Light-Duty Vehicle Trip Classification using One-Class Novelty Detection and Exhaustive Feature Extraction

Lei Zhu*
Senior Member, IEEE
Department of Systems Engineering & Engineering Management
University of North Carolina at Charlotte
Charlotte, USA
Lei.Zhu@uncc.edu

Brennan Borlaug
Center for Integrated Mobility Sciences
National Renewable Energy Laboratory
Golden, USA
brennan.borlaug@nrel.gov

Lei Lin
Member, IEEE
Goergen Institute for Data Science
University of Rochester
Rochester, USA
Lei.Lin@rochester.edu

Jacob Holden
Center for Integrated Mobility Sciences
National Renewable Energy Laboratory
Golden, USA
jacob.holden@nrel.gov

Jeffrey Gonder
Center for Integrated Mobility Sciences
National Renewable Energy Laboratory
Golden, USA
Jeff.Gonder@nrel.gov

***Abstract*—Travel mode classification within travel survey data sets, especially light-duty vehicle (LDV) trips, is foundational, though nontrivial, to emerging mobility systems, travel behavior analysis, and fuel consumption estimation. Current travel mode detection approaches require well-sampled and balanced data sets with ground truth travel mode labels. The detection approaches are rarely applied and validated on large-scale, real-world data sets, which may not satisfy the dataset requirements. This paper proposes an LDV trip detection model as a supplement to current travel mode detection methods, for the case when the training set is highly (and/or completely) unbalanced, to the extent that classical machine-learning approaches become difficult or impossible to deploy. The proposed model uses a novelty detection technique—one-class support vector machines (OCSVMs)—and a novel exhaustive feature extraction (EFE) technique on continuous time series data (i.e., Global Positioning System [GPS] speed profiles) for single-mode trip trajectories. Training and validation of the model are conducted on a large-scale, real-world data set. The proposed method accurately identifies LDV trips from a broad set of multimodal trips by leveraging a wealth of preexisting in-vehicle GPS travel data. Additional sensitivity analysis sheds light on the optimal training size and feature selection, which will benefit applications limited by highly imbalanced data. The paper also discusses performance comparison with regular machine-learning approaches, the model's robustness, and the potential to extend the proposed model to multi-modal trip prediction.**

***Keywords*—Travel mode detection, novelty detection, one-class support vector machines, Global Positioning System, GPS.**

I. INTRODUCTION

Travel surveys for different travel modes, including light-duty vehicle (LDV) travel [1], have been consistently conducted, but over time, best-in-practice data collection techniques have evolved to computer-assisted methods with Global Positioning System (GPS) components [2-4], which reduces the burden on respondents to document every detail about their daily travel and activities [5], while providing reliable spatiotemporal data to better inform not only transportation planning [6-8] and traffic demand management [9-11], but also vehicle fuel economy simulation and estimation [12-14], which require a high level of confidence in vehicle driving cycle detection [15].

GPS devices deployed in travel surveys are generally of two types: in-vehicle (instrumented) [16, 17] or on-person (wearable) [15]. Most in-vehicle devices lack an independent power supply, instead receiving power from the vehicle's accessory socket. A benefit to in-vehicle GPS devices is that researchers can be certain that the travel mode of the collected GPS data is vehicle driving. However, these in-vehicle devices are only capable of capturing vehicle travel modes and fail to record data from other modes and personal activities. Wearable devices usually include a battery pack that allows untethered mobility for the participant and enables the collection of large-scale, multi-modal GPS travel data. Although the time and spatial characteristics of travel can be recorded with wearable GPS devices (e.g., smartphones), the travel mode information is still not directly available in many cases. In order to extract vehicle driving cycles embedded in wearable data, a travel mode classification approach for wearable GPS data is desired.

In travel surveys with a wearable GPS component, one method for obtaining travel mode labels is to ask the survey participants to remember and report their travel mode during or after the trips through paper-based questionnaires and/or computer-aided information record systems. However, this extra step introduces an additional burden to the survey participants, and data quality and accuracy are not guaranteed. An alternative way of procuring travel mode information is by applying post hoc (often algorithmic) travel mode identification methods [18, 19]. There are some techniques that have been employed for inferring travel mode based on wearable GPS data. Rule-based methods [20, 21], such as fuzzy logic applications [22], predict travel mode according to a set of rules applied to a group of trip parameters such as speed, acceleration, and distance. They are easily implemented and provide reasonable results in most cases. Machine-learning approaches generally outperform rule-based methods by applying advanced learning technologies such as neural networks [23-25], Bayesian networks, and support vector machines (SVMs) to automatically classify or predict the travel modes for wearable GPS data. But most studies only applied these techniques on well-sampled, evenly distributed and relatively small data sets [26], such as on the order of hundreds of samples [22, 23]. The performance of these methods on a large-scale, real-world test set is unknown. More importantly, conventional travel mode prediction methods typically involve the classification of two or more classes [24, 27, 28], and the training and validation data sets of these categories are well-sampled, balanced, and evenly distributed. In contrast to a

balanced and evenly distributed data set, a highly imbalanced data set, in the context of binary classification, refers to a sample that is significantly dominated by the occurance of one class over the other. A one-class data set means all samples of the data set belong to one class. Current machine-learning methods have trouble handling (or cannot handle) highly imbalanced and/or one-class data, which are common in many real-world systems such as early fault and failure detection in structures and machines [29, 30] and intrusions in electronic security systems [31].

From a travel data collection perspective, procuring a large-scale, well-sampled, and balanced data set with ground truth travel mode labels is not trivial. It requires that survey participants identify their travel mode for a trip manually or through computer aids, as previously mentioned. On the other hand, unlike wearable GPS data, in-vehicle GPS data provide a substantial number of trips for the light-duty vehicle (LDV) travel mode. The vehicle driving behavior characteristics embedded in such a large amount of GPS driving data have yet to be thoroughly explored. It is found that given a large-scale homogenous data set, a novelty detection model can be trained as an extension to more classical machine-learning methods, enabling classification in circumstances that the other methods cannot.

The National Renewable Energy Laboratory (NREL)'s Transportation Secure Data Center (TSDC) hosts a variety of large-scale GPS travel survey data sets [32]. Many of the archived travel surveys use in-vehicle GPS devices, marked as LDV driving mode, whereas a smaller subset contains wearable GPS data with multi-modal travel mode information. Because most conventional travel mode classification models require at least two-class data and cannot deal with one-class data, they depend on multiclass labeled data sets that are more or less balanced when it comes to class distribution. In cases where the training set is highly imbalanced, several approaches can be taken. Assuming a labeled binary training set, one way to rebalance the data is to undersample the majority class while keeping the minority class unchanged. This solves the imbalance in the training set but may lead to a significant reduction in sample size (especially if the minority class is exceptionally small compared to the majority class). It also discards potentially useful data from the majority class. Alternatively, we can oversample the minority class while keeping the majority class unchanged. This approach grows the size of the training set at the expense of an increased likelihood of overfitting.

The oversampling and undersampling methods might mitigate the imbalanced data problem for data sets with two or more labeled classes; however, they are incapable of dealing with one-class cases. Another promising means of addressing the highly imbalanced (and/or one-class) data set problem is through novelty detection techniques. One-class classification is a novelty detection technique in which a classifier is trained on a homogenous data set to make binary classifications on a heterogeneous, multiclass test set. Novelty detection techniques, such as one-class SVM (OCSVM) [33], are typically applied when there is a large number of positive examples and too few (or no) examples of the negative class from which to construct an explicit machine-learning model. They achieve good results by learning a description of normality from the available positive instances and discriminating between positive instances and instances that lie outside the model's interpretation of normality.

This paper presents an exhaustive feature extraction (EFE)-OCSVM novelty detection approach capable of single-mode classification while only requiring a sufficient training data set from the interested travel mode—in this case LDV travel. For our application, trusted labels for non-LDV travel data are scarce relative to the abundance of accurately labeled LDV travel (from years of in-vehicle GPS survey supplements). This approach is especially valuable for extracting travel for a particular mode from an unlabeled, multimodal, wearable GPS data set when the mode of interest is well sampled.

Another significant advantage of the proposed EFE-OCSVM novelty detection approach is its ability to handle continuous time series trip data (i.e., speed profiles) by exhaustively and systematically selecting model features. Current novelty detection models either use discretized data for modeling or manually select model features from continuous data. For example, median speed, 95th percentile acceleration, and 95th percentile speed are chosen in a study [34]. The proposed approach can exhaustively extract features from time series speed data treated as a continuous function. The large list of extracted features does a better job of describing the underlying data than traditional approaches.

This study takes advantage of the unique, large-scale, real-world travel survey data with travel mode labels available within the TSDC to train and validate the performance of the proposed method. It provides an extension to classical machine-learning approaches to use when addressing highly imbalanced or homogenous training data. Additionally, a sensitivity analysis is conducted on training data set size and feature number to better inform the circumstances in which the method is appropriate. The proposed EFE-OCSVM approach can be extended to make binary predictions for any class with a sufficient training set. We also discuss the performance comparison between the proposed approach and the traditional machine-learning methods (i.e., SVM), as well as the robustness of the model. To summarize, the contributions of this study include but are not limited to:

- Proposing an EFE-OCSVM novelty detection approach, trained by homogeneous data to extract LDV travel from large-scale, real-world wearable GPS data
- Proposing a unique EFE process to systematically extract features from continuous time series data, such as speed profiles or driving cycles in this study
- Carrying out a sensitivity analysis based on a large-scale travel training data set to determine the sufficient sample size and feature number to build a viable model for future applications.

This paper is organized as follows: Section II presents a literature review of related research. Section III presents the development of the EFE-OCSVM novelty detection model and the case study setup. Section IV discusses the experimental results and major findings. The paper closes with a summary of major conclusions and suggested future research in Section V.

II. LITERATURE REVIEW

A. Regular Travel Mode Identification Approaches

Current travel mode identification methods are primarily rule-based and machine-learning-based approaches. The rule-based techniques, including fuzzy logic applications [22, 34], define a set of rules for different trip features to predict travel modes. They are easy to implement and understand. However, they may not be flexible or robust and are less generalizable than machine-learning-based approaches. Approaches that use machine learning develop a classification or prediction model by learning the training data set through various statistical and computational techniques such as neural networks [35], Bayesian networks, and support vector machines. Recently, multiple researchers have proposed using deep neural networks [23, 24] to identify travel mode. Xiao, Juan, and Zhang [36] use Bayesian networks to detect travel modes based on smartphone-based GPS data, in which the GPS segment features construct an improved Bayesian network. Zhang et al. [28] presented a regular SVM-based multistage classification method to deduce travel mode. Stenneth et al. [26] proposed a novel approach to identify travel mode according to mobile phone GPS data and transportation network data [37, 38]. They detected travel modes using four transportation network-related features fed into five classification models (Bayesian net, decision tree, random forest, naïve Bayes, and multilayer perceptron). Although the proposed approaches can accurately identify multiple travel modes, the training and test data sets are usually small (hundreds of samples) and contain evenly distributed classes. These approaches are incapable of handling homogenous training data where the negative class is absent altogether.

B. Novelty Detection

Novelty detection [39] techniques are typically used when there is a large number of positive examples and too few (or no) examples of the negative class from which to construct an explicit model. They achieve good results in these circumstances by learning a description of normality from the available positive instances, even approaching the performance of classical methods trained on well-sampled binary data. Novelty detection has gained much research attention in applications where instances of the negative (non-normal) class are impossible or entail a price that is restrictive. Novelty detection techniques are commonly used in early fault and failure detection in structures [29, 30], intrusions in electronic security systems [31], video surveillance [40], sensor networks [36], text mining [41], and more.

Generally, novelty detection technologies are classified into five categories [39]: probabilistic, distance-based, reconstruction-based, information-theoretic, and domain-based techniques. The probabilistic approaches use density measurements of the positive class to make predictions [42]. Distance-based approaches contain spatial measurement technologies, such as nearest-neighbor and clustering techniques, and assume that novel data occur far away from their nearest neighbors [43]. Reconstruction-based approaches [44] involve training a regression model based on training data. The information-theoretic approaches use information theories, such as entropy and the Kolmogorov complexity [45], to detect novel data. Domain-based methods [46] describe a domain encompassing normal (positive-class) data by defining a boundary or hyperplane around the normal data according to the distribution of the data. The most frequently used technique for determining this boundary are support vector machines. One-class SVM [33, 47] is a major approach of the domain-based methods, in which a "novelty boundary" in the feature space is determined by maximizing the margin to separate the transformed training data from the origin in the feature space. For this application, we use the OCSVM novelty detection technique to construct a model that learns a trip-level description of LDV travel so that it may detect LDV trips from an unseen data set of multimodal speed traces.

III. MODEL DEVELOPMENT AND CASE STUDY

The proposed novelty detection model was developed on GPS travel survey data from the Transportation Secure Data Center (TSDC) at the National Renewable Energy Laboratory (NREL) [32]. The proposed model development procedure is illustrated in

Fig. 1 and includes four stages: (1) **data acquisition**, (2) **data preparation**, (3) **feature extraction**, and (4) **model training**.

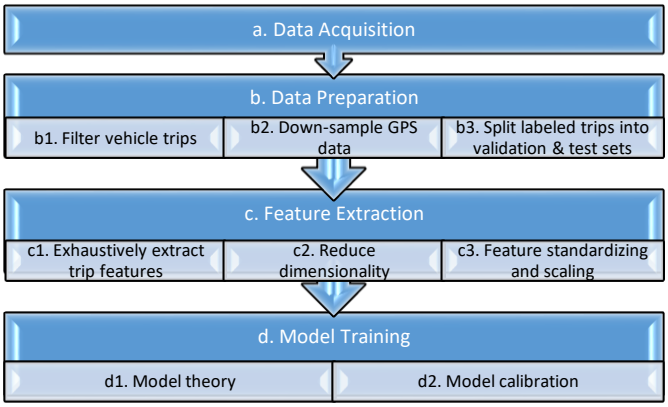


Fig. 1. Travel mode novelty detection model development

A. Data Acquisition

The GPS data used in developing the LDV mode classifier originated from the household travel studies shown in TABLE 1. These studies deployed in-vehicle GPS loggers at a frequency of 1 Hz or wearable loggers at a frequency of 0.33 Hz to track participants' travel patterns over a period of 3–7 days. In the table, only one includes labeled wearable GPS data (data set #5). There are two surveys named "California Household Travel Survey" (CHTS); however, data set #2 consists of only in-vehicle GPS data (from in-vehicle loggers), whereas data set #5 collected wearable GPS data. For in-vehicle GPS data sets, the vehicle miles traveled (VMT) represents the total driving mileage of the survey, whereas for wearable GPS data (i.e., data set #5) person miles traveled (PMT) is used to describe the total mileage for all travel modes in the survey.

TABLE 1. GPS DATA SET SUMMARY

No.	Survey Name	Year	Region	# of Vehicles /Persons	# of Trips	VMT/ PMT
1	Greater Fairbanks Transportation Survey	2013–2014	Greater Fairbanks (Alaska)	282	8,855	42,838

2	California Household Travel Survey	2010–2012	California	2,910	83,034	550,332
3	Regional Travel Survey	2011	Atlanta (Georgia)	1,653	51,370	367,709
4	Regional Household Travel Inventory	2007	Chicago (Illinois)	408	9,941	57,378
5[a]	California Household Travel Survey	2010–2012	California	7,574	160,380[b]	720,836

[a] Wearable GPS data set
[b] Refers to trip stages (portions of the trip with unique travel mode)

From TABLE 1, there is only one labeled wearable GPS data set (data set #5). The trip travel mode distribution (ratio) of data set #5 is illustrated in

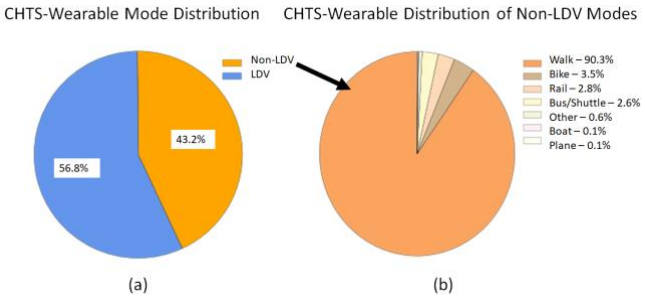


Fig. 2. From

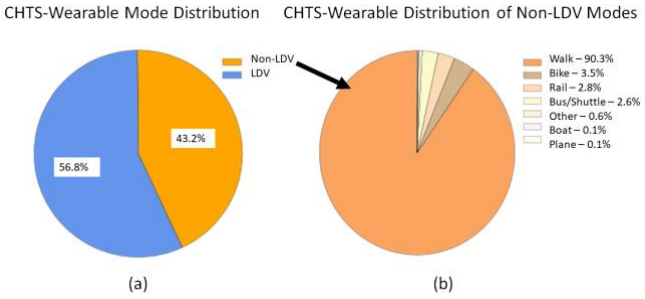


Fig. 2(a), the LDV and Non-LDV samples are roughly balanced. While from

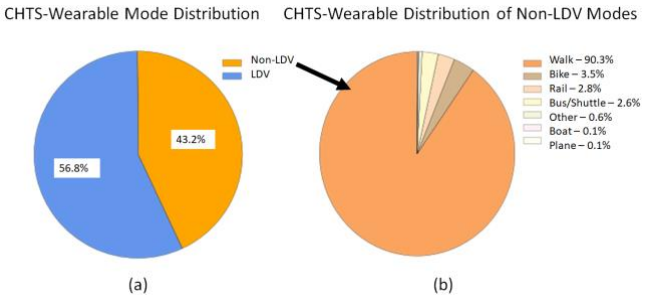


Fig. 2 (b), the walk mode dominates all other modes in the Non-LDV cases.

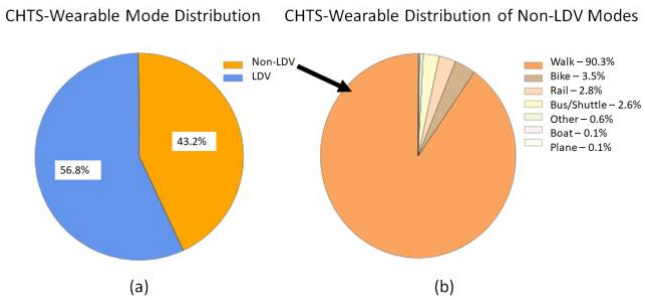


Fig. 2. CHTS wearable travel mode distribution for data set #5: (a) LDV vs. Non-LDV; (b) detailed distribution of Non-LDV modes.

The trip number of travel mode of data set #5 is illustrated in Fig. 3. From that, the LDV and non-LDV samples are roughly balanced, whereas the LDV sample size is a little bigger than the summation of all non-LDV modes' sample sizes. The walk mode dominates all other modes in the non-LDV cases.

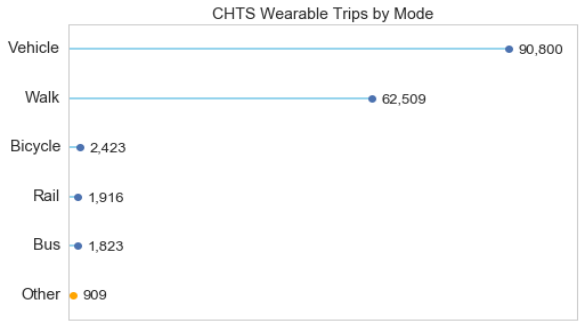


Fig. 3. CHTS wearable GPS travel mode trip counts. "Other" category represents trips with uncommon (e.g., boat, plane) or unclear travel modes (e.g., private shuttle) that were removed from the test set.

B. Data Preparation

The training set consisted of vehicle trips collected at 1 Hz frequency from in-vehicle GPS loggers. This set was filtered for nonrepresentative vehicle trips that represented outliers in the training set. A simple rule was applied to filtering and removing nonrepresentative trips with an average speed less than 5 mph and a max speed less than 10 mph. After this step, approximately 98% of vehicle trips remained in the training set.

In order to ensure that collection frequencies were consistent, vehicle points were downsampled to 0.33 Hz to match the collection frequency of the wearable loggers in the CHTS data. A beneficial side effect of the downsampling was that it significantly reduced the computational time required for feature extraction, which will be introduced in the following section.

The validation and test data sets consisted of an 80/20 split of labeled wearable trip stages (trip segments with a single mode of travel) from data set #5, where travel mode had been manually imputed through a computer-aided trip identification and analysis system, combining travel diaries and manual confirmation with a sophisticated quality control procedure. These labels were assumed to be accurate and taken as ground truth for tuning and evaluating the classifier.

C. Feature Extraction

Time series data classification is challenging because the inputs are ordered and of inconsistent lengths. To reduce complexity and speed up training, it is common for a time series

to be summarized by a reduced set of attributes/features. Selecting which features to include, however, is often difficult as it is not intuitive to determine which are most valuable for a classification task prior to evaluation. To overcome these challenges, a novel exhaustive feature extraction for time series data is used, which includes three steps as described below.

1) Exhaustively Extract Trip Features

A full spectrum of trip features reflects different trip speed profile characteristics, such as the maximum, minimum, and average speed, and the number of temporary speed peaks. Usually, it is impossible to obtain all features by hand. However, exhaustive feature extraction requires generating as many features as possible. In this study, an open-source Python package, "tsfresh," is used for extracting features from time series data. This package treats the speed traces as a continuous function and automatically calculates many distinct characteristics representative of the full signal (e.g., mean, maximum, minimum, number of peaks). The package uses a filtering procedure to avoid extracting irrelevant features algorithmically and to ensure the power and importance of each feature [48]. An exhaustive list of the features derived using tsfresh can be found in its documentation online [49].

2) Reduce Dimensionality

The previous step extracted 216 distinct descriptive features from each of the training, validation, and test trips. Features containing infinite or null values were removed, resulting in 132 distinct features for classification.

3) Feature Standardizing and Scaling

After the dimension reduction, the refined features were normalized using zero mean, unit variance scaling. This was done to maximize performance during classification.

D. Model Training

After data acquisition, data preparation, and feature extraction, the EFE-OCSVM novelty detection model for LDV travel mode can be trained.

1) Model Theory

Suppose there is a training data set $X = \{x_1, x_2, \dots, x_n\}, where\ n$ is the number of trips as observations. Let $\Phi: X \to F$ be a feature map to dot product space F from mapping data, which can be computed by some kernel functions $k(x, y) = (\Phi(x) \cdot \Phi(y))$. One popular kernel is the Gaussian kernel:

$$k(x, y) = e^{-\|x-y\|^2/2\sigma^2} \tag{1}$$

where σ is the width parameter associated with the Gaussian kernel. Then, the following quadratic programming problem is proposed to separate the data set from the origin with maximum margin:

$$\min_{w\in F, \varepsilon\in R^n, \rho\in R} \quad \frac{1}{2}\|w\|^2 + \frac{1}{vn}\sum_{i=1}^{n}\varepsilon_i - \rho \tag{2}$$

$$\text{Subject to} \quad \big(w \cdot \Phi(x_i)\big) \geq \rho - \varepsilon_i, \quad \varepsilon_i \geq 0 \tag{3}$$

where the overall margin variable is ρ and the individual error or slack variable is ε_i. The parameter w is a weight vector corresponding to all dimensions of space F. The regularization parameter $v \in (0, 1)$ significantly impacts the model performance. It serves as an upper bound on the proportion of outliers and a lower bound on the fraction of support vectors. Usually, a feasible OCSVM model needs to calibrate the parameter v for particular sample sets.

If w and ρ solve this problem, the decision function $f(x) = sgn((w \cdot \Phi(x) - \rho)$ will be positive for most examples x_i, while the support vector type regularization term $\|w\|$ will still be small.

In order to solve this problem, a Lagrangian term is constructed by introducing nonnegative multipliers α_i and β_i, which are considered as decision variables:

$$\begin{aligned} L(w, \varepsilon, \rho, \alpha, \beta) = {} & \frac{1}{2}\|w\|^2 + \frac{1}{vn}\sum_{i=1}^{n}\varepsilon_i - \rho \\ & - \sum_{i=1}^{n}\alpha_i\Big(\big(w \cdot \Phi(x_i)\big) - \rho + \varepsilon_i\Big) \\ & - \sum_{i=1}^{n}\beta_i\varepsilon_i \end{aligned} \tag{4}$$

Assuming the derivatives with respect to the primal variables w, ρ, and ε equal to zero:

$$w = \sum_{i=1}^{n}\alpha_i\Phi(x_i) \tag{5}$$

$$\alpha_i = \frac{1}{vn} - \beta_i \leq \frac{1}{vn} \tag{6}$$

$$\sum_{i=1}^{n}\alpha_i = 1 \tag{7}$$

Then, substituting (5)–(7) into the Lagrangian term (4) and the definition of kernel functions, the dual problem can be written as:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\big(x_i, x_j\big) \tag{8}$$

$$\text{Subject to} \quad 0 \leq \alpha_i \leq \tfrac{1}{vn}, \quad \textstyle\sum_{i=1}^{n}\alpha_i = 1 \tag{9}$$

The dual variable α_i will be solved by the dual problem from (8) and (9).

Then, from (3), the two inequality constraints become equalities if α_i and β_i are positive. The two equalities are $w \cdot \Phi(x_i) = \rho$ and $\varepsilon_i = 0$. After substituting (5), the primal variable ρ can be expressed as:

$$\rho = w \cdot \Phi(x_i) = \sum\nolimits_{j=1}^{n}\alpha_j k\big(x_j, x_i\big) \tag{10}$$

The final decision function for determining a sample is a positive or negative class, expressed as follows with (5) and (10):

$$f(x) = sgn(\sum_{i=1}^{n}\alpha_i k(x_i, x) - \rho) \tag{11}$$

2) Model Calibration

A designed parameter sweep was conducted to determine the optimal value for the regularization term v (nu). This was accomplished by adjusting v to maximize the overall prediction accuracy on a validation set consisting of 80% of the CHTS wearable GPS data. The remaining 20% (test set) was reserved for an unbiased performance evaluation.

Fig. 4 shows the process of v parameter tuning. From the diagram, a maximum accuracy is achieved when v = 0.05. In that case, to obtain feasible and consistent prediction accuracy results, each of the following experiments (with the exception

of classification for other, non-LDV modes) use 0.05 for parameter ν.

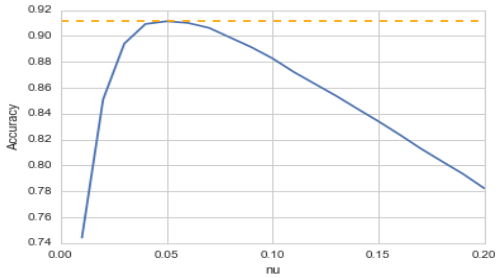


Fig. 4. Regularization parameter ν (nu) tuning

IV. Results and Discussion

A. Experiment and Results

Validation testing for the proposed LDV mode classifier was conducted in two stages. First, the model was trained on CHTS in-vehicle GPS data (data set #2) and tested on CHTS wearable GPS data (data set #5). The purpose of this validation stage was to assess the performance of the method on data from similar travel conditions. The second stage of testing involved training on all other in-vehicle GPS data gathered from data sets #1, #3, and #4 (excluding #2) and testing on CHTS wearable GPS data (data set #5). This stage was designed to test the generalization of the method across studies that differ with respect to geography and time.

The proposed model is implemented using the scikit-learn machine-learning Python package [49]. Both experiments shared the same feature space. In both stages, the model regularization parameter ν was set to 0.05. The EFE-OCSVM model prediction results for both scenarios are summarized in TABLE 2. Note that the performance estimates were generated from testing on 31,895 samples (this is 20% of all samples in the #5 wearable set). The validation set was used for hyperparameter tuning. The test set was used for performance calculations. The precision and recall were used, in addition to accuracy, to assess performance as follows:

$$Precision = \frac{\#\ of\ items\ correctly\ identified}{Total\ estimations\ to\ the\ class} \quad (12)$$

$$Recall = \frac{\#\ of\ items\ correctly\ identified}{Total\ ground\ truth\ \ items\ \ to\ the\ class} \quad (13)$$

$$Accuracy = \frac{\#\ of\ items\ correctly\ identified\ for\ all\ classes}{Total\ population} \quad (14)$$

The proposed method performs well in both stages of validation. The results indicate the robustness of the proposed model. Slightly better prediction performance is observed in stage 2. This may be due to the training data set diversity and differing data quality across studies.

TABLE 2. One-Class SVM prediction result summary

Scenario 1				
		Ground Truth		
Estimated	# of samples	Non-LDV	LDV	Precision
	Non-LDV	12,227	1,417	0.90
	LDV	1,437	16,814	0.92
	Recall	0.89	0.92	**0.92**[a]
Scenario 2				
Estimated		Ground Truth		
	# of samples	Non-LDV	LDV	Precision
	Non-LDV	63,109	5,562	0.92
	LDV	5,784	85,016	0.94
	Recall	0.92	0.94	**0.94**[a]

[a] Accuracy

B. Sensitivity Analysis of Training Sample Size for LDV Mode Prediction

The training set in validation stage 1 is large, including about 83,000 LDV driving trips and over 550,000 miles traveled, and is suitable for conducting a training sample size sensitivity analysis. Fig. 5(a) illustrates the relationship between accuracy and training sample size. The diagram indicates that as the sample size increases, the accuracy rate significantly increases up to a certain point (about 5,000, marked by the star). After that critical point, the accuracy rate may not increase much as the sample size grows. Fig. 5 (b) shows the computational efficiency (i.e., training time in seconds) increasing as the size of the training set increases. At around 5,000 trips, training is exceptionally fast and may be an acceptable target sample size for training a feasible model.

Another interesting observation is that when the size of the training set is on the order of hundreds of samples (100–1,000), the prediction accuracy rate is around 0.8–0.9 and acceptable for some applications. When the sample size drops less than 100, the classifier does not do well at all. This observation may be used to guide travel survey study data collection and indicates that roughly 5,000 samples are required to build up a feasible EFE-OCSVM model for LDV driving mode.

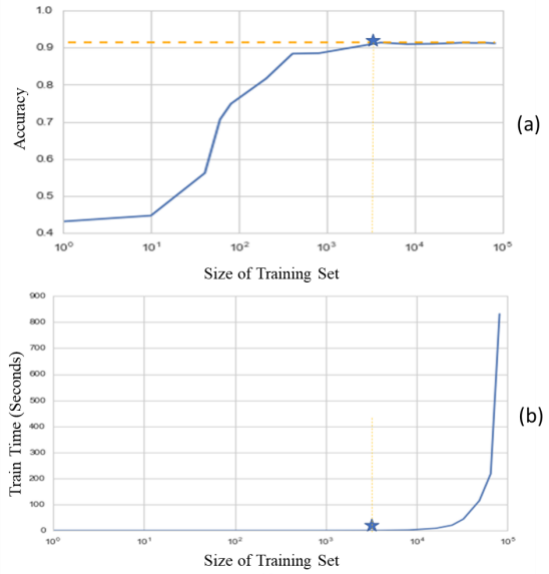


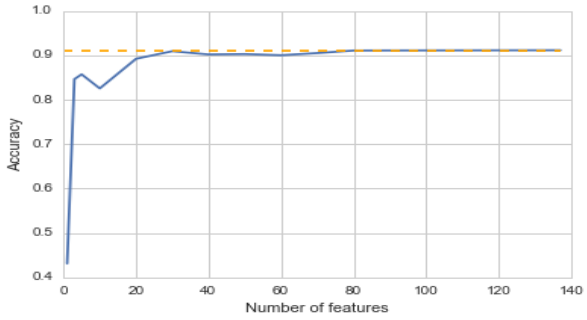
Fig. 5. Sensitivity analysis of training sample size: (a) accuracy and (b) computational efficiency

C. Feature Selection Analysis

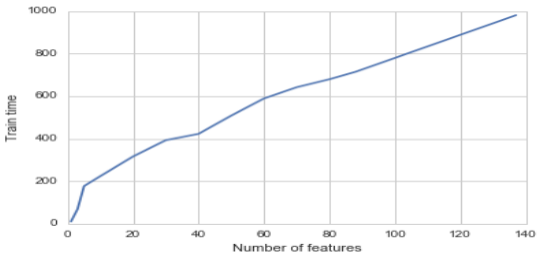
The feature importance score was derived from Randomized Logistic Regression feature scores, which is a popular feature ranking mechanism. As previously mentioned, there are 132 distinct and valid features for classification. The study of feature selection's impact on prediction accuracy and efficiency is based on the scenario 1 experiment. The features are ranked by the importance score from 1 (most important) to

132 (least important). The results are illustrated in Fig. 6. There are two observations from the diagram. First, the accuracy rate does not significantly improve when more than 30 features are used. A second observation is that somewhere between the addition of 7 and 10 features; there is a small drop in accuracy. The prediction accuracy does not strictly increase as the feature set size increases. Nevertheless, overall prediction performance improves as the number of features increases.

On the other hand, from a computational efficiency perspective, the training time steadily increases as more features are included in the model. From the observations, a proper number of selected features could be 30. Those 30 selected features include regular statistical parameters of speed series, such as mean and variance, and, importantly, other parameters which seldom appear in speed profile processing literature, such as mean autocorrelation and index mass quantile.



(a)



(b)

Fig. 6. Feature selection analysis: (a) accuracy; (b) computational efficiency

D. Experiment on Additional Travel Modes

As previously mentioned, the proposed EFE-OCSVM method is capable of classifying a specific travel mode if the sample size of that travel mode is large enough for training. In that way, an ensemble of OCSVMs can be used for classifying any travel mode in a binary classification manner. Here, LDV driving mode is tested along with other travel modes such as walk, bicycle, rail, and bus. According to Fig. 3, among the non-LDV modes, walk mode has the largest sample size (62,509 trips), whereas each of the other modes contain approximately 2,000 samples each.

The classifier performance metrics for this test included accuracy and area under curve (AUC). The definition of accuracy is referred to in (14). AUC is commonly used in situations when the sample distribution is very skewed and imbalanced. It is obtained by computing the area under the receiver operating characteristic curve. A perfect predictor has an AUC of 1, whereas a classifier that randomly guesses labels will have an AUC of 0.5.

The results of the experiments applying EFE-OCSVM methods to other travel modes within the CHTS wearable GPS data (dataset #5) are illustrated in Fig. 7. According to the result, accuracy numbers of all travel modes are very good, ranging between 0.91 and 0.99. This is mainly caused by the imbalanced data set, and the accuracy of those cases cannot be used to assess the performance and efficacy of the model. For example, for a hugely imbalanced binary data set, a naïve model that predicts all samples as belonging to the class that dominates the data set can obtain high accuracy. However, the AUC values can tell the true story of the prediction. Travel modes with large sample sizes, such as LDV and walk, have high AUC values, whereas the other modes with a small number of samples have low AUC values. The prediction results of non-LDV travel modes are reflected by the training sample size of each travel mode (i.e., the larger the sample size of the travel mode, the better the prediction performance).

One note about LDV and walk modes is that the improved prediction performance of walk mode could be explained by the consistency of low-speed characteristics of most walk mode trips.

Another note is that the ensemble of EFE-OCSVM models might be capable of classifying each class under certain conditions (i.e., if each class had a big sample size) in a binary classification manner; however, it does not mean that the ensemble of EFE-OCSVM models will outperform or replace the conventional multiclass machine-learning classifiers (e.g., SVM). The proposed EFE-OCSVM model has the unique capability of handling a highly imbalanced or one-class data set, whereas the regular methods cannot. If the test data set is well-sampled and balanced, regular classification methods will have better performance. This hypothesis will be verified in the following section.

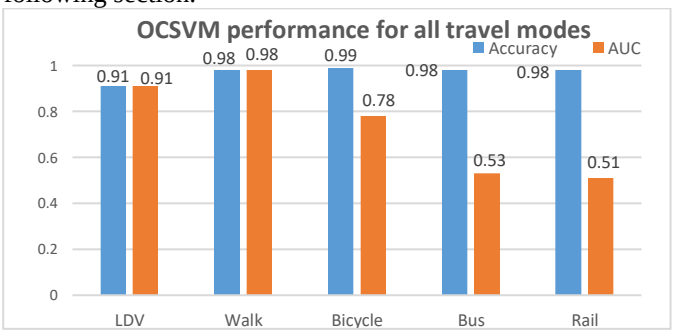


Fig. 7. OCSVM method performance for all travel modes within data set #5

E. Performance Comparison between EFE-OCSVM and Regular SVM

The EFE-OCSVM model works on homogenous LDV mode training data, which cannot be handled by regular SVM for binary classification. However, for a well-sampled, balanced data set, EFE-OCSVM approaches are not expected to beat conventional SVM methods due to the lower information level of homogenous training data. This section conducts a classification performance comparison study for EFE-OCSVM and regular binary SVM to validate the above hypothesis and investigate the performance difference.

The refined speed profile features (set of 132 features as previously referenced) are taken and used by a regular SVM for testing LDV and non-LDV mode trips within data set #5. The

80% of total samples are randomly selected and used for training, and the remaining 20% are used for testing. A regular SVM with only three features (i.e., median speed, 95th percentile speed, and 95th percentile acceleration) serves as a theoretical baseline. In order to remain consistent, the trained EFE-OCSVM was tested on an identical test set.

TABLE 3. PERFORMANCE COMPARISON OF OCSVM AND REGULAR SVMS FOR LDV MODE CLASSIFICATION

	Accuracy	AUC
EFE-OCSVM	0.92	0.92
Regular SVM (full features)	0.99	0.99
Regular SVM (three features)	0.97	0.97

TABLE 3 indicates that the proposed method can provide an acceptable and promising result, although it does not outperform the regular SVM models. The regular SVM with full features offers the best prediction, whereas the regular SVM with three typical features used by literature performs in the middle. It elucidates that the proposed method can render a reasonable and promising result (0.92 accuracy and 0.92 AUC) for the travel mode detection.

V. CONCLUSION

This study proposes a generalized innovative LDV travel mode detection approach using the OCSVM novelty detection technique. The proposed method extends widely used classical machine-learning approaches, and is suitable for large-scale, homogenous training data (where classical machine-learning approaches undoubtedly fail). The proposed method has been developed and applied on large-scale, real-world GPS data. Performance validation over two scenarios is very promising, ranging from 92%–94% accuracy in the LDV case. Additionally, it was found that approximately 5,000 training samples are sufficient to establish a viable model for LDV travel mode detection applications.

A promising prediction performance is reached when applied as a binary classifier for LDV travel on an unbalanced data set. Its ability to generalize temporally and spatially was validated by several large-scale data sets, giving confidence in its ability to generalize to additional unseen scenarios. Lastly, the experiment on identifying additional travel modes indicates that the proposed method can be easily extended to multimodal travel mode detection problems once the sample sizes for the classes of interest are large enough.

VI. ACKNOWLEDGEMENT

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding was provided by the DOE Vehicle Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

REFERENCES

[1] L. Zhu, B. Borlaug, L. Lin, J. Holden, and J. Gonder, "Identifying Light-Duty Vehicle Travel from Large-Scale Multimodal Wearable GPS Data with Novelty Detection Algorithms," in *2021 IEEE Green Technologies Conference (GreenTech)*, Denver, CO, 2021: IEEE, pp. 155-162.

[2] X. Hu, Y.-C. Chiu, Y.-L. Ma, and L. Zhu, "Studying Driving Risk Factors using Multi-Source Mobile Computing Data," *International Journal of Transportation Science and Technology*, vol. 4, no. 3, pp. 295-312, 2015.

[3] M. Wang, X. Bao, L. Zhu, and Y. L. Bao, "A Map-Matching Method Using Intersection-Based Parallelogram Criterion," *Advanced Materials Research*, vol. 403, pp. 2746-2750, 2012.

[4] L. Zhu, J. R. Holden, and J. D. Gonder, "Trajectory Segmentation Map-Matching Approach for Large-Scale, High-Resolution GPS Data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2645, no. 1, pp. 67-75, 2017, doi: 10.3141/2645-08.

[5] L. Zhu, J. Gonder, and L. Lin, "Prediction of Individual Social-Demographic Role Based on Travel Behavior Variability Using Long-Term GPS Data," *Journal of Advanced Transportation*, vol. 2017, p. 13, 2017.

[6] A. A. S. Mohamed, L. Zhu, A. Meintz, and E. Wood, "Optimum Planning for Inductively Charged On-demand Automated Electric Shuttles at Greenville, South Carolina," in *2019 IEEE Industry Applications Society Annual Meeting*, 29 Sept.-3 Oct. 2019 2019, pp. 1-10, doi: 10.1109/IAS.2019.8912343.

[7] L. Zhu, J. Wang, V. Garikapati, and S. Young, "Decision support tool for planning neighborhood-scale deployment of low-speed shared automated shuttles," *Transportation Research Record*, vol. 2674, no. 9, pp. 1-14, 2020.

[8] H. M. A. Aziz *et al.*, "An optimization-based planning tool for on-demand mobility service operations," *International Journal of Sustainable Transportation*, pp. 1-12, 2020, doi: 10.1080/15568318.2020.1847368.

[9] P. Kelly, P. Krenn, S. Titze, P. Stopher, and C. Foster, "Quantifying the difference between self-reported and global positioning systems-measured journey durations: a systematic review," *Transport Reviews*, vol. 33, no. 4, pp. 443-459, 2013.

[10] L. Zhu, J. Holden, J. Gonder, and E. Wood, "Green Routing Fuel Saving Opportunity Assessment: A Case Study on California Large-Scale Real-World Travel Data: Preprint," National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2017.

[11] X. Hu, Y.-C. Chiu, and L. Zhu, "Behavior insights for an incentive-based active demand management platform," *International Journal of Transportation Science and Technology*, vol. 4, no. 2, pp. 119-133, 2015.

[12] Y. Chen, L. Zhu, J. Gonder, S. Young, and K. Walkowicz, "Data-driven fuel consumption estimation: A multivariate adaptive regression spline approach," *Transportation Research Part C: Emerging Technologies*, vol. 83, pp. 134-145, 2017.

[13] L. Zhu, J. Gonder, E. Bjarkvik, M. Pourabdollah, and B. Lindenberg, "An Automated Vehicle Fuel Economy Benefits Evaluation Framework Using Real-World Travel and Traffic Data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 3, pp. 29-41, 2019.

[14] L. Zhu, J. R. Holden, and J. D. Gonder, "Navigation Application Programming Interface Route Fuel Saving Opportunity Assessment on Large-Scale Real-World Travel Data for Conventional Vehicles and Hybrid Electric Vehicles," *Transportation Research Record*, vol. 2672, no. 25, pp. 139-149, 2018.

[15] L. Zhu and J. D. Gonder, "A driving cycle detection approach using map service API," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 349-363, 7// 2018, doi: 10.1016/j.trc.2018.05.010.

[16] M. E. Ben-Akiva *et al.*, "Freight data collection using GPS and web-based surveys: Insights from US truck drivers' survey and perspectives for urban freight," *Case Studies on Transport Policy*, vol. 4, no. 1, pp. 38-44, 2016.

[17] V. Punzo and F. Simonelli, "Analysis and comparison of microscopic traffic flow models with real traffic microscopic data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1934, pp. 53-63, 2005.

[18] L. Gong, T. Morikawa, T. Yamamoto, and H. Sato, "Deriving personal trip data from GPS data: a literature review on the existing methodologies," *Procedia-Social and Behavioral Sciences*, vol. 138, pp. 557-565, 2014.

[19] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 3, pp. 285-297, 2009.

[20] C. Chen, H. Gong, C. Lawson, and E. Bialostozky, "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 10, pp. 830-840, 2010.

[21] H. Gong, C. Chen, E. Bialostozky, and C. T. Lawson, "A GPS/GIS method for travel mode detection in New York City," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 131-139, 2012.

[22] S. Tsui and A. Shalaby, "Enhanced system for link and mode identification for personal travel surveys based on global positioning systems," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1972, pp. 38-45, 2006.

[23] P. Gonzalez *et al.*, "Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones," in *15th World congress on intelligent transportation systems*, 2008, pp. 16-20.

[24] Y.-J. Byon, B. Abdulhai, and A. Shalaby, "Real-time transportation mode detection via tracking global positioning system mobile devices," *Journal of Intelligent Transportation Systems*, vol. 13, no. 4, pp. 161-170, 2009.

[25] L. Lin, J. C. Handley, Y. Gu, L. Zhu, X. Wen, and A. W. Sadek, "Quantifying uncertainty in short-term traffic prediction and its application to optimal staffing plan development," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 323-348, 2018/07/01/ 2018, doi: 10.1016/j.trc.2018.05.012.

[26] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2011: ACM, pp. 54-63.

[27] F. Yang, Z. Yao, and P. J. Jin, "GPS and acceleration data in multimode trip data recognition based on wavelet transform modulus maximum algorithm," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2526, pp. 90-98, 2015.

[28] L. Zhang, S. Dalyot, D. Eggert, and M. Sester, "Multi-stage approach to travel-mode segmentation and classification of gps traces," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences:[Geospatial Data Infrastructure: From Data Acquisition And Updating To Smarter Services] 38-4 (2011), Nr. W25*, vol. 38, no. W25, pp. 87-93, 2011.

[29] C. Surace and K. Worden, "Novelty detection in a changing environment: A negative selection approach," *Mechanical Systems and Signal Processing*, vol. 24, no. 4, pp. 1114-1128, 2010/05/01/ 2010, doi: 10.1016/j.ymssp.2009.09.009.

[30] L. A. Clifton, H. Yin, and Y. Zhang, "Support vector machine in novelty detection for multi-channel combustion data," in *International Symposium on Neural Networks*, 2006: Springer, pp. 836-843.

[31] V. Jyothsna, V. R. Prasad, and K. M. Prasad, "A review of anomaly based intrusion detection systems," *International Journal of Computer Applications*, vol. 28, no. 7, pp. 26-35, 2011.

[32] "Transportation Secure Data Center." (2021). National Renewable Energy Laboratory. Accessed Jan. 15, 2021: www.nrel.gov/tsdc.

[33] A. Binbusayyis and T. Vaiyapuri, "Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM," *Applied Intelligence*, pp. 1-15, 2021.

[34] N. Schüssler and K. W. Axhausen, "Processing GPS raw data without additional information," *Transportation Research Record*, vol. 2105, pp. 28-36, 2009.

[35] M. Gallo, F. Simonelli, G. De Luca, and C. Della Porta, "An artificial neural network approach for spatially extending road traffic monitoring measures," in *Environmental, Energy, and Structural Monitoring Systems (EESMS), 2016 IEEE Workshop on*, 2016: IEEE, pp. 1-5.

[36] G. Xiao, Z. Juan, and C. Zhang, "Travel mode detection based on GPS track data and Bayesian networks," *Computers, Environment and Urban Systems*, vol. 54, pp. 14-22, 2015/11/01/ 2015, doi: 10.1016/j.compenvurbsys.2015.05.005.

[37] L. Zhu and Y.-C. Chiu, "Transportation Routing Map Abstraction Approach: Algorithm and Numerical Analysis," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2528, pp. 78-85, 2015.

[38] L. Zhu, Y.-C. Chiu, and Y. Chen, "A Road Network Abstraction Approach for Traffic Analysis: Framework and Numerical Analysis," *IET Intelligent Transport Systems*, p. 18 pp., 2017.

[39] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215-249, 2014.

[40] M. Markou and S. Singh, "A neural network-based novelty detector for image sequence analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1664-1677, 2006.

[41] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *Journal of Machine Learning Research*, vol. 2, no. Dec, pp. 139-154, 2001.

[42] D. Erdogmus, R. Jenssen, Y. Rao, and J. Principe, "Multivariate density estimation with optimal marginal parzen density estimation and gaussianization," in *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, 2004: IEEE, pp. 73-82.

[43] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003: ACM, pp. 29-38.

[44] M. Augusteijn and B. Folkert, "Neural network classification and novelty detection," *International Journal of Remote Sensing*, vol. 23, no. 14, pp. 2891-2902, 2002.

[45] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004: ACM, pp. 206-215.

[46] T. Le, D. Tran, P. Nguyen, W. Ma, and D. Sharma, "Multiple distribution data description learning method for novelty detection," in *Neural Networks (IJCNN), The 2011 international joint conference on*, 2011: IEEE, pp. 2321-2326.

[47] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput*, vol. 13, no. 7, pp. 1443-71, Jul 2001, doi: 10.1162/089976601750264965.

[48] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," *arXiv preprint arXiv:1610.07717*, 2016.

[49] M. Christ. "tsfresh." https://tsfresh.readthedocs.io/en/latest/ (accessed 2017).

Lei Zhu (M' 21, SM' 22) received his Ph.D. degree in Transportation Engineering from the University of Arizona, in 2014. He is currently an assistant professor at the Department of Systems Engineering and Engineering Management at the University of North Carolina at Charlotte. His research interests include smart and sustainable mobility systems, spatial sensing technologies, advanced mobility system modeling and simulation, big data, and machine learning in transportation.

Brennan Borlaug received his B.S. degree in Mechanical Engineering from Washington State University in 2014 and M.S. degree in Information Science from the University of California at Berkeley in 2017. He has been a researcher at the National Renewable Energy Laboratory (NREL) since 2014. His research covers traffic data

processing, machine learning, and GPS data travel mode imputation.

Lei Lin (M'18) received his M.S. degree in Computer Science and Ph.D. degree in Transportation Systems Engineering from the University at Buffalo, the State University of New York, in 2013 and 2015. He is a Research Scientist at Goergen Institute for Data Science, University of Rochester with research interests of transportation big data, artificial intelligence applications in transportation, and connected and automated transportation.

Jacob Holden received B.S. and M.S. degrees in Aerospace, Aeronautical and Space Engineering from the University of Cincinnati in 2015 and 2016. He has worked as a researcher at NREL since graduating. His research interests include GPS data processing, eco-routing, and energy impacts from connected and automated vehicles.

Jeffrey Gonder received his B.S. degree in Mechanical Engineering from University of Colorado in 2001 and his M.S. degree in Mechanical Engineering from Pennsylvania State University in 2003. Following two years at ANUVU Fuel Cell Products, he launched his career at NREL in 2005 where he currently manages the Mobility, Behavior & Advanced Powertrains Group. His research interests include potential energy and mobility impacts from advanced powertrains, connected and automated vehicles, and next-generation mobility systems.