Optimized Predictive Modeling of Lane-Specific Vehicle Time Gaps for Traffic Flow and Road Management

Haiyun Sun
School of Transportation and Electrical Engineering, Dalian University of Science and Technology, China
sunhaiyun@dlust.edu.cn

Bian Ke
School of Transportation and Electrical Engineering, Dalian University of Science and Technology, China
bianke@dlust.edu.cn

Junlin Sha
School of Transportation and Electrical Engineering, Dalian University of Science and Technology, China
sjl2006ljt@163.com

Shidi Wu
School of Transportation and Electrical Engineering Dalian University of Science and Technology, China
wushidi@dlust.edu.cn

Menglu Chen
School of Transportation and Electrical Engineering Dalian University of Science and Technology, China
chen_meng_lu@yeah.net

Abstract: *Lane-specific characteristics, traffic flow, vehicle properties, and atmospheric conditions profoundly impact successive vehicle time intervals. A lane position plays an important role due to the differences in flow properties, such as speed, density, and overtaking maneuvers, which can vary significantly across different lanes. Larger or heavier vehicles, which are assigned to specific lanes, tend to require although time more space from each other due to their dimensions and braking requirements. Driver behavior, as determined by reaction times and compliance with rules, introduces variability, especially in challenging weather or road conditions, such as rain, snow, or icy roads. In addition to geographical location and distance of travel, temporal factors also influence perception and choice, including the day of the week and variations in light intensities. Effective modeling approaches and effective traffic control require a comprehensive analysis of these factors. In the current work, a dataset of independent traffic roads with all the above-mentioned variables was constructed, and time-gap prediction between two consecutive cars driving along every road lane was performed using Support Vector Regression (SVR) and Random Subspace (RSS). In addition, new optimization algorithms, referred to as the Partial Reinforcement Optimizer (PRO) and the Walrus Optimizer (WO), were utilized to enhance predictive capability, resulting in strong hybrid models. In this evaluation approach, we also employed Analysis of Variance (ANOVA's) sensitivity analysis technique to identify the most contributing feature in our data, providing us with the importance of each variable in the prediction process. The SVPR hybrid model performed the best for lane 1, with the highest R^2 of 0.997, the lowest Root Mean Square Error (RMSE) of 2.36E+07, and the lowest Ratio of RMSE to Standard deviation (RSR) of 0.059 during testing, thereby reflecting its superior predictive accuracy and minimal error. For lane 2, the hybrid SVWO model emerged as the most effective, achieving the highest R^2 of 0.989, the lowest RMSE of 3.36×10^7, and the smallest RSR of 0.107, demonstrating its robust capability in capturing lane-specific traffic dynamics. These findings highlight the potential of hybrid optimization techniques to enhance predictive performance and minimize errors in practical traffic management systems.*

Keywords: *Time gap, lane-specific factors, traffic management, machine learning, prediction*.

Received February 19, 2025; accepted October 15, 2025
https://doi.org/10.34028/iajit/23/3/1

1. Introduction

Traffic management is one of the most important aspects of urban planning and road safety. However, as cities expand and vehicle miles traveled increase, comprehending the complexities of traffic flow is becoming a more important pillar in determining the movement of people and goods. Almost every city around the world is facing increasing levels of congestion, traffic, and road safety issues, it has become critical to evaluate and optimize vehicle flow within roadways [3, 26]. One of the main variables in traffic dynamics is the time gap between consecutive vehicles. This parameter does not only reflect traffic density but also the behavior of flow stability and road safety [36]. The distance between two vehicles is known as a time gap, which is often used for analysis regarding driver behavior, traffic conditions, risk of collisions, etc., [10]. To better utilize the road, optimize the traffic, and prevent collisions, it is of high importance to get a reliable mathematical model of periods.

Although time gaps are significant concerning traffic flow analysis in every context, lane-wise characteristics add a completely different layer of complexity to the topic. They are dedicated lanes at the same level, so their arrangement must be adapted to the different behavior of the vehicles that will circulate on them and the different dynamics that different types of vehicles generate. Faster lanes are occupied by faster cars, trucks usually occupy slower lanes [2, 23]. They have limitations in braking and maneuvering, which require significanttime gaps to operate safely. Moreover, overtaking behavior is mainly restricted to some lanes, which consequently causes time gap characteristics to be distorted [27]. This and other differences indicate that the analyses are lane-specific and not a homogeneous treatment of the traffic system [16].

Driver behavior is another important factor that affects time-gap variability. How each person drives a vehicle is considered someone's driving style defined by risk degrees, experience, and exactness of traffic rules [11, 34]. Aggressive drivers create a smaller gap between their car and the car in front of them, increasing the likelihood of rear-end crashes. Conservative drivers, by contrast, create a greater gap. These differences appear more pronounced under heavy traffic or poor weather conditions such as rain, snow, or ice [8]. This again forces drivers to increase their following distances due to the poor visibility and reduced traction on the road surface, which results in big changes in time gap distribution between the lanes [31].

Environmental and temporal factors also contribute significantly to the formation of traffic patterns and vehicle gaps [21]. Weather conditions, such as rain, snow, fog, or extreme heat, affect driver response, vehicle performance, and overall road safety [28]. Adverse weather conditions add to uncertainty in traffic and are likely to require drivers to increase the following distances in order to reduce the risk of accidents. Similarly, time of day influences traffic behavior. During peak hours, greater vehicle density under normal conditions corresponds to smaller vehicle gaps, and this increases drivers' and traffic management systems' requirements [4]. On the other hand, conditions during nighttime driving, characterized by lower volumes of traffic but impaired visibility, may result in wider time gaps despite the lower density [1]. Such variations in time and environmental conditions underline the very complex nature of time gap modeling and the need to incorporate these factors into predictive models.

Recent advances in sensor technology and data collection have made it possible to gather large volumes of high-resolution, lane-level traffic data. Tools such as high-resolution cameras, inductive loop detectors, and connected vehicle systems now provide precise measurements of key traffic parameters, including time gaps, lane positions, and vehicle types [15, 25]. However, making efficient use of this kind of data has its complications. Traffic systems are inherently complex systems that are highly coupled and exhibit non-linear interactions among a large number of variables [2]. In elementary analyses, traditional statistical techniques are helpful, but when dealing with these complexities, they are insufficient. They are not able to adapt dynamically when the conditions of the traffic change, and they also have the inherent weakness of being unable to uncover new patterns or associations [18].

Machine Learning (ML) fundamentally alters traffic modeling by uncovering relationships and patterns not addressed by traditional methods. It analyzes data including traffic density, environmental conditions, and lane-specific behaviors to provide actionable insights for improving road management [17, 19]. On this basis, using huge amounts of heterogeneous resources, from weather forecasts and live traffic streams, ML enables adaptive, context-aware prediction [20]. These variations are dynamic when road conditions turn, leading to improved flowing times, optimized usage, and a more ENTER and responsive transport system [22].

Table 1 summarizes relevant studies that are related to traffic prediction. These studies highlight the variety of methods and data sources used to improve the accuracy, adaptability, and efficiency of traffic management systems. The insights from these works lay the groundwork for understanding the potential of ML in addressing complex traffic dynamics and optimizing transportation systems.

Table 1. Overview of ML applications in traffic prediction.

Authors	Year	Model/Approach	Dataset used	Results
Moumen *et al*. [22]	2023	Random forest regressor.	UK national road traffic data.	30.8% reduction in congestion achieved through adaptive traffic light control.
Meese *et al*. [20]	2024	Federated learning with blockchain.	Arterial traffic data.	Improved prediction accuracy and decentralized training in ITS environments.
Liu *et al*. [17]	2023	Gaussian process-based model.	Urban traffic flow simulation.	Captures dynamic road capacity with high accuracy in real-time applications.
Nan *et al*. [24]	2024	Digital twin with traffic prediction framework.	Cellular traffic datasets.	Efficient framework for network traffic predictions, integrating heterogeneous sources.
Maraia [19]	2023	AI-driven traffic management platform.	Live traffic data.	Enhanced efficiency in traffic and safety management using AI algorithms.
Yan *et al*. [35]	2020	Self-organizing smart city framework.	Smart transportation systems (China).	Improved traffic flow prediction accuracy and resource optimization in smart cities.
Abideen *et al*. [1]	2024	Spatio-temporal transformer-based model.	Large-scale urban data.	Accurate predictions across diverse regions and cities.
Das *et al*. [4]	2023	Multivariate association rule mining.	Urban traffic datasets.	18% reduction in crash risk by incorporating association rules in predictive models.
Dunne and Ghosh [7]	2013	Neurowavelet and ANN models.	Rainfall and urban traffic data.	Effective integration of weather data with traffic forecasts, achieving improved predictions during adverse conditions.

The main goal of this work is to create strong models to predict lane-specific vehicle time gaps using an extensive traffic dataset containing lane-specific parameters, traffic dynamics, vehicle features, and environmental conditions. This research leverages Support Vector Regression (SVR) and Random Subspace (RSS) models made more efficient by guiding them using new optimization algorithms named Partial Reinforcement Optimizer (PRO) and Walrus Optimizer (WO), to optimize predictive accuracy and minimize margins of error. The main aims of the study are:

- **The Utilized Dataset**: capturing a variety of influential factors, including lane-dependent traffic flow features and vehicular types, joint weather features, and temporal scenarios that affect vehicle time gap prediction. In addition, to improve the training of the models to achieve the most accurate outcomes and the use in reality.
- **The Data Preprocessing**: to clean and ensure dataset consistency with negligible noise and irrelevant data that helps the model to be more reliable and efficient.
- **The 5-Fold Cross-Validation and Data Splitting**: to evaluate model performance robustly while minimizing overfitting by dividing the data into 80% for training/validation and 20% for testing.
- **The Analysis of Variance (ANOVA) Sensitivity Analysis**: to quantify the influence of each feature that contributes positively/negatively towards a prediction task and help in understanding the importance of features in datasets and assisting with improving model performance.

Ultimately, the main purpose of this research is to demonstrate predictions of vehicle time gaps at the lane level accurately and reliably that can aid effective implementations in traffic management, such as adaptive traffic control, improved safety on roads, as well as efficient planning when it comes to lane usage.

2. Data Base Description and Analysis

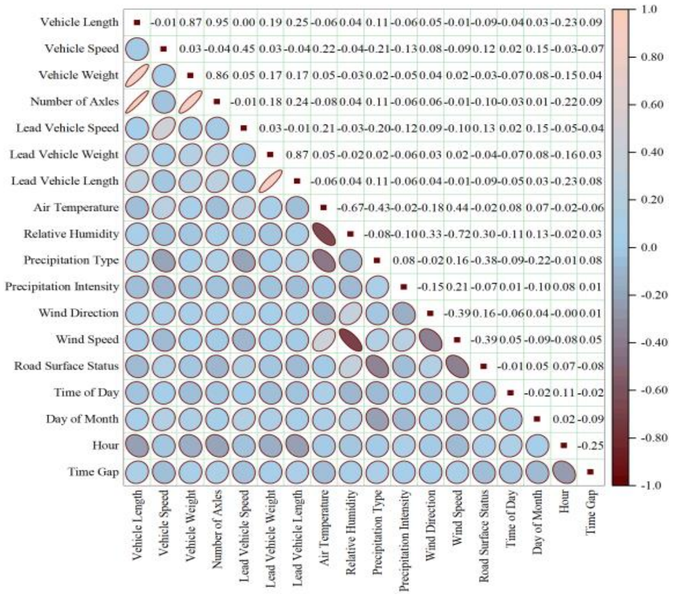
The data used in this study was sourced from [12] and is a credible comprehensive source of traffic flow and behavior data regarding vehicles totaling 311,908 vehicle observations, matched to weather conditions and road surface status data. These classes include precipitation type, road status information, and daylight conditions. It was collected over a long period, capturing a wide variety of weather scenarios ranging from dry summer conditions to adverse winter weather. lane-specific information was recorded for a two-lane road, with detailed observations of both lane 1 and lane 2, ensuring a robust foundation for the predictive modeling of vehicle time gaps under diverse environmental and traffic conditions.

Extensive data preprocessing and cleaning processes were performed to ensure the quality as well as reliability of the analysis. Data for each lane (lane 1 and lane 2) was processed individually to account for lane-specific traffic features. The dataset contained instances of missing values, which were identified and excluded during the pre-processing step as imputing missing values would have applied bias to the predictive modeling process. Consequently, the final sample sizes were 3186 for lane 1 and 2938 for lane 2.

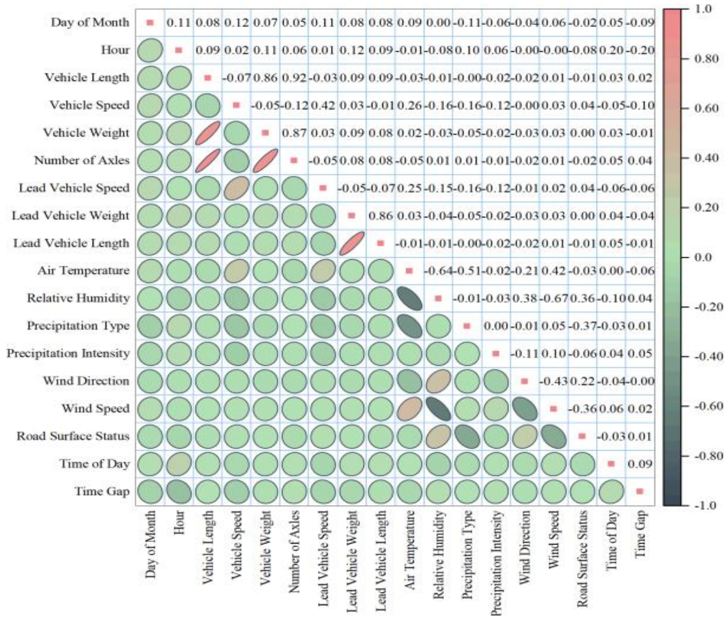
After the preprocessing step, the datasets were analyzed and modeled. Table A.1 (lane 1) and Table A.2 (lane 2), which are presented in the Appendix, provide summary statistics for the variables within the dataset. From the statistical summaries in these tables, it is evident that the two lanes exhibit distinct traffic characteristics. Lane 1 is dominated by larger and heavier vehicles traveling at slightly lower speeds, while lane 2 tends to have smaller and faster-moving vehicles. These differences emphasize the need for lane-specific modeling to accurately predict vehicle time gaps and optimize traffic flow. Additionally, the environmental variables, such as air temperature and relative humidity, remain consistent across both lanes, allowing for controlled comparisons. However, the high variability in the time gap variable suggests further investigation may be needed to account for outliers or extreme values during modeling.

Figure 1 presents the correlation matrices for the variables across both lane 1 and lane 2 datasets. The correlation matrices for both lane 1 and lane 2 datasets reveal mostly moderate correlations, with a few strong relationships standing out. In lane 1, strong correlations are observed between vehicle length and both vehicle weight and the number of axles, reflecting the predictable relationship between vehicle size and structural characteristics. Similarly, relative humidity shows a strong negative correlation with air temperature, indicating colder conditions are typically more humid. Most other correlations in lane 1 remain moderate to weak, emphasizing the complex but less direct interactions between variables.

In lane 2, the overall trends are similar, with correlations also generally moderate. Strong relationships are again seen between vehicle length and both vehicle weight and the number of axles, consistent with lane 1. A notable exception is the strong negative correlation between wind speed and relative humidity, suggesting windier conditions are associated with drier air. As with lane 1, most other variables exhibit weak or moderate correlations, highlighting a general consistency across the two lanes while also reflecting some lane-specific differences in environmental dynamics.



a) Lane 1 dataset.



b) Lane 2 dataset.

Figure 1. Correlation matrices for the variables within the dataset for both datasets (lane 1 and lane 2).

3. Methodology

3.1. Machine Learning Models

Support Vector Machine (SVM) [32, 33] is a binary classification algorithm that transforms data into a higher-dimensional space to find an optimal separating hyperplane, often using the RBF kernel for non-linear data. On the other hand, The RSS [13, 14] method builds decision trees by randomly selecting feature subsets, improving model robustness and generalization by reducing feature correlation and overfitting. Both methods enhance prediction accuracy, with SVM handling complex data and RS excelling in large, feature-rich datasets.

3.2. K-Fold Cross Validation

Cross-validation is a commonly used resampling method for assessing a model with predictive capabilities, and it is both general and computationally efficient for model assessment. One such method is k-fold cross-validation, which is very popular because of the good trade-off it provides between bias and variance. The model’s performance is thoroughly evaluated using 5-fold cross-validation in this study. The dataset was split into 5 equal parts or “folds”. In each iteration, one-fold (20% of data) was set as a testing set and the rest of the 4 folds (80% of data) were used for training. This SVR model was evaluated five times, using each fold as the training set for the other four, until all folds were used to validate the model and with the condition that every data point would serve once in every fold. The advantage of this, apart from a more robust assessment of the performance of the model, at the same time serves as a precaution against overfitting by taking the average of the results across all iterations, leading to more generalizable results [5].

3.3. Optimization Algorithms

3.3.1. Partial Reinforcement Optimizer (PRO)

The PRO is inspired by the partial reinforcement effect principles of psychological learning theory. Since its introduction, first proposed by Ferster and Skinner [9] back in 1957, the theory of PRE provides more weight for intermittent reinforcement, compared to a continuous one during learning. Offering better retention of behaviors, in turn, affects learning rate and learning strength. Reinforcement schedules: fixed-ratio, variable-ratio, fixed-interval, and variable-interval determine the timing and frequency of reinforcements. The theory also includes negative reinforcement that encourages specific behaviors by removing unpleasant circumstances and positive reinforcement that strengthens the behavior by providing favorable consequences [29].

In PRO, each learner's behavior is represented as a decision variable, and learners are modeled as solutions. The primary objective of Algorithm (1), which works with a population of learners, is to increase the objective function value through scheduling and reinforcement. Certain activities are prioritized according to their past performance using a dynamic scheduling method. Variable-interval scheduling is used to update this priority during the optimization process. The algorithm becomes more adaptive and concentrates on fruitful areas of the search space when behaviors with higher scores are more likely to be encouraged.

Decision variables are changed during the stimulation phase of the PRO optimization process to investigate novel solutions. To maintain the effectiveness and focus of the search process, stimulation factors are computed using the normalized scores of chosen choice variables. The performance of the newly created solution is assessed concerning the objective function. Positive reinforcement is used to elevate the contributing behaviors' priority if the solution gets better. On the other hand, negative reinforcement makes those behaviors less important if no improvement is seen. A rescheduling mechanism that resets priorities and behaviors to prevent stagnation is triggered when there is persistent failure to improve.

***Algorithm 1**: The PRO pseudocode.*

Initialization:
Initialize Population
Initialize Schedules
FEs, MaxFEs, SF, RR, SR
while** ($FEs \leq MaxFEs$) **do
***for** $i = 1$ to nPop **do** ▻ for all learners.*
▻ Determine Behaviors of the i^{th} learner based on the scheduler.
Calculate time parameter: $\tau \leftarrow \frac{FEs}{MaxFEs}$.
Calculate selection rate (SR): $SR \leftarrow e^{-(1-\tau)}$.
Select λ number of behaviors with highest priority in $Schedule_i$:
$\mu \subseteq \{1,2,3,\ldots,N\}$ $|\forall_j \in \mu, Schedule^j \geq Schedule^{*,\lambda}$,
$\lambda \leftarrow \{\|\mu\| \mid \|\mu\| = [U(1, N \times SR)]\}$
▻ Stimulate the selected behaviors of the i^{th} learner to get response,
Update Beta and SF
Calculate $X^{\mu}_{i,new}$: $X^{\mu}_{i,new} \leftarrow X^{\mu}_i + SF_i \times S^{\mu}_i$
Apply bound constraints ▻ Evaluate the i^{th} learner response
$f'_X \leftarrow F(X_{i,new})$
▻ Apply Positive or negative reinforcement according to the response.
if** (f'_X is better than f_X) **then
Accept new behavior $X^{\mu}_{i,new}$
Apply positive reinforcement: $Schedul\ e^{\mu}_i \leftarrow Schedul\ e^{\mu}_i + (Schedul\ e^{\mu}_i \times RR)$
else
Reject new behavior $X^{\mu}_{i,new}$
Apply negative reinforcement: $Schedul\ e^{\mu}_i \leftarrow Schedul\ e^{\mu}_i - (Schedul\ e^{\mu}_i \times RR)$
end if
Update the best solution (X_{best})
Conduct rescheduling process
$FEs = FEs + 1$
end for
end while
***return** the best solution (X_{best})*

The reinforcement process is mathematically described to dynamically alter the timetable. A behavior's score is raised in proportion to a predetermined reinforcement rate for positive reinforcement and lowered in proportion to a predetermined reinforcement rate for negative reinforcement. These techniques allow PRO to adapt according to the details of the situation by balancing exploration and exploitation. To better navigate more complex search spaces, the algorithm applies this stochastic method of selecting actions for reinforcement. Taheri *et al.* [29] give a detailed mathematical analysis and framework for this optimizer. Also, Algorithm (1) and Figure 2 present Pseudocode and framework for PRO.

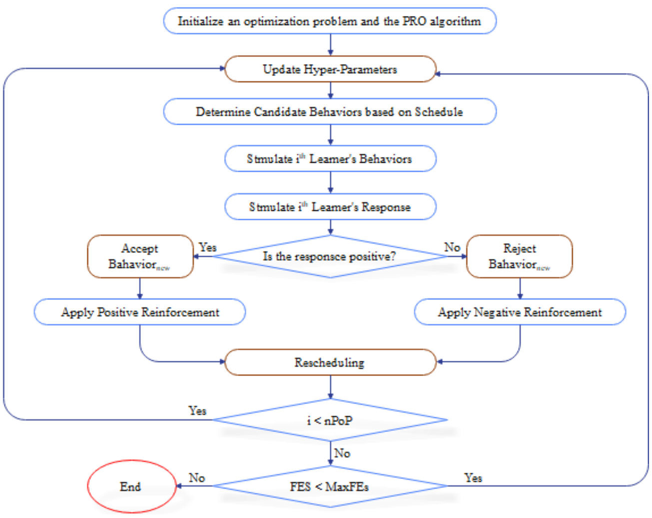


Figure 2. Flowchart of PRO structure.

3.3.2. Walrus Optimization Algorithm (WO)

The intelligent behavior of walruses in the wild serves as the model for the WO algorithm. The three main behaviors of walruses-feeding under the guidance of a dominant member, moving to new locations with seasonal changes, and avoiding or defending against predators such as killer whales and polar bears-form the basis for WO. Mathematical modeling of such behaviors is the investigation of the exploration, migration, and exploitation stages of the optimization process. WO technique achieves an effective balance between global exploration and local exploitation by applying these biological strategies to navigate towards optimal solutions [30].

The algorithm focuses on a broad search approach throughout the exploration process, inspired by the feeding nature of the walrus. The best solution, which is the strongest walrus, guides the population to promising areas within the search space. The method supports the exploration of various types of areas as the position of each walrus is adjusted depending on the position of the leader and random fluctuations. It is defined mathematically by equations involving randomly selected coefficients and leader position minus the current position to create new positions.

The migration step represents the periodic migration of walruses to new areas. The process helps diversify the search process of the algorithm by encouraging the migration of the walruses to whole new areas in the search space. A walrus changes its position whenever it picks another individual from the population randomly to be its migration target. This stage enhances the performance of the algorithm in efficiently searching new regions and breaking local optima. Lastly, the exploitation stage focuses on optimizing solutions in the neighborhood regions through the simulation of how walruses react to threats from predators.

This phase uses a neighborhood search method by establishing a neighborhood around every walrus and looking for improvements. The walrus moves to a new position if there exists a more appealing location in the neighborhood. This stage intensifies the search in promising areas to guarantee that WaOA converges toward the global optimum. Trojovsky and Dehghani [30] provide a rigorous mathematical analysis and formulation of this optimizer. Also, Algorithm (2) and Figure 3 present pseudocode and framework for WO.

***Algorithm 2**: The WO pseudocode.*

Input: Algorithm parameters (population size N, maximum iteration T)
1: Initialize the population and define the related parameters
2: Evaluate the fitness values and obtain the best solution
*3: **While** $t \leq T$*
*4: **If** $|Danger\_signal| \geq 1$ {Exploration phase}*
5: Update new position of each walrus: $X_{i,j}^{t+1} = X_{i,j}^{t} + Migration\_step$
*6: **Else** {Exploitation phase}*
*7: **If** $Safety\ signal \geq 0.5$ // Breeding behavior //*
*8: **For** each male walrus*
9: Update new position based on Halton sequence
*10: **End For***
*11: **For** each female walrus*
12: Update new position:
$Female_{i,j}^{t+1} = Female_{i,j}^{t} + \alpha.\left(Male_{i,j}^{t} - FeMale_{i,j}^{t}\right) + (1-\alpha).\left(X_{best}^{t} - FeMale_{i,j}^{t}\right)$
*13: **End For***
*14: **For** each juvenile walrus*
15: Update new position: $Juvenile_{i,j}^{t+1} = \left(O - Juvenile_{i,j}^{t}\right).P$
*16: **End For***
*17: **Else** // Foraging behavior //*
*18: **If** $|Danger\ signal| \geq 0.5$ // Gathering behavior //*
19: Update new position of each walrus: $X_{i,j}^{t+1} = X_{i,j}^{t}.R - \left|X_{best}^{t} - X_{i,j}^{t}\right|.r_4^2$
*20: **Else** // Fleeing behavior //*
21: Update new position of each walrus: $X_{i,j}^{t+1} = (X_1 + X_2)/2$
*22: **End If***
*23: **End If***
*24: **End If***
25: Update the walrus position
26: Calculate the fitness value and update the current best solution
27: $t = t + 1$
*28: **End While***
Output: the best solution

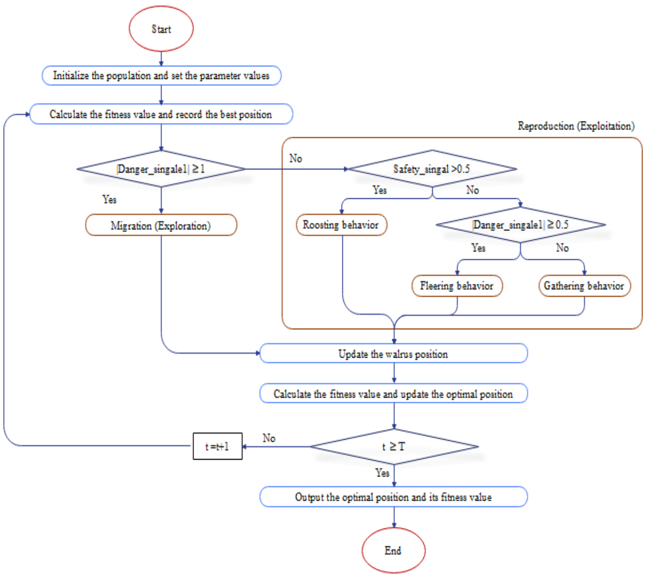


Figure 3. Flowchart of WO structure.

3.4. Performance Indices of Models

There exist many performance indices to evaluate predictive models, each with its way of quantifying the precision, reliability, and resilience of predictive methods. They give insight into how well the models capture patterns in data and generalize to unseen samples. Using multiple indices allows for a more holistic performance assessment on diverse metrics. The metrics used to compare the original and predicted data in this study are coefficient correlation (R^2), RMSE, and the Ratio of RMSE to RSR. The formula for each is as follows from Equations (1), (2), and (3).

$$R^2 = \left(\frac{\sum_{i=1}^{n}(b_i - \bar{b})(m_i - \bar{m})}{\sqrt{\left[\sum_{i=1}^{n}(b_i - \bar{b})^2\right]\left[\sum_{i=1}^{n}(m_i - \bar{m})^2\right]}}\right)^2 \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - b_i)^2} \quad (2)$$

$$RSR = \frac{RMSE}{St.D} \quad (3)$$

where,

- The sample size is denoted by n.
- The predicted value is represented by b_i.
- $\bar{m}$ and $\bar{b}$ respectively stand for the measured and mean predicted values.
- The measured value is denoted by m_i.

3.5. Sensitivity Analysis

The present study applies the ANOVA method to find out which predictors are most sensitive to variable inputs. The approach has several advantages: it improves prediction accuracy, reduces runtime, and lowers computational requirements. In most cases, ANOVA satisfies its basic assumptions and is a simple, widely used technique for comparing means of groups. It gives efficient inference in cases where the sizes of the groups are not equal and intuitively illustrates the relationship between the variables. A powerful tool for doing an in-depth analysis of data, it can be used for more than two groups without raising the possibility of type I errors [6].

4. Methodology Flowchart

The methodology of this study consists of several interconnected stages, as depicted in the flowchart presented in Figure 4:

- Stage 1: Data preparation: in the first stage, a database was compiled in which for lane 1 3186 samples were provided, and for lane 2 2938 samples, in which 18 input parameters and one output target (the time gap in lanes). The data was subject to careful preprocessing to produce clean data of high quality and accuracy, which was then split into training (80%) and testing (20%) portions for the model development and evaluation.
- Stage 2*:* Training models: SVR and RSS models were trained using the training subset. Using a 5-fold cross-validation technique applied during the training process for robust and unbiased evaluation. By systematically splitting the dataset into five subsets, each subset would serve as a validation fold in turn. The results from cross-validation guided how to best configure each model using metrics like R^2 and RMSE.
- Stage 3: Development of hybrid models: two new metaheuristic algorithms were used to optimize the SVR and RSS models: PRO and WO. Then, these four hybrid models were established: Support Vector Regression optimized by the Partial Reinforcement optimizer (SVPR), Support Vector Regression optimized by the Walrus Optimizer (SVWO), Random Subspace optimized by the Partial Reinforcement Optimizer (RSPR), and Random Subspace optimized by the Walrus Optimizer (RSWO). However, these hybrid models leveraged the strengths of the base models and optimizers for better prediction accuracy and lower errors.
- Stage 4: Model validation: the testing subset was employed to validate the performance of the hybrid models. Various evaluator metrics, including R^2, RMSE, and RSS, were used to assess the models' predictive capabilities.
- Stage 5: Sensitivity analysis: an ANOVA sensitivity analysis was conducted to evaluate the influence of the 18 input parameters on the model's prediction performance.

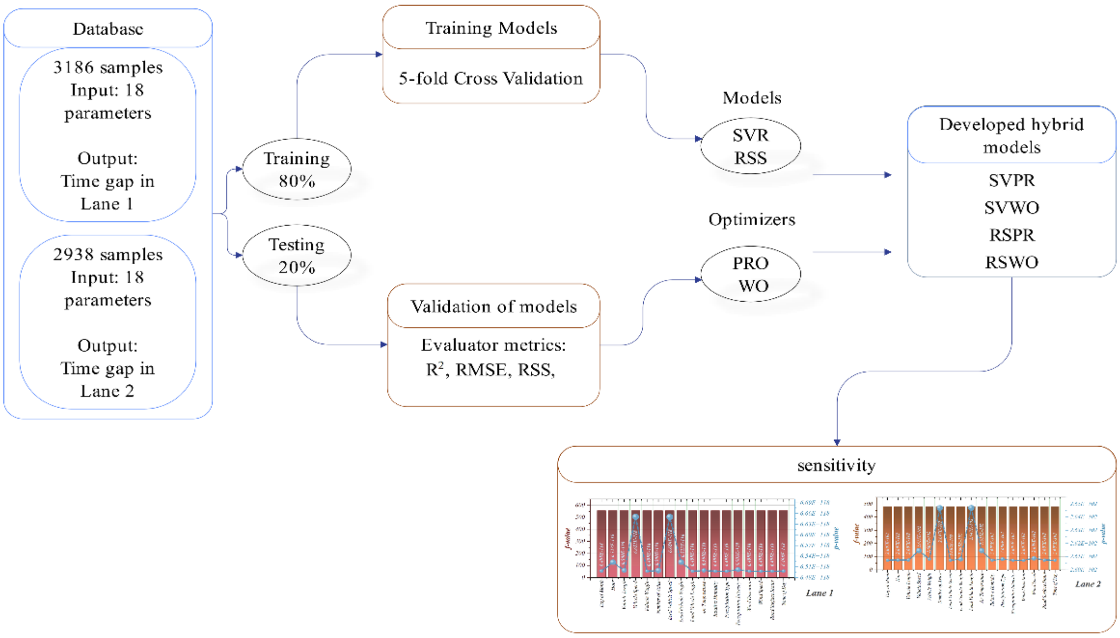


Figure 4. Methodology flowchart of the study.

5. Results

5.1. Cross-Validation Results

Tables 2 and 3 have shown different fold selections for the SVR and RSS models, respectively, based on their performance metrics in cross-validation. For lane 1, SVR obtained its highest R^2 value of 0.980 and the lowest RMSE of 3.85E+07 at fold 2, the best testing fold for the model. On the other hand, RSS had the best performance in fold 3 with an R^2 value of 0.936 and an RMSE of 6.99E+07, showing that fold 3 was most suitable for this model. That would mean each model has their peak performance on a different fold, further increasing the importance of this technique.

For lane 2, fold 5 was taken as the best testing fold for SVR with the highest R^2 value of 0.900 and lowest RMSE as 8.71E+07. In contrast, RSS had its best performance in fold 3; hence, the R^2 value was 0.943, with an RMSE of 7.63E+07. The results prove how k-fold cross-validation can help decide the most appropriate data partitioning for model evaluation.

Table 2. Results of fold selection (train-test split) for the models: lane 1-time gap prediction.

Model	Index values	K-fold				
		K1	K2	K3	K4	K5
SVR	R2	0.974	0.980	0.930	0.957	0.970
	RMSE	6.13E+07	3.85E+07	6.14E+07	6.10E+07	5.63E+07
RSS	R2	0.838	0.887	0.936	0.853	0.832
	RMSE	9.47E+07	7.53E+07	6.99E+07	1.18E+08	1.16E+08

Table 3. Results of fold selection (train-test split) for the models: lane 2-time gap prediction.

Model	Index values	K-fold				
		K1	K2	K3	K4	K5
SVR	R^2	0.892	0.850	0.832	0.859	0.900
	RMSE	8.74E+07	9.58E+07	9.50E+07	9.09E+07	8.71E+07
RSS	R^2	0.861	0.913	0.943	0.880	0.851
	RMSE	8.69E+07	9.39E+07	7.63E+07	8.10E+07	9.24E+07

5.2. Convergence Analysis and Hyperparameters

Tables 4 and 5 summarize the convergence behavior of the developed hybrid models (SVPR, SVWO, RSPR, and RSWO) over 200 iterations for lane 1 and 2.

Concerning lane 1, the initial RMSE values reveal that all models start with high error magnitudes. By the final iteration, all models demonstrate substantial reductions in RMSE, with SVPR achieving the lowest error of 2.13E+07, followed by SVWO, while RSPR and RSWO models show slightly higher RMSEs. Across iteration ranges, SVPR shows a steady and consistent error decline, with the most notable reduction occurring after iteration 160. SVWO and RSPR follow a similar pattern, reducing error gradually in earlier ranges (1-80) and more substantially after iteration 120. RSWO, despite starting with a significantly higher RMSE, demonstrates strong improvements in later stages (160-200), reflecting a slower but effective convergence process.

Table 4. Convergence results for the hybrid models over optimization iterations: lane 1-time gap prediction.

Hybrid	Initial RMSE	Average RMSE (in iteration range)					Final RMSE
		1-40	40-80	80-120	120-160	160-200	
SVPR	6.10E+07	5.78E+07	5.09E+07	4.57E+07	3.64E+07	2.56E+07	2.13E+07
SVWO	6.03E+07	5.78E+07	5.18E+07	4.71E+07	4.16E+07	3.61E+07	3.19E+07
RSPR	6.09E+07	5.75E+07	5.17E+07	4.70E+07	4.11E+07	3.50E+07	3.23E+07
RSWO	1.59E+08	1.46E+08	1.15E+08	9.90E+07	8.73E+07	7.26E+07	5.02E+07

Table 5. Convergence results for the hybrid models over optimization iterations: lane 2-time gap prediction.

Hybrid	Initial RMSE	Average RMSE (in iteration range)					Final RMSE
		1-40	40-80	80-120	120-160	160-200	
SVPR	1.58E+08	1.52E+08	1.38E+08	1.26E+08	9.82E+07	7.13E+07	6.52E+07
SVWO	1.26E+08	1.20E+08	1.05E+08	9.11E+07	5.32E+07	3.90E+07	3.50E+07
RSPR	2.82E+08	2.73E+08	2.38E+08	2.12E+08	1.79E+08	1.19E+08	8.41E+07
SWO	1.56E+08	1.51E+08	1.35E+08	1.18E+08	8.72E+07	7.57E+07	6.73E+07

Concerning lane 2, the initial RMSE values of the models are relatively higher, in comparison to lane 1, revealing higher error magnitudes of the models in the initial stages. By the final iteration, SVWO achieves the lowest RMSE of 3.50E+07, followed by RSPR. SVPR and RSWO converge to intermediate final errors, moderately higher than SVWO mode. Among all models, SVWO exhibits the sharpest error reduction, particularly between iterations 80-160, reflecting its efficiency in optimization. RSPR, although starting with the highest initial RMSE, steadily reduces error over the first 120 iterations before slowing down in later stages. SVPR follows a consistent convergence trajectory, similar to SVWO but less effective overall. RSWO shows slower initial progress but achieves significant error reduction in the final 40 iterations.

Table 6. Optimal hyperparameter values for the hybrid models at final step of optimization iterations: lane 1-time gap prediction.

Models name	Hyperparameters
SVPR	C=1.340 epsilon=0.090
SVWO	C=1 epsilon=0.097
RSPR	C=1.289 epsilon=0.097
RSWO	C=1 epsilon=0.893

Table 7. Optimal hyperparameter values for the hybrid models at final step of optimization iterations: lane 2-time gap prediction.

Models name	Hyperparameters
SVPR	C=1.467 epsilon=0.097
SVWO	C=1.236 epsilon=0.094
RSPR	C=1.377 epsilon=0.092
RSWO	C=1.326 epsilon=0.799

Tables 6 and 7 present the hyperparameters tuning for the developed models for lane 1 and lane 2, respectively. The hyperparameters, especially the regularization parameter and error tolerance, have an important role in the shape of convergence and predictive performance of the models. The variation in these parameters across the models reflects the distinct strategies employed by the optimization techniques to balance model complexity and error minimization. For lane 1, the hyperparameters differ slightly among models, which implies that the learning dynamics for each hybrid have subtly adjusted the hyperparameters. Similarly, for lane 2, fine-tuning emphasizes the importance of finding the right values for these parameters so that effective learning can be achieved for different traffic conditions.

5.3. Investigation of the Performance of ML Models

Tables 8 and 9 show the performance results between base models and their hybrid counterparts for lane 1 and lane 2 lane time gap predictions, respectively. Three crucial metrics are used to evaluate the models in the building and validation and testing stages: RMSE, and RSR. The hybrid approaches including SVPR, SVWO, RSPR, and RSWO show the optimization process serves as a key step to increase the performance of basic models with different traffic flow complexities.

Table 8. Performance evaluators value for the base models and hybrid counterparts: lane 1-time gap prediction.

Models name	Index values								
	R^2_{Train}	$R^2_{validation}$	R^2_{Test}	$RMSE_{Train}$	$RMSE_{validation}$	$RMSE_{Test}$	RSR_{Train}	$RSR_{validation}$	RSR_{Test}
SVR	0.977	0.980	0.990	3.82E+07	3.73E+07	4.27E+07	0.152	0.142	0.106
SVPR	0.993	0.995	0.997	2.13E+07	1.96E+07	2.36E+07	0.085	0.074	0.059
SVWO	0.984	0.988	0.993	3.19E+07	2.95E+07	3.54E+07	0.127	0.112	0.088
RSS	0.942	0.880	0.909	7.02E+07	7.11E+07	6.55E+07	0.246	0.356	0.322
RSPR	0.987	0.974	0.976	3.33E+07	3.26E+07	3.20E+07	0.117	0.163	0.157
RSWO	0.970	0.937	0.951	5.02E+07	5.08E+07	4.68E+07	0.176	0.254	0.230

Table 8 highlights the performance of models concerning lane 1, in which the hybrid SVPR model stands out, achieving the highest R^2 score of 0.997, the lowest RMSE of 2.36E+07, and the smallest RSR of 0.059 in the testing phase, indicating its superior predictive accuracy and error reduction. SVWO follows as the next best performer, with a test R^2 of 0.993, an RMSE of 3.54E+07, and an RSR of 0.088, highlighting its robustness though slightly less effective than SVPR.

The base SVR model, while demonstrating decent predictive performance (R^2=0.990, RMSE=4.27E+07, RSR=0.106), is outperformed by its hybrid counterparts, particularly SVPR and SVWO, which benefit from advanced optimization techniques. Among the RSS-based models, RSPR performs reasonably well, with R^2 and RMSE of 0.976 and 3.20E+07, respectively, showing improvement over the base RSS model, which struggles with the lowest test R^2 of 0.909 and the highest RMSE of 6.55E+07. RSWO performs moderately well but with its lower test R^2 and RMSE, lags behind SVPR and SVWO.

Table 9 reveals distinct trends compared to lane 1. In the test phase, SVWO delivers the most impressive results, achieving the highest R^2 score of 0.989, the lowest RMSE of 3.36E+07, and the smallest RSR of 0.107, demonstrating its strong ability to model the traffic flow in lane 2. SVPR, while less effective, also performs well, with a test R^2 of 0.961, an RMSE of 6.32E+07, and an RSR of 0.200, reflecting its capability to balance error minimization and model complexity.

Table 9. Performance evaluators value for the base models and hybrid counterparts: lane 2-time gap prediction.

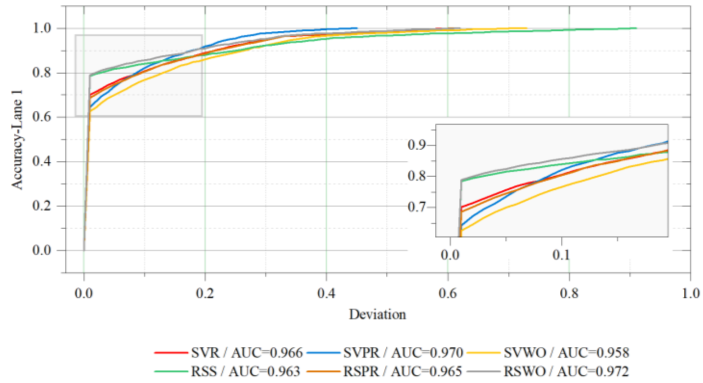
Models name	Index values								
	R^2_{Train}	$R^2_{validation}$	R^2_{Test}	$RMSE_{Train}$	$RMSE_{validation}$	$RMSE_{Test}$	RSR_{Train}	$RSR_{validation}$	RSR_{Test}
SVR	0.935	0.968	0.951	7.61E+07	8.22E+07	7.14E+07	0.263	0.181	0.227
SVPR	0.952	0.978	0.961	6.52E+07	6.77E+07	6.32E+07	0.225	0.149	0.200
SVWO	0.986	0.995	0.989	3.50E+07	3.21E+07	3.36E+07	0.121	0.071	0.107
RSS	0.899	0.870	0.883	1.08E+08	1.43E+08	6.37E+07	0.343	0.376	0.381
RSPR	0.936	0.919	0.860	8.41E+07	1.11E+08	7.10E+07	0.266	0.291	0.424
RSWO	0.958	0.947	0.904	6.73E+07	8.87E+07	5.67E+07	0.213	0.233	0.339

The base SVR model, despite its simplicity, performs reasonably well in lane 2, achieving a test R^2 of 0.951, an RMSE of 7.14E+07, and an RSR of 0.227, though it falls short of the hybrid models, particularly SVWO. The RSS-based models, however, struggle significantly in lane 2. RSPR shows weaker predictive performance with the lowest test R^2 of 0.860 and a high RSR of 0.424, while RSWO improves slightly, though it still underperforms compared to SVWO and SVPR. The base RSS model demonstrates the weakest overall performance, with the highest RMSE and the largest RSR of 0.381 in the test phase.

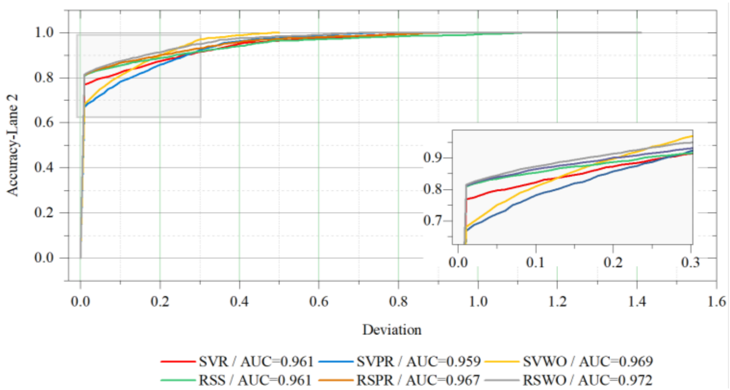
The results demonstrated that across training and validation, hybrid models demonstrate superior performance relative to their base counterparts, indicating the significance of optimization techniques such as PRO and WO in improving predictive precision and minimizing error. This highlights the power of hybridization for learning complex lane-specific traffic dynamics, particularly in lane 2, where a higher diversity of vehicle types, speeds, and overtaking behavior can pose greater challenges for modeling. Real-world applications such as adaptive traffic control systems, will greatly benefit from such precise models of signalized intersections (as lane-specific predictions are required to develop the split, for instance). Nevertheless, applying these models in practice needs to address the uncertainties infused by the variability related to driver behavior and environmental conditions in addition to practical concerns like data quality, computational efficiency, and scalability necessary for successful integration of these models in real-time traffic management systems.

5.4. Visual Analysis of the Models Performance

Figure 5 presents the Receiver Operating Characteristic (REC) curves of the developed models over lane 1 and lane 2, linking accuracy with deviation tolerance. The curves show how well each model predicts as the deviation increases, whereas the higher curves indicate better predictive accuracy. Inset sections in both figures provide a detailed view of model performance in the critical low-deviation range, emphasizing their behavior in high-precision scenarios.



a) Vehicle time gap prediction for lane 1 dataset.



b) Vehicle time gap prediction for lane 2 dataset.

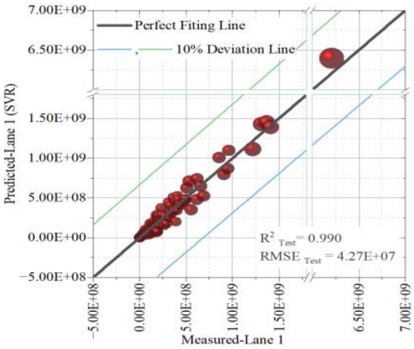
Figure 5. REC curves and the predictive accuracy of the developed prediction models based on AUC values (predictions for both lane 1 and lane 2 datasets).

Analyzing the Area Under Curve (AUC) values, which are the quantitative measures for the models' overall predictive accuracy, one may notice that the hybrid model RSWO outperforms all others for both lanes, with an AUC of 0.972 for lane 1 and lane 2. The SVPR comes closest in lane 1 with an AUC of 0.970, and the base model SVR achieves an AUC of 0.966, which reflects its competitive but slightly lower accuracy. In lane 2, SVWO and RSPR perform well with AUC values of 0.969 and 0.967, respectively, again showing the consistent advantage of hybrid models over base ones in capturing lane-specific traffic dynamics.

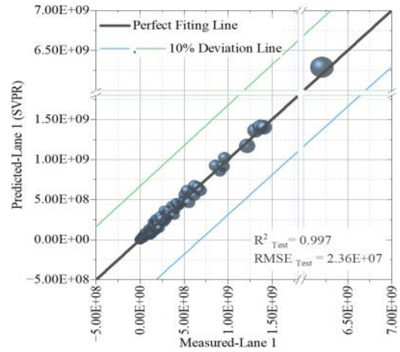
Scatter plots and error distributions in Figure 6 display the performance of different base and hybrid approaches in predicting the vehicle time gaps of lane 1 and lane 2. Scatter plots show predictions vs actual values, while error distributions show the remaining error frequency and size, including z-score outlier detection.

In each of the scatter plots a discernible trend is apparent, that hybrid models outperform base models as seen in the fact that they lie close to the perfect-fit line, have higher values, and have lower RMSE scores. The predictive accuracy and minimal error of SVPR and SVWO are indicative of the efficiency of the applied optimization techniques. Compared to this, base models (especially with the RSS model), have larger errors, less agreement with the perfect-fit line, and lower accuracy.

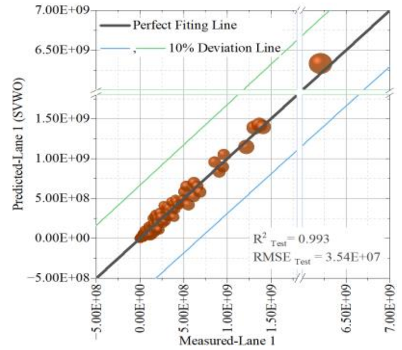
These results are confirmed by the error distributions with fitted curves appearing as histograms. For hybrid models, the residuals are indeed centered around zero which indicates that they are not biased and have a good error profile. Hybrid models also produce narrower distributions of error, with fewer extreme errors, and larger ranges of error in base models show greater variability. The few outliers were found with a z-score limit of ±3, demonstrating the resilience of the hybrid models in limiting extreme prediction errors.



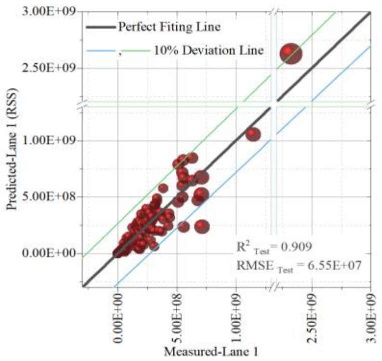
a) Scatter plot for predictions by SVR-lane 1 dataset.



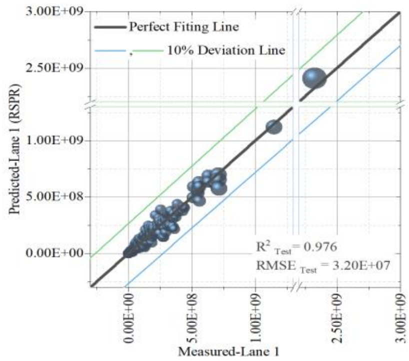
b) Scatter plot for predictions by SVPR-lane 1 dataset.



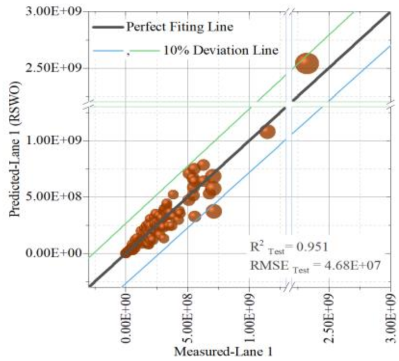
c) Scatter plot for predictions by SVWO-lane 1 dataset.



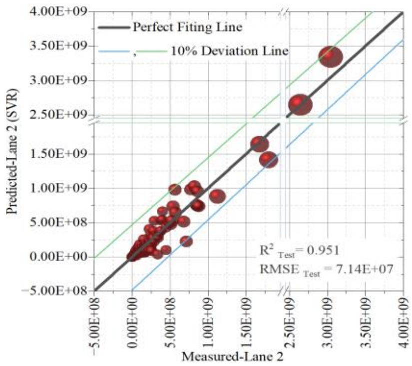
d) Scatter plot for predictions by RSS-lane 1 dataset.



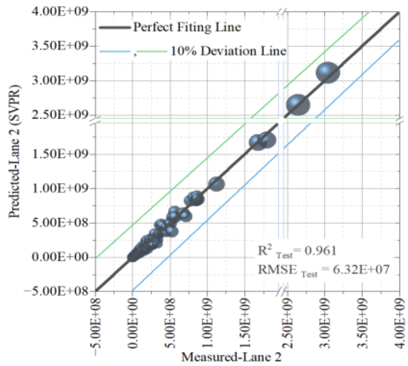
e) Scatter plot for predictions by RSPR-lane 1 dataset.



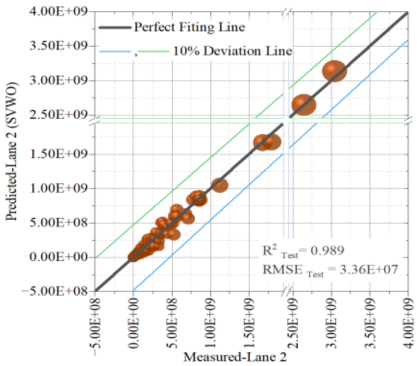
f) Scatter plot for predictions by RSWO-lane 1 dataset.



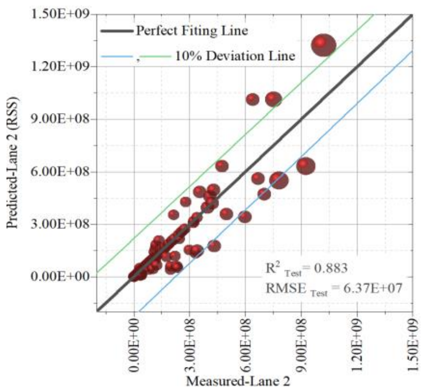
g) Scatter plot for predictions by SVR-lane 2 dataset.



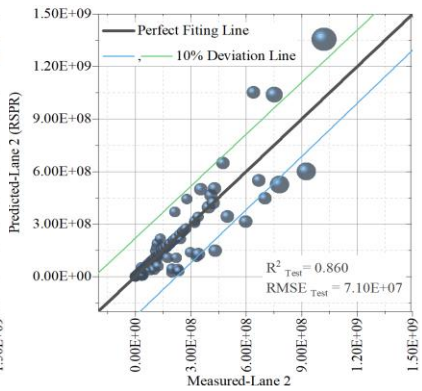
h) Scatter plot for predictions by SVPR-lane 2 dataset.



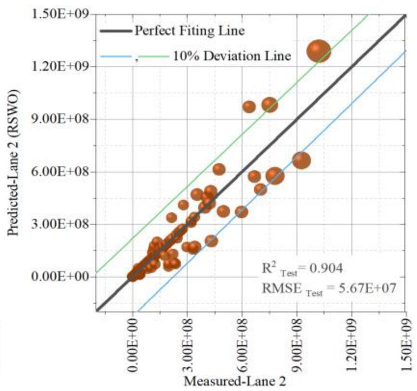
i) Scatter plot for predictions by SVWO-lane 2 dataset.



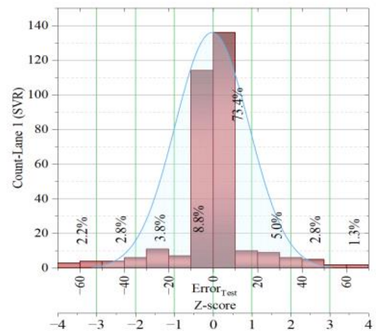
j) Scatter plot for predictions by RSS-lane 2 dataset.



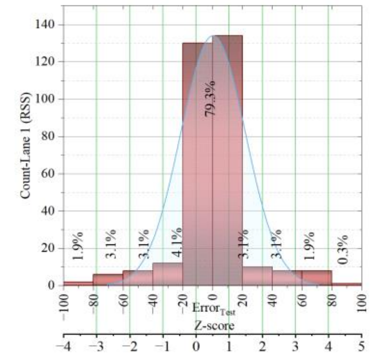
k) Scatter plot for predictions by RSPR-lane 2 dataset.



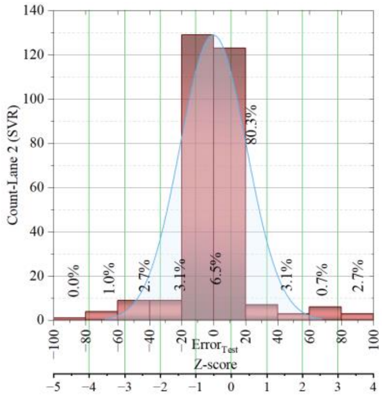
l) Scatter plot for predictions by RSWO-lane 2 dataset.



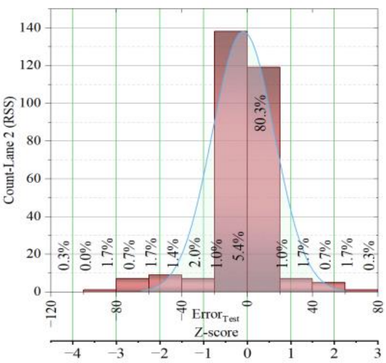
m) Error distribution plot for predictions by SVR-lane 1 dataset. n) Error distribution plot for predictions by SVPR-lane 1 dataset. o) Error distribution plot for predictions by SVWO-lane 1 dataset.



p) Error distribution plot for predictions by RSS-lane 1 dataset. q) Error distribution plot for predictions by RSPR-lane 1 dataset. r) Error distribution plot for predictions by RSWO-lane 1 dataset.



s) Error distribution plot for predictions by SVR-lane 2 dataset. t) Error distribution plot for predictions by SVPR-lane 2 dataset. u) Error distribution plot for predictions by SVWO-lane 2 dataset.

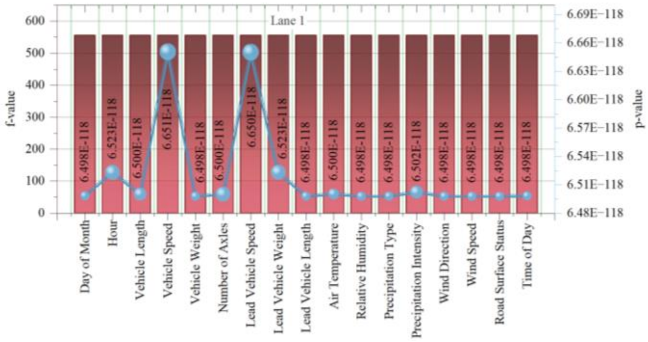


v) Error distribution plot for predictions by RSS-lane 2 dataset. w) Error distribution plot for predictions by RSPR-lane 2 dataset. x) Error distribution plot for predictions by RSWO-lane 2 dataset.

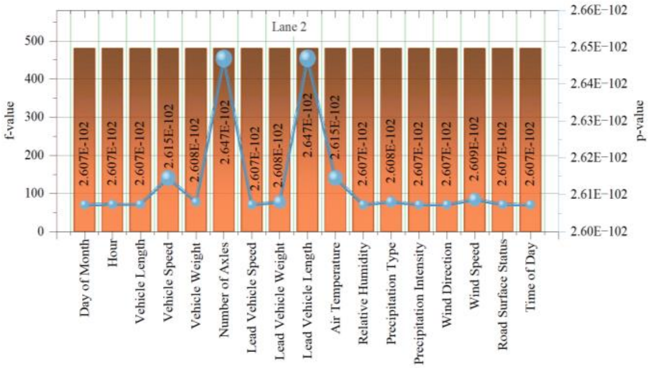
Figure 6. Performance comparison of base and hybrid models (for lane 1 and lane 2 datasets): scatter plots and error distributions with z-score outliers.

5.5. Sensitivity Analysis Results

The results of ANOVA sensitivity, as depicted in Figure 7 show that f-values for lanes are similar for all variables, and the p-values fluctuate slightly. The p values within the range of 7.43671e-03 to 2.28107e-02, demonstrate they are statistically significant at the 5% level in predicting Normalized Mutual Information (NMI) based on the variables used in this experiment. Thus, the analysis suggests that all metrics are significant predictors, negating feature selection.



a) Sensitivity of variables in lane 1 dataset.



b) Sensitivity of variables in lane 2 dataset.

Figure 7. ANOVA sensitivity analysis results for lane-specific time gap prediction (lane 1 and lane 2).

6. Discussion and Perspectives

The study has brought forth a new method for predicting lane-specific vehicle time gaps using hybrid ML models, optimized with novel algorithms. High predictive accuracy with minimal error was exhibited by these models, with great potential for applications in real-world traffic management. This discussion now looks at the broader applicability of the findings, especially within the context of digital highways, accident prevention, and road safety enhancement.

The hybrid models suggested in this research, i.e., SVPR for lane 1 and SVWO for lane 2, were extremely precise in the prediction of the time headways of vehicles. With the inclusion of a wide variety of factors such as vehicle characteristics, surroundings, and time variation, these models can be regarded as an active response to prevailing traffic issues. The capacity of these systems to predict conditions makes them especially well-suited to be included in real-time adaptive traffic control systems. Traffic lights, for instance, can be controlled in real-time based on prediction periods, thus eliminating congestion, evening out traffic flow, and reacting to emergent circumstances such as heightened vehicle concentration during rush hours or poor weather conditions.

In urban and suburban settings, where smooth traffic flow is of prime importance, these models can reduce travel time, and fuel usage, and improve safety. These forecasting systems also fit into smart city planning efforts that focus on maximizing urban infrastructure and transport networks. Furthermore, their ability to provide real-time, data-driven information allows local government departments to plan and manage traffic more effectively, thereby ensuring safer and more efficient roads.

Concepts of digital highways with Internet of Things (IoT) sensors, networked cars, and real-time data processing are in line with the models developed here. The accurate prediction of lane-by-lane time gaps will improve traffic management, reduce congestion, and optimize the effective usage of capacity on these highways. In addition, predictive modeling will allow traffic management authorities to recognize emerging trends and inform strategic planning for infrastructure and resource investment in future needs.

Some of the most glaring, yet substantial, benefits of such predictive models include road safety improvement. Such models can detect inconsistent or irregular time gaps, thus detecting accident-prone stretches. Additionally, real-time alerts can warn drivers in risky areas, prompting better safety habits, such as safe follow distances or slower speeds. These measures will take a long way in averting many rear-end crashes and other common accidents from tailgating situations.

Moreover, these models can identify patterns and anomalies typical of unsafe driving behavior, including consistently low time gaps that indicate aggressive driving. These are employable in framing targeted interventions, which may involve enforcement or public safety campaigns for encouraging safer driving behavior. For example, bigger vehicles needing increased braking distances can be directed into the right lanes, thus minimizing the occurrence of accidents resulting from wrong lane usage.

Such predictive models offer lane-specific actionable insights, thus enabling traffic management systems to optimize lane usage and overtaking strategies under high-speed conditions. This is particularly important for reducing accidents caused by inadequate changes in lanes or reckless driving maneuvers in congested areas. Such system integration with traffic management approaches offers a complete toolkit for responding to the challenges of contemporary transportation concerning road safety and operational efficiency.

7. Conclusions

This paper focuses on predicting lane-specific vehicle time gaps on a two-lane highway. Time gaps are amongst the most significant factors to increase road safety and traffic efficiency. Time gaps affect traffic flow stability, risk of collisions, and lane usage planning. The study considers the inherent complexity of traffic systems. Traffic systems involve nonlinear interactions between vehicle behavior, environmental, and time factors. In order to manage such complexity, advanced machine learning methods are used. It makes use of hybrid models with the assistance of modern optimization methods. This is a shift from the conventional method of traffic control. It offers authentic, data-driven data for utilitarian use. The original dataset had 311,908 records. It was reduced to 3,186 samples for lane 1 and 2,938 samples for lane 2 upon cleansing. The dataset had a high range of variables. These are vehicle characteristics that are addressed (e.g., speed, weight, and length), environmental conditions (e.g., rainfall, temperature, and humidity), and traffic flows (e.g., time of day and road condition). These led to enriching the dataset and making it holistic for predictive modeling. The study used two models: SVR and RSS. These were also improved by combining them with the PRO and the WO. A 5-fold cross-validation approach was employed to measure model performance and prevent overfitting. For lane 1, SVR performed best in fold 2 while RSS performed best in fold 3. For lane 2, SVR performed best in Fold and RSS performed best in fold 3. These results underscore the importance of cross-validation in identifying optimal model configurations. The hybrid models exhibited exceptional predictive performance. For lane 1, the SVPR model achieved the highest R^2 of 0.997, the lowest RMSE of 2.36E+07, and the smallest RSR of 0.059 during testing, highlighting its superior accuracy and error minimization. For lane 2, the SVWO model emerged as the best performer, attaining an $R^2$of 0.989, the lowest RMSE of 3.36E+07, and the smallest RSR of 0.107, demonstrating its robust capability to capture lane-specific traffic dynamics. These results reflect the effectiveness of the hybrid optimization techniques in solving increasingly sophisticated, nonlinear traffic dynamics. Additionally, through ANOVA analysis, it was established that all the categories of variables, vehicle attributes, weather, and time dynamics, play significant roles in improving predictability accuracy. The analysis verified the use of all variables without any further requirement for feature selection.

8. Future Directions

While the models obtain outstanding performances, there is still room for further development and improvement.

- **Integration with Additional Data Sources**: real-time weather data, real-time traffic feeds, and driving behavior can be integrated to enhance the accuracy and robustness of the models. Multimodal data fusion can enable better predictions in highly dynamic and complex traffic flow situations.
- **Advancements in Optimization Algorithms**: future work studies can use more sophisticated optimization algorithms to improve convergence and precision. This strategy is especially advantageous when trying to produce big and varied datasets.
- **Scalability and Computational Efficiency**: practical real-world usage requires scalable solutions with the ability to process high amounts of traffic data in real-time. Future work should include studying the computational requirements of these models and investigating the application of distributed computing or edge computing methods for seamlessly integrating these models into existing traffic management systems.
- **Application to Autonomous Vehicles**: the application of these models to autonomous vehicle technology has the potential to profoundly change the interaction dynamics and driving behaviors of autonomous vehicles when it is interacted with human drivers in a mixed-traffic setting.

Funding

This work was supported by the Discipline Team Cultivation Project of Dalian University of Science and Technology under Grant KYXK2025001

References

[1] Abideen Z., Sun X., and Sun C., "Traffic Flow Prediction: A 3D Adaptive Multi-Module Joint Modeling Approach Integrating Spatial-Temporal Patterns to Capture Global Features," *Journal of Forecasting*, vol. 43, no. 7, pp. 2766-2791, 2024. DOI:10.1002/for.3147

[2] Ayu Z., Sari P., and Mauramdha G., "Analysis of Traffic Flow Due to Lane Changes by Heavy Vehicles," *in Proceedings of the IOP Conference Series: Earth and Environmental Science*, Lagos, pp. 1-11, 2024. DOI:10.1088/1755-1315/1294/1/012026

[3] Brahimi M., Karatzas S., Theuriot J., and Christoforou Z., "Drones for Traffic Flow Analysis of Urban Roundabouts," *International Journal for Traffic and Transport Engineering*, vol. 9, no. 3, pp. 62-71, 2020. DOI: 10.5923/j.ijtte.20200903.02

[4] Das S., Dutta A., Tamakloe R., and Khan M., "Analyzing the Time-Varying Patterns of Contributing Factors in Work Zone-Related Crashes," *Journal of Transportation Safety and Security*, vol. 16, no. 4, pp. 655-682, 2023. DOI:10.1080/19439962.2023.2246020

[5] Dey L., Naik B., Poojita O., and Pothireddi K., "SMM4H 2024: 5 Fold Cross Validation for Classification of Tweets Reporting Children's Disorders," *in Proceedings of the 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, pp. 55-57, 2024. DOI: 10.18653/v1/2024.smm4h-1.12

[6] Ding H., Feng P., Chen W., and Lin H., "Identification of Bacteriophage Virion Proteins by the ANOVA Feature Selection and Analysis," *Molecular BioSystems*, vol. 10, no. 8, pp. 2229-2235, 2014. DOI:10.1039/c4mb00316k

[7] Dunne S. and Ghosh B., "Weather Adaptive Traffic Prediction Using Neurowavelet Models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 370-379, 2013. DOI:10.1109/TITS.2012.2225049

[8] Faouzi N., Billot R., Nurmi P., and Nowotny B., "Effects of Adverse Weather on Traffic and Safety: State-of-the-Art and a European Initiative," *in Proceedings of the 15th International Road Weather Conference*, Quebec City, pp. 1-7, 2010. https://sirwec.org/wp-content/uploads/2022/04/Quebec-D-45.pdf

[9] Ferster C. and Skinner B., *Schedules of Reinforcement*, Prentice-Hall, 1957. DOI:10.1037/10627-000

[10] Formosa N., Quddus M., Ison S., and Timmis A., "A New Modeling Approach for Predicting Vehicle-Based Safety Threats," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18175-18185, 2022. DOI:10.1109/TITS.2022.3156763

[11] Gao H., Qu T., Hu Y., and Chen H., "Personalized Driver Car-Following Model-Considering Human's Limited Perception Ability and Risk Assessment Characteristics," *in Proceedings of the 6th CAA International Conference on Vehicular Control and Intelligence*, Beijing, pp. 1-6, 2022. DOI:10.1109/CVCI56766.2022.9965180

[12] Hjelkrem O. and Ryeng E., "Driver Behaviour Data Linked with Vehicle, Weather, Road Surface, and Daylight Data," *Data Brief*, vol. 10, pp. 511-514, 2017. DOI: 10.1016/j.dib.2016.12.036

[13] Ho T., "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832-844, 1998. DOI:10.1109/34.709601

[14] Kilimci Z. and Omurca S., "Enhancement of the Heuristic Optimization Based on Extended Space Forests Using Classifier Ensembles," *The International Arab Journal of Information Technology*, vol. 17, no. 2, pp. 188-195, 2020, https://doi.org/10.34028/iajit/17/2/6

[15] Kumar M. and Anwar S., "Deep Learning Model for UAV aided Traffic Analysis and Vehicle Classification," *in Proceedings of the 5th International Conference on Data Intelligence and Cognitive Informatics*, Tirunelveli, pp. 1219-1224, 2024. DOI:10.1109/ICDICI62993.2024.10810974

[16] Laarej A., Sansalvador J., Lakouari N., and Ezzahraouy H., "Study of Energy Dissipation and Satisfaction Rates in Mixed Traffic Flow with Lights: A Two-Lane Cellular Automaton Approach," *in Proceedings of E3S Web of Conferences*, Les Ulis, pp. 1-12, 2024. DOI:10.1051/e3sconf/202458204001

[17] Liu Z., Lyu C., Wang Z., Wang S., Liu P., and Meng Q., "A Gaussian-Process-Based Data-Driven Traffic Flow Model and its Application in Road Capacity Analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1544-1563, 2023. DOI:10.1109/TITS.2022.3223982

[18] Lopukhova E., Abdulnagimov A., Voronkov G., and Grakhova E., "Machine Learning-Driven Calibration of Traffic Models Based on a Real-Time Video Analysis," *Applied Sciences*, vol. 14, no. 11, pp. 1-22, 2024. DOI: 10.20944/preprints202404.1799.v1

[19] Maraia L., "Data-Driven Traffic Management," *Traffic Technology International*, vol. 2023, no. 3, pp. 45, 2023. DOI:10.12968/S1356-9252(24)40014-2

[20] Meese C., Chen H., Li W., Lee D., and et al., "Adaptive Traffic Prediction at the ITS Edge with Online Models and Blockchain-Based Federated Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 10725-10740, 2024. DOI:10.1109/TITS.2024.3391053

[21] Mitsakis E., Stamos I., Papanikolaou A., Aifadopoulou G., and Kontoes H., "Assessment of Extreme Weather Events on Transport Networks: Case Study of the 2007 Wildfires in Peloponnesus," *Natural Hazards*, vol. 72, no. 1, pp. 87-107, 2013. DOI:10.1007/s11069-013-0896-3

[22] Moumen I., Abouchabaka J., and Rafalia N., "Adaptive Traffic Lights Based on Traffic Flow Prediction Using Machine Learning Models," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5813-5823, 2023. DOI:10.11591/ijece.v13i5.pp5813-5823

[23] Mwalimo D., Wainaina M., and Kaluki W., "Mixed Vehicular Traffic Flow Model on an Inclined Multilane Road," *International Journal of Innovative Science and Research Technology*, vol. 5, no. 7, pp. 331-342, 2020. DOI:10.38124/IJISRT20JUL276

[24] Nan H., Li R., Zhu X., Ma J., and Niyato D., "An Efficient Data-Driven Traffic Prediction Framework for Network Digital Twin," *IEEE Network*, vol. 38, no. 1, pp. 22-29, 2024. DOI:10.1109/MNET.2023.3335952

[25] Pi Y., Duffield N., Behzadan A., and Lomax T., "Lane-Specific Speed Analysis in Urban Work Zones with Computer Vision," *Traffic Injury Prevention*, vol. 24, no. 3, pp. 242-250, 2023. DOI:10.1080/15389588.2023.217352

[26] Po L., Rollo F., Bachechi C., and Corni A., "From Sensors Data to Urban Traffic Flow Analysis," *in Proceedings of the IEEE International Smart Cities Conference*, Casablanca, pp. 478-85, 2019. DOI:10.1109/ISC246665.2019.9071639

[27] Roy R. and Saha P., "Analysis of Vehicle-Type-Specific Headways on Two-Lane Roads with Mixed Traffic," *Transport*, vol. 35, no. 6, pp. 588-604, 2021. DOI:10.3846/transport.2020.14136

[28] Sathiaraj D., Punkasem T., Wang F., and Seedah D., "Data-Driven Analysis on the Effects of Extreme Weather Elements on Traffic Volume in Atlanta, GA, USA," *Computers Environment and Urban Systems*, vol. 72, pp. 212-220, 2018. DOI: 10.1016/j.compenvurbsys.2018.06.012

[29] Taheri A., RahimiZadeh K., Beheshti A., Baumbach J., and et al., "Partial Reinforcement Optimizer: an Evolutionary Optimization Algorithm," *Expert Systems with Applications*, vol. 238, no. Part F, pp. 1-20, 2024, DOI:10.1016/j.eswa.2023.122070

[30] Trojovsky P. and Dehghani M., "A New Bio-Inspired Metaheuristic Algorithm for Solving Optimization Problems Based on Walruses Behavior," *Scientific Reports*, vol. 13, no. 8770, 2023. https://doi.org/10.1038/s41598-023-35863-5

[31] Tselentis D. and Papadimitriou E., "Driver Profile and Driving Pattern Recognition for Road Safety Assessment: Main Challenges and Future Directions," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, no. 1, pp. 83-100, 2023. DOI:10.1109/OJITS.2023.3237177

[32] Vapnik V., "The Nature of Statistical Learning Theory," *Springer Science and Business Media*, 2013. DOI: 10.1007/978-1-4757-3264-1

[33] Vasudevan P. and Ekambaram C., "HYAQP: A Hybrid Meta-Heuristic Optimization Model for Air Quality Prediction Using Unsupervised Machine Learning Paradigms," *The International Arab Journal of Information Technology*, vol. 21, no. 5, pp. 953-966, 2024. https://doi.org/10.34028/iajit/21/5/15

[34] Wei T., Zhu T., Bai H., Zhao L., and Wang X., "Effects of Driver Gender, Driving Experience, and Visibility on Car-Following Behavior," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2679, no. 1, pp. 2166-2182, 2024. DOI:10.1177/03611981241258988

[35] Yan J., Liu J., and Tseng F., "An Evaluation System Based on the Self-Organizing System Framework of Smart Cities: A Case Study of Smart Transportation Systems in China," *Technological Forecasting and Social Change*, vol. 153, no. 1, pp. 119371, 2020. DOI: 10.1016/j.techfore.2018.07.009

[36] Zhang X., Zhao Z., and Li J., "ARDE-N-BEATS: An Evolutionary Deep Learning Framework for Urban Traffic Flow Prediction," *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 2391-2403, 2023. DOI:10.1109/JIOT.2022.3212056

Haiyun Sun received her Master's degree from Dalian Jiaotong University. Now, she is working at Dalian University of Science and Technology. Her research interests include Traffic Management and Safety.

Bian Ke received the Master's degree from Dalian Jiaotong University. Now, she is working at Dalian University of Science and Technology. Her research interests include Traffic Management and Safety.

Junlin Sha received the Master's degree in transportation planning and management from Jilin University. Now, she is working at Dalian University of Science and Technology. Her research interests include Urban Public Transport Operation and Management.

Shidi Wu was born in Jinzhou City, Liaoning Province in 1984. He received a master's degree in transportation planning and management from Jilin University. Now, he is working at Dalian University of Science and Technology. His research interests include Traffic and Transportation Engineering.

Menglu Chen received the Master's degree from Dalian University of Technology. Now, she is working at Dalian University of Science and Technology. Her research interests include Urban Public Transport Operation and Management.

Appendix A.1.

Table A1. Statistical properties of the variables related to lane 1's dataset.

Variables	Indicators											
	Min	**Mean**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**Max**	**St. dev**
Vehicle length	126	865.4	502	525	542	556	571	595	737	1672	1751.5	2246	530.408
Vehicle speed	14	77.273	64	70	73	76	78	80	82	85	89	117	10.141
Vehicle weight	136	5139	1064.5	1392	1574.5	1719	1889	2107	2601	8698	17110	36877	6874.16
Number of axles	2	2.864	2	2	2	2	2	2	2	5	6	7	1.477
Lead vehicle speed	14	77.359	64.5	70	73	76	78	80	82	85	89	117	10.080
Lead vehicle weight	72	5123	1070	1393	1575	1719	1888	2104	2576	8679	17086	36877	6868
Lead vehicle length	108	861.8	503	525	541	555	570	594	722.5	1667	1746	2264	528.8
Air temperature	-3.4	1.823	-0.7	0.3	0.8	1.2	1.5	1.9	2.4	2.9	4.4	11.4	2.414
Relative humidity	51	86.017	71	77	82	87	90	92	93	94	96	97	10.138
Precipitation type	0	0.492	0	0	0	0	0	1	1	1	1	1	0.500
Precipitation intensity	0	1.037	1	1	1	1	1	1	1	1	1	2	0.206
Wind direction	6	150.70	17	34	51	96	152	169	191	264	343	360	113.31
Wind speed	0	3.064	0	0	0	1	2.2	3	4.3	6	8.4	14	3.297
Road surface status	0	2.035	0	1	1	2	3	3	3	3	3	3	1.166
Time of day	0	0.898	0	0	1	1	1	1	1	1	2	2	0.565
Day of month	1	13.131	1	2	3	5	17	17	18	25	30	31	10.462
Hour	0	13.513	5	8	11	13	15	16	17	18	20	23	5.913
Time gap	0.130	1.13E+08	2.14E+05	1.74E+06	3.31E+06	7.84E+06	2.21E+07	4.56E+07	9.36E+07	1.66E+08	3.08E+08	6.21E+09	2.71E+08

Table A.2. Statistical properties of the variables related to the lane 2's dataset.

Variables	Indicators											
	Min	**Mean**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**Max**	**St. dev**
Vehicle length	1	13.506	1	2	3	4	12	18	19	28	30	31	11.303
Vehicle speed	0	12.785	7	9	10	11	13	15	16	17	19	23	5.18
Vehicle weight	109	786.51	473	499	515	531	547	572	634.7	1068.6	1710	2099	485.40
Number of axles	19	80.675	68	73	77	79	81	83	86	88	92	128	10.59
Lead vehicle speed	71	4279.3	1022.4	1307	1464.2	1602	1741.5	1903.6	2253.9	5255	13259.9	36129	6025.7
Lead vehicle weight	2	2.680	2	2	2	2	2	2	2	3	5	7	1.31
Lead vehicle length	19	80.713	68	74	77	79	81	83	86	88	92	128	10.51
Air temperature	71	4285.6	1014.1	1312	1467	1599.8	1738.5	1903.6	2253.9	5263.6	13238.3	36129	6041.8
Relative humidity	109	788.11	474	499	515	531	548	572	627.9	1089.6	1711	2179	487.61
Precipitation type	-3.4	1.917	-1.4	0.3	0.7	1.3	1.9	2.3	2.8	3.2	4.7	11.4	2.48
Precipitation intensity	51	85.633	70	76	82	86	89	91	93	94	96	97	10.03
Wind direction	0	0.549	0	0	0	0	1	1	1	1	1	1	0.50
Wind speed	0	1.029	1	1	1	1	1	1	1	1	1	2	0.19
Road surface status	6	153.22	17	28	40	62	146	180	218.5	304	349	360	123.05
Time of day	0	3.362	0	0	0.9	1.9	2.7	3.4	4.4	6.3	8.5	14	3.20
Day of month	0	1.772	0	0	1	2	2	3	3	3	3	3	1.25
Hour	0	0.731	0	0	0	1	1	1	1	1	2	2	0.65
Time gap	-0.8	1.26E+08	2.66E+05	2.28E+06	6.47E+06	1.82E+07	3.64E+07	6.46E+07	1.01E+08	1.64E+08	2.87E+08	6.74E+09	3.12E+08