A Deep Ensemble Neural Network Approach for FHWA Axle-based Vehicle Classification using Advanced Single Inductive Loops

Yiqiao Li*
Ph.D. Candidate
Department of Civil and Environmental Engineering
Institute of Transportation Studies
4000 Anteater Instruction and Research Building (AIRB)
University of California, Irvine
Irvine, CA 92697
yiqial1@uci.edu

Andre Y.C. Tok, Ph.D.
Testbeds Manager and Assoc. Project Scientist
Institute of Transportation Studies
4000 Anteater Instruction and Research Building (AIRB)
University of California, Irvine
Irvine, CA 92697
ytok@uci.edu

Stephen G. Ritchie, Ph.D.
Professor of Civil Engineering and
Director, Institute of Transportation Studies
4000 Anteater Instruction and Research Building (AIRB)
University of California, Irvine
Irvine, CA 92697
sritchie@uci.edu

Word Count: 5,580 words + 6 table (250 words per table) * 6 = 7,080 words

*Corresponding Author

Submitted [August 1st, 2020]

ABSTRACT

The Federal Highway Administration (FHWA) vehicle classification scheme is designed to serve various transportation needs such as pavement design, on-road emission estimation, and transportation planning. Many transportation agencies rely on Weigh-In-Motion and Automatic Vehicle Classification sites to collect these essential vehicle classification count data. However, the spatial coverage of these detection sites across the highway network are limited due to high installation and maintenance costs.

One cost-effective approach investigated by researchers has been the use of single inductive loop sensors as an alternative to obtain FHWA vehicle classification data. However, most datasets used to develop such models are skewed since many classes belonging to larger truck configurations are rarely observed in the roadway network. This increases the challenge to accurately classify under-represented classes, even though many of these minority classes may pose disproportionately adverse impacts on pavement infrastructure and the environment. As a consequence, previous models have been unable to adequately classify under-represented classes, and the overall performance of the models are often masked by excellent classification accuracy of majority classes, such as passenger vehicles and five-axle tractor trailers. To resolve the challenge of imbalanced datasets in the FHWA vehicle classification problem, this paper describes a study that developed a bootstrap aggregating (bagging) deep neural network (DNN) model on a truck-focused dataset using single inductive loop signatures. The proposed method significantly improved the model performance on several truck classes, especially minority classes such as Classes 7 and 11 which were overlooked in previous research studies.

Keywords: FHWA vehicle classification, Single inductive loops, Dropout, Bootstrap aggregated deep neural network

1. INTRODUCTION

FHWA's Traffic Monitoring Guide outlines a standardized classification scheme to serve various transportation needs. The FHWA vehicle classification scheme categorizes vehicles into thirteen classes based on tire and axle combination while partially taking into account general body configuration (*1*). This classification scheme has been widely used for pavement design to account for the dissimilar pavement impacts attributed to physical vehicle characteristics such as axle loads, spacing and tire configuration (*2*). In addition, aggregated FHWA vehicle classes have been used as input for on-road emission estimation models (*3*) as well as in freight forecast modeling (*4*, *5*)

Since truck size and weight laws vary by states and trucks configurations populations vary across states (*6*), some states agencies modify the FHWA's 13 categories to meet their transportation application needs (*1*). For example, Class 9 type 32 trucks in California are distinguished from other Class 9 trucks in the FHWA 13-Category Scheme and form a standalone class labeled as Class 14 (*7*). Since the data used in this paper was collected in California at the statewide level, the model was focused on classifying vehicles into the California-modified FHWA scheme (FHWA-CA). Table 1 provides a brief description for each class in the FHWA-CA classification scheme.

Table 1 FHWA-CA classification scheme definitions (*8*)

FHWA-CA Class	Vehicle Description	Class Includes	# of axle
1	Motorcycle	Motorcycles	2
2	Passenger Vehicles	All cars, Cars with one-axle trailers, Cars with two-axle trailers	2, 3 or 4
3	Other two-axle four-tire Single-unit Vehicle	Pickups and vans, Pickups and Vans with one- and two- axle trailers	2, 3 or 4
4	Bus	Two- and three-axle buses, Bus with trailer	2 or 3 (tractor)
5	Two-axle, Six-tire, single-unit trucks	Two-axle trucks, two-axle trucks with trailer	2 (tractor)
6	Three-axle single-unit trucks	Three-axle trucks, Three-axle tractors without trailers	3
7	Four or more axle single-unit trucks	Four-, five, six- and seven-axle single-unit trucks	4 or more
8	Four or fewer axle single-trailer trucks	Two-axle trucks pulling one- and two-axle trailers, Two-axle tractors pulling one- and two-axle trailers, Three-axle tractors pulling one-axle trailers	3 or 4
9	Five-axle single-trailer trucks	Two-axle tractors pulling three-axle trailers, Three-axle tractors pulling two-axle trailers, Three-axle trucks pulling two-axle trailers	5
10	Six or more axle single-trailer trucks	Three-axle tractors pulling three-axle trailers	6 or more
11	Five or fewer axle multi-trailer trucks	Multiple configurations (Multi-unit trucks)	4 or 5
12	Six-axle multi-trailer trucks	Multiple configurations (Multi-unit trucks)	6
13	Seven or more axle multi trailer trucks	Multiple configurations (Multi-unit trucks)	7 or more
14	Single and tandem axle on tractor, single and single axle on trailer	Single and tandem axle on tractor, single and single axle on trailer	5
15	Unclassified vehicles	Multiple configurations	2 or more

Weigh-In-Motion (WIM) and Automatic Vehicle Classification (AVC) sites using axle sensor technologies can directly capture vehicle axle configuration information. Hence, these types of systems have been commonly used to report vehicle classification counts according to FHWA-based schemes. However, those sensors are not comprehensively deployed in the transportation highway network due to their high installation and maintenance costs. Conversely, inductive loop sensors are much more widely deployed in many jurisdictions as they have a much lower installation and maintenance cost. Studies have investigated the use of inductive vehicle signatures to classify vehicles based on the FHWA classification scheme (*9*, *10*). However, a closer evaluation of these efforts showed that the performance of these models was skewed towards non-trucks – the high accuracy in classifying passenger vehicles classes and Class 9 trucks conceals the deficiencies of the models in identifying other FHWA truck-related classes. Even though trucks generally account for approximately 5 to 20 percent in traffic streams, the adverse impact of misclassifying trucks could be significant. For instance, implementation of a biased model may underestimate the pavement damage caused by trucks, since pavement structures are disproportionately impacted by heavy trucks (*2*). From a planning perspective, unreliable truck counts may result in a flawed understanding of truck activities and misinformed policy decisions to manage future demand for truck movements and operations. Furthermore, the class bias issue also occurs within truck-related classes. Certain types of trucks in the FHWA classification scheme – such as Classes 7, 11, 12 and 13 – are not as frequently on the roadway network. On the other hand, Class 9 trucks are the most common multi-unit truck configuration observed along most corridors and show great variability in their body types. However, the basic assumption of canonical machine learning algorithms for classification problem is that the number of training instances in considered classes are relatively similar (*13*). Consequently, the imbalanced dataset poses a natural difficulty for many classification algorithms to correctly classify minority classes, since they are naturally biased towards majority classes. Notwithstanding, some of the minority classes remain prominent components in both pavement design (*2*) and freight forecast modeling (*4*, *5*).

To address this gap, this paper details the development of an accurate and transferable FHWA vehicle classification model using a truck-focused dataset. The model shows significant improvement over previous signature-based FHWA vehicle classification models (*9*, *10*) for all truck classes in terms of F1 scores, especially on minority classes such as Classes 7 and 11 which were overlooked by previous studies (*9*, *10*). The algorithm development comprised three steps. First, the training and testing dataset was split using stratified sampling to retain the representativeness of the training instances for minority classes. Then, a deep neural network (DNN) with dropout layers was constructed to reduce the generalization error. Finally, a bootstrap aggregating ensemble (bagging) was developed to address the classification challenge with an imbalanced dataset. In this step, bootstrap resampling was applied on the training set to approximate the feature distribution of the under-represented classes to facilitate a better understanding of those classes with limited training instances to learn from. The bagging DNN successfully improved the model performance on minority classes without compromising the accuracy of majority classes. The spatial and temporal transferability of the model was empirically tested using independent datasets.

2. LITERATURE REVIEW

Axle Sensor Methods

FHWA axle-based vehicle classification data has been traditionally collected via axle-based sensors, such as road tube arrays (*5*), piezoelectric sensors (*14*), WIM systems (*15*) and wireless accelerometer sensors (*16*), all of which have the ability to capture axle numbers and spacing configuration. Department of transportation agencies across the U.S. rely on existing classification sites equipped with such axle-based sensors for reporting FHWA vehicle classification counts. Efforts have been devoted to test the performance of classification sites and investigating methods to enhance the classification accuracy on these systems, which yield their own limitations on distinguishing classes with overlapping axle configurations (*14*, *15*). For example Classes 2 through 5 include two-axle vehicles (*15*) and the first axle-spacing of Classes 3, 5 and 8 share overlapping range values (*14*). Kwigizile et al. improved the correct classification rate for overlapping classes by breaking down the 13 FHWA classes into 28 detailed subclasses and using a probabilistic neural network to assign axle spacing values from a calibrated WIM site to those predefined subclasses (*15*). This model reduced the error rate of the calibrated WIM site from 9.5 percent to 6.2 percent. The author highlighted that the misclassification came from systematic errors due to overlapping axle configurations across FHWA classes. By introducing weight value as one of the input variables, the error rate further reduced to 3.0 percent. However, weight values vary spatially and temporally and may affect the transferability of their model. Bitar et al. adopted a probabilistic approach to improve the accuracy of the classification site equipped with piezoelectric sensors (*14*). In their study, a comprehensive classification error analysis was conducted on the overlapping axle configuration across FHWA scheme's categories. Their hypothesis was that axle spacing distributions are different for classes that may share similar axle configurations. Hence, the axle spacings associated with each class were fitted into Gaussian distributions. Subsequently, the optimal class boundary thresholds for the overlapping axle configuration were determined according to the estimated axle-spacing distributions. The error rate was significantly reduced from the original sensor outputs especially for Classes 3, 6, and 7 (*14*). However, axle-based sensors do not provide any truck general body type related information defined by the FHWA classification scheme. Therefore, it is a challenge for axle detectors to accurately differentiate trucks with overlapping axle-spacing distribution, even though they have distinctly different body types. For example, the error rate of their model on Class 4 (bus or bus with a trailer) was higher compared to the original sensor measurements. The first-spacing probability density distribution of Classes 4 and 5 had a significant overlap and did not present a clear class decision boundary, although those distributions came from two distinct body types -- buses in Class 4 and single-unit trucks in Class 5.

In addition, a prototype wireless accelerometer system, which detects the pavement vibration when vehicles traverse the detection area, have also been explored for estimating axle-based classification (*16*). The sensor identified the axle configuration through locating the vibration peaks. The accelerometer-based classification was evaluated using calibrated WIM classification results. This wireless sensor was able to achieve the same level of accuracy as the current WIM

system on estimating FHWA classes. However, accelerometer sensors would experience the same limitations as other axle detectors in distinguishing classes with overlapping axle spacing ranges.

The detection sites equipped with WIM systems or piezoelectric sensors are sparsely deployed in the roadway network due to their high installation cost. Hence, researchers investigated existing inductive loop sensors as an alternative, since they are widely deployed across the U.S and their installation and operation cost are relatively cheap.

Single Inductive Signature Methods

Unlike conventional inductive loop detector cards which sample at a rate of 240 Hz (*17*), advanced loop detectors capture detailed inductance changes when a vehicle traverses an inductive loop sensor. The resulting high-resolution waveform generated by the inductance change measurements is known as an "inductive vehicle signature" (*18*). However, the exact axle locations cannot be directly identified from inductive vehicle signatures (Figure 1), which is one challenge in classifying vehicles into the FHWA scheme.

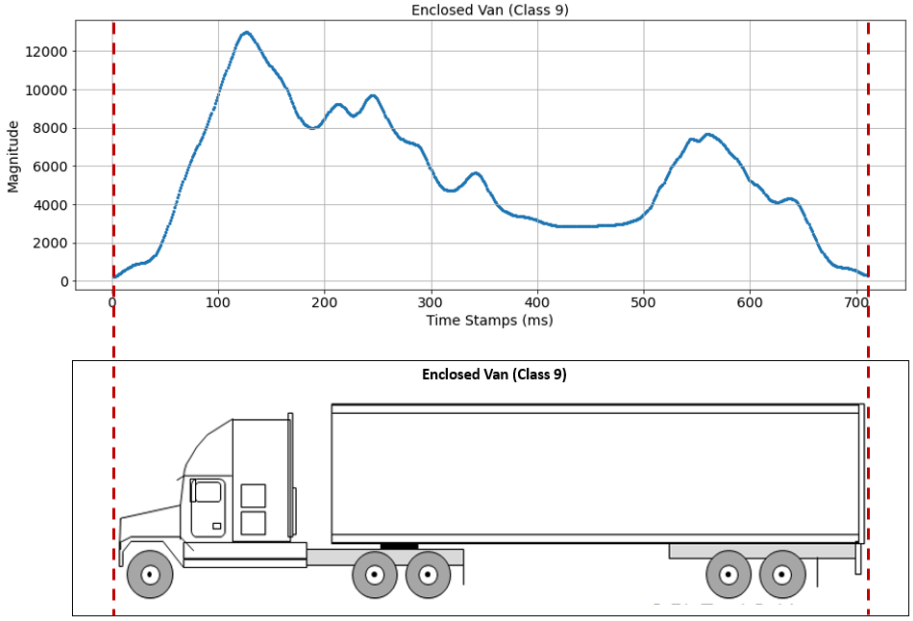


Figure 1 Class 9 Enclosed Van and its corresponding raw signature

Jeng and Ritchie made the first attempt to classify vehicle on the basis of FHWA scheme using single inductive loop signature data (*9*). The piecewise slope rates (PSR) of each interpolated signature were used as a reduced representation of each signature pattern. The PSRs of each signature were separated into five groups by visual observation of PSR plots of all vehicle classes (*19*). Unfortunately, due to the data limitation, Class 10 to Class 13 trucks, which have disproportionately severe negative impacts on pavement structure, were not considered in their model development process. Later, Jeng et al. enriched their dataset with multi-unit trucks and proposed a new vehicle classification algorithm (*10*). The new algorithm composed of two steps. First, vehicle signatures were transformed and reconstructed with wavelet transformation. Then, the transformed vehicle signatures were grouped into FHWA classes using K nearest neighbor.

Even though, the overall accuracy was 92 percent, the weakness on classifying minority classes was obscured by the high performance on predicting the majority classes (Class 2 and Class 3), which overlooked the poorer performance on several truck classes (Class 6, 7, 8, 11, 12, 13). Hence, one of the focus of this paper is to address the effect of dataset imbalance on the performance of minority classes.

3. DATA DESCRIPTION

The vehicle signature data used in this paper were collected at 20 different detection sites across California in 2012, 2013, and 2016. The geographical distribution of the 20 data collection sites is shown in Figure 2.

Figure 2 Data Collection site for model training, hyperparameter tuning and transferability testing

The selected detector sites experienced high truck volumes, a wide variety of truck types, and with the data collection effort spanning various traffic conditions. A total of 44,438 vehicle signature records were processed primarily at the truck lanes in each facility, with a resulting vehicle class distribution as shown in Table 2.

Table 2 Vehicle Class Distribution from ground truth

FHWA-CA Scheme	Counts	Imbalance Rate	Number of Body Types
1	2	0.0001	1
2	2,946	0.1494	4
3	8,203	0.4159	2
4	772	0.0391	3
5	7,055	0.3577	32
6	1,535	0.0778	27
7	279	0.0142	9
8	1,463	0.0742	24
9	19,724	1.0000	40
10	138	0.0070	16
11	1,518	0.0770	21
12	268	0.0136	12
13	3	0.0002	1
14	764	0.0387	14

The imbalance rate presented in Table 2 was the metric used to understand the quantitative relationship between majority and minority classes. It is defined as the ratio of each minority class to the majority class (Class 9). One key challenge in the model development was to accommodate the imbalanced dataset. When training an imbalanced dataset, models are generally prone to enhance the prediction accuracy of majority class (*13*). This will lead the majority classes and the overall model to achieve relatively high prediction accuracy at the expense of minority classes (*13*).

Typically, the issue of imbalanced datasets in classification is addressed at either the data or algorithm level (*20*). At the data level, minority classes are typically oversampled, while majority classes are undersampled to balance the dataset. However, undersampling may compromise the generality of the model. For instance, Class 9 is a majority class comprising heterogeneous body types with distinct signature waveforms, as shown in Figure 3.

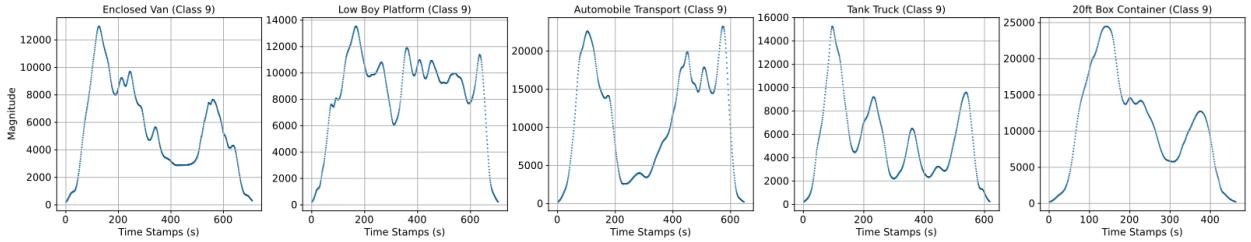


Figure 3 Signatures of different truck body types

Undersampling may cause information to be lost by removing unique vehicle configurations that would have helped the model to better capture the characteristics of diverse vehicles found in this class. On the other hand, synthetic data methods which are generally used in oversampling could create overlapping instances between the minority class and the majority class, and further reduce the prediction accuracy for the majority class (*20*). Therefore, this study investigated algorithm-level enhancements to improve the performance of the signature-based FHWA classification model.

According to Table 2, both Classes 1 and 13 have training instances extremely small and less than thirty, which are empirically considered as insignificant sample sizes. Moreover, this research primarily focused on the FHWA-CA truck-related classes. Therefore, Classes 1 (motorcycle) and 15 (unclassified vehicle) were excluded in the modeling process. Since Class 13 share similar body types as well as axle configurations with Class 12 vehicles, they were combined in the model. Classes 1 and 15 (Unclassified vehicle) may be included and Classes 12 and 13 may be split in the future with further enrichment of the dataset.

4. MODEL DEVELOPMENT

Prior to model development, stratified sampling was initially used to partition the training and testing datasets with a 70-30 split, respectively, in order to ensure that a sufficient number of samples for each class can be observed in both the training and testing sets.

4.1 Feature Extraction

First, raw signatures were processed using cubic spline interpolation to eliminate noises and obtain a set of feature vectors with the same dimension and normalized on both the horizontal (time) and vertical (magnitude) axis (*19*). This yielded a vector of 31 magnitude features equally spaced along the normalized time domain with 30 degrees of freedom. Subsequently, 30 differences were derived from 31 magnitude values and then further normalized along y axis, which forced magnitude and difference values fallen into the same scale. Finally, the last value of the vector of 31 magnitudes was dropped to retain the independence of the feature vector. This resulted in 60 (30 magnitudes and 30 differences) independent features that were used as inputs for the vehicle classification model. The feature extraction process is presented in Figure 4.

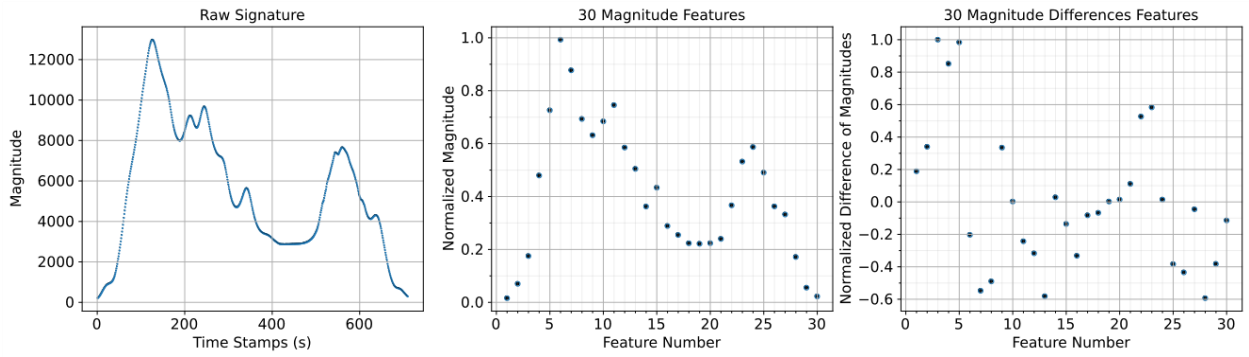
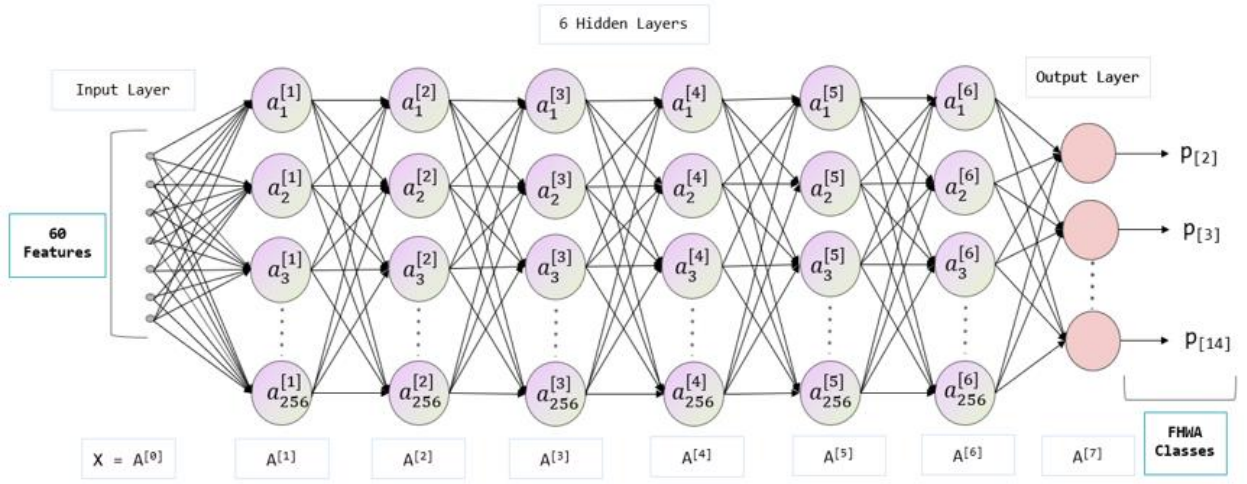


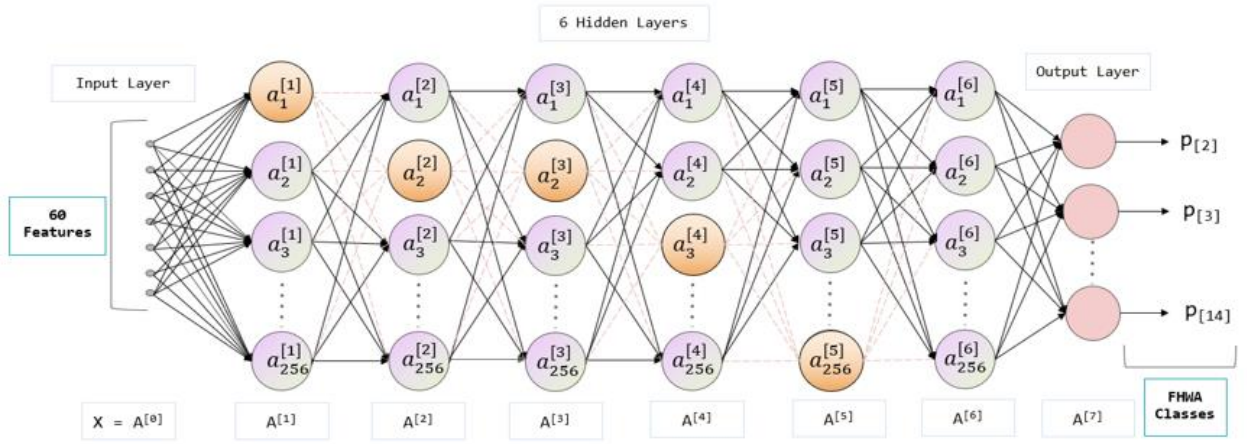
Figure 4 Preprocessing and Feature Extraction

4.2 Deep Neural Network Architecture

A deep neural network model with dropout regularization was developed to classify vehicles based on the FHWA-CA scheme. The model was constructed with 6 hidden layers, 256 neurons on each hidden layer (shown in Figure 5). The Rectified Linear Unit (ReLU) (*21*) was used as the activation function on each hidden layer, while the Softmax activation function was applied on the output layer to represent the probability distribution over the 12 FHWA categories (where Classes 12 and 13 were combined). To avoid gradients vanishing and exploding, He (*22*) and Xavier (*23*) weight initialization methods were applied to the hidden layers with ReLU and Softmax activation functions, respectively. The deep neural network model was trained with a minibatch size of 100 and learning rate of 0.001. The Adam optimizer (*24*) was adopted to solve this highly non-linear optimization problem.



(a) The Deep Neural Network Architecture



(b) The Deep Neural Network with Dropout Layers

Figure 5 Model Structure

Balancing bias and variance of the deep neural network model is an essential task during the hyperparameter tuning process. Bias represents the expected deviation from the true value of the function while variance measures the deviation from the expected estimator which is caused by the unseen dataset. The deep neural network model is typically prone to overfit the training set as models increase in complexity, especially for the majority classes. This results in a trained model with low bias and high variance as shown in Figure 6a, where the training error tends to decrease (low bias) and the testing error tends to increase (high variance). Dropout regularization – a computationally inexpensive but powerful regularization method for deep neural network models – was implemented at each layer in the deep neural network to prevent overfitting (*25*). 30 percent of the neurons within each hidden layer were randomly dropped out while the remaining 70 percent were retained (Figure 5b) during the training process. A comparison of early stopping without and with dropout regularization is shown in Figure 6c and Figure 6d, respectively. The training process needed to be terminated significantly earlier at 5 epochs without dropout regularization. This resulted in a poorer test data prediction accuracy of 0.87, compared with 0.91 for the latter.

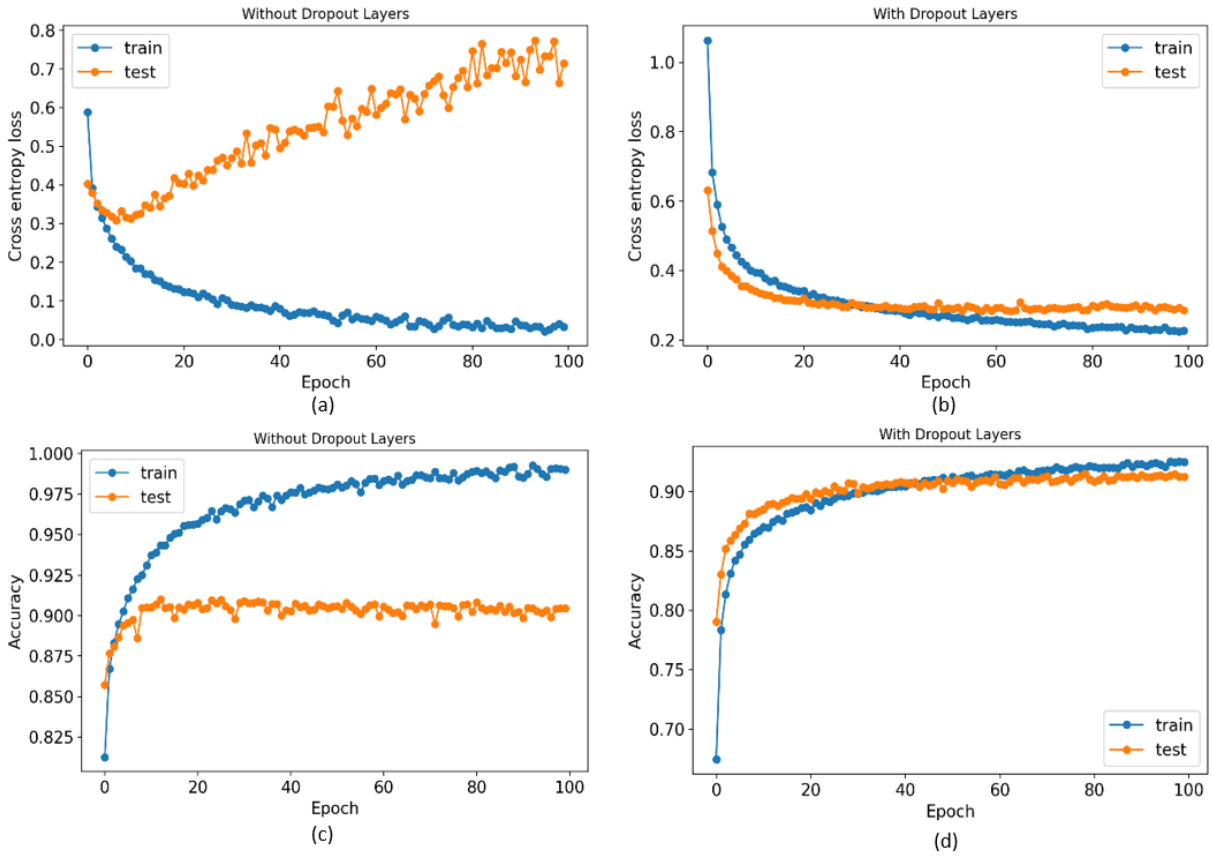


Figure 6 Learning Curve

4.3 Bootstrap Aggregating

As Table 1 shows, the labeled FHWA classes yielded an imbalanced class distribution. The number of instances belonging to a certain class, such as Class 9, was significantly higher than any other labeled classes in the dataset. The objective function of the designed neural network model is to minimize the global error rate. The entire cost function for a multi-class classification problem given *m* training instances labeled with *n* classes can be written as:

$$J = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} log\hat{y}^{(i)} + \left(1 - y^{(i)}\right)\log\left(1 - \hat{y}^{(i)}\right)\right) \quad (1)$$

Where, $y^{(i)}$ and $\hat{y}^{(i)}$ represents the true class and the predicted class of training instances i respectively.

As Equation 1 shows, the cost function does not handle the class distribution in the dataset. Without enough training instances to approximate the feature distributions of minority classes, the model is inclined to compromise the performance of minority to achieve low global error rate resulting in the poor performance on minority classes.

In order to have a better understand of feature distributions within each class, bootstrap aggregating ensemble was adopted for the model development. Bootstrap aggregation (Bagging) is an ensemble strategy used to enhance the generalizability of the model through the combination of several models trained by multiple bootsrap samples (*26*). The basic idea of Bagging comes from the bootstrapping resampling techique, which is used to approximate the empirical distribution of the observed data, especially for datasets with small class samples. Stratified bootstrapping was applied on the training set and ten sets of bootstrapped samples were formed by resampling the stratified training instances with replacement. The bootstrap samples were fed into the DNN model with the same model struture. Subsequently, the prediction scores from ten models were averaged and the class corresponding to the highest averaged prediction score was considered as the final decision. The bagging DNN model structure is shown in Figure 7.

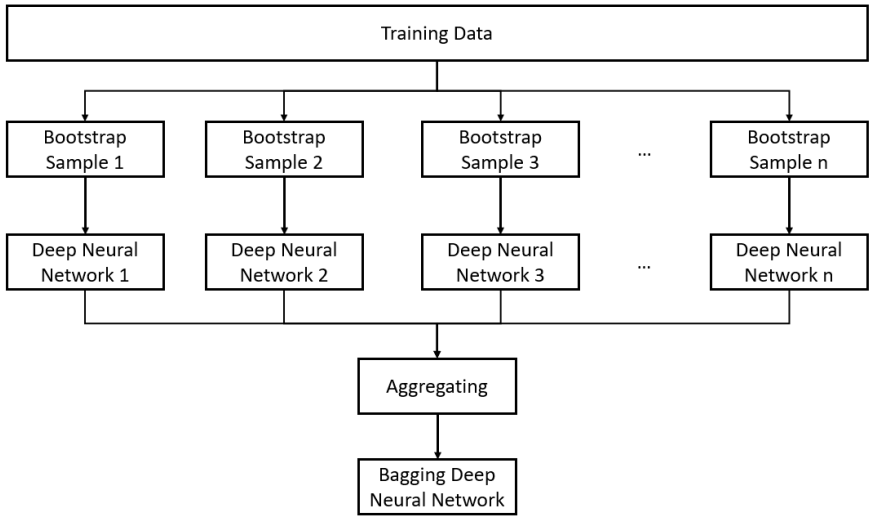


Figure 7 Illustration of Bagging DNN

5. Model Results

5.1 Evaluation Metrics

Determining appropriate performance measures is an essential task for evaluating the model built upon an imbalanced dataset. The accuracy measure (Equation 2) – which is derived from a confusion matrix (Table 3) – presents the percentage of total instances being correctly classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Table 3 Generic Confusion Matrix

		Predicted Class	
		Positives	Negatives
Actual Class	Positives*	True Positives (*TP*)	False Negatives (*FN*)
	Negatives*	False Positives (*FP*)	True Negatives (*TN*)

Note: For the illustration purpose, the positive cases were assumed to be the minority class and negative cases were the majority class

The accuracy measure is determined by both *TP* and *TN* (Table 3) on the numerator of the fraction. If the *TN*, which represents the correctly classified instance from the majority class, is disproportionally larger than *TP*, the accuracy value will still be large, even though the performance on the minority class remains poor. Therefore, this accuracy measurement is sensitive to class skews, and is not a reasonable metric for selecting models developed on an imbalanced dataset (*27*).

This section discusses three metrics that have generally been used to evaluate the performance of classification models built with imbalanced datasets. Precision (Equation 3) and recall (Equation 4) are two basic metrics, which are directly calculated from the confusion matrix (Table 3). Both precision and recall do not involve the true negative value, which represents the number of majority instances being correctly classified. Hence, these two metrics evaluate the model performance of majority and minority classes independently. The F1 score in Equation 5 is the harmonic mean of precision and recall. This metric accounts for precision and recall simultaneously while evaluating the models. In order to evaluate the models developed using imbalanced datasets, F1 score was primarily used in this paper for the model comparison.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\ score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{5}$$

5.2 Results Analysis

Figure 8 shows the recall distribution – which is also referred to as "Correct Classification Rate" (CCR) in previous research (*28, 29*) – for the DNN models built with 10 sets of bootstrapped samples. Minority classes such as Classes 4, 7, 10, and 12 & 13 resulted in relatively high

prediction variances, since limited training instances were used to learn from the features for each single model. Therefore, the bagging ensemble model was needed.

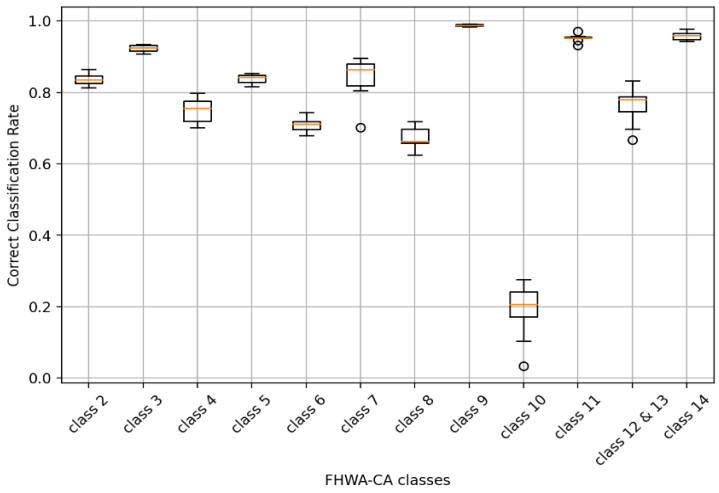


Figure 8 Correct Classification Rate across all Classes

Table 4 shows the F1 score for each class of three model structures. Dropout technique enhanced the generality of the DNN model and improved F1 score on testing set for all classes. With bagging ensemble, the model performance improved across most of the minority classes without compromising the model performance on the majority classes. The overall accuracy of the final model was 0.92 and the average F1 score was 0.83. This bagging DNN was applied on a spatially and temporally independent detection site and was still able to achieve an accuracy of 0.87 and average F1 score of 0.72.

Table 4 Test Result Comparison

	F1 Score of Single DNN Without Dropout	F1 Score of Single DNN With Dropout	F1 Score of Bagging DNN	Test Samples
Class 2	0.85	0.85	0.87	847
Class 3	0.87	0.88	0.89	2,057
Class 4	0.71	0.81	0.83	268
Class 5	0.82	0.85	0.87	1,579
Class 6	0.72	0.80	0.80	362
Class 7	0.77	0.86	0.84	77
Class 8	0.69	0.73	0.78	427
Class 9	0.98	0.98	0.99	4,318
Class 10	0.22	0.35	0.36	29
Class 11	0.96	0.94	0.96	276
Class 12 & 13	0.78	0.79	0.81	66
Class 14	0.96	0.97	0.97	175
Accuracy	**0.89**	**0.91**	**0.92**	**10,481**
Average F1 score	**0.78**	**0.82**	**0.83**	**10,481**

The performance of the bagging DNN approach was also compared with a state-of-the-art Wavelet-KNN -based classification algorithm developed by Jeng et al.(*10*). Jeng et al. evaluated

their model performance on each class using CCR. For the overall model performance, the accuracy value was selected as the evaluation metrics in their paper. However, such accuracy is a bias towards the majority class, which was class 2 in their dataset. Therefore, the F1 scores of their model were recalculated for a fair comparison. As Table 5 shows, the bagging DNN model achieves the same level of accuracy in terms of the accuracy metric. Considering the F1 score, the bagging DNN outperform the previous model (*10*). Except for Class 2, the F1 score for all classes are significantly higher than the previous approach (*10*). This indicates that the imbalanced dataset issue was well-managed by the bagging DNN model.

Since, trucks have disproportionately negative impact on pavement structures (*30*), having an accurate prediction results on truck-related classes (from class 5 to class 14) is essential to effective pavement design. Therefore, these two models were also evaluated using weight average, where truck-related classes were assumed to be at least two times more important than passenger vehicles. As Table 5 shows, the bagging DNN model was superior to the state-of-the-art signature-based FHWA vehicle classification algorithm on predicting truck-related class. The bagging ensemble model achieved a weighted average F1 score of 0.82, where the Wavelet-KNN model only had a value of 0.47.

Table 5 Model Comparison

	Bagging Deep Neural Network			**Wavelet-KNN (*10*)**		
	Recall (CCR)	**F1 Score**	**Testing Samples**	**Recall (CCR)**	**F1 Score**	**Testing Samples**
Class 1	N/A	N/A	N/A	0.83	0.81	74
Class 2	0.85	0.87	847	0.97	0.97	11,177
Class 3	0.93	0.89	2,057	0.78	0.79	1,568
Class 4	0.78	0.83	268	0.59	0.10	17
Class 5	0.86	0.87	1,579	0.67	0.72	543
Class 6	0.75	0.80	362	0.48	0.39	65
Class 7	0.87	0.84	77	0.67	0.13	3
Class 8	0.71	0.78	427	0.46	0.33	48
Class 9[1]	0.99	0.99	4,318	0.86	0.91	754
Class 10	0.17	0.36	29	0.67	0.11	3
Class 11	0.97	0.96	276	0.58	0.26	14
Class 12 & 13[2]	0.76	0.81	66	0.75	0.50	4
Class 14	0.97	0.97	175	N/A	N/A	N/A
Accuracy	**0.92**		**10,481**	**0.92**		**14,270**
F1 score	**0.83**		**10,481**	**0.50**		**14,270**
Weighted Average F1 Score (1:2)	**0.82**		**10,481**	**0.47**		**14,270**

Note: [1] Class 9 in the Wavelet-KNN model was combined with Class 14.[2] Classes 12 and 13 were split in the Wavelet-KNN model with CCR of 1.00 and 0.50 respectively, and with F1 score of 0.45 and 0.50 respectively.

5.3 Error Analysis

As Table 6 shown below, the model presented in this study achieved an F1 score greater than 0.80 for most classes, with the exception of Classes 8 and 10. According to the confusion matrix in Table 6, 8.7 percent of Class 8 vehicles are misclassified as Class 3 and 4.7 percent of Class 8

vehicles are misclassified as Class 5. This is mainly caused by the overlapping body types across Classes 3, 5 and 8.

Table 6 Confusion Matrix for Test Set

	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Class 11	Class 12&13	Class 14	Testing Samples	F1 Score
Class 2	716	122	0	8	0	0	1	0	0	0	0	0	847	0.87
Class 3	75	1921	1	44	0	0	13	3	0	0	0	0	2057	0.89
Class 4	1	6	208	39	7	1	4	2	0	0	0	0	268	0.83
Class 5	6	165	12	1359	23	1	10	3	0	0	0	0	1579	0.87
Class 6	1	0	8	72	271	9	1	0	0	0	0	0	362	0.80
Class 7	0	0	0	3	7	67	0	0	0	0	0	0	77	0.84
Class 8	0	37	4	20	0	0	303	63	0	0	0	0	427	0.78
Class 9	0	1	0	8	0	0	18	4286	3	0	0	2	4318	0.99
Class 10	0	0	0	0	0	0	3	20	5	0	0	1	29	0.36
Class 11	0	0	0	1	1	0	0	1	0	267	6	0	276	0.96
Class 12 & 13	0	0	0	0	0	0	1	1	0	13	50	1	66	0.81
Class 14	0	0	0	0	0	0	0	4	1	1	0	169	175	0.97

Note: Yellow cell indicate correct classifications by class. Grey cells highlight significant misclassifications.

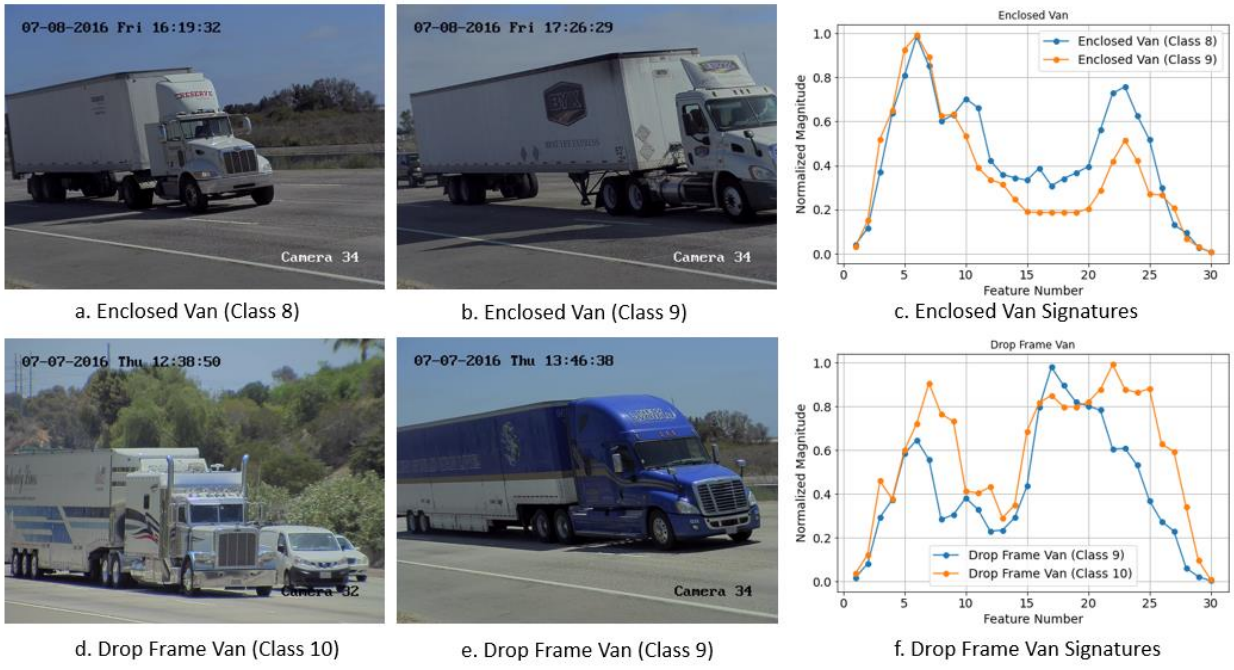
These three classes are also hard to be distinguished using current classification sites (*8*). As Figure 9 and Figure 10 present, some Classes 3, 5 and 8 trucks share very similar body types and axle configurations on their drive unit. Therefore, it remains a challenge to classify these vehicles accurately using either inductive loops or any axle sensors.

Figure 9 Class 3 vs Class 5 (*8*)

Figure 10 Class 3 vs Class 8 (*8*)

The main differences across Classes 8, 9 and 10 lie in the number of the axles that the truck-trailer combination has. Since inductive loop signatures do not have ability to directly capture the axle number of each truck-trailer, correctly distinguishing these classes has been a challenge for signature-based models (*10*). Nevertheless, the bagging DNN model significantly improved the performance of such classes over the Jeng et al.'s approach (*10*). However, misclassified vehicles were still observed among those classes due to the overlapping body type across FHWA classes. As Figure 11 shows, a Class 8 enclosed van (Figure 11a) was misclassified as a Class 9 (Figure 11b) vehicle, where Class 9 and Class 8 enclosed vans share similar shapes. Likewise, the bagging

DNN was found to misclassify a Class 10 drop frame van in Figure 12c into a Class 9, which is also a common axle configuration amongst drop frame vans.



a. Enclosed Van (Class 8) b. Enclosed Van (Class 9) c. Enclosed Van Signatures

d. Drop Frame Van (Class 10) e. Drop Frame Van (Class 9) f. Drop Frame Van Signatures

Figure 11 Overlapping body type across FHWA classes

Compared to conventional axle detectors, inductive loop signature data have demonstrated the ability to distinguish the body type of trucks with relatively high accuracy (*31*). Therefore, errors associated with overlapping axle configurations of different vehicle body types (refer to Figure 12) which are a major source of confusion at classification sites, were better managed by the signature-based bagging DNN model proposed in this study.

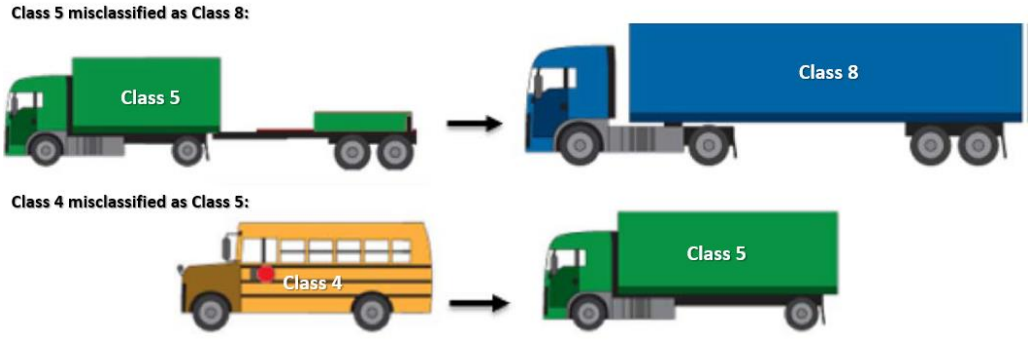


Figure 12 Error cases for piezoelectric sensors (*14*)

6. Conclusion and Discussion

This study proposed an accurate vehicle classification model to classify vehicle based on FHWA-CA scheme using single inductive loops. The proposed model was developed on a truck-focused dataset and utilized bagging ensemble technique to resolve the classification challenges of accommodating an imbalanced dataset which is typically observed in the FHWA classification scheme. The modeling process involved three major steps. First, the dataset was partitioned using the stratified sampling approach to retain a proportional number of samples from minority classes for both training and testing set. Then, a DNN model was constructed to assign signatures to their corresponding FHWA-CA classes. Dropout regularization was applied during the fine-tuning process, which successfully alleviated overfitting of the DNN model. Finally, a bagging ensemble technique was used to address the imbalanced dataset issue. This bagging DNN model significantly outperformed a state-of-the-art FHWA classification model using inductive loop signature data (*10*) for all truck-related classes. The bagging DNN model was able to achieve an F1 score of 0.83, where the comparable model obtained a value of 0.50. From the error analysis, it was observed that the majority error cases came primarily from the overlapping body types across FHWA classes. For instance, both Classes 8 and 9 shared body types such as semi-trailer enclosed van. In addition, semi-trailer drop frame vans exists in both Classes 9 and 10. According to the error analysis, the proposed model likely inferred the FHWA classes through their corresponding body types. Notwithstanding, the overlapping axle configurations with different general body types, which is a common type of the errors at classification sites (*14*) were still generally well-managed in the signature-based bagging DNN model.

Inductive loop sensors remain the most widely deployed detector infrastructure in the State of California and the United States, and inductive signature-based classification models have been widely deployed in the State of California (*32*). This research demonstrates that the improvements in the inductive signature-based model described in this paper is a cost-effective solution that can provide accurate classification performance across truck categories according to the FHWA scheme, while concurrently addressing common body configuration confusion issues experienced by axle-based detection systems.

Acknowledgements

This study was made possible through funding received by the University of California Institute of Transportation Studies from the State of California through the Public Transportation Account and the Road Repair and Accountability Act of 2017 (Senate Bill 1). The authors would like to thank the State of California for its support of university-based research, and especially for the funding received for this project. The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California. This paper does not constitute a standard, specification, or regulation.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Yiqiao Li, Andre Tok, Stephen G. Ritchie; data collection: Andre Tok; analysis and interpretation of results: Yiqiao Li, Andre Tok; draft manuscript preparation: Yiqiao Li, Andre Tok, Stephen G. Ritchie. All authors reviewed the results and approved the final version of the manuscript.

Reference:

1. Federal Highway Administration. *Traffic Monitoring Guide FHWA*. 2013.
2. Gillespie,T.D,Karamihas,S.M. & Sayer, M. W. *Effects of Heavy-Vehicle Characteristics on Paveemnt Response and Performance (NCHRP Report 353)*. 1993.
3. Guensler, R., S. Yoon, H. Li, and J. Jun. *Heavy-Duty Diesel Vehicle Modal Emission Model (HDDV-MEM) Volume I: Modal Emission Modeling Framework*. 2005.
4. Chase, K. M., P. Anater, and T. Phelan. Freight Demand Modeling and Data Improvement. *Freight Demand Modeling and Data Improvement*, No. December, 2017. https://doi.org/10.17226/22734.
5. Beagan, D., D. Tempesta, and K. Proussaloglou. *Quick Response Freight Methods*. 2019.
6. Federal Highway Administration. Compilation of Existing State Truck Size and Weight Limit Laws. https://ops.fhwa.dot.gov/freight/policy/rpt_congress/truck_sw_laws/app_a.htm. Accessed Jul. 20, 2020.
7. Quinley, R. WIM Data Analyst's Manual: FHWA Report IF-10-018. 2010, p. 183.
8. FHWA (Federal Highway Administration). Verification, Refinement, and Applicability of Long-Term Pavement Performance Vehicle Classification Rules: Chapter 2. Introduction to Vehicle Classification. https://www.fhwa.dot.gov/publications/research/infrastructure/pavements/ltpp/13091/002.cfm. Accessed May 16, 2020.
9. Jeng, S. T., and S. G. Ritchie. Real-Time Vehicle Classification Using Inductive Loop Signature Data. *Transportation Research Record*, No. 2086, 2008, pp. 8–22. https://doi.org/10.3141/2086-02.
10. Jeng, S. T., L. Chu, and S. Hernandez. Wavelet-k Nearest Neighbor Vehicle Classification Approach with Inductive Loop Signatures. *Transportation Research Record*, No. 2380, 2013, pp. 72–80. https://doi.org/10.3141/2380-08.
11. Li, Y., A. Y. C. Tok, and S. G. Ritchie. Individual Truck Speed Estimation from Advanced Single Inductive Loops. *Transportation Research Record*, Vol. 2673, No. 5, 2019, pp. 272–284. https://doi.org/10.1177/0361198119841289.
12. Leevy, J. L., T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya. A Survey on Addressing High-Class Imbalance in Big Data. *Journal of Big Data*, 2018, p. 30.
13. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence*, Vol. 5, No. 4, 2016, pp. 221–232. https://doi.org/10.1007/s13748-016-0094-0.
14. Bitar, N., and H. H. Refai. A Probabilistic Approach to Improve the Accuracy of Axle-Based Automatic Vehicle Classifiers. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18, No. 3, 2017, pp. 537–544. https://doi.org/10.1109/TITS.2016.2580058.
15. Kwigizile, V., R. N. Mussa, and M. Selekwa. Connectionist Approach to Improving

Highway Vehicle Classification Schemes - The Florida Case. *Journal of the Transportation Research Board*, No. 1917, 2005, pp. 182–189.

16. Ma, W., D. Xing, A. McKee, R. Bajwa, C. Flores, B. Fuller, and P. Varaiya. A Wireless Accelerometer-Based Automatic Vehicle Classification Prototype System. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 1, 2014, pp. 104–111. https://doi.org/10.1109/TITS.2013.2273488.

17. Coifman, B., and S. Neelisetty. Improved Speed Estimation from Single-Loop Detectors with High Truck Flow. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, Vol. 18, No. 2, 2014, pp. 138–148. https://doi.org/10.1080/15472450.2013.801708.

18. Sun, C., and S. G. Ritchie. Individual Vehicle Speed Estimation Using Single Loop Inductive Waveforms. *Journal of Transportation Engineering*, Vol. 125, No. December, 1999, pp. 531–538.

19. Jeng, S. T. *Real-Time Vehicle Reidentification System for Freeway Performance Measurements*. University of California, Irvine, 2007.

20. Aouatef Mahanil; Ahmed Riad Baba Ali. Classification Problem in Imbalanced Datasets. In *Intech*, p. 13.

21. Nair, V., and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning*, No. 3, 2010, pp. 807–814. https://doi.org/10.1.1.165.6419.

22. He, K., X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2015 Inter, 2015, pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.

23. Glorot, X., and Y. Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *Journal of Machine Learning Research*, Vol. 9, 2010, pp. 249–256.

24. Kingma, D. P., and J. L. Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.

25. Mele, B., and G. Altarelli. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Physics Letters B*, Vol. 299, No. 3–4, 2014, pp. 345–350. https://doi.org/10.1016/0370-2693(93)90272-J.

26. Breiman, L. Bagging Predictors: Technical Report No. 421. *Department of Statistics University of California*, No. 2, 1994, p. 19.

27. Joshi, M. V. On Evaluating Performance of Classifiers for Rare Classes. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2002, pp. 641–644. https://doi.org/10.1109/icdm.2002.1184018.

28. Hernandez, S. V., A. Tok, and S. G. Ritchie. Integration of Weigh-in-Motion (WIM) and Inductive Signature Data for Truck Body Classification. *Transportation Research Part C:*

Emerging Technologies, Vol. 68, 2016, pp. 1–21. https://doi.org/10.1016/j.trc.2016.03.003.

29. Sahin, O., R. V. Nezafat, and M. Cetin. Methods for Classification of Truck Trailers Using Side-Fire Light Detection and Ranging (LiDAR) Data. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, Vol. 0, No. 0, 2020, pp. 1–13. https://doi.org/10.1080/15472450.2020.1733999.

30. Gillespie, T., and S. Karamihas. Heavy Truck Properties Significant to Pavement Damage. *Vehicle-Road Interaction*, 2009, pp. 52-52–12. https://doi.org/10.1520/stp13248s.

31. Hernandez, S. V. *Integration of Weigh-in-Motion and Inductive Signature Data for Truck Body Classification*. University of California, Irvine.

32. Tok, A., K. (Kate) Hyun, S. Hernandez, K. Jeong, Y. (Ethan) Sun, C. Rindt, and S. G. Ritchie. Truck Activity Monitoring System for Freight Transportation Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2610, No. 1, 2017, pp. 97–107. https://doi.org/10.3141/2610-11.