsonCorr

[38] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Evanston, IL, USA: Routledge, 1988. [Online]. Available: https://www.taylorfrancis.com/books/9780203771587

[39] P. Ping, W. Qin, Y. Xu, C. Miyajima, and T. Kazuya, "Spectral clustering based approach for evaluating the effect of driving behavior on fuel economy," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, Houston, TX, USA, May 2018, pp. 1–6. doi: 10.1109/I2MTC.2018.8409675.

[40] M. Stoer and F. Wagner, "A simple min-cut algorithm," *J. ACM*, vol. 44, no. 4, pp. 585–591, 1997.

[41] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 11, no. 9, pp. 1074–1085, Sep. 1992.

[42] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[43] D. Wagner and F. Wagner, "Between min cut and graph bisection," in *Proc. Int. Symp. Math. Found. Comput. Sci.*, 1993, pp. 744–750.

[44] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.

[45] G. Schofield, J. R. Chelikowsky, and Y. Saad, "A spectrum slicing method for the Kohn–Sham problem," *Comput. Phys. Commun.*, vol. 183, no. 3, pp. 497–505, 2012.

[46] *Apache Spark*. Accessed: Jun. 19, 2019. [Online]. Available: https://spark.apache.org/

[47] C. Lanczos, "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators," *J. Res. Nat. Bureau Standards*, vol. 45, no. 4, pp. 255–282, 2012.

[48] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, Feb. 2003.

[49] P. Ping, Y. Sheng, W. Qin, C. Miyajima, and K. Takeda, "Modeling driver risk perception on city roads using deep learning," *IEEE Access*, vol. 6, pp. 68850–68866, 2018.

[50] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[51] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arxiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[52] S. Han, Y. Wang, H. Yang, W. J. Dally, J. Kang, H. Mao, Y. Hu, X. Li, Y. Li, D. Xie, H. Luo, and S. Yao, "ESE: Efficient speech recognition engine with sparse LSTM on FPGA," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays (FPGA)*, 2017, pp. 75–84.

[53] J. Morton, T. A. Wheeler, and M. J. Kochenderfer, "Analysis of recurrent neural networks for probabilistic modeling of driver behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1289–1298, May 2017.

[54] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[55] TensorFlow. (2019). *TensorFlow Tutorials*. [Online]. Available: https://tensorflow.google.cn/tutorials/

[56] TensorFlow. (2019). *Class LSTMCell*. [Online]. Available: https://www.tensorflow.org/versions/r1.13/api_docs/python/tf/nn/rnn_cell/LSTMCell?hl=en#class_lstmcell

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980v8*. [Online]. Available: https://arxiv.org/abs/1412.6980v8

[58] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[59] R. J. Kate and R. J. Mooney, "Semi-supervised learning for semantic parsing using support vector machines," in *Proc. NAACL-Short Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2007, pp. 81–84.
[60] M. Song and D. Civco, "Road extraction using SVM and image segmentation," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 12, pp. 1365–1371, 2004.
[61] G. Guo, S. Z. Li, and K. Chan, "Face recognition by support vector machines," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2000, pp. 196–201.
[62] *Support Vector Machines for Binary Classification*. Accessed: Jun. 19, 2019. [Online]. Available: https://ww2.mathworks.cn/help/stats/support-vector-machines-for-binary-classification.html
[63] C. X. Ling, J. Huang, and H. Zhang, "AUC: A better measure than accuracy in comparing learning algorithms," in *Proc. 16th Conf. Can. Soc. Comput. Stud. Intell. AI*, vol. 2671, Jun. 2003, pp. 329–341. [Online]. Available: https://www.researchgate.net/publication/221442229_AUC_A_Better_Measure_than_Accuracy_in_Comparing_Learning_Algorithms
[64] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r," *Law Hum. Behav.*, vol. 29, no. 5, pp. 615–620, 2005.

PENG PING received the B.S. degree in automation from the Beijing University of Chemical Technology, Beijing, China, in 2010, and the M.S. degree in automation from the Nanjing University of Science and Technology, Nanjing, China, in 2013. He is currently pursuing the Ph.D. degree with Southeast University, Nanjing. From 2013 to 2015, he was an R&D Engineer with the Cloud Switch Group, Huawei Technologies Company, Ltd. Since 2017, he has been a Joint Ph.D. Student with Nagoya University. His research interests include vehicle safety, data-mining, cloud computing, and eco-driving.

WENHU QIN received the Ph.D. degree from Southeast University, Nanjing, China, in 2005. He is currently a Professor with the School of Instrument Science and Engineering, Southeast University, where he has been on the faculty, since 1997. He directs the Vehicle Safety and Virtual Reality Laboratory, Southeast University. He has authored or coauthored over 30 journal papers, ten conference papers, and a book. He holds the three patents. His research interests include vehicle safety, virtual reality, crowd simulation, and road traffic accident reconstruction.

YANG XU received the B.S. degree in instrument science from the East China University of Technology, Nanchang, China, in 2012, and the M.S. degree in instrument science from the Hefei University of Technology, Hefei, China, in 2015. He is currently pursuing the Ph.D. degree with Southeast University, Nanjing, China. Since 2018, he has been a joint Ph.D. Student with The University of Queensland. His research interests include machine learning, data science, bio-medical signal processing, and human–computer interaction.

CHIYOMI MIYAJIMA received the B.E., M.E., and Ph.D. degrees in computer science from the Nagoya Institute of Technology, Japan, in 1996, 1998, and 2001, respectively, where she was a Research Associate with the Department of Computer Science, from 2001 to 2003. She was a Designated Associate Professor with the Graduate School of Information Science, Nagoya University, Japan, from 2003 to 2016, where she was an Associate Professor with the Institutes of Innovation for Future Society, from 2016 to 2018. Since 2018, she has been an Associate Professor with Daido University, Nagoya, Japan. Her research interest includes the analysis and the modeling of driver behavior.

KAZUYA TAKEDA received the B.E.E., M.E.E., and Ph.D. degrees from Nagoya University, Japan, in 1983, 1985, and 1994, respectively. Since 1985, he has been with Advanced Telecommunication Research Laboratories and KDD R&D Laboratories, Japan. In 1995, he started a research group for signal processing applications at Nagoya University, where he is currently a Professor with the Institutes of Innovation for Future Society. His research interests include investigating driving behavior using data centric approaches and utilizing signal corpora of real driving behavior. He is also a member of the Board of Governors of the IEEE Intelligent Transportation Systems Society.