IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 14, 2021, accepted July 16, 2021, date of publication July 21, 2021, date of current version August 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3099029

A Universal Automated Data-Driven Modeling Framework for Truck Traffic Volume Prediction

AMIRSAMAN MAHDAVIAN[1], (Graduate Student Member, IEEE), ALIREZA SHOJAEI[2], MILAD SALEM[3], (Graduate Student Member, IEEE), HALUK LAMAN[1], NAVEEN ELURU[1], AND AMR A. OLOUFA[1]

[1]Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA
[2]Myers-Lawson School of Construction, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[3]Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816, USA

Corresponding author: Amirsaman Mahdavian (amirsaman@knights.ucf.edu)

ABSTRACT Knowledge of the truck traffic volumes on state and interstate highways is critical for highway authorities and federal organizations. Increased urbanization, population growth, and economic development have led to an increased demand for freight travel. Several planning applications demand reliable and accurate truck traffic prediction. A review of the available literature indicated that limited research had been performed on the development and utilization of a universal automatic framework for truck traffic volume prediction. As a result, there is a gap to incorporate inclusive predictors, a broad dataset, a comprehensive feature selection approach, and a robust cross-validation method that utilizes both linear and non-linear algorithms. The present study uses a hyperparameter optimization framework to select the appropriate feature selection method and modeling approach among a comprehensive list of available state of the art approaches. Distinct from models based on individual case studies, the proposed framework allows for greater customization and minimized MAPE error. The developed framework automates much of the traffic count forecasting process, and the resulting method is less labor-intensive and may be utilized without the need for experienced data analysts. Florida's interstate highways historical traffic data were used to test the feasibility of the proposed framework. The results of the Florida Case Study revealed the superiority of non-linear models in the generalization and prediction of traffic volumes over linear models. The random forest algorithm results on the test dataset in this study demonstrate this model's ability to predict truck traffic with 86% accuracy. Spatial variables were the most significant variable group, followed by road characteristics.

INDEX TERMS Data-driven modeling, truck traffic, traffic volume, prediction model, regression analysis, forecasting, machine learning.

I. INTRODUCTION

The extent of truck road travel in the U.S. has substantially increased due to various disruptive effects. These include the impact of technology and social and demographic changes, urbanization and globalization, environmental and energy trends, economic and workforce changes, and political and fiscal trends. Despite other transportation modes, trucks remain the principal mode of freight transportation, and about 69% of the total national tonnage is transported by truck. A growing economy and the evolution of time-sensitive freight services have significantly increased the number of trucks on the nation's highways. According to the Texas A&M Transportation Institute [1], truck trips can be expected to increase from 557,000 daily trips in 2014 to over one million daily trips by 2040. These higher truck volumes will have a substantial impact on the level of congestion and air quality in many regions. Therefore, the rapid growth in truck traffic has become a crucial issue for traffic managers, decision-makers, and road users.

The differences in size and operation between trucks and cars means that trucks could potentially harm the surrounding traffic, leading to an increase in crash severity, driver frustration, and vehicular emissions. These factors result in a greater need for accurate truck traffic prediction, which can be crucial in the design and management of road pavement and bridges, reconditioning and reconstruction of highway pavement, planning for truck movements, environmental

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

impact analysis, and investment policies. The growing importance of truck trips in both engineering and planning has created a need for truck volume estimation. Various prediction methods have been generated using data typically available for planning applications. Metropolitan Planning Organizations (MPOs), transportation planners, decision-makers, and researchers have strived to address the issue of forecasting truck traffic to estimate future highway needs.

Analysis of the current literature detailing truck traffic volume prediction has been limited to a few algorithms, methodologies, and a selective subset of variables in each publication. As a result, there is a gap for utilizing a universal automated framework. The framework for this study is tested through a comprehensive dataset, inclusive of Florida highway traffic data from 2001 to 2017. Fifty-nine independent variables were used as inclusive predictors, with the use of both liner and non-liner algorithms (5 and 4 algorithms, respectively) trained and tested utilizing a robust cross-validation method. The most appropriate selection method and modeling approach to reduce Mean Absolute Percentage Error (MAPE), error were automatically selected using a hyperparameter optimization framework, also known as a grid search. There is an apparent demand in the literature for the development of a universal automated framework. Previous attempts to select an optimal method of truck volume forecasting have not been comprehensive in multiple areas of analysis, including the methods of feature selection, the algorithms used, and additional aspects of the truck volume forecasting pipeline such as the number of predictors. Furthermore, optimization attempts tailored toward a single case study are less advantageous than a universal automated framework that can become customized based on specific parameters.

The proposed framework addresses the constraints of previous modeling attempts in truck traffic estimation and can be readily used to aid decision makers and contractors in the planning stages of project development. Additionally, the results presented in this study confirm the high degree of prediction accuracy of this model. Others can utilize this framework through the methodology outlined in this study. While the dataset discussed is optimized for the State of Florida, new users may incorporate their local projections for the efficient modeling of truck traffic prediction. The development of a more generalized, universal automated framework that can be applied to multiple scenarios provides a substantial benefit and offers customization on a case-by-case basis.

II. LITERATURE REVIEW

Traffic volume prediction has gained growing attention due to the accelerated advancement and implementation of intelligent transport systems (ITS). With the widespread use of traffic sensors and recently developed traffic sensor technologies, there is a copious amount of traffic data available, bringing the transportation sector into the age of big data. Moreover, a growth in the field of connected, automated, shared, and electric (CASE) vehicles plays prominent roles in the number of trips by trucks [56], [57], and [58]. As a result, transportation management is currently experiencing a transformation in the attempt to employ more data-driven methods. The current practices in truck trip activity estimation and freight modeling, fall short in meeting today's need. The level of data complexity makes it essential to reevaluate the traffic volumes forecast dilemma based on deep-structured non-linear models with a considerable amount of traffic data. Several techniques have been developed to estimate freight movements, and can be broadly classified under two categories, commodity-based and vehicle-based approaches.

Generally, the traffic flow forecast can be classified into three classes, short-term forecasting, medium-term forecasting, and long-term forecasting. The one with the period from five to thirty minutes is usually noted as the short-term forecast, that from thirty minutes to a few hours is the medium-term forecast, and that a day or longer is the long-term prediction. Traffic volume forecasting that is a type of long-term prediction can be obtained through forecasting methods such as econometric regressions, travel-demand modeling, and neural network modeling [2].

Short- and mid-term prediction models: Since the 1980s, scholars have started to investigate short- and mid-term traffic flow prediction, which is believed to be useful for real-time traffic control [3]. The neural network (NN) algorithm has been frequently applied in civil engineering projects [59] and more specifically for traffic flow prediction from most beginning research to today due to their strength in handling non-linearity and universal approximability of unknown functions that exist in traffic behavior. Zheng *et al.* [4] mixed NNs and Bayesian inference to predict future traffic flow. Besides the NN methods, there are many other prediction approaches such as the Kalman filter [5], time series models [6], [7], the support vector regression (SVR) [8], the k-nearest neighbor [9], the hybrid models [10], [11] and the gradient boosting tree regression [12]. The comprehensive information on existing models can be obtained in Vlahogianni *et al.* [13] and Lippi *et al.* [14].

Different states have developed their own models for forecasting freight movements, most of which are commodity-based. In 2000, the Indiana state authorities built a database of commodity flows within the state, employing the Commodity Flow Survey from 1997 to forecast the freight movement for the entire state.

There are two widely used approaches in estimating a regional level freight trip generation; the first approach is the vehicle-based model, where the number of each type of vehicle is generated via a conducting mode split for classification along with trip generation. The second approach is the commodity-based model, in which average payload factors are estimated to convert the tons or value of commodity into the number of trucks. These models were generated to forecast intra-urban area movements of trucks and commodities by mode, and the additional movements generated by

the state, regional, and national movements of freight and commodities into urban areas.

A typical vehicle-based method produces truck trip estimates using land-use and socio-economic data [15]. Based on the object for which a model is used, models are categorized into several subgroups. These subgroups can be listed as; Traffic Count-Based Models, GIS-based Models, Self-Calibrating Gravity Models, Partial Matrix Techniques [16], Heuristic Models [17], Facility Forecasting Techniques [18], and others. The traditional four-step model has been of the most common technique, which is performed as a combination of the techniques mentioned above based on the needs and preferences of the state agencies. It is being utilized to predict the number of internal and external trips made inside an area by type, time of day, zonal Origin-Destination pair, mode of travel employed to make the trip, the routes took, or others.

A. MATHEMATICAL VEHICLE-BASED TRUCK TRAFFIC PREDICTION MODELS

Mathematical models are used to forecast truck traffic over particular network links and nodes [19]. These models are generally large in size and complexity, make various assumptions, and need adaptations of robust linear and non-linear programming algorithms to simplify the calculations [20]. Over the last decades, many mathematical traffic volume prediction models have been developed to support traffic management and enhance transportation efficiency. The development of a traffic counts model can be perceived as a temporal and spatial method. With the spread of ITS detectors, real-time traffic data became available. Traffic flow prediction based on traffic counts along with capacity and environmental factors are being used to forecast short- and mid-term traffic patterns [21]–[23]. Furthermore, long-term traffic forecasting via Average Daily Traffic (ADT), Monthly Average Daily Traffic (MADT), and Average Annual Daily Traffic (AADT) predictions of corridors or segments for both each type of vehicles represent another approach where historical averaged traffic counts have been used with exploratory variable groups [24]–[26]. These forecast methods can be categorized into three classes, including naïve, parametric, and non-parametric methods. Following is a brief review of each method.

1) PARAMETRIC MODELS

The structure of parametric models is predetermined, and the parameters of the model must be determined by utilizing data. The intrinsic knowledge of traffic processes in traffic simulation models can be captured in these structures. Overall, a lower quantity of data is required compared to non-parametric models. The traffic simulation models utilize the origin-destination (OD) traffic matrix considering the theory of network equilibrium. Traffic simulation models consist of macroscopic, microscopic, and mesoscopic modeling. In macroscopic modeling, the global variables of a roadway network are analyzed, including mean speeds, densities, and traffic flows. In macroscopic models, also named as kinematic wave models, trip generation rates and multiple linear regression models are commonly used methodologies. This approach was named as the LWR model and was first introduced by Lighthill and Whitham [27]. In microscopic modeling, the interactions between private vehicles are simulated based on the longitudinal (car-following) and lateral (lane changing) behavior of vehicles in a network system. Kometani and Sasaki [28] introduced the first car following model derived by Newton's equations. Lastly, in mesoscopic modeling, there is a blend of macroscopic and microscopic modeling [29].

2) NON-PARAMETRIC MODELS

The non-parametric title does not imply that these models completely lack parameters. Instead, it signifies that the features and number of the parameters are not fixed in the beginning and are adjustable. In non-parametric models, the form and selected parameters need to be determined by investigating the data. Furthermore, no awareness of the underlying methods is needed [30]. Usually, more data are required for the analysis process of non-parametric modeling compared to parametric models. The dynamic, complex, and non-linear characteristic of the traffic flow makes it a suitable phenomenon for non-parametric methods. Polson *et al.* [31] stated that transitions within the traffic-free flow, recovery time, breakdown probability, and average travel time, reflects a sharp non-linearity in the traffic flow which makes its predictions more complicated. Polson *et al.* [31] and Oswald *et al.* [32] claimed that the non-parametric models' superior capability to capture temporal-spatial relationships and non-linear patterns, make them more accurate for traffic forecasting compared to the parametric models.

K-Nearest Neighbor (KNN) approach is among the most well-known non-parametric modeling methods where the k events of the historical database, which are most similar to the current traffic situation, are used to forecast the desirable data point. Based on their distance of the nearest events to the current situation, the results are calculated using a simple average or weighted average method. Smith *et al.* [33], Rice and Van Zwet [34], Bajwa [35] and You and Kim [36] have indicated that KNN is a computationally fast technique that can outperform the naïve method; however, no studies have found it to be more accurate than more advanced non-parametric methods. Locally weighted regression is another non-parametric method that Nikovski *et al.* [37] and Zhong [38] reported shows excellent results in terms of forecasting accuracy and calculation time. Polson *et al.* [31] stated that the Bayesian network method can help handle large-system level transportation network problems.

Random Forest (RF), *Decision Tree (DT)*, and *Support Vector Regressor (SVR)* are among other non-parametric models used for traffic flow prediction. DT allows for the creation of a highly interpretable model on the traffic data, which can be used for finding common patterns shared between different traffic data points [39], [40]. Liu and Wu [41]

proposed using RF for traffic flow prediction due to its robustness and practicality and showed the generalization capabilities of this model. SVR has been leveraged for modeling traffic data and has shown superior performance when compared to linear models [42], [43].

Finally, *Neural Networks (NNs)* are the most extensively utilized models in traffic prediction because they are capable of modeling non-linear and dynamic processes proficiently. Even if the underlying relationships in a dataset are not clear, a neural network-based model is competent in generalizing accurate forecasts due to its non-parametric and non-linear characteristics [44]. However, neural networks were regarded as a black box and difficult to fully understand since they contain many nodes, elaborate structures, and non-linear functions [14], and [45].

B. LEADING PREDICTORS FOR TRUCK VOLUME PREDICTION

Al-Deek *et al.* [46] reported that the primary factors affecting truck volume were found to be the amount and direction of cargo vessel freight and the weekday of operation. Tsapakis *et al.* [26] developed 12 models based on regression and Bayesian analysis using data taken from 67 continuous data recorders to predict the AADT for heavy-duty trucks. Roadway functional class, population density, and spatial location had the highest importance factor in their developed daily truck traffic prediction models. Golias *et al.* [47] presented a statistical approach using a stepwise linear regression to create predictive models for estimating truck volumes. Number of employees estimated sales volume, and the number of establishments based on the standard industrial classification for the region are considered to be good predictors of truck volumes [47]. Lu *et al.* [48] developed a truck volume prediction model; results revealed that both linear and compound growth models fit the truck traffic growth trends well. Growth rates estimated from less than six years of data may have considerable variation, which can lead to significant errors in pavement response prediction. Roadway characteristics and socio-economic factors cannot be used to predict truck traffic growth rates with high accuracy directly. However, some factors are significantly associated with traffic growth, which can assist pavement designers in selecting appropriate defaults for traffic growth rates. These factors include population density, population density growth rate, land use, and highway functional classification [48].

C. THE CURRENT STATE OF PRACTICE AT FDOT

With the latest updates made in January 2020, the Florida Statewide Model (FLSWM) [49] for travel demand forecasting is a traditional four-step model with a freight demand modeling component named FreightSIM. In the four-step model developed using the Citilabs Cube Voyager and Avenue software platform, trips are generated from the 2010 OD Survey in Florida at the census block level, and traffic counts from 2001 to 2015 are used for validation and calibration at the TAZ level. Trip distribution is performed by gravity models combined with multinomial logit models for destination choice. To forecast truck traffic, the analysis modules used in the FreightSIM model includes sound synthesis, supplier firm selection, distribution channels, shipment size and frequency, modes and transfers, and finally, freight trip assignment that is integrated into the overall highway assignment as truck traffic. Also, the input/output database from the U.S. Bureau of Economics, port tonnage information, employment data from County Business Patterns (CBP), and freight flow from the freight analysis framework version 4 (FAF4) is utilized in the calibration and validation of the FreightSIM model. FAF is a national framework developed by the Bureau of Statistics and Federal Highway Administration (FHWA) to provide a comprehensive understanding of the overall picture of freight movements among the US and forecast from 2020 through 2045 for both optimistic and pessimistic growth scenarios. In the FDOT model, some of the socioeconomic variables, and freight-related economic variables, with the 2010 origin-destination (OD) survey of Florida, were employed to predict future traffic counts.

The truck counts prediction model on state highways developed in this study may assist transportation planners and decision-makers in inserting highly accurate traffic counts into their four-step or activity-based models. In doing so, they would be able to increase the robustness of predictions and quantify more accurate truck traffic in order to assist with near-, mid-, and long-term planning solutions. A review of the literature demonstrates there has not been substantial research thus far on the development of a universal automated framework for truck volumes prediction models. The model presented in this study incorporates a broad dataset (Florida highways between 2001 and 2017) and inclusive predictors (59 independent variables) utilizing both the linear (5 algorithms) and non-linear (4 algorithms) algorithms, employing a robust cross-validation method. Furthermore, the pipeline of this study incorporated a hyperparameter optimization framework (or grid search) to identify the best feature selection method, and the modeling approach in order to decrease the MAPE error.

This study aimed to build a model to fill these identified gaps to help contractors and planners enhance the truck count estimation. The results of this study demonstrate the high accuracy of the developed framework that could be easily generalized and employed by other users. By following the step-by-step methodology described in this research, and utilizing data related to their local predictors and projects, users can optimize this truck volumes prediction models accordingly. The final model and leading factors may vary from the ones selected for the tested predictors and dataset optimized for the state of Florida in this study.

III. METHODOLOGY

The main goal of this study was to develop a framework to generates a highly accurate long-term truck traffic prediction model for U.S. highways using an extensive pool of independent variables. This research utilized historical monthly

average daily truck traffic (MADTT) data and employed machine learning methods. This study aims to address the following objectives:

- Evaluating the prediction accuracy of multiple machine learning algorithms that consider multiple linear and non-linear relationships between variables to forecast the truck traffic volumes.
- Investigating the impact of the socio-economic, energy market, U.S. economy, and construction market on the truck traffic patterns.
- Examining the impact of road characteristics on the truck traffic volumes.
- Assessing the impact of spatio-temporal predictors on the MADTT.

In this study, various machine learning models were trained on national highways to forecast their MADTT. The primary data for this research were obtained from the Florida Department of Transportation's (FDOT) historical traffic database, which contains the past 17 years (from 2001 to 2017) of traffic data for the 259 sites. The database contains the historical monthly average daily truck traffic (MADTT) for highways, interstate, county, location, max speed, number of lanes, K factor, D factor, and truck percentage information for each site under this study.

Since the periodic pattern is crucial to the modeling procedure, the data preparation is designed to include this pattern implicitly in the inputs that are given to the model. This inclusion ranges from including the time of the data sampling, to many inputs which themselves change periodically. Since the temporal level of dependent and independent variables should be matched in the modeling, and the highest-resolution of the accessible independent data were the monthly level, authors were limited to use the monthly level rather than other higher-resolution truck traffic data such as weekly, daily or hourly level information. However, for the periodic pattern of the traffic data, a study by Sun *et al.* [54] employed a two-layer fast Fourier transform (FFT)-based traffic prediction scenario in which the discrete wavelet transform (DWT) with two different threshold values were adopted to decompose the high-frequent-noise of the traffic data.

The dataset utilized in this study contains 52,836 monthly data points for six interstate highways of Florida, including four primary and two auxiliary interstate highways. Table 1 presents detailed information regarding the highways and the number of their studied sites used in the project. On average, the MADTT data of a site studied about every 5.75 miles of the road on the Florida highways.

Also, in this study, the authors used data from both telemetered traffic monitoring sites (TTMS) and portable traffic monitoring sites (PTMS). TTMS is continuous traffic monitoring sites that send traffic data to the Transportation Statistics (TranStat) office by wireless communications or phone. PTMS are traffic monitoring sites that have loops and axle sensors in the road with leads running back into a cabinet located on the shoulder, to achieve higher accuracy. As shown in FIGURE 1, 259 sites studied in this paper cover the majority of the Florida interstates. 211 PTMS sites (green colored sites) and 48 TTMS sites (red-colored sites) data were collected from the FDOT's database.

TABLE 1. Interstate highways and sites under the study in this project.

Interstate ID	Interstate type	Length (mi)	Number of cosites / interstate highway	Road length/site
I95	Primary	382.0	82	4.6
I10	Primary	362.2	52	6.9
I75	Primary	470.6	68	6.9
I4	Primary	132.2	47	2.8
I275	Auxiliary	60.6	9	6.7
I 110	Auxiliary	6.3	1	6.3
		Total = 1414.3	Total = 259	Mean = 5.7

FIGURE 1. 259 sites included in this study on the Florida interstates map.

A. STATISTICAL ANALYSIS

The dataset utilized in this study contains 52836 trucks monthly data points for six interstate highways of Florida, including four primary and two auxiliary interstate highways. Table 2 depicts the statistical information of the directional monthly traffic flow for cars for all the dataset of this study (The term ''directional'' refers to the 2 different datasets, one is for the direction of south to north of the road, and the other one is for the north to south direction.). Statistical analysis was used to identify the range for most of the data. Means and medians describe central tendency, and percentiles help identify the range for most of the data.

B. PREDICTOR VARIABLES

This research utilized a pool of 59 candidate variables as predictors, obtained from relevant resources and the previous research, to develop the truck traffic model more precisely. The pool of candidate variables selected in previous studies mentioned in the literature review section, includes population density, growth rate, land use, highway functional classification, spatial, jobs (number of employees, and income),

TABLE 2. MADTT data statistical description.

Item	N/E Trucks	S/W Trucks	Total Trucks
Mean	1,134,867	1,344,004	2,478,871
Std	97,695	102,439	199,560
Minimum value	943,280	1,157,397	2,100,677
First quartile	1,061,814	1,263,920	2,324,526
Median value	1,105,466	1,316,157	2,419,644
Third quartile	1,215,878	1,433,383	2,651,906
Maximum value	1,359,654	1,567,772	2,926,388

crude oil price, and GDP. However, this project utilized seven categories of independent variables (59 variables) including construction market variables (5 variables, such as building permits, and construction spending), energy market variables (4 variables, such as crude oil price, and electricity price), socioeconomics variables (11 variables, such as population, and employees), U.S. economy variables (27 variables, such as CPI, GDP, and DJI), road characteristics variables (4 variables, such as the number of lanes, and max speed), temporal variables (3 variables), and spatial variables (4 variables, such as county name, and interstate ID). The spatial variables were added to the second developed model of this study. FIGURE 13 in Appendix A shows the specific factors of each type of predictor category.

This project collected the local and global data related to continuous socioeconomic variables for Florida, such as income, household size, licensed drivers, labor force, and length of the paved road. Concerning the U.S. economy, while Gross Domestic Products (GDP) mirrors the national income and economic health of the U.S., the Consumer Price Index (CPI) is broadly used to represent inflation at the national level. Moreover, factors representing interest rates at the national level are crucial macroeconomic indicators; the prime loan rate and federal funds rate are two popularly employed measures representing interest rates. Additionally, the evaluation of stock market indices as leading indicators of construction cost is another possible measure since they are widely accessible.

Regarding the construction market, construction spending is a measure of the value of new construction activities, including non-residential projects. The employment level in construction is a valuable measure to describe the U.S. and FL labor force in the construction sector of the economy. Ultimately, the number of new privately-owned housing units with authorized construction (housing permits) provides useful information about expected construction activity in the near future. Regarding energy prices, this category of variables has been widely neglected as one of the possible leading indicators in truck traffic prediction models. ''Crude oil prices,'' ''gas prices,'' ''natural gas prices,'' and ''electricity prices'' were used as a measure for drawing energy price levels. This study also considers road characteristics variables to examine their impact on the highway truck traffic patterns. Three variables were recognized as road characteristics variables, including ''max speed,'' ''number of lanes,'' and ''toll roads.''

C. MODEL DEVELOPMENT

The pipeline for this work consists of data preprocessing, feature selection, model creation, parameter optimization, and evaluation using the Scikit-learn [50] library for machine learning in Python [51] programming language. Throughout the preprocessing phase, the data is standardized and divided into training, test, and validation datasets. After, the training and validation sets are fed to a feature selection module that identified the necessary features within the data and eliminated other independent variables (predictors). FIGURE 2 presents the pipeline of this project. The central feature of the workflow is the loop between feature selection, modeling, and hyperparameter optimization modules that would automatically canvas the variations of features and modeling methods. It delivers the best-performing model with the best subset of features based on the input dataset.

As shown in the flowchart in Figure 3, feature selection is applied to the normalized and partitioned data:

$$X_{selected} = \bigcup_{j=0}^{n} X_j S(X_j)$$

where S represents the function that would decide if a feature column is selected or not in a binary fashion. The selected data is the model using linear and nonlinear modeling. At inference time, the outcome for a given datapoint is calculated as:

$$Y_i = F(X_i S(X_i))$$

where F represents the trained model.

D. DATA PREPROCESSING AND PARTITIONING

At the data preprocessing stage, all independent variables (predictors) were transformed into a number. Then, the numeric data was standardized to normal distributions with an average of 0 and a standard deviation of 1 to support the regularizations of the models. Following standardization, the prepared data were divided into training, test, and validation datasets. As the data under this research is time series, exploring the integrity and temporal continuity of the data was essential. As a result, randomly splitting the dataset into different parts for validation would not be appropriate. As shown in FIGURE 3, the evaluation method employed in this study relied on the nested cross-validation expanding window method. In this method, the training dataset has a training subset, and a validation set in the inner loop (yellow dashed box) starting with three years of serial data for each dataset. The training set was increased by three years in each split. The testing dataset consisted of the next three successive years of the dataset after the validation dataset. For the inner loop, each split went through a research pipeline presented in FIGURE 2. Then, concerning the outer loop, after employing the outcomes of each split, the error was averaged. This method ensures that final model is robust and is not an overfit model or a randomly accurate one.

These models were trained on the data and evaluated on the test set. In order to find the optimum feature selection

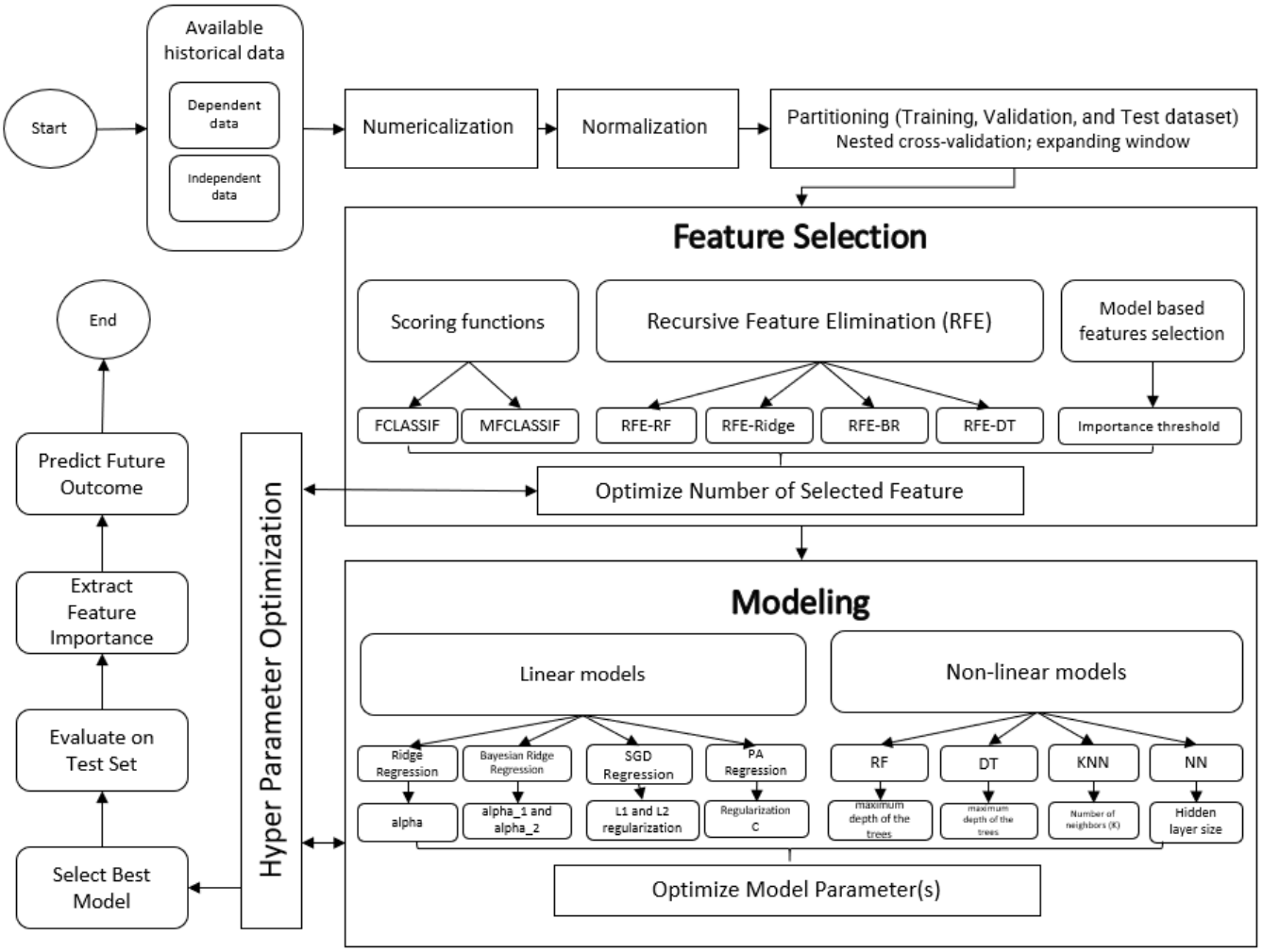

FIGURE 2. The pipeline of the study.

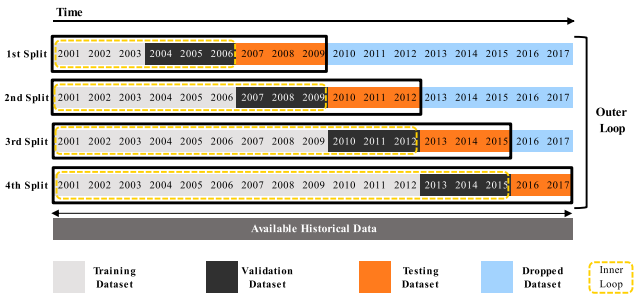

FIGURE 3. Nested cross-validation; expanding window.

tool, model, and parameters, a grid search was performed in the parameter optimization stage based on the validation set.

E. PERFORMANCE MEASUREMENT SCALES

To examine the performance of the feature selections and modeling approaches, three measures of error, including the Mean Absolute Percentage Error (MAPE), R-Squared, and Mean Absolute Error (MAE) were considered. In cases such as this study investigating the truck volume dataset, the aim was to produce the best possible forecast while understanding the possible error in those estimates. The MAPE provides the most suitable mean for evaluating the error in this research and the models were evaluated using this metric on the test set.

F. FEATURE SELECTION

Feature selection is the method of selecting the most suitable predictors and dropping redundant variables from the pool of possible predictors. Depending on the model's structure, feature selection can enhance a model's precision. This technique can be carried out by observing the participation of each candidate variable to the models' accuracy, and then reducing useless and repetitive variables while keeping the most useful ones. In some cases, unnecessary features can lower a model's accuracy. For each parameter set, the cross-validation method presented earlier served to train, validate, and test the model. In this study, three approaches were applied to determine the leading predictors affecting the truck volume prediction models. First, valuable features were determined via a model utilizing SelectFromModel function from Scikit-learn [50]. Several modeling techniques capable

of implicit feature selection, including Ridge Regression (Ridge), Bayesian Ridge (BR) Regression, Random Forest (RF), and Decision Tree (DT), were employed in this section. The importance threshold considered for the selection parameter of this step changes between 0.25, 0.5, 0.75, 1, 1.25, 1.5, and 1.75. Secondly, the Recursive Feature Elimination (RFE in Scikit Learn [50]) was carried. In this manner, the least essential features were dropped gradually until the most suitable features were discovered. The models which were utilized to determine the importance of features are the same as the former step (RFE-RF, RFE-Ridge, RFE-BR, and RFE-DT). In the RFE step, the number of ultimately selected features varies between 1, 3, 5, 10, 20, 30, 40, 50, and 60. Thirdly, a scoring function was employed to find the ''K'' best features in the dataset (SelectKBest in Scikit Learn [50]). The scoring functions employed in this work were ANOVA F-value (FCLASSIF) and Mutual Information (MFCLASSIF). The number of ultimately selected features of this step also fluctuates between 1, 3, 5, 10, 20, 30, 40, 50, and 60. These feature selection approaches were implemented inside a grid search and eventually compared to find the best set of parameters.

G. MODELING APPROACHES

Multiple machine learning algorithms were used in this research, particularly those based on the non-linear relationships among variables to predict the MADTT. The models (SelectFromModel function from Scikit-learn [50]) employed in this research were Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Neural Network (NN). Furthermore, linear regression models, including Linear Regression (Linear), Stochastic Gradient Descent (SGD) Regression, and Passive-Aggressive (PA) Regression, Ridge Regression (Ridge), and Bayesian Ridge (BR), were used as a benchmark to determine the level of improvement of employing non-linear models. This selection of models allowed us to compare models with various levels of linearity or non-linearity, while having control over parametric models. These machine learning methods were applied to the data using the earlier discussed nested cross-validation approach to split the data to train, validate, and test the model. An expanding data window was used for training the model, validating on the next three consecutive years after training the dataset, and then testing it on three consecutive years of data. For the RF and DT algorithms in this research, the model parameter (MP), which is the maximum depth of the trees, varies between 5, 20, 50, 75, 100, and 200. Regarding the K-Nearest Neighbors algorithm employed, the model parameter, Number of neighbors (K), changes between 1, 3, 5, 7, 10, and 16. Concerning the Neural Network models, the MP, which represents the number of nodes employed in this study, varies between 16, 64, and 256 In the linear algorithms, for the Ridge Regression, the MP represents the regularization strength (alpha) and varies between 0.1, 1, 10, 100, 10000 and 1e6. For Bayesian Ridge Regression, the model parameter shows the shape and inverse scale parameters of the prior gamma distribution (alpha_1 and alpha_2) and varies between 0.1, 1, 10, 100, 10000 and1e6. Regarding the Stochastic Gradient Descent Regression, the MP represents the elastic net mixing parameter of L1 and L2 regularization (L1 ratio), and fluctuates between 0, 0.15, 0.3, 0.5, 0.75 and 1. Ultimately, for Passive Aggressive Regression, MP shows the maximum step size (regularization C), and changes between 0.1, 1, 10, 100, 10000, and 1e6. Table 3 presents the various models and the associated modeling parameters employed in this study.

TABLE 3. Modeling parameters of the study.

Non-linear models	RF	maximum depth of the trees	5	20	50	75	100	200
	DT	maximum depth of the trees	5	20	50	75	100	200
	K-NN	Number of neighbors (K)	1	3	5	7	10	16
	NN	number of neurons	16	64	256			
linear models	Ridge Regression	alpha	0	1	10	100	10000	1000000
	Bayesian Ridge Regression	alpha_1 and alpha_2	0	1	10	100	10000	1000000
	SGD Regression	L1 and L2 regularization	0	0.2	0.3	0.5	0.75	1
	PA Regression	(regularization C)	0	1	10	100	10000	1000000

As demonstrated in the feature selection and the modeling approach sections, the developed hyper-parameter optimization grid includes a wide range of values for the parameters, from reasonably low values to reasonably high values, so that it could be applied to various datasets with differing characteristics.

IV. RESULTS AND DISCUSSION

A. MODEL WITHOUT SPATIAL VARIABLES

A comparison of the accuracy of various models on the test dataset using the grid search is presented in FIGURE 4. It is evident that non-linear models outperform the linear models, including Linear Regression, Ridge Regression, BR, SGD, and PA. Among non-linear models, RF, KNN, and DT model perform better than NN model. The error presented in FIGURE 5 is the average of the error of the four splits for the mixed trucks (summation of both directions of the truck traffic) of the dataset described in FIGURE 3. The MAPE error (the performance measure used in this study) on test dataset presents a reliable value of about 22.27%.

1) THE SELECTED MODEL OF THIS STUDY FOR THE CURRENT TERM

As shown in FIGURE 4, empirically, RF, KNN, and DT show the best results among the non-linear models. However, theoretically, the KNN model is only capable of predicting the

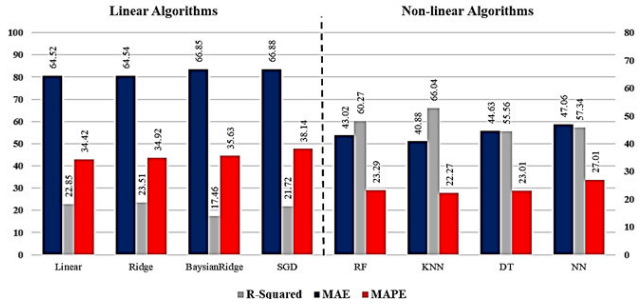

FIGURE 4. **Comparison of different models' best performance on the test dataset for mixed trucks.**

data from the training dataset, which results in biased results. The KNN model finds the K nearest instances to the instance in question and predicts the output by averaging the output of those instances. Through these instances, the model can be interpreted. However, the essential features are not highlighted. Moreover, the model does not learn from data and has to search the data for each prediction. This disadvantage has a silver lining as it makes updating the data and model easier. Concerning DT, this model creates a decision tree based on splitting features. At its leaves is the regression output. The decision-making process and the results are interpretable. However, it can overfit if many features are present since the decision-making handles sparse data at the leaves. However, RF implements many decision trees (500 trees) on the data. It does so by randomly choosing groups of data to train on. Since RF implements many decision trees, it becomes less prone to overfitting while keeping the advantages of decision trees. Thus, the RF model presents an appropriate model, both empirically and theoretically, and was selected for current term prediction in this study.

FIGURE 5 illustrates the result of the four best feature selection approaches utilized in this study on the validation data set. It was found that all four feature selection approaches can provide appropriate modeling of the data, demonstrating the success of the grid search process in finding suitable training parameters for each feature selection method. However, *RFE Ridge* has the lowest MAPE on the validation dataset among various feature selection approaches.

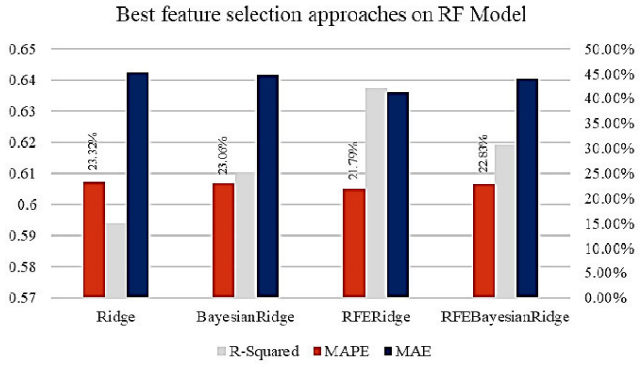

FIGURE 5. **Comparison of the feature selection of RF models on validation dataset for mixed trucks.**

A comparison of the accuracy of the RF algorithm on the four splits of the data is presented in FIGURE 6. It illustrates that split 4 (mentioned in FIGURE 3), the split that covers all the dataset has a lower MAPE error (18.24%) on the test dataset, compared to other splits. The MAPE error of split 4 of the RF models on the validation dataset and the test dataset does not differ considerably, which shows that the developed model is robust.

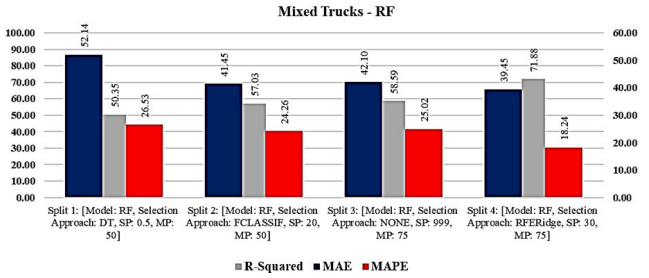

FIGURE 6. **RF model's performance on test dataset for mixed trucks.**

FIGURE 7 presents the comparison of ground truth and prediction via plotting them against each other. The prediction closely follows the ground truth, and the points are located around the 45-degree line.

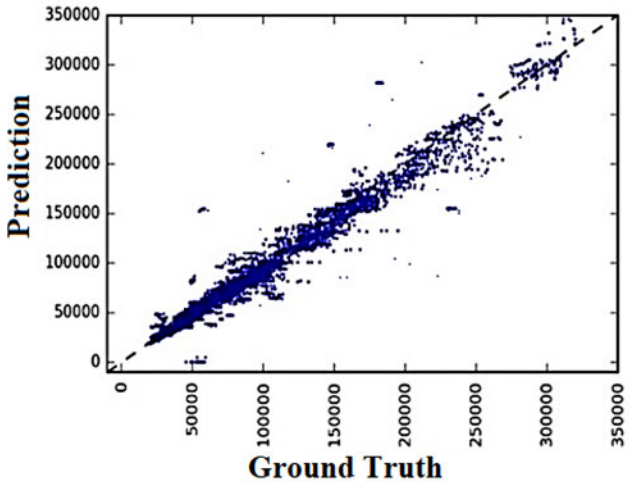

FIGURE 7. **Comparison of ground truth and prediction via plotting them against each other within the validation dataset using RF algorithm.**

FIGURE 8 depicts the model optimization of the mixed trucks on the 4th split on the validation dataset. The optimum feature selection and modeling approach for this case was found to be *RFE Ridge* and *RF*, respectively. For finding the best selection parameter, the number of features that are ultimately selected is changed between 10 to 40. The same approach is taken for optimizing the *RF* model by alternating the maximum depth of the trees from 5 to 200. The *RF* model, with the depth of 75 trained on 30 selected features, has the lowest MAPE of 18.44% on the validation dataset.

FIGURE 9 illustrates the feature categories importance for mixed trucks. Socioeconomic variables, with 49% feature importance, ranked first among the seven categories of

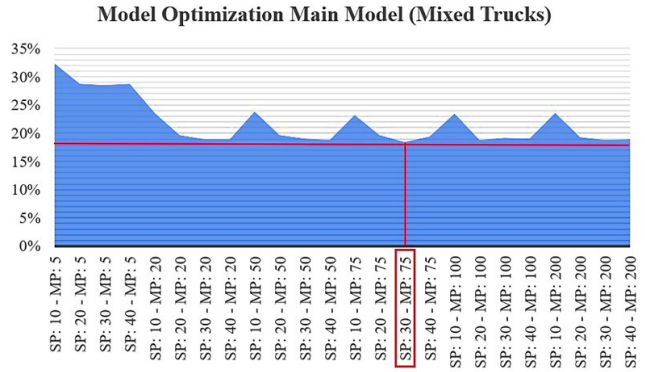

FIGURE 8. Model optimization for mixed trucks.

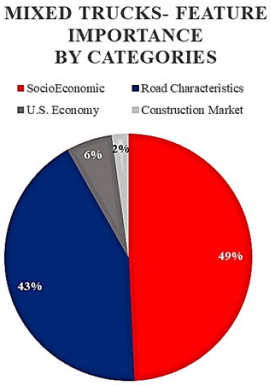

FIGURE 9. Feature importance categorical from best performing models for trucks' RF model.

this study. Road characteristics (43% feature importance) and U.S. economy related variables (6% feature importance) were ranked second and third in this study.

FIGURE 10 illustrates top six important features that were performing better than other parameters for mixed trucks. 'Number of Lanes', which depicts the capacity of roadway, has the most important influence on the truck prediction model with a 31% importance. Moreover, 'length of paved roads centerline miles' ranked second with a 27% importance. Concerning socioeconomic variables, features such as number of 'licensed drivers' and 'population' are essential variables for truck prediction model.

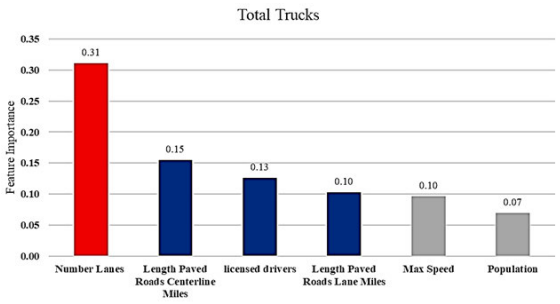

FIGURE 10. Top six important features of the best performing models for mixed trucks.

2) SELECTED MODEL FOR LONG-TERM PREDICTIONS

It is important to note that, the RF model is only capable of interpolating and using the dataset values, which makes the algorithm a suitable option for current and current- term, short-term and mid-term modeling. Concerning the better generalization capabilities of NN, they give an edge to NN models to be used for future projections (long-term studies). The NN algorithm is capable of extrapolating and generating prediction values by changing the hidden layer size to predict the mid- and long-term MADTT. This model is trained by finding the bias and weights of artificial neural network through stochastic gradient descent. It possesses a layer of nodes, each of which has a non-linear activation function. FIGURE 11 shows the four splits results of cross-validation utilized in this study for the NN algorithm; demonstrates that split 3 has a lower MAPE error (with 80% prediction accuracy) and performs better compared with other splits.

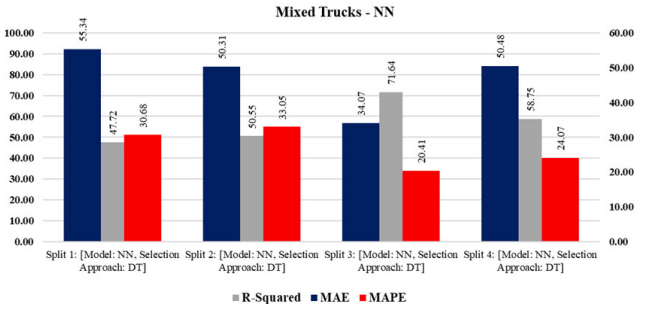

FIGURE 11. Best model for mid- and long-term planning: NN.

The model optimization of mixed trucks on the 3rd split on the validation dataset for the NN models show that the DT feature selection approach with importance threshold of 0.75. The NN model algorithm with 256 nodes with a hidden layer has the lowest MAPE of 20.41% on the validation dataset. Moreover, the MAPE error on the test dataset is 24.06% which is reasonable.

B. FORECAST OF DIRECTIONAL TRUCK TRAFFIC VOLUME – CASE STUDIES

In this section, the developed directional NN models (with spatial variables) were deployed to forecast the directional truck volumes for 2018 to 2050. The pool of 59 independent variables in this study contained seven categories including the energy market variables, construction market variables, U.S. economy variables, and socioeconomics variables (excluding population, licensed drivers, length paved road line miles, and centerline miles), where NN was used to predict the future values.

A variety of univariate modeling techniques were used to predict the future values of the independent variables to feed the model as an input. Two general types of univariate modeling were used to predict the time-series predictors of this study, namely Auto Regressive Moving Average (ARMA) and Smoothing. The ARMA is the most common classification of models used in forecasting univariate time series.

This type of model is represented as an ARMA (p,q), where p is the AR order, and q is the MA order. The order of the AR and MA was chosen via an autocorrelation correlogram function (ACF) and a partial autocorrelation correlogram function (PACF). On the other hand, the Smoothing method includes simple, exponential, double exponential smoothing methods, and Holt-Winters (Linear, Seasonal additive, multiplicative additive). A pool of x independent variables of this study contained seven categories. Including the energy market variables, construction market variables, U.S. economy variables, and seven variables of socioeconomics variables (excluding population, licensed drivers, length paved road line miles and centerline miles) which ARMA and Smoothing methods were employed to predict the future values. Regarding two other socioeconomic variables, (population and licensed drivers) the results of the Rayer *et al.* [52] were utilized. Furthermore, about the last two socioeconomic variables, the length of paved roads of Florida Highways (length of the paved roads lane miles and length of the paved roads centerline miles) were considered to be fixed during the future years. Ultimately, the road characteristics variables and spatial variables were considered fixed throughout the projection period. Finally, one site from interstate highways *I75* and *I4*, was selected to show the results for the projection period 2018 to 2050 using the truck directional NN model (with spatial variables) and projected predictors.

1) CASE STUDY #1: I4, ORANGE COUNTY, SITE ID: 753051

FIGURE 12 depicts the total, North/Eastbound and South/Westbound historical, and projected truck traffic employing the directional NN model (with spatial variables) developed by this research. The historical traffic data covers 2001 (beginning month 1) to 2017 (ending month 204) monthly average daily truck traffic (MADTT) of passenger vehicles. The projected values are MADTT between 2018 (beginning month 205) to 2050 (ending month 600).

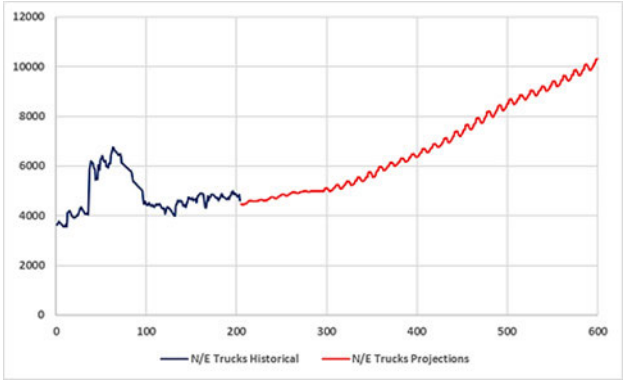

FIGURE 12. The truck traffic projections of case study #1.

C. DISCUSSION

Precise prediction of traffic flow is a crucial component of the Intelligent Transportation System (ITS) [53]. This study generated and optimized a framework containing feature selection (three-step approach) to assist the training of the traffic flow prediction models with high accuracy. With its' high prediction accuracy, the proposed methodology presents a promising potential complementary tool to be utilized in the calibration and validation of the existing truck volume prediction models. In contrast to the study by Lu *et al.* [48] that have shown that both linear and compound growth models were fit the truck traffic growth trends well, this study has shown that linear models are not able to predict the MADTT accurately. Additionally, this study confirms the results of studies by Polson *et al.* [31], Oswald *et al.* [32], Rilett and Park [44], and Liu and Wu [41] that claimed that the non-parametric models' superior capability to capture temporal-spatial relationships and non-linear patterns make them more accurate for truck traffic forecasting compared to the parametric models.

By analyzing the results on the test and validation dataset, it can be concluded that non-linear models outperform linear models. This can be viewed in the notable gap within the performances of linear models on the truck dataset versus that of the non-linear models. Overall, four models, including DT, RF, NN, and KNN, were evaluated. The generalization capabilities of RF give it an edge for current and near-term MADTT projections. The RF algorithm results on the test dataset of the study demonstrate the ability of the model to predict the MADTT with 82% accuracy. However, by adding the spatial related variables (county, interstate, site id, and Euclidean geometry of each site), the accuracy of the model improved to 86%, illustrating the importance of considering the location-related features for the truck traffic prediction models. Regarding the important features of the RF model (with spatial variables), the "spatial variables" category ranked first with a 48% importance, and second, "road characteristics," with a 26% importance. Both have a significant role in the truck counts prediction model. Furthermore, the NN developed model (with spatial variables) for the long-term predictions shows the capability of the model to predict the MADTT with an 80% accuracy.

V. CONCLUSION

A literature review indicated that the majority of truck-traffic modeling studies encompass one or two linear or nonlinear algorithms. In these studies, one model's success over another was inconsistent and varied depending on the specific case study being discussed, and the results could not be directly utilized outside of the prospective case study being reviewed. These findings suggest that truck traffic forecasting is dependent on the interplay between local and global variables that may be either linear or non-linear based on multiple factors such as location, project type, and the level of analysis. This complication can be overcome by using a universal framework for truck volume forecasting that is more generalized to optimize the process and the final outcome based on specific input data characteristics. To that end, the analysis performed in this study was all-inclusive of the reviewed methods for feature selection and modeling approach. It utilized a broad dataset of the variables discussed to confirm that new users

could efficiently utilize the framework generated. By following the proposed method, regardless of the location, type, or scope of the project involved, users can input their data to identify the critical truck volume forecasting factors related to their project in a much more automated way than previously explored. This provides advantages over existing models that utilize assumptions and methodologies specific to a certain case study. Also, the proposed framework increases the number of predictors involved to allow for more accurate forecasting; and automating the methodology reduces the time and expertise required to forecast the complexities of the truck traffic network.

In this study, a data-driven methodology was employed to identify the top features and modeling approach. This allowed for the inclusion of all available linear and non-linear models, the independent variables and parameters involved in feature selection and modeling approach selection. The resulting framework generated is more comprehensive and can be appropriately utilized by new users. By following this framework, a user may identify the feature selection methods, algorithms, and set of features most suitable to their unique project and dataset automatically. This assertion is possible because the framework developed in this model incorporates not only the approaches previously highlighted in the literature but also contains improvements and enhancements to create a more complex model based on the number of the employed features and the feature selection methods employed in an automated fashion. The framework of this study was then validated using the historic data gathered from 259 traffic sites, spanning the course of 17 years. The results of this Florida dataset analysis demonstrate how this model can be successfully applied. The selection features and models used were chosen through a data-driven method in order to prevent bias, and the results indicate which features may be classified as high importance in the process of truck volume prediction based on this dataset. The framework in this study then is not only more comprehensive than a stand-alone case study-based approach, but it can be used for more accurate generalization. The generated framework not only incorporates all the approaches executed in the reviewed literature but also goes beyond them in terms of the variations of features and modeling methods.

This framework can help planners in obtaining truck volume on state roadways to quantify truck traffic in order to assist with long-term planning solutions, such as roadway expansions (by calculating the level of service to find the critical links which need investments for expansion of the road (adding lanes or constructing new roads [53]) or an additional bridge, plan development for pavement designs, prediction and planning for future truck trips, environmental impact analysis and the examination of highway investment policies. Transportation planners would be able to plan for the critical links on the U.S. highways currently facing overcapacity issues and investigate the optimized solutions for enhancing the traffic network considering the existing investment gap. The results could also be used to attract private sector partnerships to foster economic development and improve safety and mobility by developing a suitable request for proposals and decent incentives accurately and on time. As a result, the quality of life of citizens could be increased by avoiding traffic congestion, enhancing air quality, and decreasing the number of crashes.

The limitations of this study include the sample size (this study utilized 259 sites and 17 years of historical truck traffic data), data type (this study employed monthly level historical truck traffic data – it would be better to use weekly, daily, or even hourly data), examining several other essential variables in developing the automated, connected, trucks' scenarios, and finally including the environmental and energy trends related variables as predictors to study the traffic with even better accuracy. For future work, aside from the limitations mentioned above, scholars should investigate automated and connected trucks, platooning, and, more importantly, the importance of considering the managed lanes for trucks on the truck traffic counts on the highways. Finally, the classification on trucks (medium-sized vs. heavy-duty vehicles) and loaded and unloaded vehicle information from weight in motion (WIM) data could also be added to the model.

APPENDIX A
INDEPENDENT VARIABLES (PREDICTORS)

See FIGURE 13.

APPENDIX B
MODEL WITH SPATIAL VARIABLES

It is essential to consider the impact of spatial variables related to the location of the input data of each site. To test the importance of the spatial variables on the developed truck traffic prediction model, this study added four spatial variables into the prediction model's predictors pool. Table 4 depicts the spatial variables considered in this paper among the previous candidate variables.

By comparing the different models' best performance on the test set and the average error of the four splits, non-linear models outperform linear models. However, the MAPE error of the model with spatial candidate variables (added to the previous dataset: all 59 predictors) shows a better performance compared to the model of section A in the manuscript (without spatial variables; 55 predictors). The comparison of these models confirms a 4% improvement in the accuracy of the MADTT by adding the spatial variables shown in Table 5.

A. THE SELECTED MODEL (WITH SPATIAL VARIABLES) OF THIS STUDY FOR THE CURRENT TERM

FIGURE 14 illustrates the optimum feature selection and modeling approach for this case were found to be Bayesian Ridge and RF, respectively. For finding the best selection parameter, the number of features that are ultimately selected was changed between the importance threshold of 0.25 and 1.75. The same approach was taken for optimizing the RF model by alternating the maximum depth of the trees from 5 to 200. The RF model, with a depth of 100 trained on

A1. Socioeconomic candidate variables				
#	Candidate variable	Acronyms	Scope	Source
1	Population		FL Counties	U.S. Bureau of Census
2	Number of Licensed Drivers		FL Counties	FL Department of Highway Safety and Motor Vehicles
3	Length Paved Roads (Centerline Miles)		FL Counties	Florida Department of Transportation
4	Length Paved Roads (Line Miles)		FL Counties	Florida Department of Transportation
5	Number of Household Estimates	HHEUS	FL Counties	U.S. Bureau of Census
6	Civilian Labor Force	CLFFL	FL State	U.S. Bureau of Labor Statistics
7	All Employees	AEFL	FL State	U.S. Bureau of Labor Statistics
8	Unemployment Rate	URUS	U.S.	U.S. Bureau of Labor Statistics
9	Change in Labor Market Conditions Index	CLMCIUS	U.S.	U.S. Bureau of Labor Statistics
10	Average Hourly Earnings Labor Employees: Construction	AHEPNECUS	U.S.	U.S. Bureau of Labor Statistics
11	Average Weekly Hours of All Employees Construction	AWHAECFL	FL State	U.S. Bureau of Labor Statistics
A2. Economy candidate variables				
#	Candidate variable	Acronyms	Scope	Source
1	Gross Domestic Products	GDP	U.S.	U.S. Bureau of Economic Analysis
2	Industrial Production	IP	U.S.	Federal Reserve System
3	Inflation Rate	IRUS	U.S.	World Bank
4	Consumer Price Index FL	CPIFL	FL State	U.S. Bureau of Labor Statistics
5	CPI for Urban Consumers: New vehicles	CPIAUCNV	U.S.	U.S. Bureau of Labor Statistics
6	CPI for All Urban Consumers: Used cars and trucks	CPIAUCUCT	U.S.	U.S. Bureau of Labor Statistics
7	Price Pressures Measure	PPMUS	U.S.	U.S. Bureau of Labor Statistics
8	Bank Prime Loan Rate	BPLRUS	U.S.	Federal Reserve System
9	30-Year Conventional Mortgage Rate	30YCMR	U.S.	Federal Reserve System
10	Leading Index for U.S.	LIUS	U.S.	Federal Reserve Bank
11	Leading Index for Florida	LIFL	FL State	Federal Reserve Bank
12	Producer Price Index for Commodities	PPIACO	U.S.	Federal Reserve Bank
13	Effective Federal Funds Rate	EDDRUS	U.S.	Federal Reserve Bank
14	M1	M1	U.S.	Federal Reserve Bank
A2. Economy candidate variables				
#	Candidate variable	Acronyms	Scope	Source
15	M2	M2	U.S.	Federal Reserve Bank
16	Gold Prices	GP	U.S.	Yahoo Finance
17	Silver Prices	SP	U.S.	Yahoo Finance
18	Durable Goods Orders	DGOUS	U.S.	Yahoo Finance
19	Dow Jones Index Adj Close	DJI	U.S.	Yahoo Finance
20	S&P 500 Index	S&P500	U.S.	Yahoo Finance
21	St. Louis Fed Financial Stress Index	StLFFSI	U.S.	Yahoo Finance
22	Wilshire 5000 Total Market Full Cap	W5000TMFCI	U.S.	Yahoo Finance
23	NASDAQ Composite Index, Index	NASDAQ	U.S.	Yahoo Finance
24	Canada / U.S. Foreign Exchange Rate	CANUSER	U.S.	Yahoo Finance
25	China / U.S. Foreign Exchange Rate	CHUSER	U.S.	Yahoo Finance
26	Mexico / U.S. Foreign Exchange Rate	MEXUSER	U.S.	Yahoo Finance
27	U.S. / Euro Foreign Exchange Rate	USEUER	U.S.	Yahoo Finance
A3. Construction market candidate variables				
#	Candidate variable	Acronyms	Scope	Source
1	New Private Housing Units (Building Permits)	NPHUABPFL	FL State	U.S. Bureau of Census
2	Construction Spending Nonresidential	CSNUS	U.S.	U.S. Census Burea
3	Construction Spending Highway	TCSHSUS	U.S.	U.S. Census Burea
4	Construction Employees FL	CEFL	FL State	U.S. Bureau of Labor Statistics
5	Construction Employees U.S.	AECHCEUS	U.S.	U.S. Bureau of Labor Statistics
A4. Energy market candidate variables				
1	Electricity Price	ELECFL	FL State	U.S. Energy Information Administratio
2	Crude Oil Price	COP	U.S.	
3	Natural Gas Prices	NGP	U.S.	
4	Gas Price FL	GASPFL	FL State	
Road characteristics candidate variables				
1	Max Speed	Cosite		Florida Department of Transportation
2	Number Lanes			
3	Toll Road			

FIGURE 13. The predictors employed in the primary model of this study.

TABLE 4. Spatial candidate independent variables.

	Spatial variables	
1	County Name	Florida Department of Transportation (FDOT)
2	Interstate ID	
3	Cosite ID	
4	Euclidean Geometry	

selected features with an importance score higher than 1.5, had the lowest MAPE of 12.06% on the validation dataset.

FIGURE 15 depicts the categorical feature importance derived from the best performing models for mixed trucks. Spatial variables' category had the most significant impact on the truck traffic model with a value of 48%. Road characteristics, with the value of 26%, had the second rank.

TABLE 5. Comparison of the developed RF models on test dataset.

Models	Label Name	Fold	Selection Approach	Model	R-squared	MAPE Test
Primary model	Total Trucks	4	RFE Ridge	RF	0.72	18.23%
Model with spatial variables included	Total Trucks	4	Bayesian Ridge	RF	0.80	14.53%

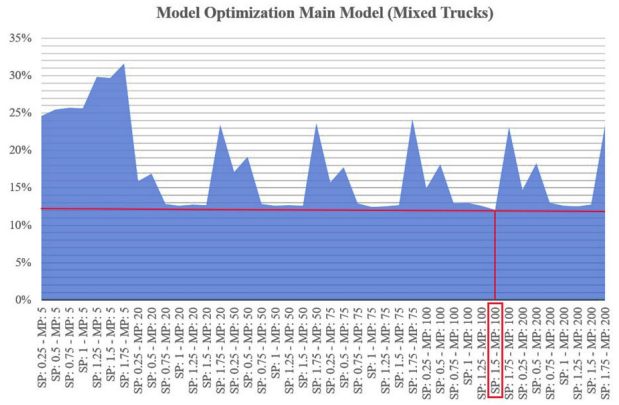

FIGURE 14. Model optimization for mixed trucks (with spatial variables).

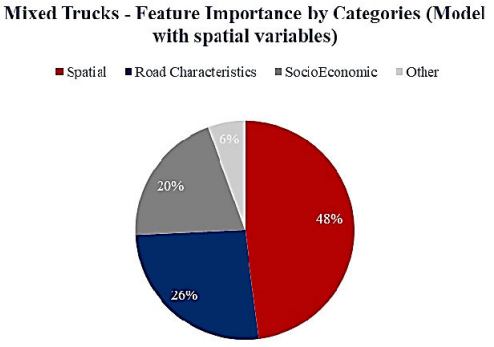

FIGURE 15. Categorical feature importance derived from the best performing models for trucks (for the model with spatial variables).

FIGURE 16 depicts the top six important features that are performing better than other parameters for the mixed trucks of the model with spatial variables. The "site" or "co-site", which depicts the ID of the studied location, had the most

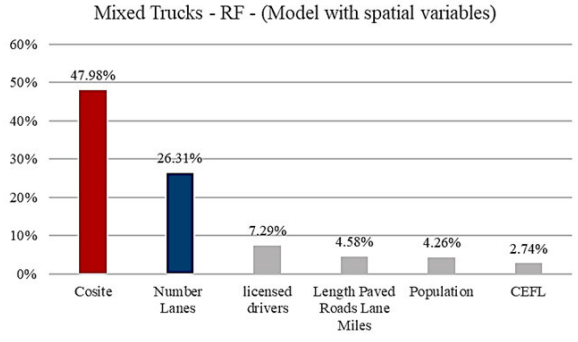

FIGURE 16. Feature importance derived from the best performing models for mixed trucks (for the model with spatial variables).

important influence on the truck prediction model. Moreover, the ''number of lanes'' shows the second important feature with a 26% importance.

Additionally, this study developed separate models for each direction (North/Eastbound and South/Westbound) of the truck traffic flow. The model optimization of the North/Eastbound of truck traffic on the 4th split on the validation dataset for the RF models depicted that the 'RFERF' feature selection approach, with 30 chosen features and the RF model algorithm with 75 trees, has the lowest MAPE of 12.50% on the validation dataset. On the other hand, the model optimization of the South/Westbound of the truck traffic on the 4th split on the validation dataset for the RF models showed that the 'RFE Bayesian Ridge' feature selection approach, with 20 selected features and the RF model algorithm with 50 trees, had the lowest MAPE of 11.96% on the validation dataset.

B. SELECTED MODEL (WITH SPATIAL VARIABLES) FOR LONG-TERM TRUCK TRAFFIC PROJECTIONS

A comparison of the generated NN models for the framework of this study, shown in Table 6, confirms a 4% improvement in the accuracy of the MADTT by adding the spatial variables.

TABLE 6. Comparison of the NN developed models on test dataset.

Models	Label Name	Fold	Selection Approach	Model	R-squared	MAPE Test
Model without spatial variables	Total Trucks	4	DT	NN	0.60	24.06%
Model with spatial variables included	Total Trucks	4	RF	NN	0.72	20.03%

A comparison of the accuracy of the NN model with spatial variables on the four splits of the cross-validation is shown in FIGURE 17. It is apparent that split 4 outperforms the other splits of the data. Split 4 has a MAPE error of 20.03% on the test dataset.

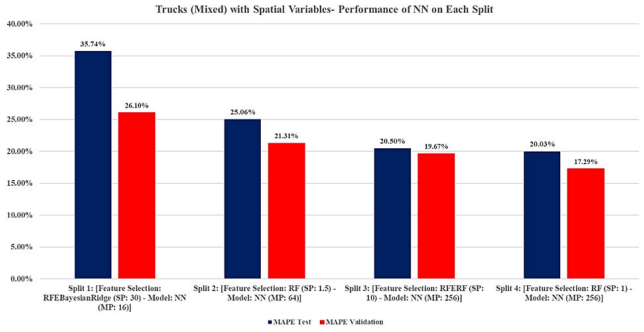

FIGURE 17. NN model's (with spatial variables) performance on test dataset for total trucks.

The model optimization of the North/Eastbound of the truck traffic of (with spatial variables) on the 4th split on the validation dataset for the NN models showed that the RFERF feature selection approach, with 10 selected features and the NN model algorithm with 64 nodes in the hidden layer, has the lowest MAPE of 17.77% on the validation dataset. On the other hand, the model optimization of the South/Westbound of the truck traffic (with spatial variables) on the 4th split on the validation dataset for the NN models showed that the RF feature selection approach with importance threshold of 0. 5. The NN model algorithm with 256 nodes in the hidden layer had the lowest MAPE of 17.46% on the validation dataset.

APPENDIX C
CASE STUDY #2: I75, MARION COUNTY, SITE ID: 360437

See FIGURES 18–20.

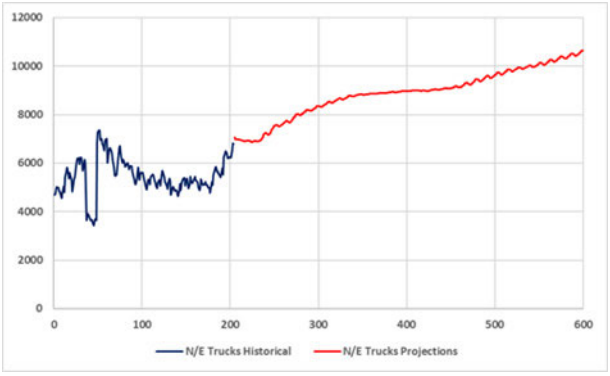

FIGURE 18. The N/E truck traffic projections of case study #2.

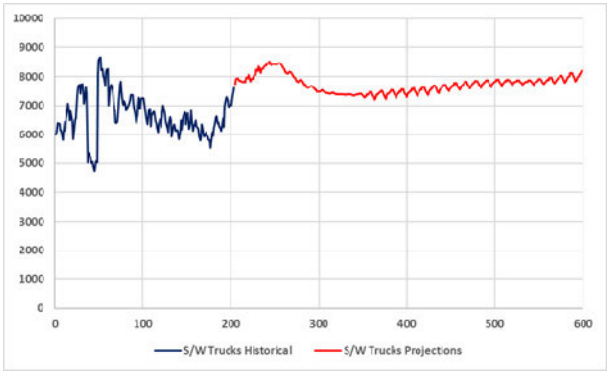

FIGURE 19. The S/W truck traffic projections of case study #2.

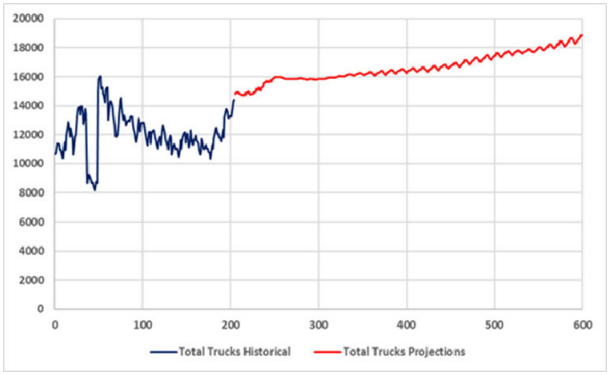

FIGURE 20. The mixed truck traffic projections of case study #2.

ACKNOWLEDGMENT

The Florida Department of Transportation, provided all the data required for this study, including the historical Truck traffic data, from 2001 to 2017. The authors want to appreciate the following FDOT individuals: Jeremy Dilmore, Dr. Martin Markovich, and Joey Gordon.

REFERENCES

[1] *Transportation Planning Implications of Automated/Connected Vehicles on Texas Highways*, Texas A&M Transp. Inst., Dallas, TX, USA, 2017.
[2] M. Fu, J. A. Kelly, and J. P. Clinch, "Estimating annual average daily traffic and transport emissions for a national road network: A bottom-up methodology for both nationally-aggregated and spatially-disaggregated results," *J. Transp. Geography*, vol. 58, pp. 186–195, Jan. 2017.
[3] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. B, Methodol.*, vol. 18, no. 1, pp. 1–11, 1984.
[4] W. Zheng, D.-H. Lee, and Q. Shi, "Short-term freeway traffic flow prediction: Bayesian combined neural network approach," *J. Transp. Eng.*, vol. 132, no. 2, pp. 114–121, Feb. 2006.
[5] J. Guo, W. Huang, and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 50–64, Jun. 2014.
[6] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *Eur. Transp. Res. Rev.*, vol. 7, no. 3, pp. 1–9. 2015.
[7] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1836, no. 1, pp. 143–150, Jan. 2003.
[8] J. Wang and Q. Shi, "Short-term traffic speed forecasting hybrid model based on chaos-wavelet analysis-support vector machine theory," *Transp. Res. C, Emerg. Technol.*, vol. 27, pp. 219–232, Feb. 2013.
[9] Y. Wu, H. Tan, J. Peter, B. Shen, and B. Ran, "Short-term traffic flow prediction based on multilinear analysis and K-nearest neighbor regression," in *Proc. 15th COTA Int. Conf. Transp. Prof.*, 2015, pp. 556–569.
[10] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 557–569, Feb. 2016.
[11] A. Allström, J. Ekström, D. Gundlegård, R. Ringdahl, C. Rydergren, A. M. Bayen, and A. D. Patire, "Hybrid approach for short-term traffic state and travel time prediction on highways," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2554, no. 1, pp. 60–68, Jan. 2016.
[12] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015.
[13] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 3–19, Jun. 2014.
[14] M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 871–882, Jun. 2013.
[15] M. P. Boile, S. Benson, and J. Rowinski, "Freight flow forecasting–An application to New Jersey highways," *J. Transp. Res. Forum*, vol. 39, no. 2, pp. 159–170, 2000.
[16] R. Mustafa and M. Zhong, "Applying GIS-based traditional travel demand model for improved network-wide traffic estimation: New brunswick case study," in *Proc. ICTIS*, Jun. 2011, pp. 816–822.
[17] D. Janssens, G. Wets, T. Brijs, and K. Vanhoof, "The development of an adapted Markov chain modelling heuristic and simulation framework in the context of transportation research," *Expert Syst. Appl.*, vol. 28, no. 1, pp. 105–117, Jan. 2005.
[18] R. Cervero, "Induced travel demand: Research design, empirical evidence, and normative policies," *J. Planning Literature*, vol. 17, no. 1, pp. 3–20, 2002, doi: 10.1177/088122017001001.
[19] M. Meyer and E. Miller, *Urban Transportation Planning*, 2nd ed. New York, NY, USA: McGraw-Hill, 2001.
[20] T. L. Friesz, "Strategic freight network planning models," in *Handbook of Transportation Modeling*. Pergamon, 2000, pp. 527–537. [Online]. Available: https://tmip.org/content/strategic-freight-network-planning-models
[21] L. N. N. Do, H. L. Vu, B. Q. Vo, Z. Liu, and D. Phung, "An effective spatial-temporal attention based neural network for traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 108, pp. 12–28, Nov. 2019.
[22] P. Duan, G. Mao, W. Yue, and S. Wang, "A unified STARIMA based model for short-term traffic flow prediction," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 1652–1657.
[23] Y. Kim and J. Hong, "Urban traffic flow prediction system using a multifactor pattern recognition model," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2744–2755, Oct. 2015.
[24] E. Bagheri, M. Zhong, and J. Christie, "Improving AADT estimation accuracy of short-term traffic counts using pattern matching and Bayesian statistics," *J. Transp. Eng.*, vol. 141, no. 6, Jun. 2015, Art. no. A4014001.
[25] H. J. Roh, S. Sharma, P. K. Sahu, and S. Datla, "Analysis and modeling of highway truck traffic volume variations during severe winter weather conditions in Canada," *J. Mod. Transp.*, vol. 23, no. 3, pp. 228–239, 2015.
[26] I. Tsapakis, W. H. Schneider, and A. P. Nichols, "A Bayesian analysis of the effect of estimating annual average daily traffic for heavy-duty trucks using training and validation data-sets," *Transp. Planning Technol.*, vol. 36, no. 2, pp. 201–217, Mar. 2013.
[27] M. J. Lighthill and G. B. Whitham, "On kinematic waves II. A theory of traffic flow on long crowded roads," *Proc. Roy. Soc. London. A. Math. Phys. Sci.*, vol. 229, no. 1178, pp. 317–345, 1955.
[28] E. Kometani, "Dynamic behavior of traffic with a nonlinear spacing-speed relationship," in *Proc. Theory Traffic Flow, (Proc. Symp. TTF (GM))*, 1959, pp. 105–119.
[29] W. Burghout, H. N. Koutsopoulos, and I. Andréasson, "Hybrid mesoscopic–microscopic traffic simulation," *Transp. Res. Rec.*, vol. 1934, no. 1, pp. 218–225, 2005.
[30] H.-F. Yang, T. S. Dillon, E. Chang, and Y.-P. P. Chen, "Optimized configuration of exponential smoothing and extreme learning machine for traffic flow forecasting," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 23–34, Jan. 2019.
[31] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, pp. 1–17, Jun. 2017.
[32] R. K. Oswald, W. T. Sherer, and B. L. Smith, "Traffic flow forecasting using approximate nearest neighbor nonparametric regression," Center Transp. Stud., Univ. Virginia, Charlottesville, VA, USA, Tech. Rep. UVA-CE-ITS_01-4, 2001.
[33] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 10, no. 4, pp. 303–321, Aug. 2002.
[34] J. Rice and E. Van Zwet, "A simple and effective method for predicting travel times on freeways," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 3, pp. 200–207, Sep. 2004.
[35] S. Bajwa, "Short-term travel time prediction using traffic detector data," Univ. Tokyo, Tokyo, Japan, Tech. Rep. FHWA/TX-08/0-5141-1, 2003.
[36] J. You and T. J. Kim, "Implementation of a hybrid travel time forecasting model with GIST," in *Proc. 3rd Bi-Annu. Conf. Eastern Asia Soc. Transp. Stud.*, Taipei, Taiwan, Sep. 1999, pp. 231–256.
[37] D. Nikovski, N. Nishiuma, Y. Goto, and H. Kumazawa, "Univariate short-term prediction of road travel times," in *Proc. IEEE Intell. Transp. Syst.*, Sep. 2005, pp. 1074–1079.
[38] M. Zhong, S. Sharma, and P. Lingras, "Refining genetically designed models for improved traffic prediction on rural roads," *Transp. Planning Technol.*, vol. 28, no. 3, pp. 213–236, Jun. 2005.
[39] H. Crosby, P. Davis, and S. A. Jarvis, "Spatially-intensive decision tree prediction of traffic flow across the entire UK road network," in *Proc. IEEE/ACM 20th Int. Symp. Distrib. Simulation Real Time Appl. (DS-RT)*, London, U.K., Sep. 2016, pp. 116–119.
[40] W. Alajali, W. Zhou, S. Wen, and Y. Wang, "Intersection traffic prediction using decision tree models," *Symmetry*, vol. 10, no. 9, p. 386, Sep. 2018.
[41] Z. Liu, Z. Li, K. Wu, and M. Li, "Urban traffic prediction from mobility data using deep learning," *IEEE Netw.*, vol. 32, no. 4, pp. 40–46, Jul. 2018.
[42] J. Ahn, E. Ko, and E. Y. Kim, "Highway traffic flow prediction using support vector regression and Bayesian classifier," in *Proc. Int. Conf. Big Data Smart Comput. (BigComp)*, Hong Kong, Jan. 2016, pp. 239–244.
[43] M. Deshpande and P. R. Bajaj, "Short term traffic flow prediction based on neuro-fuzzy hybrid sytem," in *Proc. Int. Conf. ICT Bus. Ind. Government (ICTBIG)*, Indore, India, Nov. 2016, pp. 1–3.
[44] L. R. Rilett and D. Park, "Direct forecasting of freeway corridor travel times using spectral basis neural networks," *Transp. Res. Rec.*, vol. 1752, no. 1, pp. 140–147, 2001.
[45] Y. Zhang, Y. Zhang, and A. Haghani, "A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 65–78, Jun. 2013.
[46] H. M. Al-Deek, G. Johnson, A. Mohamed, and A. El-Maghraby, "Truck trip generation models for seaports with container and trailer operation," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1719, no. 1, pp. 1–9, Jan. 2000.

[47] M. Golias, M. Boile, and K. Ozbay, "Estimating truck volumes on state highways—A statistical approach," in *Proc. TRB Annu. Meeting*, 2005, pp. 1–21.
[48] Q. Lu, Y. Zhang, and J. T. Harvey, "Growth of truck traffic volume for mechanistic-empirical pavement design," *Int. J. Pavement Eng.*, vol. 10, no. 3, pp. 161–172, 2009.
[49] T. Corkery, "FLSWM version 7.0: New 2045 horizon-year model incorporates extensive structural improvements," Dept. Florida Transp., Forecasting Trends Office, 2020. [Online]. Available: https://www.fsutmsonline.net/index.php?/model_pages/comments/flswm_version_7.0_new_2045_horizon _yearmodelincorporates_extensive_structur
[50] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[51] Python. (2020). *Python 3.8.3*. [Online]. Available: https://www.python.org/downloads/
[52] S. Rayer and Y. Wang, "Projections of Florida population by county, 2020–2045, with estimates for 2019," College Liberal Arts Sci., Bur. Econ. Bus. Res., Univ. Florida, Gainesville, FL, USA, Tech. Rep. 53-186, 2020.
[53] A. Mahdavian, A. Shojaei, M. Salem, J. S. Yuan, and A. A. Oloufa, "Data-driven predictive modeling of highway construction cost items," *J. Construct. Eng. Manage.*, vol. 147, no. 3, 2021, Art. no. 04020180, doi: 10.1061/(ASCE)CO.1943-7862.0001991.
[54] X. Yang, Y. Zou, J. Tang, J. Liang, and M. Ijaz, "Evaluation of short-term freeway speed prediction based on periodic analysis using statistical models and machine learning models," *J. Adv. Transp.*, vol. 2020, Jan. 2020, Art. no. 9628957.
[55] P. Sun, N. AlJeri, and A. Boukerche, "A fast vehicular traffic flow prediction scheme based on Fourier and wavelet analysis," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6, doi: 10.1109/GLOCOM.2018.8647731.
[56] A. Mahdavian, A. Shojaei, and A. A. Oloufa, "Assessing the long-and mid-term effects of connected and automated vehicles on highways' traffic flow and capacity," in *Proc. Int. Conf. Sustain. Infrastruct.: Leading Resilient Communities Through 21st Century*. Reston, VA, USA: American Society of Civil Engineers, 2019, pp. 263–273.
[57] A. Mahdavian, A. Shojaei, S. Mccormick, T. Papandreou, N. Eluru, and A. A. Oloufa, "Drivers and barriers to implementation of connected, automated, shared, and electric vehicles: An agenda for future research," *IEEE Access*, vol. 9, pp. 22195–22213, 2021.
[58] A. Shojaei and A. Mahdavian, "Revisiting systems and applications of artificial neural networks in construction engineering and managements," in *Proc. Interdependence Between Struct. Eng. Construct. Manage. Conf.*, 2019. [Online]. Available: https://www.researchgate.net/profile/Amirsaman-Mahdavian/publication/334446578_REVISITING_SYSTEMS_AND_APPLICATIONS_OF_ARTIFICIAL_NEURAL_NETWORKS_IN_CONSTRUCTION_ENGINEERING_AND_MANAGEMENT/links/5d29fd2d299bf1547cb43586/REVISITING-SYSTEMS-AND-APPLICATIONS-OF-ARTIFICIAL-NEURAL-NETWORKS-IN-CONSTRUCTION-ENGINEERING-AND-MANAGEMENT.pdf
[59] A. Shojaei and A. Mahdavian, "Revisiting systems and applications of artificial neural networks in construction engineering and managements," Tech. Rep., 2019.

AMIRSAMAN MAHDAVIAN (Graduate Student Member, IEEE) received the Bachelor of Science degree in civil engineering from the Amirkabir University of Technology (Polytechnic of Tehran), and the Ph.D. degree in transportation management from the University of Central Florida (UCF). He is currently a main part of the Data Science Team, Civil Engineering Department Research Group, UCF. He worked as a Business Analyst in transportation and construction projects in the industry for a period of four years. He received several honors for his contributions, including at UCF.

ALIREZA SHOJAEI is currently an Assistant Professor with the Department of Building Construction, Myers-Lawson School of Construction, Virginia Tech. His research program is focused on computer-based modeling and simulation. Most of his current work focuses on digital innovation and the application of emerging technologies in the built environment and project management and economics in the construction industry. His track record of publications and grant proposals include information modeling, data sensing and analytics, machine learning, and smart design and construction.

MILAD SALEM (Graduate Student Member, IEEE) received the Bachelor of Science degree in electrical engineering from the Sharif University of Technology, Tehran. He is currently pursuing the Ph.D. degree in computer engineering with the University of Central Florida. He is also a member of the Transilico Research Group, University of Central Florida. His main research interests include machine learning and semi-supervised learning, specifically transferring learned knowledge between different tasks.

HALUK LAMAN received the master's and Ph.D. degrees in civil engineering with a specialty in traffic systems engineering from the University of Central Florida. He is in traffic impact analysis, data analytics with advanced statistical models in travel demand models and traffic safety, ITS technology deployment, and evaluation. His background includes working with ArcGIS for macro level transportation planning projects and VISSIM for traffic micro-simulations. He has also been involved in more than 50 transportation engineering projects in the last eight years.

NAVEEN ELURU is currently a Professor with the Department of Civil, Environmental and Construction Engineering, University of Central Florida. His research is primarily geared towards the formulation and development of discrete choice models that allow us to better understand the behavioral patterns involved in various decision processes. He has conducted research on several topics, including transportation planning, transportation safety, land-use modeling, integrated demand supply models, public health, and environmental sciences. He has published extensively in premier transportation safety journals, such as *Accident Analysis and Prevention* and *Analytic Methods in Accident Research*.

AMR A. OLOUFA is currently a Professor and the Director of construction management with the University of Central Florida. He has a long history of substantial contributions in transportation and construction research and education and has served in many leadership capacities, as evidenced by his role, among others, as the editorial board of 20 journal articles. He has managed 19 transportation management related projects with funding in excess of $6 million supported mostly by FDOT and CATSS. His research interests include traffic analysis, truck diversion, software development, simulation, big data, data analytics, and ITS.