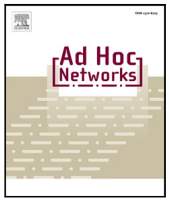
Ad Hoc Networks 161 (2024) 103508

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc



Survey paper

Towards optimal tuned machine learning techniques based vehicular traffic prediction for real roads scenarios

Raneem Qaddoura [a], Maram Bani Younes [b,c,*], Azzedine Boukerche [c]

[a] School of Computing and Informatics Al Hussein Technical University, Amman, Jordan
[b] Information Technology, Philadelphia University, Amman, Jordan
[c] PARADISE Research Laboratory, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Keywords:
Machine learning
Traffic characteristics
Traffic prediction
SUMO
KNN
Grid search
Evolutionary algorithms
Swarm intelligence
Optimization algorithms
Metaheuristic algorithms

ABSTRACT

Smart cities have been widely investigated, and several algorithms, techniques, and protocols have recently been developed to serve smart city environments. Intelligent traffic management is one of the main research fields in this area. It aims to control and monitor the traffic over downtown and highway scenarios. The real-time traffic characteristics on road networks highly affect safety conditions and driving behaviors. Detecting hazardous traffic conditions and congested areas early helps drivers make the best decision to avoid dangerous situations. After extensive training, machine learning algorithms are mainly used for prediction in this field. In the intelligent traffic control target, trained machines are beneficial in obtaining instant and accurate predictions regarding traffic congestion levels and their distributions. Several parameters control the efficiency of the selected machines. This includes the dataset quality, the regression algorithm's parameters, and how they have been tuned. In this work, we first create high-quality datasets representing the moving vehicles for different main roads in various countries, including Bank St in Canada, North Rodeo Drive (NRD) in Los Angeles, USA, Queen Rania Al Abdullah (QRA) road in Jordan, and Sheikh Zayed Road (SZR) in Saudi Arabia. Then, we test the most popular regression algorithms and evaluate their performance in terms of temporal prediction of traffic characteristics and traffic congestion in different geographical road scenarios. We optimized these regression algorithms and tuned the used parameters using the grid search approach. After that, more advanced metaheuristic optimization algorithms were used to improve the prediction accuracy of the selected machine. Experimentally, we have proved the enhancements in predicting traffic characteristics for the tested road scenarios. Both approaches show enhanced results compared to using the regression algorithm without tuning the parameters. It shows advanced results for both approaches for the coefficient of determination measure, named R squared (R^2), with approximately 0.07, 0.04, and 0.08 enhanced values for Bank St., NRD, and SZR roads, respectively. In contrast, it shows approximately 0.95, 0.49, and 3.62 enhanced values for Bank St., NRD, and SZR roads, respectively, for the Root Mean Squared Error ($RMSE$) measure.

1. Introduction

Intelligent traffic management systems are essential to developing and implementing smart cities [1]. The real-time traffic distribution and existing traffic congestion over the road networks highly affect the decisions of these systems and drivers during their trips. Early prediction of traffic congestion or accidents on the road network gives time to the intelligent control system and/or drivers to make the most suitable decisions. Machine learning algorithms are widely used for general predictions. The datasets are mainly required to train these machines, and they affect their predictions' accuracy and quality. Moreover, different training algorithms could be used for these machines. Based on the investigated research problem's characteristics and the datasets' main features, they have different behaviors and abilities to predict.

Real-time traffic on the road network is one of the most popular fields of research that have been widely investigated recently. These studies aim to enhance the safety and efficiency conditions over the road networks [2–5]. Traffic density, speed, estimated travel times, existing context, configured platoons, and other traffic parameters have been investigated on the road networks. This estimates the traffic characteristics and their special context there [2,4]. These predictions and estimations are essential to developing safe and efficient applications on the road network. The latter applications aim mainly to assist fresh drivers and control autonomous vehicles in making early decisions

* Corresponding author at: Information Technology, Philadelphia University, Amman, Jordan.
E-mail addresses: raneem.qaddoura@htu.edu.jo (R. Qaddoura), mbani047@uottawa.ca (M.B. Younes), boukerch@uottawa.ca (A. Boukerche).

https://doi.org/10.1016/j.adhoc.2024.103508
Received 3 March 2024; Received in revised form 31 March 2024; Accepted 11 April 2024
Available online 16 April 2024
1570-8705/© 2024 Elsevier B.V. All rights reserved.

regarding their trips. The most optimal path on the downtown [6] and the safe speed at road intersections [7] or exit/entrance points on highways [8,9] are examples of the most critical decisions. These decisions can be recommended for traveling vehicles based on the real-time traffic characteristics in the area of interest.

Different advanced techniques and algorithms are used to evaluate the traffic metrics and parameters on the road network. This includes image processing mechanisms [10,11], vehicular ad-hoc network communication (VANET) based applications [7,9], and artificial intelligence-based mechanisms [12–15]. The image processing and VANET-based techniques use images and the gathered traffic data to measure the real-time traffic parameters in the area of interest. They can only analyze the images or the gathered traffic data to assess real-time traffic characteristics. On the other hand, machine learning mechanisms can predict traffic distribution and other accurate characteristics in future periods. Predicting the future traffic distribution and acting accordingly can be very efficient over the road network in terms of reducing estimated travel time and making the most appropriate decision. However, the wrong predictions in these scenarios can lead to fatal and drastic accidents. Thus, more investigations and enhancements are required to guarantee high prediction accuracy in this field of research. This includes obtaining high-quality datasets and advanced tuning of the parameters in the algorithm used.

The tuning process can be achieved using different techniques, including random search, grid search, bayesian search, and metaheuristic optimization. The random search approach works by experimenting with random values of the hyperparameters in a defined, bounded search space. The grid search approach works by experimenting with the different combinations of the parameters' values in the search space and evaluating the quality of the results to find the best set of parameters' values. Bayesian search uses a probability model for the objective function considering past performances. Metaheuristic algorithms have been used to optimize problems through high-level procedures. These algorithms consider generic optimizers that can be implemented for any problem [16,17]. The metaheuristic algorithms find the most suitable solutions for defined problems, which are known for simplicity and powerfulness in solving optimization problems [18–20]. This is by encoding all candidate solutions and searching for the best solution independently of the problem type [21]. The metaheuristics mechanisms can be population-based, single-point, nature-inspired, or non-nature-inspired. Moreover, metaheuristic algorithms can be used for tuning machine learning solutions.

The main challenges affecting the machine learning-based approaches to predict the traffic characteristics over road networks are summarized in two points. The availability of high-quality datasets that can be used to train machine learning algorithms correctly. Besides properly tuning and setting the parameters of these algorithms to enhance their efficiency [13,22]. In this work, first, we use the popular urban mobility simulator, Simulation of Urban MObility (SUMO) [23,24] to generate the traffic mobilities over real road scenarios in Jordan, Saudi Arabia, Canada, and the United States of America (USA). Large datasets were obtained from these traces to test the solutions of the considered machine-learning algorithms. Two approaches are considered in this study, including the grid search optimization approach [25] and metaheuristic optimization approach [21,26], to tune the parameters of the considered algorithms. Moreover, these considered algorithms' performance and the best algorithm's optimized version have also been experimentally tested. The contributions of this work can be summarized in the following points:

- Generating high-quality datasets that reflect the traffic characteristics of vehicles for real road scenarios located in different geographical areas (i.e., Jordan, Saudi Arabia, Canada, and the USA).
- Using different optimization mechanisms to tune the parameters of the regression algorithms to enhance their predictions and performance quality.
- Comparing the performance of the selected regression algorithms and the optimized versions experimentally.

The novelty of this work can be seen in the use of metaheuristic algorithms to tune the parameters of the regression algorithm on custom-generated datasets simulating the traffic of selected roads in Jordan, Saudi Arabia, Canada, and the USA.

The rest of this paper is organized as follows: Section 2 illustrates previous traffic prediction techniques, declaring their main functionalities and limitations. Section 3 presents different tuned machine learning approaches developed to predict traffic characteristics over road networks. Section 4 illustrates the generated datasets for real roads on various geographical areas using SUMO. It also discusses the tested experiments and the results they obtained to compare the tuned algorithms and the original regression algorithms. Finally, Section 5 summarizes the entire paper.

2. Related work

Predicting traffic distribution and detecting congestion on road networks have been extensively investigated in the literature [2,4,5]. This is essential to help provide real-time applications and efficient services for drivers and autonomous vehicles that rely on real-time traffic characteristics [6,7]. Table 1 summarizes the main technologies for evaluating traffic, detecting congestion, and predicting traffic distributions over road networks. Early researchers have installed camcorders in the area of interest and obtained videos that help to monitor the traffic distribution and mobility there for long periods of time [10,11]. Then, advanced algorithms have been used to process, analyze, and extract the specific characteristics from the gathered images [10,22]. Traffic density, context, and violations are the main parameters to be extracted and predicted from these images. These mechanisms require a long processing time. Moreover, unclear videos, lousy weather conditions, and bad camera quality lead to low prediction accuracy. On the other hand, the detected characteristics of the investigated area of interest are not real-time due to the high processing time delay [10,11].

Furthermore, researchers and developers have installed sensors and inductive loop detectors on the road networks to investigate the context and characteristics of the available traffic there [2]. The setup and maintenance processes of these pieces of equipment are highly expensive. Moreover, there is low accuracy in predicting the traffic parameters due to external possible interference. This is like other objects on the road networks or internal errors in the sensing tools. The technology of VANETs has also been used to evaluate and estimate the real-time traffic parameters on the road network. A periodic message among traveling vehicles that summarizes the main attributes of each vehicle is used [4,5]. These messages are collected in each area of interest using the connecting vehicles. Then, the gathered data are statistically analyzed to evaluate and/or predict the traffic parameters on the road network. Vehicles should have a wireless transceiver to communicate and participate in these protocols. Global Positioning System (GPS) receivers are also needed in traveling vehicles to determine their relative locations to other existing vehicles and objects and digital maps to trace their route [2,3].

Machine learning and artificial intelligence mechanisms have been widely used to predict the traffic characteristics on road networks [12,13,27]. Grid search approach [28] and other metaheuristic approaches [29] can be used to tune and optimize the parameters of the machine learning algorithms. Evolutionary algorithms (EA) and swarm intelligence (SI) algorithms are two metaheuristic algorithms for solving optimization problems with nature-inspired behavior. EA includes well-regarded optimization algorithms like Genetic algorithm (GA) [30], Differential Evolution (DE) [31,32] and Evolution Strategy (ES) [33]. In contrast, SI includes Particle Swarm Optimization (PSO) [34], FireFly Algorithm (FFA) [35], and Ant Colony Optimization (ACO) [36].

Other recent metaheuristic algorithms are also proposed, such as Grey Wolf Optimizer (GWO), Multi-Verse Optimizer (MVO), Whale

Table 1
Technologies of predicting and detecting traffic congestion on the road networks.

Technology	Main objectives	Requirements	Advantages	Limitations
Image processing [10,11]	Evaluate traffic and detect congestion.	Camcorders and advanced processing algorithms.	Evaluate and detect.	It requires long processing time (Not real-time detection) and bad accuracy for bad videos.
Sensors [2]	Evaluate traffic and detect congestion.	Sensors and inductive loop detectors.	Fast and accurate evaluation.	Setup and maintenance processes are costly.
VANETs [4,5]	Evaluate traffic, detect congestion, and make short-term predictions.	Wireless transceivers are in each vehicle, as are RSUs, GPS systems, and digital maps.	Accurate and real-time traffic evaluations.	Very short-term prediction of the traffic distribution with low accuracy.
Machine learning [12,13,27]	Evaluate traffic, detect congestion, and predict traffic distribution.	Good dataset and tuned algorithm.	Long-term prediction for traffic distribution	The availability of datasets and tuning of the training algorithm.

Optimization Algorithm (WOA), and Harris Hawks Optimization (HHO) [37,38]. GWO, MVO, WOA, and HHO are advanced metaheuristic algorithms proposed and co-authored by Mirjalili et al. [39]. The GWO algorithm [39] imitates grey wolves' hunting mechanism and leadership hierarchy. It has tested against different metaheuristic algorithms such as PSO, DE, ES, Evolutionary Programming, and Gravitational Search algorithms (GSA), and the results of GWO were competitive and promising. MVO [40] is based on the multi-verse concept and mainly the three concepts: the white hole, black hole, and wormhole. MVO has also experimented against metaheuristic algorithms such as PSO, GWO, GA, and GSA. Results showed that MVO outperformed other algorithms in some scenarios. WOA [41] imitates humpback whales' social behavior, it was tested against state-of-art metaheuristic algorithms, and results showed that it is a competitive algorithm that can produce better solutions. Moreover, HHO imitates the cooperative behavior and chasing styles of Harris' hawks in nature [42]. It has experimented with metaheuristic algorithms, including GA, PSO, GWO, and others, on 29 benchmark problems and several real-world engineering problems. It showed promising results compared to the other metaheuristic algorithms.

According to [43], real-world datasets are gathered for governmental purposes, data science competitions, or academic research. The datasets are either public, semi-public, or private. The paper also shows that different prediction workflows are observed in different studies, including statistical, machine learning, and deep learning models. Machine learning and deep learning models are based on random forest (RF), LightGBM, Gaussian progress regression (GPR), multiple linear regression (MLR), feed-forward neural networks (FFNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

The general regression machine learning mechanisms have been used previously to predict the traffic parameters in real road scenarios [12,27]. Here, we extend the previous studies by enhancing the quality of the prediction results. This work mainly focuses on tuning the parameters of the regression algorithm using grid search and metaheuristic approaches. Then, the performance of the modified versions is evaluated experimentally. In addition, this work includes generating high-quality datasets for real road scenarios for moving vehicles located in Jordan, Saudi Arabia, Canada, and the USA.

The main focus of this study differs from the previous works because predicting the position and speed of the vehicles using metaheuristic algorithms for tuning the parameters of the regression algorithm was not found in the previous studies, as far as we know. In addition, the simulated roads and traffic generated new datasets that can be used for further studies. The datasets can be found in the following link: https://github.com/RaneemQaddoura/Traffic-Real-Road

3. Tuned machine learning approaches

In this section, we delve into the two approaches in this study: the grid search approach and the metaheuristic optimization approach. The primary objective of both methodologies is to construct a model capable of taking the characteristics of a vehicle at various timestamps as input and, in turn, predicting the future values of position, speed, x-coordinate, and y-coordinate. This dynamic interplay is visually represented in Fig. 1. Notably, the predictive outcomes for speed, position, x-coordinate, and y-coordinate pertain to the subsequent 1–5 min, while the input values correspond to the specific timestamp under consideration.

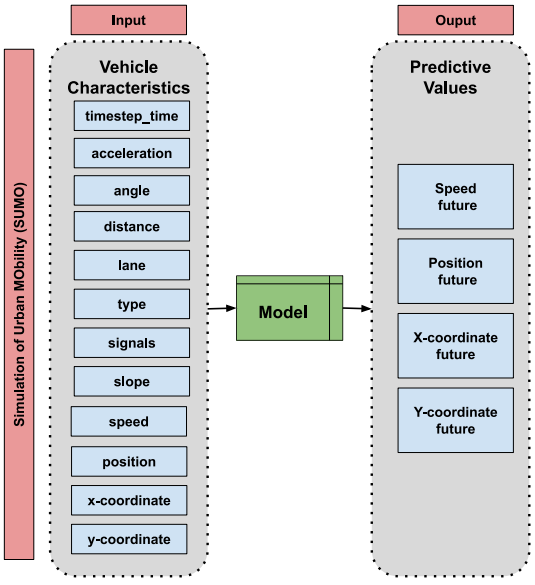


Fig. 1. Input and output values of the ML model.

3.1. Grid search approach

First, the grid search approach is used for accurately predicting the traffic parameters of real road scenarios, as shown in Fig. 2. The process includes gathering the basic travel vehicle data on real road scenarios. Besides, it consists of the machine learning technique with parameter tuning of the K-Nearest Neighbor (KNN) regression algorithm [44].

The KNN regression algorithm is selected because it produces better results than other regression algorithms for this type of problem, such as the Linear Regression (LR) and the Decision Tree (DT) algorithm, as early discussed [12,27]. The learning and prediction processes are performed to train and test the generated datasets. This is mainly for the coming 1, 2, 3, 4, and 5 min. The prediction features include each dataset's position, speed, x-coordinate, and y-coordinate labels. The grid search approach, which we refer to by KNN_G, is used to tune the parameters of the KNN algorithm. This mainly considers the 'n_neighbors', 'weights', and 'metric' parameters.

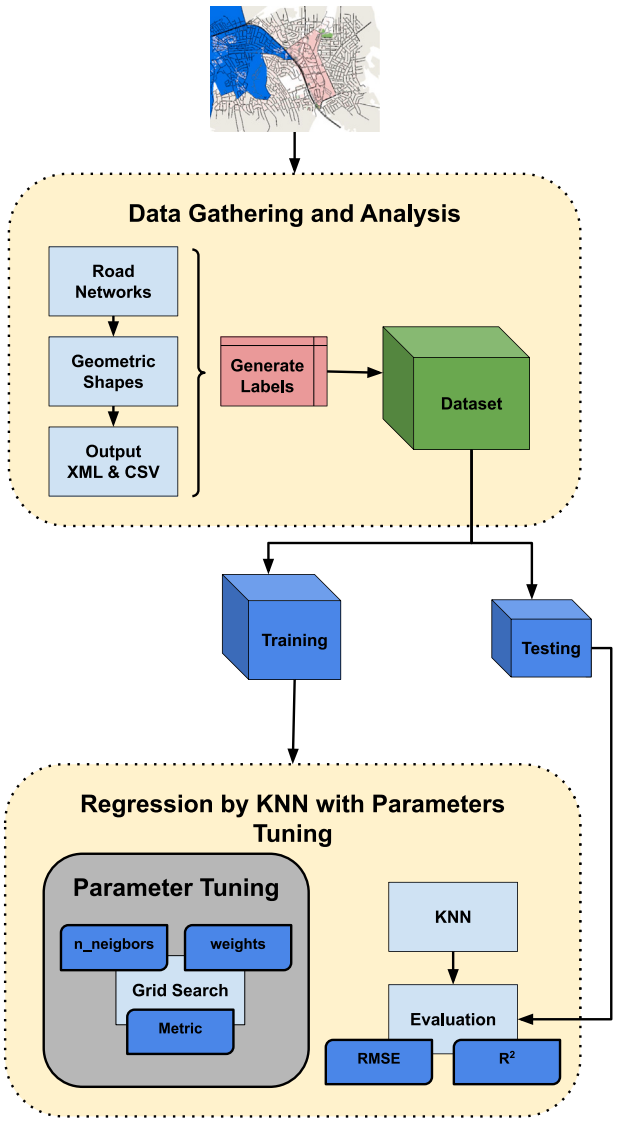


Fig. 2. Tuned traffic prediction machine learning with grid search approach.

After testing the different combinations of these parameters, the set that archives the best performance is selected. It is then considered for the learning process to generate the regression model from the training datasets. The prediction process requires applying the model to the testing part of the dataset. The accuracy of the prediction results is evaluated using the Root Mean Squared Error ($RMSE$) and R Squared (R^2) measures. Finally, the results are compared with other general regression algorithms (i.e., LR, DT, and KNN).

After that, based on the predicted parameters of each traveling vehicle during the next 1, 2, 3, 4, and 5 s, the number of vehicles in each predefined area for future periods can be calculated. Thus, this helps to predict the general traffic state of that area by showing the number of vehicles and reflecting the level of traffic congestion for that area.

3.2. Metaheuristic optimization

The metaheuristic optimization approach applied in this study is presented in Fig. 3. The map is processed in the data gathering and analysis step by generating the road networks, geometric shapes, and

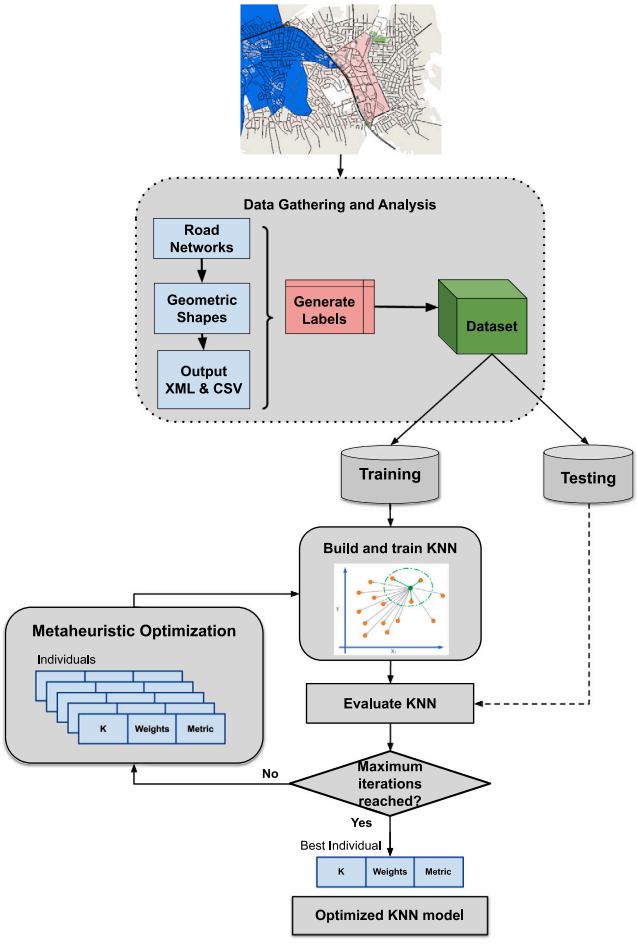


Fig. 3. Tuned machine learning-based traffic prediction with metaheuristic optimization.

output XML & CSV. Then, the labels are generated, indicating the speed, position, x-coordinate, and y-coordinate for the coming 1 min. The generated dataset is split into training and testing portions. The first KNN regression model has random values for the K, weights, and metric parameters. It is evaluated using two fitness functions, $F1$ and $F2$, based on the R^2 and $RMSE$ measures, respectively. The two fitness functions are multi-objective functions that consider the four labels, position, speed, x-coordinate, and y-coordinate, which are observed by Eqs. (1) and (2). The objective of the two fitness functions is to minimize their values, and thus, the complement of the R^2 measure ($R^{2'}$) is considered in Eq. (1).

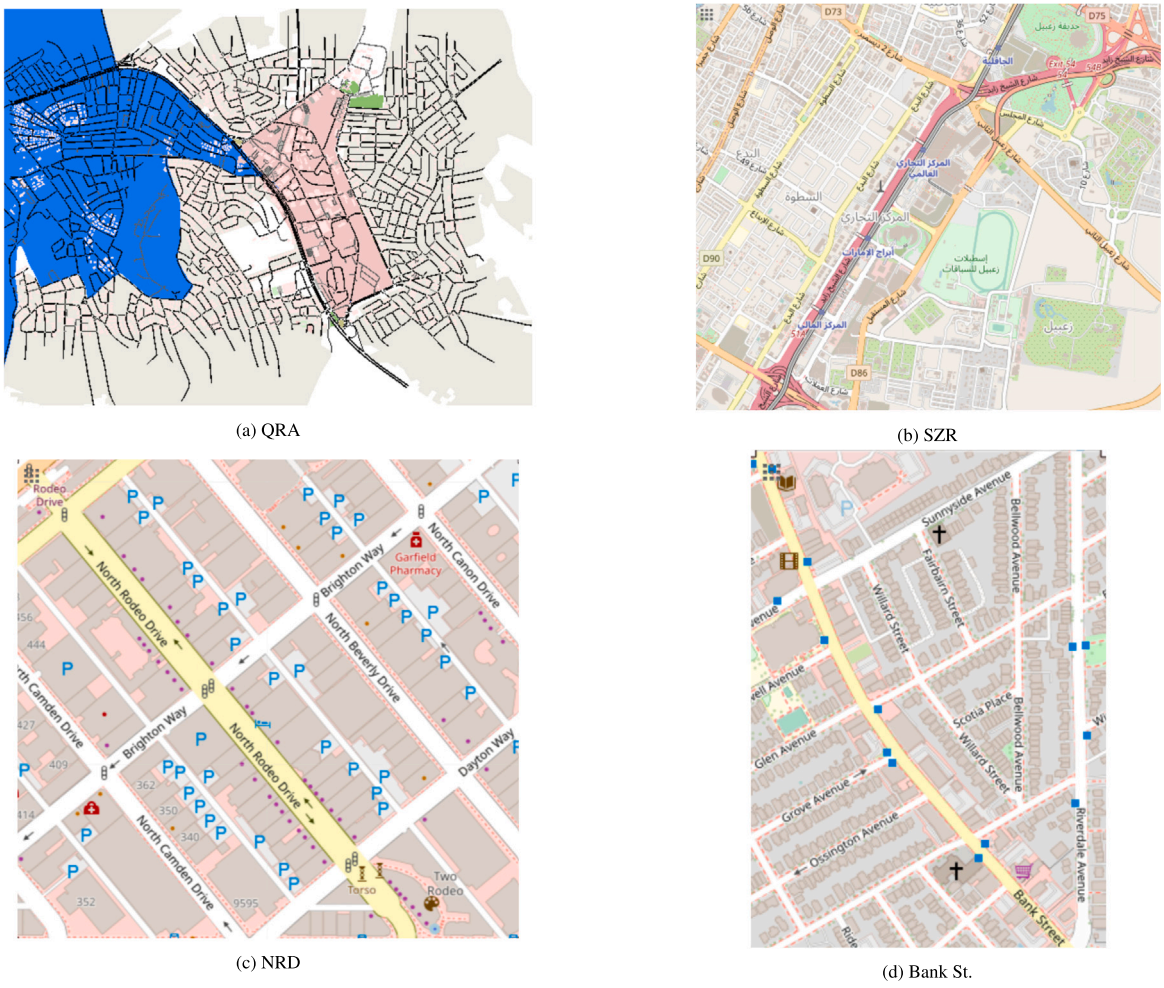
$$F1 = (R^{2'}_{Speed} + R^{2'}_{pos} + R^{2'}_{x} + R^{2'}_{y})/4 \quad (1)$$

$$F2 = (RMSE_{Speed} + RMSE_{pos} + RMSE_{x} + RMSE_{y})/4 \quad (2)$$

Afterward, each iteration's metaheuristic optimization process updates the individual containing the parameters' values. The KNN regression model is rebuilt using the training portion and evaluated with the fitness function using the testing portion. The best individual is selected at each iteration and updated using the metaheuristic algorithm until the number of iterations is reached, generating the final best individual.

4. Performance evaluation

First, This section generates datasets for real road scenarios in four countries: Jordan, Saudi Arabia, the USA, and Canada. Then, it discusses the results obtained from applying the two approaches discussed



(a) QRA

(b) SZR

(c) NRD

(d) Bank St.

Fig. 4. Road network scenarios.

in the aforementioned section. The first approach, considering the grid search approach, works with the QRA road of Jordan for 50,000 s, where the aim is to predict the position, speed, x-coordinate, and y-coordinate values of vehicles for the coming 1, 2, 3, 4, and 5 min. On the other hand, the second approach, considering metaheuristic multi-objective optimization, works with three different roads: Bank St. in Canada, NRD in the USA, and SZR in Saudi Arabia, for 900 s in the coming 1 min. Twenty iterations with twenty individuals are considered for the experiments. The experiments are repeated ten times for each metaheuristic algorithm.

4.1. Generating datasets for real roads on different geographical areas

The Simulation of Urban MObility (SUMO) package [45] has been mainly used to generate and model traffic distribution datasets. First, we selected different real road scenarios in four countries (Jordan, Saudi Arabia, USA, and Canada). The 'netconvert' application generates the selected road network scenarios. We used the SUMO GUI (i.e., the 'polyconvert' application) to determine the geometrical shapes of the selected road network scenarios. The 'randomTrips.py' script is run to generate and determine the trip of each vehicle on the road networks. After that, the output of the generated random trips is stored in XML file format. The output files are converted into CSV files using the 'xml2csv.py' script available in SUMO packages. CSV files contain the attributes of the traveling vehicles during the pre-determined trip interval. The main extracted attributes in these CSV files are the vehicle acceleration, angle, distance, lane, position, signals, slope, speed, type, and location (i.e., x and y coordinates). These attributes are considered the main features of the classification models.

Each tested road network consists of several streets with single-lane or multi-lanes. Speed limit, length, and shape data are the main attributes assigned to each street lane. Connections among located streets are the second component of the road network. Each connection links incoming and outgoing streets. The main attributes of each connection are incoming lanes, outgoing lanes, direction, and state. Nodes are the last component of road networks. The node component's main attributes are the location (i.e., x-coordinate, y-coordinate), the lanes that end at the intersection, the lanes within the intersection, and the intersection shape.

In this work, four main roads have been selected from four different countries: Queen Rania Al Abdullah (QRA) road in Jordan, Sheikh Zayed Road (SZR) in Saudi Arabia, North Rodeo Drive (NRD) in Los Angeles, USA, and Bank St in Canada with the areas surrounding them which are represented by Figs. 4(a), 4(b), 4(c) and 4(d), respectively. As discussed earlier, the datasets generated from SUMO consist of many streets, lanes, nodes, and connections. The main features of the investigated road network scenarios are summarized in Table 2. In this context, the first approach employs the grid search technique on the QRA road in Jordan for 50,000 s to predict vehicle position, speed, x-coordinate, and y-coordinate values for the next 1 to 5 min. The second approach, utilizing metaheuristic optimization, focuses on three distinct roads – Bank St. in Canada, NRD in the USA, and SZR in Saudi Arabia – spanning 900 s for predictions in the upcoming 1 min. The features and connections of the road networks are summarized in Table 2, providing a concise overview of the diverse scenarios considered in our study.

Table 2
The main features of the investigated roads.

Road	Streets	Lanes	Nodes	Conn.
QRA	19 006	19 832	4804	26 695
SZR	26 083	30 241	6345	30 023
NRD	1091	1375	397	1229
Bank St.	2868	3030	838	3341

Note that the aforementioned durations of 50,000 s and 900 s pertain to the generation of the datasets, serving as comprehensive snapshots of road conditions. Subsequently, these datasets become the basis for predicting vehicle parameters. The prediction task specifically targets the forthcoming 1, 2, 3, 4, and 5-min intervals for the QRA road dataset and the upcoming 1-min interval for the multi-road datasets. This distinction in timeframes clarifies that the extensive dataset generation periods are separate from the specific prediction times, ensuring the relevance of our experimental setup to real-world scenarios.

The features that are considered for training the machine learning models include the time of each step of the experiments, which are set to be every one second in addition to the id of the vehicle and the acceleration, angle, distance, lane, position, signals, slope, speed, x-coordinate, and y-coordinate of the vehicle.

The labels of each row in the dataset are automatically generated, including the position, speed, x-coordinate, and y-coordinate for the coming period, using the following rows. For example, for the prediction of the position, speed, x-coordinate, and y-coordinate for the coming 1 min, the four different labels are generated for each vehicle at each point in time for the coming 1 min by copying the position, speed, x-coordinate, and y-coordinate features' values from a following row with a timestamp of the coming 1 min for the same vehicle.

4.2. Grid search approach

Recalling Fig. 2, we employ the grid search approach to meticulously tune the parameters of the K-Nearest Neighbor (KNN) regression algorithm by exploring various combinations. This exhaustive exploration results in identifying the optimal parameter set for our model. The specific values considered for each parameter during the experiments include n_neighbors (2, 4, 6, and 8), weights ('Uniform' and 'Distance'), and metrics ('Euclidean', 'Manhattan', 'Chebyshev', and 'Cosine').

The results delineate the configuration for the K-Nearest Neighbor (KNN) regression algorithm, specifically favoring four n_neighbors, 'distance' weight, and 'Manhattan' metric. Comprehensive performance metrics, including R^2 and $RMSE$ values, are presented for the multi-label evaluation of KNN_G, LR, KNN, and DT in Tables 3 through 7. Moreover, the evaluation of the predicted results for different tested periods (i.e., 1, 2, 3, 4, and 5 min) are presented in Tables 3, 4, 5, 6, and 7. From these aforementioned tables, we can infer that the grid search approach has improved the results of the KNN for all the labels. Notably, for the 1-min speed prediction, the grid search yields a remarkable enhancement of approximately 33% in R^2 and 35% in $RMSE$, as highlighted in Table 3. Furthermore, KNN_G consistently outperforms the original LR and DT algorithms, showcasing improvements of 58% and 17% in R^2 for the 3-min prediction, as evidenced in Table 5.

Additionally, Figs. 5 and 6 provide visual insights into $RMSE$ and R^2 values for each label. KNN_G demonstrates superior performance across all time values, evidenced by the consistently lowest $RMSE$ values in Fig. 5. Conversely, linear regression (LD) exhibits the poorest $RMSE$ results for all labels, with deteriorating values as time progresses. This trend is particularly noticeable in the x-coordinate label. In addition, the prediction of the larger values of time produced better values of $RMSE$ for the position and speed labels. This might be justified by the unexpected traffic generated in the street, which leads to insignificant changes in the position and significant changes in the vehicle's speed for near periods because the vehicle will unexpectedly slow down. The R^2 analysis in Fig. 6 further underscores KNN_G's efficacy, exhibiting high values for improved label predictions. In contrast, linear regression consistently lags behind other algorithms, deteriorating performance as time values increase.

4.3. Metaheuristic optimization

Table 8 shows the different average values of the fitness function across the Bank St., NRD, and SZR roads. WOA, MVO, HHO, PSO, and GWO are used to optimize the parameters' values of the KNN regression algorithm. MVO, PSO, and GWO observe the best fitness value for $F1$ for the bank St. and SZR, while it is observed for WOA, MVO, and GWO for the NRD. On the other hand, PSO shows better fitness value for $F2$ for Bank St., while MVO and PSO have the best value for SZR. In the case of NRD, all the optimizers have performed identically with the lowest value of $F2$ except for HHO, which has a slightly higher value. HHO showed the worst value for both fitness functions for almost all the roads.

The comparison of the KNN regression grid search approach with other well-known regression techniques, including KNN regression, LR, and DT, is summarized in Table 9. It is well observed that the KNN regression with grid search performs better than the other algorithms for both fitness functions.

The optimization task's convergence curve shows the fitness function's values change across the different iterations. Since the metaheuristic algorithm optimizes the solution from one iteration to another, the convergence curve shows smaller values when progressing through the iterations. Fig. 7 shows the change of values for $F1$ and $F2$ for the three roads in the study. HHO shows the highest values across the iterations for the two functions except for $F2$ for Bank St., indicating the worst values compared to the other metaheuristic algorithms. WOA for $F2$ of Bank St. has the worst value compared to the different algorithms, showing slow convergence. It is also observed that the lowest values of the fitness functions are observed in early iterations and are kept still for later iterations, indicating that the metaheuristic algorithms can find the best values without the need for many iterations.

In summary, it is widely recognized that optimizing parameter tuning leads to improved results, as observed in both the Grid Search and meta-heuristic approaches. SZR road performed better using metaheuristic algorithms for both fitness functions, while the NRD road performed better using the Grid Search approach. In contrast, metaheuristic algorithms showed better results for the Bank St. road for F1 than the Grid Search approach, while it showed the opposite for F2. In addition, the experiments can be generalized to include other roads in different countries. They can also solve other problems under different circumstances, such as different levels of traffic, holidays and yearly vacations, and accidents.

4.4. Implications of the results

Traffic prediction models are vital tools in urban planning and public policy decision-making because they enable preventive actions to address traffic-related difficulties and provide insights into future traffic patterns. Creating realistic traffic scenarios is possible using urban mobility simulators, such as SUMO, which facilitates planning future transportation systems and evaluating current infrastructure. This work shows the usefulness of machine learning algorithms in predicting traffic mobilities by analyzing huge datasets collected from several geographical regions, including Saudi Arabia, Jordan, Canada, and the USA.

The application of metaheuristic and grid search optimization techniques to optimize algorithm parameters highlights the adaptability and efficiency of these strategies in improving prediction accuracy.

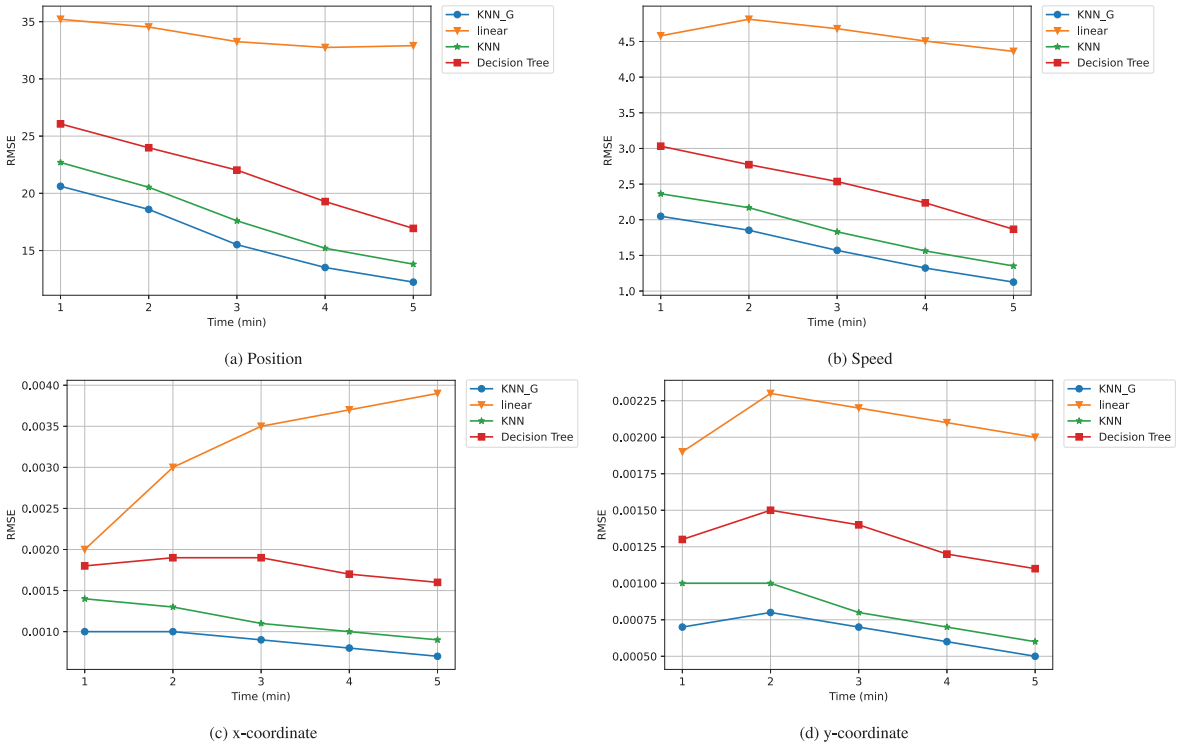


Fig. 5. RMSE values for the position, speed, x-coordinate, and y-coordinate labels using different regression techniques.

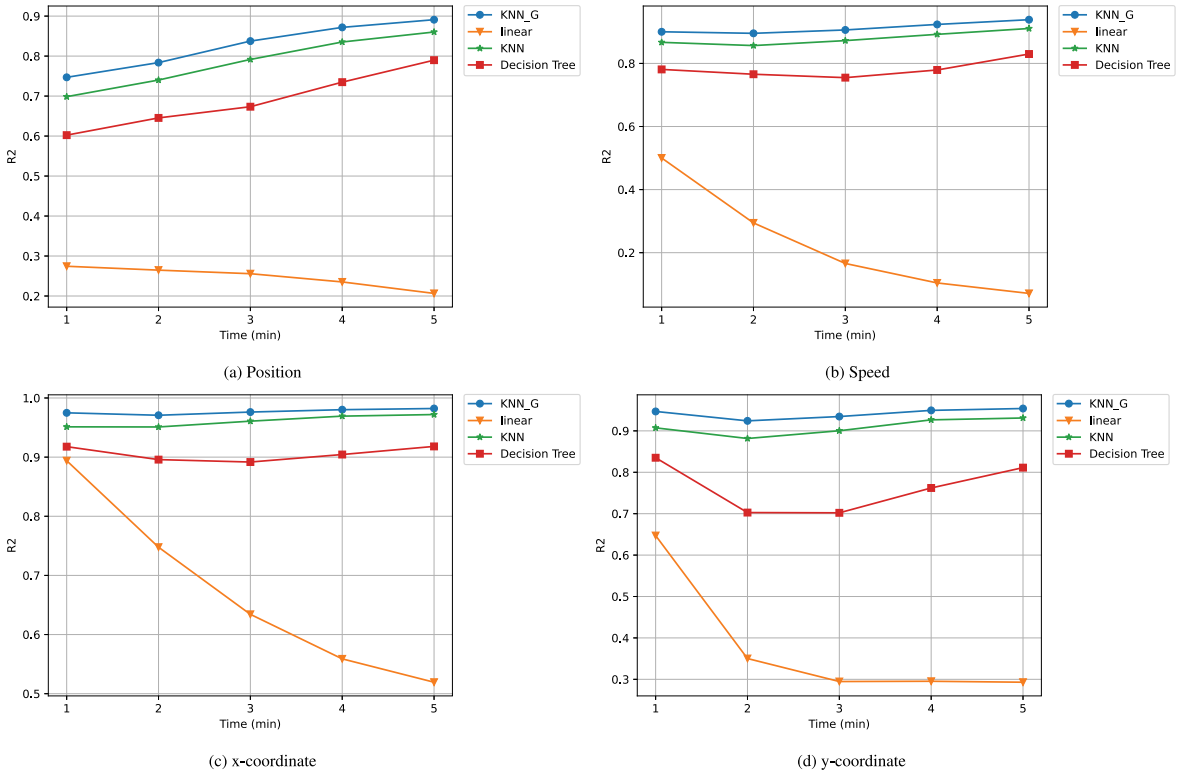


Fig. 6. R^2 values for the position, speed, x-coordinate, and y-coordinate labels using different regression techniques.

Table 3
The R^2 and $RMSE$ values for the prediction of 1 min.

Algorithm	R^2_{pos}	R^2_{speed}	R^2_x	R^2_y	$RMSE_{pos}$	$RMSE_{speed}$	$RMSE_x$	$RMSE_y$
DT	0.602329	0.78109	0.91769	0.83509	26.06679	3.03179	0.00179	0.00129
LR	0.27439	0.50029	0.89399	0.64719	35.20979	4.58009	0.00209	0.00189
KNN	0.69849	0.86679	0.95119	0.90729	22.69749	2.36439	0.00139	0.00109
KNN_G	0.74689	0.90029	0.97489	0.94689	20.6175	2.048899	0.000199	0.00069

Table 4
The R^2 and $RMSE$ values for the prediction of 2 min.

Algorithm	R^2_{pos}	R^2_{speed}	R^2_x	R^2_y	$RMSE_{pos}$	$RMSE_{speed}$	$RMSE_x$	$RMSE_y$
DT	0.64394	0.76569	0.89579	0.70269	23.98189	2.77259	0.00189	0.00149
LR	0.26469	0.29439	0.74769	0.35019	34.53219	4.81179	0.00299	0.00229
KNN	0.74009	0.85659	0.95109	0.88159	20.52919	2.16909	0.00129	0.00109
KNN_G	0.78369	0.89539	0.97079	0.92419	18.59419	1.85239	0.00109	0.00079

Table 5
The R^2 and $RMSE$ values for the prediction of 3 min.

Algorithm	R^2_{pos}	R^2_{speed}	R^2_x	R^2_y	$RMSE_{pos}$	$RMSE_{speed}$	$RMSE_x$	$RMSE_y$
DT	0.67339	0.75489	0.89169	0.70209	22.02859	2.53619	0.00189	0.00139
LR	0.25579	0.16579	0.63419	0.29469	33.25349	4.67959	0.00349	0.00219
KNN	0.79159	0.87229	0.96079	0.90029	17.59789	1.83079	0.00109	0.00079
KNN_G	0.83749	0.90609	0.97619	0.93459	15.51569	1.57069	0.00089	0.00069

Table 6
The R^2 and $RMSE$ values for the prediction of 4 min.

Algorithm	R^2_{pos}	R^2_{speed}	R^2_x	R^2_y	$RMSE_{pos}$	$RMSE_{speed}$	$RMSE_x$	$RMSE_y$
DT	0.73479	0.77919	0.90439	0.76209	19.28019	2.23709	0.00169	0.00119
LR	0.23519	0.10409	0.55889	0.29529	32.74339	4.50649	0.00369	0.00209
KNN	0.83519	0.89219	0.96929	0.92659	15.19909	1.56309	0.00099	0.00069
KNN_G	0.87179	0.92359	0.98019	0.94949	13.51819	1.32259	0.00079	0.00059

Table 7
The R^2 and $RMSE$ values for the prediction of 5 min.

Algorithm	R^2_{pos}	R^2_{speed}	R^2_x	R^2_y	$RMSE_{pos}$	$RMSE_{speed}$	$RMSE_x$	$RMSE_y$
DT	0.78979	0.82979	0.91809	0.81119	16.93429	1.86649	0.00159	0.00109
LR	0.20649	0.07089	0.51959	0.29309	32.90189	4.36139	0.00389	0.00209
KNN	0.86019	0.91069	0.97209	0.93109	13.81029	1.35179	0.00089	0.00059
KNN_G	0.89109	0.93859	0.98219	0.95409	12.24139	1.12539	0.00069	0.00049

Table 8
R^2 and $RMSE$ average values using different metaheuristic algorithms with KNN regression.

Algorithm	Bank St.		NRD		SZR	
	$F1$	$F2$	$F1$	$F2$	$F1$	$F2$
KNN-WOA	0.1447	4.9873	0.0840	2.6874	0.1847	19.8645
KNN-MVO	0.1438	4.9535	0.0840	2.6874	0.1846	19.8621
KNN-HHO	0.1456	4.9623	0.0843	2.6968	0.1857	19.8874
KNN-PSO	0.1438	4.9534	0.0842	2.6874	0.1846	19.8621
KNN-GWO	0.1438	4.9535	0.0840	2.6874	0.1846	19.8645

Table 9
$F1$ and $F2$ average values comparison.

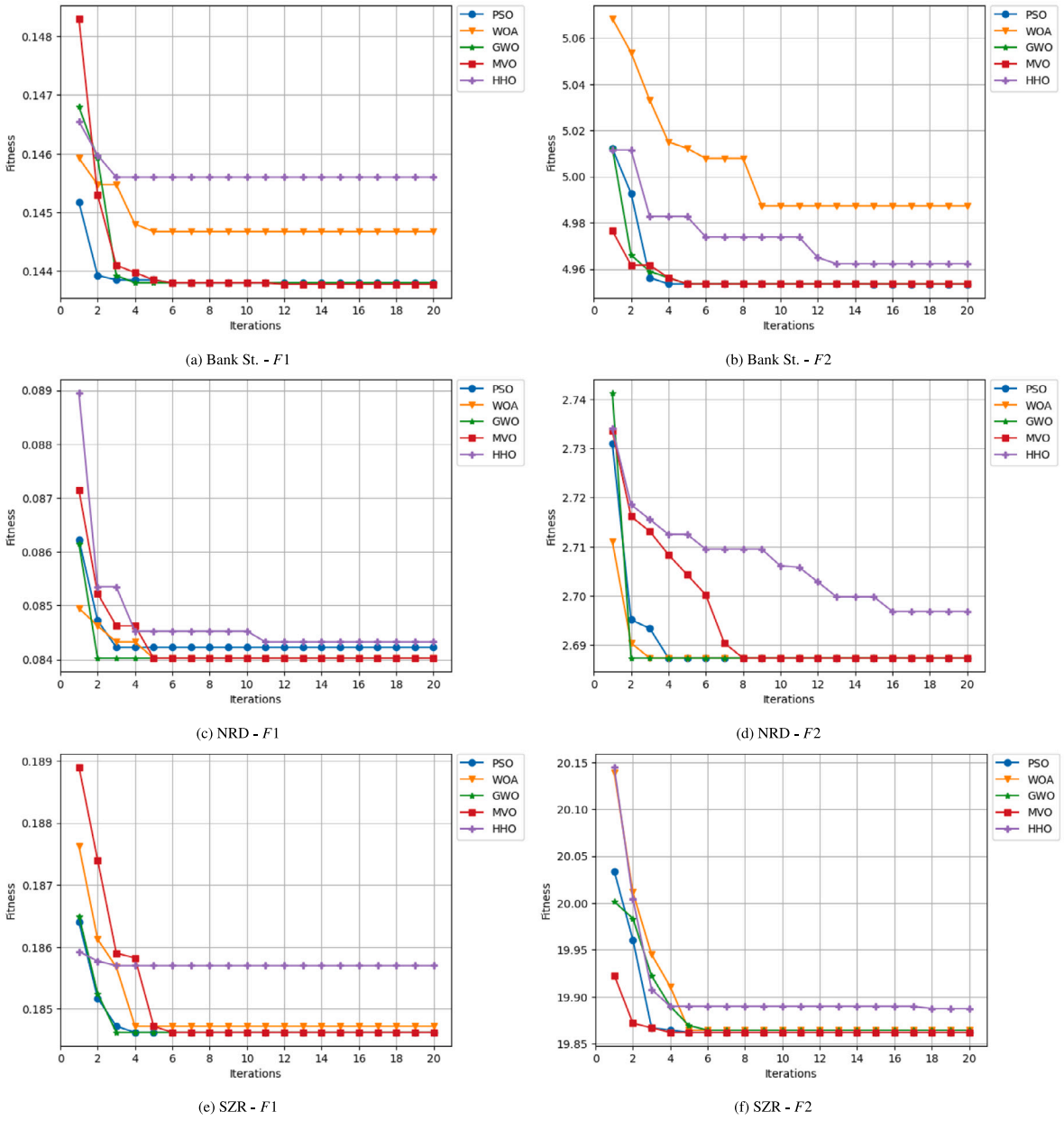
Algorithm	Bank St.		NRD		SZR	
	$F1$	$F2$	$F1$	$F2$	$F1$	$F2$
KNN_G	0.1475	4.9250	0.0828	2.6560	0.1850	20.0383
KNN	0.2144	5.8791	0.1240	3.1494	0.2617	23.4795
LR	0.6171	11.0062	0.5144	7.5359	0.5800	44.4655
DT	0.3049	6.4248	0.2092	4.0146	0.3108	24.0443

By assessing the effectiveness of diverse algorithms and determining the most optimum variant, policymakers acquire a significant understanding of the efficacy of various predictive models. These results can help guide decisions on developing urban infrastructure, managing traffic, and allocating funds for transportation projects. Stakeholders can also improve the quality of life for urban people by anticipating traffic dynamics, optimizing road networks, reducing congestion, and promoting sustainable urban growth by extrapolating study findings.

5. Conclusion and future work

Using the SUMO package, this work generated four high-quality datasets for QRA, SZR, Bank St., and NRD roads in Jordan, Saudi Arabia, Canada, and the USA. This paper also studied two approaches for predicting the position, speed, x-coordinate, and y-coordinate labels for the coming times for the vehicles of road networks. The experiments are conducted using two approaches. The first approach experimented with the QRA road using the grid search for selecting the best set of parameter values for the KNN regression algorithm to detect the speed, position, x-coordinate, and y-coordinate of vehicles for the coming 1, 2, 3, 4, and 5 min. The second approach experimented with the other roads using metaheuristic optimization of the KNN parameters for the coming 1 min. The second approach considered a multi-objective approach to minimize the summative values of $R^{2'}$ and $RMSE$ for the different predicted values presented in two different fitness functions.

The results showed that the first approach of experiments indicated that the grid search is optimizing the prediction quality in terms of R^2 and $RMSE$ for the different predicted labels and future times. On the other hand, metaheuristic optimization was another choice for selecting a good set of parameter values to increase the model's prediction quality. The convergence curves of the fitness values showed that the metaheuristic optimizer could find the best set of parameter



(a) Bank St. - $F1$ (b) Bank St. - $F2$

(c) NRD - $F1$ (d) NRD - $F2$

(e) SZR - $F1$ (f) SZR - $F2$

Fig. 7. Convergence curves for the different roads for $F1$ and $F2$ values.

values in early iterations. Both approaches show promising results for optimizing the regression model by tuning the parameters of the regression algorithm. Metaheuristic algorithms showed better results than the Grid Search approach for SZR road for F1 and F2 and Bank St. for F1. In contrast, F1 for Bank St and F1 and F2 for NRD road show the best results for the Grid Search approach.

In future work, we aim to investigate other roads with longer time durations for moving vehicles. We would also like to investigate the effect of the congestion based on the predicted values of speed, position, x-coordinate, and y-coordinate. We plan to investigate the effect of the different circumstances on the generated model's quality and study the vehicles' behavior at road intersections or exit/entrance points on highways. We will also extend our testing to more realistic traffic conditions, including peak hours, working days, and weekends. Further, we recognize the importance of validating our models in practical applications like vehicular networks and real traffic management scenarios, which will be a focal point in our future endeavors.

CRediT authorship contribution statement

Raneem Qaddoura: Writing – original draft, Validation, Software, Methodology. **Maram Bani Younes:** Writing – review & editing, Validation, Data curation, Conceptualization. **Azzedine Boukerche:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A. Khanna, R. Goyal, M. Verma, D. Joshi, Intelligent traffic management system for smart cities, in: Futuristic Trends in Network and Communication Technologies: First International Conference, FTNCT 2018, Solan, India, February 9–10, 2018, Revised Selected Papers 1, Springer, 2019, pp. 152–164.

[2] M.B. Younes, A. Boukerche, A performance evaluation of an efficient traffic congestion detection protocol (ECODE) for intelligent transportation systems, Ad Hoc Netw. 24 (2015) 317–336.

[3] R. Al Mallah, A. Quintero, B. Farooq, Distributed classification of urban congestion using VANET, IEEE Trans. Intell. Transp. Syst. 18 (9) (2017) 2435–2442.

[4] A. Boukerche, Y. Tao, P. Sun, Artificial intelligence-based vehicular traffic flow prediction methods for supporting intelligent transportation systems, Comput. Netw. 182 (2020) 107484.

[5] Y. Yang, Y. Xu, J. Han, E. Wang, W. Chen, L. Yue, Efficient traffic congestion estimation using multiple spatio-temporal properties, Neurocomputing 267 (2017) 344–353.

[6] M.B. Younes, A. Boukerche, G. Rom'an-Alonso, An intelligent path recommendation protocol (ICOD) for VANETs, Comput. Netw. 64 (2014) 225–242.

[7] M.B. Younes, A. Boukerche, Intelligent traffic light controlling algorithms using vehicular networks, IEEE Trans. Veh. Technol. 65 (8) (2015) 5887–5899.

[8] M.B. Younes, A. Boukerche, Safety and efficiency control protocol for highways using intelligent vehicular networks, Comput. Netw. 152 (2019) 1–11.

[9] V. Sadatpour, F. Zargari, M. Ghanbari, A collision aware opportunistic routing protocol for VANETs in highways, Wirel. Pers. Commun. 109 (1) (2019) 175–188.

[10] P. Jadhav, P. Kelkar, K. Patil, S. Thorat, Smart traffic control system using image processing, Int. Res. J. Eng. Technol. (IRJET) 3 (3) (2016) 1207–1211.

[11] V. Pandit, J. Doshi, D. Mehta, A. Mhatre, A. Janardhan, Smart traffic control system using image processing, Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS) 3 (1) (2014) 2278–6856.

[12] R. Qaddoura, M.B. Younes, Temporal prediction of traffic characteristics on real road scenarios in Amman, J. Ambient Intell. Humaniz. Comput. 14 (7) (2022) 1–16.

[13] A. León-Luque, C. Barajas, C. Peña-Guzmán, et al., Determination of the optimal dosage of Aluminum Sulfate in the coagulation-flocculation process using an Artificial Neural Network, Int. J. Environ. Sci. Dev. 7 (5) (2016) 346–350.

[14] Z. Wang, P. Sun, Y. Hu, A. Boukerche, A novel hybrid method for achieving accurate and timeliness vehicular traffic flow prediction in road networks, Comput. Commun. 209 (2023) 378–386.

[15] J. Wang, A. Boukerche, Non-parametric models with optimized training strategy for vehicles traffic flow prediction, Comput. Netw. 187 (2021) 107791.

[16] M. Abdel-Basset, L. Abdel-Fatah, A.K. Sangaiah, Metaheuristic algorithms: A comprehensive review, in: Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, Vol. 2018, Elsevier, 2018, pp. 185–231.

[17] A. Boukerche, J. Wang, A performance modeling and analysis of a novel vehicular traffic flow prediction system using a hybrid machine learning-based model, Ad Hoc Netw. 106 (2020) 102224.

[18] A.G. Hussien, F.A. Hashim, R. Qaddoura, L. Abualigah, A. Pop, An enhanced evaporation rate water-cycle algorithm for global optimization, Processes 10 (11) (2022) 2254.

[19] R. Zheng, A.G. Hussien, R. Qaddoura, H. Jia, L. Abualigah, S. Wang, A. Saber, A multi-strategy enhanced African vultures optimization algorithm for global optimization problems, J. Comput. Des. Eng. 10 (1) (2023) 329–356.

[20] P. Sun, N. Aljeri, A. Boukerche, Machine learning-based models for real-time traffic flow prediction in vehicular networks, IEEE Netw. 34 (3) (2020) 178–185.

[21] R. Obiedat, O. Harfoushi, R. Qaddoura, L. Al-Qaisi, A. Al-Zoubi, An evolutionary-based sentiment analysis approach for enhancing government decisions during COVID-19 pandemic: the case of jordan, Appl. Sci. 11 (19) (2021) 9080.

[22] D. Jutury, N. Kumar, A. Sachan, Y. Daultani, N. Dhakad, Adaptive neuro-fuzzy enabled multi-mode traffic light control system for urban transport network, Appl. Intell. 53 (6) (2023) 7132–7153.

[23] P.A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, E. Wießner, Microscopic traffic simulation using SUMO, in: The 21st IEEE International Conference on Intelligent Transportation Systems, IEEE, 2018, URL https://elib.dlr.de/124092/.

[24] P. Barbecho Bautista, L.F. Urquiza-Aguiar, M. Aguilar Igartua, STGT: SUMO-Based Traffic Mobility Generation Tool for Evaluation of Vehicular Networks, Association for Computing Machinery, New York, NY, USA, 2021, http://dx.doi.org/10.1145/3479240.3488523.

[25] B. Shekar, G. Dagnew, Grid search-based hyperparameter tuning and classification of microarray cancer data, in: 2019 Second International Conference on Advanced Computational and Communication Paradigms, ICACCP, IEEE, 2019, pp. 1–8.

[26] R. Obiedat, R. Qaddoura, A.-Z. Ala'M, L. Al-Qaisi, O. Harfoushi, M. Alrefai, H. Faris, Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution, IEEE Access 10 (2022) 22260–22273.

[27] R. Qaddoura, M. Bani Younes, A. Boukerche, Predicting traffic characteristics of real road scenarios in Jordan and Gulf region, in: Proceedings of the 17th ACM Symposium on QoS and Security for Wireless and Mobile Networks, 2021, pp. 115–121.

[28] G. Hafeez, K.S. Alimgeer, I. Khan, Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid, Appl. Energy 269 (2020) 114915.

[29] A. Kaveh, K.B. Hamedani, S.M. Hosseini, T. Bakhshpoori, Optimal design of planar steel frame structures utilizing meta-heuristic optimization algorithms, in: Structures, Vol. 25, Elsevier, 2020, pp. 335–346.

[30] J.H. Holland, Genetic algorithms, Sci. Am. 267 (1) (1992) 66–73.

[31] R. Storn, K. Price, Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces, J. Glob. Optim. 11 (4) (1997) 341.

[32] R. Storn, On the usage of differential evolution for function optimization, in: Proceedings of North American Fuzzy Information Processing, Ieee, 1996, pp. 519–523.

[33] H.-G. Beyer, H.-P. Schwefel, Evolution strategies–a comprehensive introduction, Nat. Comput. 1 (2002) 3–52.

[34] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4, IEEE, 1995, pp. 1942–1948.

[35] X.-S. Yang, Firefly algorithms for multimodal optimization, in: Stochastic Algorithms: Foundations and Applications: 5th International Symposium, SAGA 2009, Sapporo, Japan, October 26-28, 2009. Proceedings 5, Springer, 2009, pp. 169–178.

[36] P. Korošec, J. Šilc, A distributed ant-based algorithm for numerical optimization, in: Proceedings of the 2009 Workshop on Bio-Inspired Algorithms for Distributed Systems, 2009, pp. 37–44.

[37] R. Qaddoura, H. Faris, I. Aljarah, P.A. Castillo, Evocluster: An open-source nature-inspired optimization clustering framework in python, in: Applications of Evolutionary Computation: 23rd European Conference, EvoApplications 2020, Held As Part of EvoStar 2020, Seville, Spain, April 15–17, 2020, Proceedings 23, Springer, 2020, pp. 20–36.

[38] R. Qaddoura, H. Faris, I. Aljarah, P.A. Castillo, EvoCluster: an open-source nature-inspired optimization clustering framework, SN Comput. Sci. 2 (2021) 1–12.

[39] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey wolf optimizer, Adv. Eng. Softw. 69 (2014) 46–61.

[40] S. Mirjalili, S.M. Mirjalili, A. Hatamlou, Multi-verse optimizer: a nature-inspired algorithm for global optimization, Neural Comput. Appl. 27 (2) (2016) 495–513.

[41] S. Mirjalili, A. Lewis, The whale optimization algorithm, Adv. Eng. Softw. 95 (2016) 51–67.

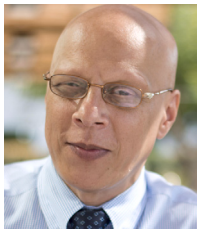
[42] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, H. Chen, Harris hawks optimization: Algorithm and applications, Future Gener. Comput. Syst. 97 (2019) 849–872.

[43] W. Jiang, Cellular traffic prediction with machine learning: A survey, Expert Syst. Appl. 201 (2022) 117163.

[44] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Amer. Statist. 46 (3) (1992) 175–185.

[45] D. Krajzewicz, J. Erdmann, M. Behrisch, L. Bieker, Recent development and applications of SUMO-Simulation of Urban MObility, Int. J. Adv. Syst. Meas. 5 (3&4) (2012).

Maram Bani Younes (Member, IEEE) received the master's and bachelor's degrees from the University of Jordan, Amman, Jordan, the Ph.D. degree in computer science from the University of Ottawa, Ottawa, ON, Canada, in 2015. She is currently an Associate Professor with the Department of Cybersecurity and Information Security, Philadelphia University, Jordan, and a Research Associate with the School of Information Technology and Engineering, University of Ottawa. Her research interests include wireless networks, wireless ad hoc, and sensor networks, green applications, security protocols, vehicular networks, and traffic efficiency for the vehicular network. She was the recipient of the some prestigious research prizes, such as, the best paper award in Globecom 2012, best Researcher award in Paradise Research Laboratory 2015, excellence in scientific research at Philadelphia University 2020, and Rising Star Award from MSWIM 2022.



Azzedine Boukerche [FIEEE, FEiC, FCAE, FAAAS, FAAIA] is a Distinguished University Professor and holds a Senior Canada Research Chair Tier-1 position with the University of Ottawa. He is a Research Fellow at he Center for Applied Mathematics and Bioinformatics (CAMB) at Gulf University for Science and Technology, Kuwait. He is the founding director of the PARADISE Research Laboratory and the DIVA Strategic Research Center, and NSERCCREATE TRANSIT at the University of Ottawa. He has received the C. Gotlieb Computer Medal Award, Ontario Distinguished Researcher Award, Premier of Ontario Research Excellence Award, G. S. Glinski Award for Excellence in Research, IEEE Computer Society Golden Core Award, IEEE CS-Meritorious Award, IEEE TCPP Leadership Award, IEEE ComSoc ComSoft and IEEE ComSoc ASHN Leadership and Contribution Award, and the University of Ottawa Award for Excellence in Research. He serves as an Editor-in-Chief for ACM ICPS and Associate Editor for several IEEE transactions and ACM journals and is also a Steering Committee Chair for several IEEE and ACM international conferences. His current research interests include sustainable sensor networks, autonomous and connected vehicles, wireless networking and mobile computing, wireless multimedia, QoS service provisioning, performance evaluation and modeling of largescale distributed and mobile systems, and large scale distributed and parallel discrete event simulation. He has published extensively in these areas and received several best research paper awards for his work. He is a Fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the American Association for the Advancement of Science.