Engineering Science & Technology
https://ojs.wiserpub.com/index.php/EST/

UNIVERSAL WISER PUBLISHER

Research Article

AI-Assisted Identification and Prediction of Dynamic Vehicular Communication Network States Using Taxi GPS Trajectories

Yixuan Li[1], **Hong Zhang**[1,2*]

[1]Transportation Institute, Inner Mongolia University, Hohhot, 010020, China
[2]Inner Mongolia Engineering Research Center for Intelligent Transportation Equipment, Hohhot, China
E-mail: zh-hong@imu.edu.cn

Received: 14 June 2026; **Revised:** 30 July 2026; **Accepted:** 24 August 2026

Abstract: Empirical monitoring of potential Vehicle-to-Vehicle (V2V) connectivity remains less studied than traffic-flow prediction and simulation-based vehicular networking. This study develops a taxi-Global Positioning System (GPS)-driven framework for identifying and predicting dynamic vehicular communication-network states in Hohhot, China. Taxi positions were aggregated into 672 15-min snapshots and converted into potential V2V proximity networks under communication radii of 50, 100, and 150 m. Standardized topology features were clustered into sparse-fragmented, transitional, and relatively-connected states, and the next-state prediction was evaluated using Persistence, Random Forest, and eXtreme Gradient Boosting (XGBoost). Increasing the radius raised the mean degree from 3.2930 to 9.3370, whereas the largest connected component ratio remained below 0.05, indicating local connectivity but global fragmentation. State self-transition probabilities were 0.8680, 0.8716, and 0.7713 for 50, 100, and 150 m, respectively. XGBoost performed best at 50 m, Random Forest at 100 m, and Persistence at 150 m. Under a common supplementary horizon-aligned protocol, average best Macro F1 decreased as the prediction horizon increased from 15 to 30, 60, and 90 min. The framework therefore provides a lightweight basis for short-term communication-state monitoring, while more realistic wireless constraints and fully inductive validation remain future work.

***Keywords*:** vehicular communication network, taxi Global Positioning System (GPS) trajectory, network state identification, short-term prediction

1. Introduction

Connected vehicles and Intelligent Transportation Systems (ITS) increasingly depend on the ability of vehicles to exchange information with nearby road users and infrastructure. Vehicle-to-Vehicle (V2V) communication supports cooperative awareness, decentralized traffic-information sharing, and edge-assisted mobility services. In dense urban traffic, however, potential communication opportunities vary with vehicle movement, traffic demand, road geometry, and the assumed communication range. Understanding the evolution of the resulting network is therefore important for communication-state monitoring.

A Vehicular Ad Hoc Network (VANET) can be represented as a dynamic graph in which vehicles are nodes and communication opportunities are edges. Such a representation makes it possible to measure local connectivity, fragmentation, isolated vehicles, and the evolution of connected components. These indicators are relevant for connected vehicle applications

Copyright ©2026 Hong Zhang, et al.
DOI: https://doi.org/10.37256/est.81202710314
This is an open-access article distributed under a CC BY license
(Creative Commons Attribution 4.0 International License)
https://creativecommons.org/licenses/by/4.0/

because they describe whether information can remain local, propagate through a larger component, or become limited to small disconnected groups. For urban digital twins and edge-assisted transportation management, network-state awareness can provide a compact description of communication readiness without simulating every packet-level process.

Recent research on vehicle-to-everything communication emphasizes that V2V, vehicle-to-infrastructure, and vehicle-to-pedestrian links are core elements of intelligent connected vehicle systems [1]. Studies on connected and automated transportation systems also highlight that reliable network-state awareness and cybersecurity are necessary for maintaining cooperative services under uncertain communication conditions [2].

From a network perspective, the topology of urban vehicular communication can change rapidly with traffic movement and vehicle density. Recent digital-twin-based work has modeled V2V time-varying complex networks using topological indicators, indicating that dynamic graph representations can support communication-network analysis [3].

Vehicle trajectory data provide a practical empirical basis for observing urban mobility and constructing movement-based network representations. Recent vehicle-trajectory studies show that logged vehicle trajectories can support data-driven safety assessment and transportation applications [4]. Trajectory-data-driven network representations have also been used for traffic-state prediction, showing that vehicle movement records can be transformed into network-level features for predictive modeling [5].

Most machine-learning-based transportation studies predict traffic flow, speed, travel time, or road-network conditions rather than communication-network states. Reviews of traffic forecasting and graph-based learning confirm that the dominant targets remain mobility and traffic variables [6], [7], while graph-based transport models mainly serve traffic forecasting or system-level surrogate modeling [8], [9]. Consequently, their prediction scores are not directly comparable with a three-class V2V communication-state task.

Digital-twin research for ITS emphasizes compact, data-driven state representations and predictive monitoring [10], [11]. Nevertheless, two gaps remain: many vehicular-network studies rely on simulated mobility or protocol assumptions, and empirical trajectory studies usually predict traffic states rather than potential V2V communication states. Evidence from medium-sized cities is particularly limited.

This study addresses these gaps by constructing a taxi-Global Positioning System (GPS)-driven potential V2V proximity network for Hohhot, identifying interpretable network states, and evaluating their short-term predictability. The novelty does not lie in proposing a new machine-learning algorithm; it lies in the integrated trajectory-to-network-to-state framework, the use of real urban trajectories, and the validation of state identification and prediction under multiple proximity radii.

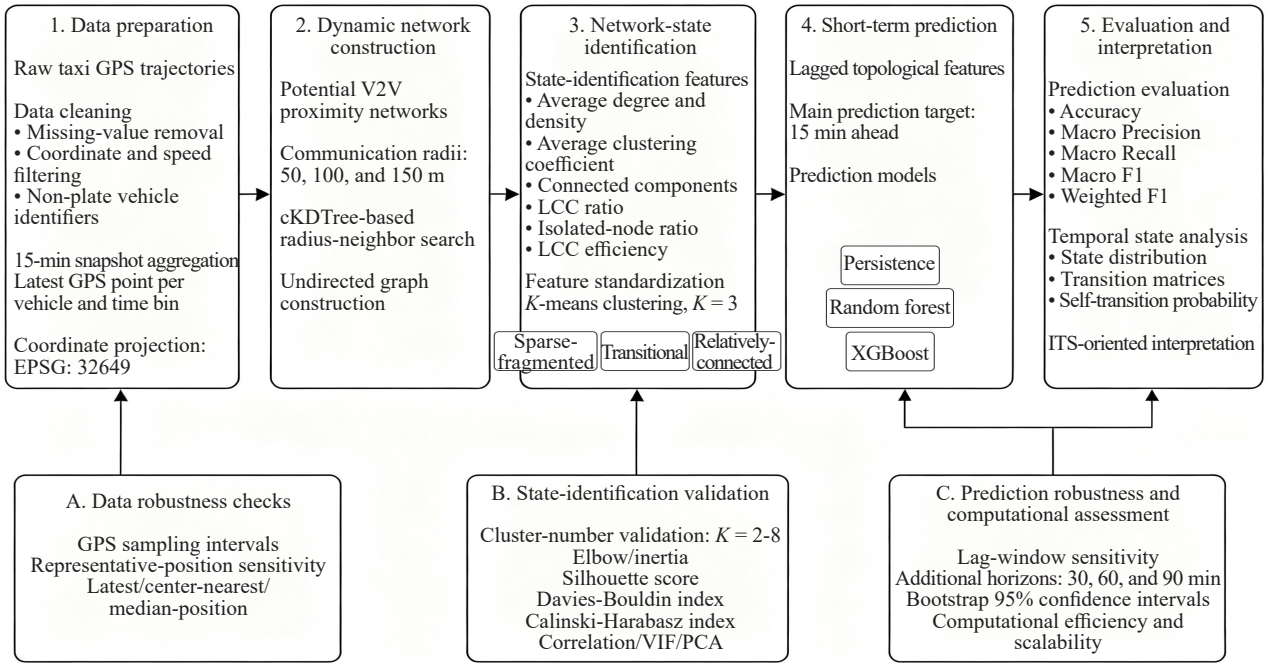


Figure 1. Methodological workflow of the proposed taxi-GPS-driven vehicular communication-network state identification and prediction framework

The main contributions are as follows. First, one week of real taxi GPS trajectories is transformed into 672 dynamic potential V2V proximity-network snapshots under 50, 100, and 150 m radii. Second, interpretable sparse-fragmented, transitional, and relatively-connected states are identified from standardized topology features, with the three-state choice evaluated using K = 2-8 clustering diagnostics and Principal Component Analysis (PCA). Third, next-state and multi-horizon prediction are evaluated using Persistence, Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), together with chronological holdout evaluation, bootstrap uncertainty analysis, representative-position sensitivity, and computational-efficiency assessment.

The overall methodological workflow, including the main analysis stages and supplementary validation procedures, is illustrated in Figure 1.

2. Materials and methods

2.1 *Study area and taxi GPS data*

The empirical data consist of taxi Global Positioning System (GPS) trajectories from the Hohhot metropolitan area, China, covering 2026-05-17 to 2026-05-23. The source files contained 79,434,576 records. After removing records with missing vehicle identifiers, timestamps, or coordinates, abnormal coordinates, negative speeds, and speeds above 150 km/h, 79,196,449 records remained. A total of 5,571 taxis appeared in the final snapshot dataset, and 5,570 vehicles contributed at least one valid inter-record sampling interval. For intervals of 0-3,600 s, the mean, median, and 95th-percentile (P95) sampling intervals were 26.66, 30.00, and 31.00 s, respectively. The cleaned records were grouped into 15-min bins, producing 672 time bins and 2,324,160 vehicle-bin representative records. For each taxi and time bin, the latest GPS point was retained. Plain-text license plate numbers were not used in the final modeling, and only aggregated network-level results are reported. The dataset and preprocessing statistics are summarized in Table 1.

Table 1. Taxi GPS dataset and preprocessing statistics

Item	Value	Description
Study area	Hohhot metropolitan area	Taxi GPS records within the main urban area
Study period	2026-05-17 to 2026-05-23	Seven consecutive days
Taxis in snapshot dataset	5,571	5,570 contributed valid sampling intervals
Raw GPS records	79,434,576	Source Parquet records
Cleaned GPS records	79,196,449	After field, coordinate, and speed filtering
Final vehicle-bin records	2,324,160	Latest point per taxi per 15-min bin
Sampling interval	Mean 26.66 s; median 30 s; P95 31 s	Computed for 0 < interval ≤ 3,600 s
Temporal resolution	15 min	96 snapshots per day
Number of time bins	672	7 days × 96 snapshots
Communication radii	50, 100, 150 m	Distance-based sensitivity settings
Coordinate reference system	EPSG: 32649	Projected coordinates in meters
Privacy handling	No plain-text plate numbers in modeling	Only aggregated results reported

The data represent taxi movements in the main urban area rather than all road users. Accordingly, the constructed graphs describe taxi-based potential V2V communication opportunities and do not imply complete coverage of private vehicles, buses, or all wireless links in the city.

Ethics approval and data protection. The taxi GPS data were derived from routine operational monitoring and

were authorized for academic research use by the competent transportation authority in Hohhot. The research team did not directly recruit or interact with individual taxi drivers, and no research-specific data were collected from drivers. Individual informed consent was not separately obtained by the research team because no research-specific data were collected directly from drivers. The retrospective use and analysis of these pre-existing data were reviewed and approved by the Academic and Ethics Committee of the Department of Mechanical Engineering, Transportation Institute, Inner Mongolia University (Approval No. IMU-TI-ME-2026-018; May 29, 2026). Vehicle identifiers were de-identified or encoded for analysis, and only aggregated network-level results are reported.

2.2 *Dynamic V2V network construction*

Each taxi is treated as a node in a time-specific undirected graph. Two taxis are linked when their projected Euclidean distance does not exceed the radius R. V2V effective range is not a fixed protocol-independent quantity. The 3rd Generation Partnership Project (3GPP) evaluation framework uses scenario-specific channel and system assumptions [12], while recent 5.9-GHz field measurements show that path loss, shadowing, and multipath vary substantially among highway, suburban, and urban environments and with obstruction density [13]. Accordingly, 50, 100, and 150 m were selected as conservative short-, intermediate-, and relatively extended urban proximity thresholds for sensitivity analysis. They are not interpreted as guaranteed Dedicated Short-Range Communications (DSRC) or Cellular Vehicle-to-Everything (C-V2X) coverage limits; instead, they allow the study to quantify how progressively relaxed proximity assumptions affect topology, state identification, and predictability. Longitude and latitude were projected to European Petroleum Survey Group (EPSG): 32649 coordinates so that distances could be calculated in meters.

For each 15-min snapshot and radius, cKDTree radius-neighbor search was used to avoid exhaustive pairwise distance enumeration. The resulting graph is a potential V2V proximity network rather than a packet-level wireless simulation. It is undirected because the available GPS data contain positions but not transmitter power, antenna orientation, channel state, or packet-delivery information; therefore, an edge denotes mutual spatial proximity rather than confirmed symmetric bidirectional connectivity.

2.3 *Topological indicators*

For each network snapshot, the number of nodes, number of edges, average degree, density, average clustering coefficient, number of connected components, Largest Connected Component (LCC) ratio, isolated-node ratio, and LCC efficiency were calculated. The clustering inputs were selected to represent complementary aspects of local connection intensity, normalized density, clustering, fragmentation, component reach, and within-LCC communication efficiency. The definitions of the network topology indicators are summarized in Table 2.

Table 2. Definitions of network topology indicators

Indicator	Definition
Number of nodes	Number of taxis appearing in a 15 min snapshot
Number of edges	Number of taxi pairs whose projected distance is not greater than R
Average degree	Mean number of V2V neighbors per taxi
Density	Ratio of observed edges to all possible undirected edges
Average clustering coefficient	Mean local clustering coefficient of nodes
Connected components	Number of connected components in the snapshot
LCC ratio	Largest connected component size divided by the number of nodes
Isolated node ratio	Proportion of nodes with degree zero
LCC efficiency	Efficiency computed inside the largest connected component only

It is important to clarify the efficiency terminology. The efficiency indicator used in this manuscript is LCC efficiency, computed inside the largest connected component. It is not whole-network global efficiency with disconnected node pairs treated as zero. This distinction is necessary because the taxi-based V2V networks are highly fragmented at the studied radii.

2.4 *AI-assisted network state identification*

K-means clustering was applied separately for each radius to standardized values of average degree, density, average clustering coefficient, number of connected components, LCC ratio, isolated-node ratio, and LCC efficiency. To assess the number of clusters, $K = 2$-8 was evaluated using inertia/Sum of Squared Errors (SSE), silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. No single metric selected $K = 3$ at every radius: $K = 2$ generally produced better silhouette and Davies-Bouldin values, whereas $K = 3$ substantially reduced SSE and preserved an operationally meaningful transitional state. $K = 3$ was therefore retained as a transparent, domain-informed compromise rather than a universally metric-optimal solution.

The three clusters were labeled sparse-fragmented, transitional, and relatively-connected according to their multivariate profiles in average degree, LCC ratio, isolated-node ratio, and LCC efficiency. The term relatively-connected is deliberately comparative because the LCC ratio remains low even in the strongest state. Standardization and *K*-means were fitted to all 672 snapshots to define a descriptive taxonomy for the observed week. The subsequent chronological prediction therefore evaluates forecasting conditional on this fixed taxonomy; it is not a fully inductive train-only clustering pipeline. Temporal dependence is analyzed separately through state-transition matrices and lagged-feature prediction.

2.5 *Short-term network state prediction*

The primary target is the communication-network state with a 15-minute lead time. Three models were evaluated: a Persistence baseline, RF [14], and XGBoost [15]. Persistence assigns the most recently observed state to the next time bin, whereas RF and XGBoost use lagged topological features. Supplementary experiments further evaluate 30-, 60-, and 90-min-ahead targets.

The primary trainable-model comparison uses three historical time bins (45 min), matching the original modeling protocol used for the primary radius-specific comparison. Sensitivity analysis compares 1, 3, 4, 6, and 8 bins, corresponding to 15, 45, 60, 90, and 120 min of history. Predictor variables include lagged node count, edge count, average degree, density, average clustering coefficient, number of connected components, LCC ratio, isolated-node ratio, and LCC efficiency.

2.6 *Model evaluation and validation*

Model performance was assessed using Accuracy, Macro Precision, Macro Recall, Macro F1, and Weighted F1. A nonparametric bootstrap over the 96 test-day observations was used to obtain descriptive 95% confidence intervals for Accuracy and Macro F1. Because adjacent 15-min states are temporally dependent, these intervals are used to quantify uncertainty and to discourage overinterpretation of small differences; they are not treated as formal proof of pairwise statistical significance. The primary chronological split used 573 pre-test observations from 2026-05-17 to 2026-05-22 for model fitting and 96 observations from 2026-05-23 for testing, with no random temporal shuffling.

Model settings were fixed before test evaluation rather than optimized on the held-out test day. The primary comparison, lag-window sensitivity analysis, and supplementary horizon analysis are reported as distinct protocols so that their results are not conflated. The fixed hyperparameter settings of the prediction models are summarized in Table 3.

Table 3. Hyperparameter settings of the prediction models

Model	Parameter	Value	Selection
Persistence	Rule	Next state = current state	Deterministic baseline
RF	n_estimators	300	Fixed before test evaluation
RF	max_depth	None	Fixed before test evaluation
RF	class_weight	balanced	Fixed before test evaluation
RF	random_state	42	Reproducibility
XGBoost	n_estimators	300	Fixed before test evaluation
XGBoost	max_depth	4	Fixed before test evaluation
XGBoost	learning_rate	0.05	Fixed before test evaluation
XGBoost	subsample/colsample_bytree	0.8/0.8	Fixed before test evaluation
XGBoost	objective/eval_metric	multi:softprob/mlogloss	Multiclass classification

3. Results

3.1 *Descriptive statistics of dynamic V2V network snapshots*

The final dataset contains 672 dynamic network snapshots for each communication radius. The mean number of nodes is approximately 3,458.57 across radii because the node set is defined by taxi availability in each time bin. The mean number of edges increases from 6,300.52 at 50 m to 12,576.34 at 100 m and 17,793.56 at 150 m. Similarly, the average degree increases from 3.2930 to 6.5763 and 9.3370. These results confirm that the radius setting has a strong influence on local V2V opportunities.

The mean LCC ratio remains low across all radii: 0.032233 at 50 m, 0.035541 at 100 m, and 0.041461 at 150 m. This indicates that the taxi-based communication network is locally connected but globally fragmented. The result is consistent with the spatial dispersion of taxis and the short communication ranges considered in the analysis. The corresponding average network statistics are summarized in Table 4. The spatial distribution of taxi GPS trajectory density is shown in Figure 2.

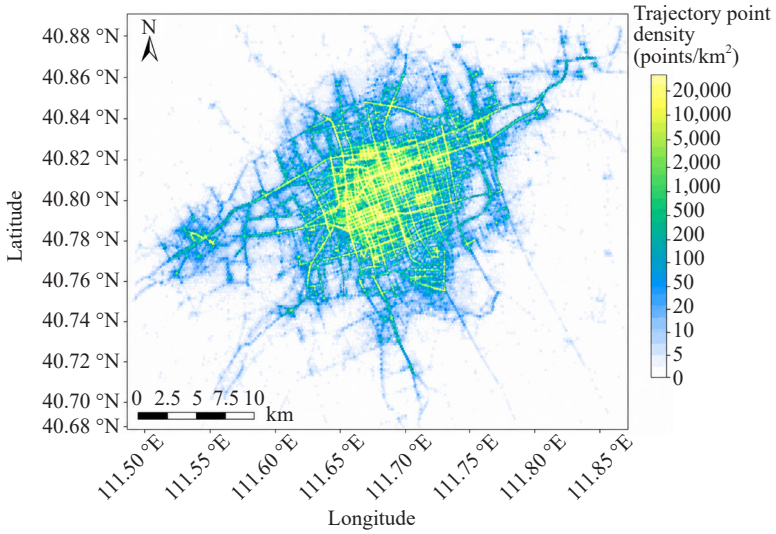


Figure 2. Spatial distribution of taxi GPS trajectory density in the study area

Table 4. Average network statistics under different communication radii

Radius (m)	Mean nodes	Mean edges	Mean degree	Mean LCC ratio	Mean LCC efficiency
50.00	3,458.57	6,300.52	3.2930	0.0322	0.5445
100.00	3,458.57	12,576.34	6.5763	0.0355	0.6764
150.00	3,458.57	17,793.56	9.3370	0.0415	0.5925

3.2 *Effect of communication radius on network topology*

The radius monotonicity check showed that the number of edges, average degree, and LCC ratio increase monotonically from 50 m to 100 m and 150 m for every time bin. This provides an internal consistency check for the network construction process. The LCC efficiency indicator is not required to be monotonic because it is computed within the largest connected component, whose membership can change as the radius increases.

As shown in Figures 3 and 4, temporal curves show daily variation in edges and average degree, reflecting changing taxi activity and spatial concentration. The boxplots further show that larger communication radii consistently shift local connectivity indicators upward while leaving the network fragmented at the city scale. The distributions of the key network indicators across communication radii are further illustrated in Figure 5.

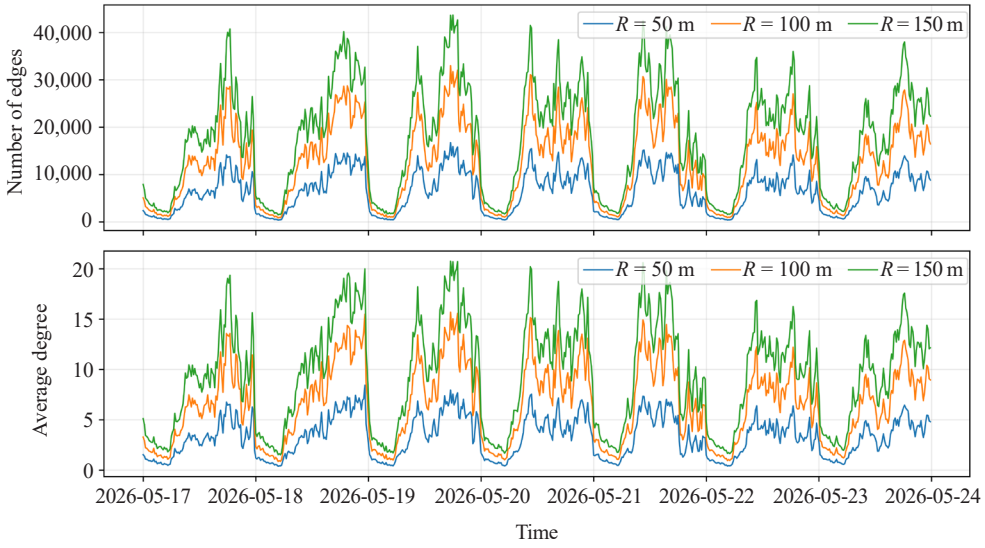


Figure 3. Temporal evolution of the number of edges and average degree under different communication radii

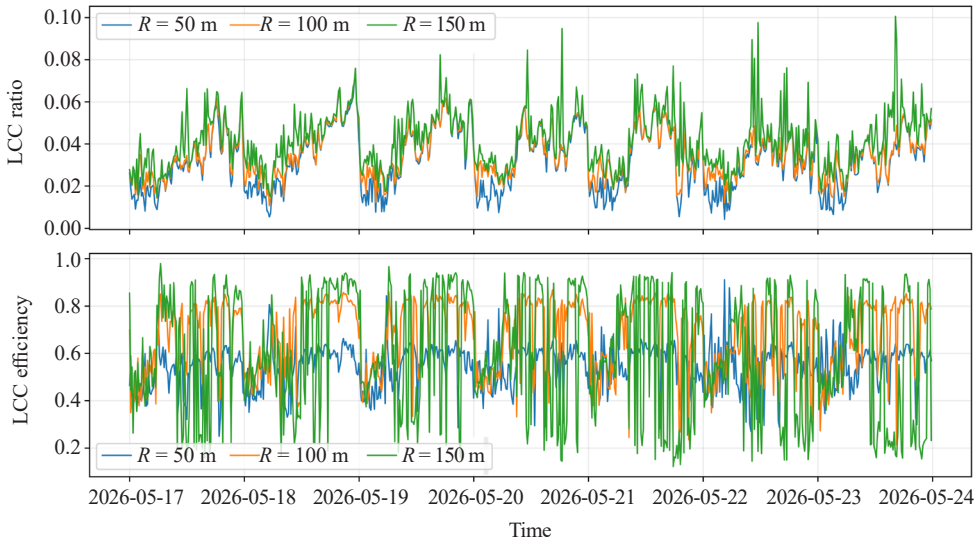


Figure 4. Temporal evolution of LCC ratio and LCC efficiency under different communication radii

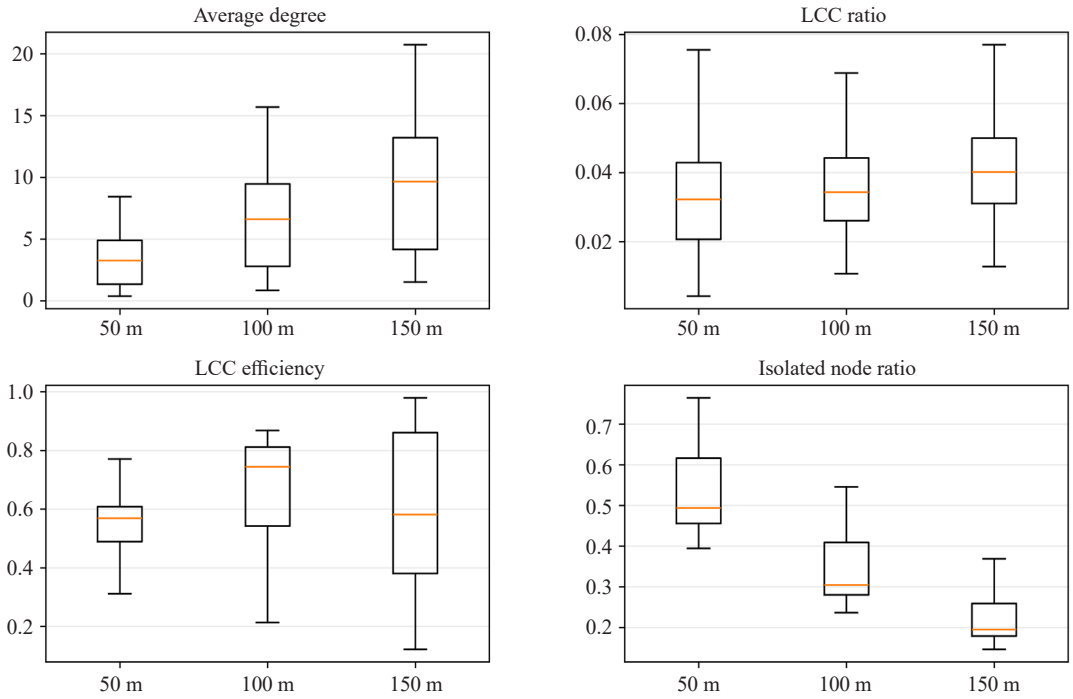


Figure 5. Boxplots of key network indicators under different communication radii

3.3 *AI-assisted network state identification*

The *K*-means results divide the snapshots into sparse-fragmented, transitional, and relatively-connected states. Across the three radii, moving from $K = 2$ to $K = 3$ reduced SSE by approximately 25%-38%. $K = 2$ retained higher silhouette scores and lower Davies-Bouldin values, while the Calinski-Harabasz comparison was mixed. Thus, the quantitative diagnostics do not establish $K = 3$ as uniquely optimal; rather, they support its use as a domain-informed compromise that preserves a transitional state. The radius-specific cluster profiles are summarized in Table 5. Table 6 reports the direct $K = 2$ versus $K = 3$ comparison, and the complete $K = 2$-8 diagnostics are provided in Supplementary Table S1.

Table 5. Cluster profiles of the identified network states

Radius (m)	Network state	Average degree	LCC ratio	Isolated node ratio	LCC efficiency	Sample count
50	Relatively-connected	5.5390	0.0473	0.4646	0.6061	239
50	Sparse-fragmented	0.9238	0.0165	0.6769	0.5224	193
50	Transitional	2.9617	0.0299	0.4919	0.5010	240
100	Relatively-connected	10.7823	0.0482	0.2879	0.8077	241
100	Sparse-fragmented	1.7924	0.0239	0.4692	0.5330	181
100	Transitional	5.9853	0.0317	0.3071	0.6536	250
150	Relatively-connected	12.4713	0.0445	0.1929	0.8318	277
150	Sparse-fragmented	2.9574	0.0289	0.3045	0.5689	196
150	Transitional	11.2574	0.0496	0.1804	0.2825	199

The distribution of states varies with communication radius. Larger radii make relatively-connected snapshots more frequent, although the overall LCC ratio remains small. This finding suggests that short-range V2V opportunities are sensitive to communication radius but remain spatially localized in the taxi-only network. The distribution of the identified network states across communication radii is shown in Figure 6.

The state labels describe multivariate cluster profiles and do not imply that every indicator is strictly ordered. At 150 m, for example, the transitional cluster has a slightly higher mean LCC ratio and a slightly lower isolated-node ratio than the relatively-connected cluster, but its LCC efficiency is markedly lower (0.2825 versus 0.8318). This pattern indicates a larger but internally less efficient dominant component and motivates the radius-relative interpretation of the labels.

Table 6. Direct comparison of $K = 2$ and $K = 3$ clustering diagnostics

R (m)	K	SSE	Silhouette	Davies-bouldin	Calinski-harabasz
50	2	2,010.29	0.5187	0.7334	897.77
50	3	1,276.67	0.4222	0.8624	898.01
100	2	1,957.17	0.5294	0.6879	940.33
100	3	1,207.74	0.4298	0.8775	968.33
150	2	2,264.44	0.4752	0.7849	721.81
150	3	1,688.35	0.3839	1.1183	597.47

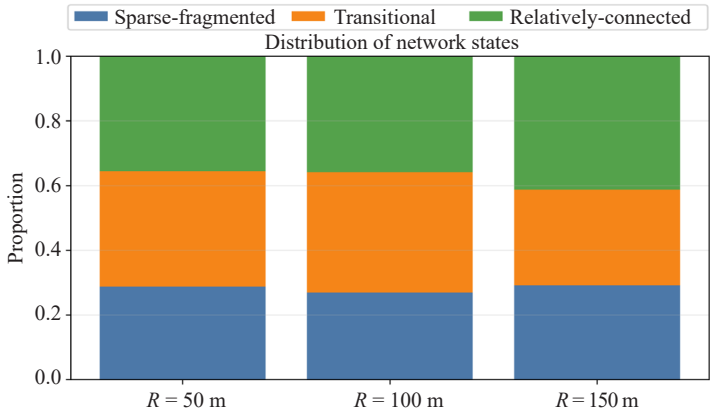


Figure 6. Distribution of network states under different communication radii

3.4 *State transition and temporal persistence*

The state transition matrices show clear temporal inertia. The mean self-transition probabilities are 0.8680 at 50 m, 0.8716 at 100 m, and 0.7713 at 150 m. In practical terms, a snapshot is likely to remain in the same state in the next 15 min interval, especially for 50 m and 100 m. The corresponding state transition matrices are shown in Figure 7.

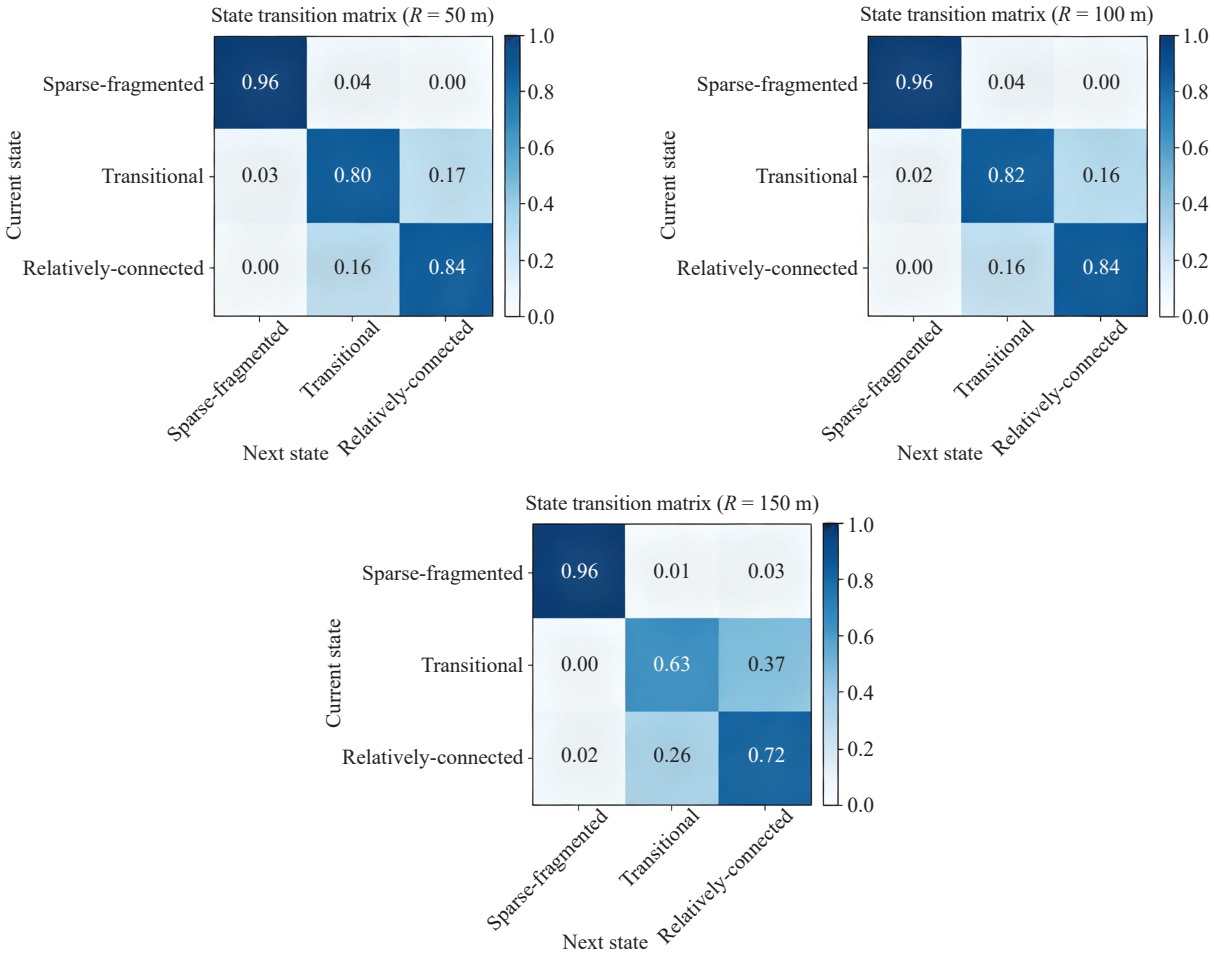


Figure 7. State transition matrices for different communication radii

This inertia explains why the Persistence baseline is competitive. It also indicates that short-term network-state monitoring can exploit recent state information without requiring complex sequence models.

3.5 *Short-term network state prediction*

Prediction results differ by radius. At 50 m, XGBoost achieves the highest Macro F1 of 0.9014 (descriptive bootstrap 95% Confidence Interval (CI): 0.8423-0.9590), compared with 0.8911 (0.8264-0.9473) for Persistence. At 100 m, RF performs best with a Macro F1 of 0.8928 (0.8282-0.9492), compared with 0.8714 (0.7976-0.9289) for Persistence. At 150 m, Persistence achieves the highest Macro F1 of 0.8520 (0.7856-0.9136). The overlapping intervals, limited test-day sample, and temporal dependence indicate that these small numerical differences should not be interpreted as formal evidence of model superiority.

These results suggest that lightweight machine-learning models can add predictive value at 50 and 100 m, while the strong temporal persistence of the 150 m state sequence makes the baseline difficult to outperform. The analysis therefore emphasizes effect size, uncertainty, and consistency across radii and baselines rather than significance claims based on a single test day. The prediction performance comparison is illustrated in Figure 8.

The supplementary multi-horizon experiment uses a common direct-horizon feature construction for 15-, 30-, 60-, and 90-min targets. Its 15-min reference is therefore not identical to the primary three-bin-history comparison in Table 7, and Table 8 is used only to assess degradation within the same horizon-aligned protocol. Under this protocol, the average best Macro F1 decreases from 0.8715 at 15 min to 0.7786, 0.7040, and 0.6838 at 30, 60, and 90 min, respectively. The results support extension to longer horizons but show that reliability declines as the forecast horizon increases. Complete model-by-horizon results are provided in Supplementary Table S2.

Table 7. Prediction performance of Persistence, RF, and XGBoost

Radius (m)	Model	Accuracy	Macro precision	Macro recall	Macro F1	Weighted F1
50	Persistence	0.8854	0.8924	0.8900	0.8911	0.8853
50	Random forest	0.8438	0.8532	0.8511	0.8520	0.8436
50	XGBoost	0.8958	0.9000	0.9038	0.9014	0.8958
100	Persistence	0.8646	0.8729	0.8701	0.8714	0.8644
100	Random forest	0.8854	0.8968	0.8912	0.8928	0.8866
100	XGBoost	0.8438	0.8598	0.8524	0.8548	0.8457
150	Persistence	0.8438	0.8521	0.8521	0.8520	0.8440
150	Random forest	0.8333	0.8466	0.8406	0.8421	0.8349
150	XGBoost	0.7708	0.8014	0.7851	0.7841	0.7715

Table 8. Best Macro F1 in the supplementary horizon-aligned prediction experiment

Horizon	50 m	100 m	150 m	Mean best macro F1
15 min	Persistence: 0.8911	Persistence: 0.8714	Persistence: 0.8520	0.8715
30 min	Persistence: 0.7919	Persistence: 0.7916	Persistence: 0.7523	0.7786
60 min	XGBoost: 0.7252	XGBoost: 0.7284	Persistence: 0.6585	0.7040
90 min	XGBoost: 0.7295	RF: 0.6773	XGBoost: 0.6446	0.6838

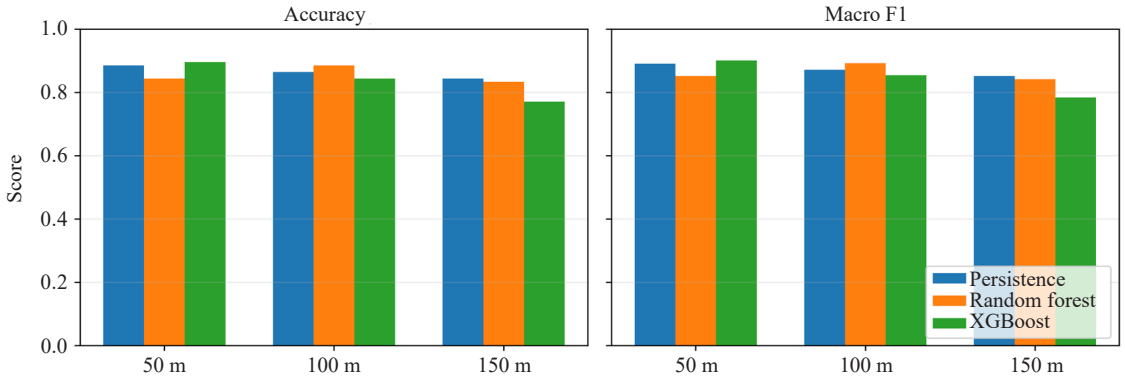


Figure 8. Prediction performance comparison among Persistence, RF, and XGBoost

3.6 *Lag-window sensitivity and feature importance*

The primary radius-specific comparison in Table 7 uses a three-bin (45-min) history. The lag-window sensitivity analysis shows that six historical bins (90 min) provide the highest mean Macro F1 for the trainable models across radii (0.8613), whereas the three-bin setting remains competitive and more parsimonious (0.8545). The 90-min result is therefore reported as a sensitivity finding rather than being retrospectively substituted for the pre-specified primary comparison. History length is distinct from prediction horizon: the principal target remains 15 min ahead. The complete lag-window sensitivity results are summarized in Table 9. The lag-window sensitivity results are illustrated in Figure 9.

Feature-importance and redundancy analyses indicate that vehicle activity scale and local connectivity dominate prediction. Strong correlations were observed among edge count, average degree, and density, as expected from their mathematical relationships. Nevertheless, the retained indicators describe different physical aspects of network scale, local intensity, fragmentation, and within-LCC efficiency. After standardization, the first two principal components explain 89.81%, 89.75%, and 83.07% of the variance at 50, 100, and 150 m, respectively, supporting a compact low-dimensional representation of the state structure. The detailed PCA explained-variance results are reported in Supplementary Table S5.

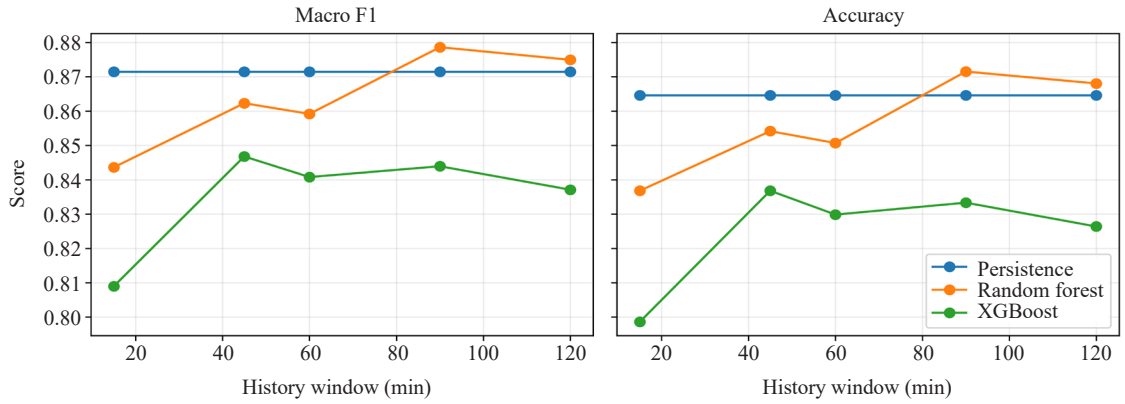


Figure 9. Lag-window sensitivity analysis

Table 9. Lag-window sensitivity analysis

Lag window	History length (min)	Model	Mean accuracy	Mean macro F1	Mean weighted F1
1	15	Persistence	0.8646	0.8715	0.8646
1	15	Random forest	0.8368	0.8436	0.8351
1	15	XGBoost	0.7986	0.8090	0.7983
3	45	Persistence	0.8646	0.8715	0.8646
3	45	Random forest	0.8542	0.8623	0.8550
3	45	XGBoost	0.8368	0.8468	0.8377
4	60	Persistence	0.8646	0.8715	0.8646
4	60	Random forest	0.8507	0.8592	0.8519
4	60	XGBoost	0.8299	0.8408	0.8311
6	90	Persistence	0.8646	0.8715	0.8646
6	90	Random forest	0.8715	0.8786	0.8725
6	90	XGBoost	0.8333	0.8439	0.8345
8	120	Persistence	0.8646	0.8715	0.8646
8	120	Random forest	0.8681	0.8749	0.8687
8	120	XGBoost	0.8264	0.8371	0.8276

4. Discussion

The results demonstrate the value of using real taxi GPS data to study dynamic potential V2V communication opportunities. Ahmed et al. used GPS-derived trip characteristics and RF to classify road traffic conditions and reported 81% accuracy [16], while Yasuda et al. transformed observed trajectories into a network representation for deep-learning-based traffic-state prediction [5]. These studies demonstrate the value of trajectory-derived representations, but their targets are road traffic conditions rather than communication-network readiness. Because the state definitions, datasets, and evaluation metrics differ, direct numerical comparison or superiority claims would be inappropriate. In the present one-week dataset, best 15-min-ahead Macro F1 values of approximately 0.85-0.90 indicate nontrivial predictability of the fixed three-state taxonomy. The strong Persistence baseline reveals temporal inertia, while the gains from RF and XGBoost at 50 and 100 m indicate that lagged topology features contain information beyond the previous state. External-city and longer-period validation is still required before operational reliability can be claimed.

A consistent pattern is that the taxi-based V2V network is locally connected but globally fragmented. Increasing the communication radius substantially increases the number of edges and average degree, but the LCC ratio remains low. This means that short-range communication may support local information sharing among nearby taxis, while citywide propagation through taxi-only V2V links is limited at the studied radii.

The strong temporal inertia of network states provides an important interpretation for the model results. Persistence performs well because the state in a 15 min interval often remains unchanged in the following interval. Random Forest and XGBoost add value when lagged topological features contain additional information beyond the previous state, particularly at 50 m and 100 m. At 150 m, the smoother state process makes the baseline more competitive.

The engineering implication is that lightweight state prediction can support rolling city-level communication-readiness monitoring after each completed 15-min window. Potential uses include early warning of sparse-fragmented conditions, adaptive Roadside Unit (RSU) activation, mobile-relay or unmanned-aerial-vehicle dispatch, edge-computing resource reservation, and communication-state updates in urban traffic digital twins. The results also suggest that model complexity can be adapted to the communication radius: RF or XGBoost can be used where

topology features add predictive value, whereas the low-cost Persistence rule may be sufficient when the state process is smoother. These are prospective supervisory monitoring and resource-planning applications, not evidence of an operational deployment or suitability for sub-second safety-critical control.

Table 10. Computational-efficiency summary under the tested environment

Module	Observed workflow span/scale	Interpretation
Cleaning and snapshot preparation	≈ 186 s for seven days	Offline preprocessing
Network construction and feature extraction	≈ 120 s for 672 snapshots × 3 radii	cKDTree-based processing
K-means clustering	≈ 119 s	Includes I/O and output generation
Prediction training and inference	≈ 118 s	Includes fitting, evaluation, and outputs
Mean graph-metric runtime per snapshot	0.0044/0.0062/0.0080 s	R = 50/100/150 m
Test environment	Intel Core Ultra 9 285K; 31.47 GiB RAM; Windows	CPU processing; GPU not used

Computational efficiency was evaluated on a Windows workstation equipped with an Intel Core Ultra 9 285K Central Processing Unit (CPU) and 31.47 GiB of Random-Access Memory (RAM); the Graphics Processing Unit (GPU) was not used. Logged full-week workflow spans were approximately 186 s for cleaning and snapshot preparation, 120 s for network construction and feature extraction, 119 s for clustering, and 118 s for model fitting, evaluation, and output generation. These stage spans include data Input/Output (I/O), serialization, and result writing and should not be interpreted as pure algorithmic kernel time. Mean recorded graph-metric runtime per snapshot was 0.0044, 0.0062, and 0.0080 s at 50, 100, and 150 m. The measured times are substantially shorter than the 15-min update interval under the tested data scale, supporting potential supervisory near-real-time monitoring. They do not constitute evidence of an operational deployment, which would additionally require live ingestion, communication latency, fault tolerance, and heterogeneous-hardware testing. The complete per-snapshot runtime statistics are provided in Supplementary Table S4. The computational-efficiency results under the tested environment are summarized in Table 10.

A representative-position sensitivity analysis was conducted to assess boundary effects from selecting the latest point in each bin. Agreement between center-nearest and latest labels was 0.9271, 0.9211, and 0.7768 at 50, 100, and 150 m; agreement between median-position and latest labels was 0.8780, 0.8795, and 0.7470. Thus, the main state patterns are robust at 50 and 100 m, whereas the 150 m case is more sensitive and should be interpreted with greater caution. Complete agreement and Cohen's kappa values are reported in Supplementary Table S3.

Several limitations should be acknowledged. The analysis includes taxis only and models distance-based potential communication opportunities rather than confirmed wireless links. Euclidean proximity omits channel fading, non-line-of-sight obstruction, interference, Medium Access Control (MAC) contention, latency, packet loss, transmitter and antenna heterogeneity, and directional asymmetry. The state taxonomy was standardized and clustered using the full observed week before chronological prediction, so the forecasting experiment is conditional on a fixed descriptive taxonomy rather than a fully inductive train-only clustering pipeline. In addition, observation-level bootstrap intervals do not fully account for temporal dependence, and correlated topology indicators may give related dimensions additional weight in *K*-means. These limitations require the state labels and uncertainty intervals to be interpreted as empirical summaries rather than externally validated physical states or formal significance tests.

Future work should integrate buses, private vehicles, RSUs, and Vehicle-to-Infrastructure (V2I) links; evaluate channel fading, packet loss, communication latency, and protocol-level behavior; and test transferability across cities and longer time periods. A fully inductive pipeline should fit the scaler and clustering model within each training fold and assign test snapshots using training-derived centroids. Longer time series would also permit blocked or moving-block bootstrap procedures that better preserve temporal dependence. Packet-level simulation or field communication measurements could then connect the present proximity-network states to service reliability and infrastructure-assisted communication strategies.

5. Conclusions

This study developed and evaluated a taxi-GPS-driven framework for identifying and predicting dynamic potential V2V communication-network states in Hohhot. The contribution is the integrated empirical state-monitoring framework rather than a new machine-learning algorithm.

First, increasing the communication radius from 50 m to 150 m raises the number of edges and average degree, but the LCC ratio remains low. The taxi-based V2V network is therefore best interpreted as locally connected and globally fragmented.

Second, *K*-means clustering identifies three interpretable states. The three-state solution is supported as a balance between clustering diagnostics and the need to represent a transitional condition, while PCA shows that the first two components capture most of the feature variance.

Third, RF and XGBoost improve prediction at selected radii, but the overlapping descriptive bootstrap intervals require small differences to be interpreted cautiously. Within the separate supplementary horizon-aligned protocol, the average best Macro F1 decreases from 0.8715 at 15 min to 0.7786, 0.7040, and 0.6838 at 30, 60, and 90 min, confirming that performance declines as the forecast horizon increases.

The findings suggest that lightweight communication-state awareness may support connected-vehicle applications and urban traffic digital twins, subject to taxi-only coverage, distance-based connectivity, fixed-taxonomy evaluation, and the need for external validation.

Supplementary materials

Supplementary materials associated with this article are available online and can be accessed via the journal website: https://ojs.wiserpub.com/index.php/EST/article/view/10314/4598.

Acknowledgements

The authors would like to thank the data provider and project collaborators for their support in data preparation and research implementation.

Author contributions

Yixuan Li: Conceptualization, methodology, software, formal analysis, data curation, visualization, writing-original draft preparation. Hong Zhang: Conceptualization, supervision, project administration, funding acquisition, writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (62362053), the Natural Science Foundation of Inner Mongolia Autonomous Region of China (2026SHZR3462), and the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region of China (NJYT23060).

Data availability statement

The raw taxi GPS data used in this study are not publicly available due to third-party restrictions and privacy considerations. The processed network indicators and analysis code can be made available from the corresponding author upon reasonable request, subject to approval by the data provider.

AI tool disclosure

During the preparation of this manuscript, the authors used OpenAI ChatGPT to assist with language drafting. The authors reviewed, revised, and verified all contents, interpretations, and conclusions, and take full responsibility for the final manuscript.

Conflict of interest

The authors declare no competing financial interest.

References

[1] X. Zhang, J. Li, J. Zhou, S. Zhang, J. Wang, Y. Yuan, et al., "Vehicle-to-everything communication in intelligent connected vehicles: A survey and taxonomy," *Automotive Innovation*, vol. 8, pp. 13-45, 2025.

[2] A. Abdo, H. Chen, X. Zhao, G. Wu, and Y. Feng, "Cybersecurity on connected and automated transportation systems: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1382-1401, 2024.

[3] H. Zhang, Z. Zheng, Y. Li, and L. Lu, "Digital twin-based analysis of V2V time-varying complex networks," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 7, pp. 10225-10239, 2025.

[4] X. Yan, S. Feng, D. J. LeBlanc, C. Flannagan, and H. X. Liu, "Evaluation of automated driving system safety metrics with logged vehicle trajectory data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 8913-8925, 2024.

[5] S. Yasuda, H. Katayama, W. Nakanishi, and T. Iryo, "Trajectory data-driven network representation for traffic state prediction using deep learning," *International Journal of Intelligent Transportation Systems Research*, vol. 22, no. 1, pp. 136-145, 2024.

[6] L. Cheng, D. Lei, and S. Tao, "Recent advances in deep learning for traffic probabilistic prediction," *Transport Reviews*, vol. 44, no. 6, pp. 1129-1135, 2024.

[7] S. Rahmani, A. Baghbani, N. Bouguila, and Z. Patterson, "Graph neural networks for intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8846-8885, 2023.

[8] S. Narayanan, N. Makarov, and C. Antoniou, "Graph neural networks as strategic transport modelling alternative - A proof of concept for a surrogate," *IET Intelligent Transport Systems*, vol. 18, no. 11, pp. 2059-2077, 2024.

[9] W. Ju, Y. Zhao, Y. Qin, S. Yi, J. Yuan, Z. Xiao, et al., "COOL: A conjoint perspective on spatio-temporal graph neural network for traffic forecasting," *Information Fusion*, vol. 107, p. 102341, 2024.

[10] C. Ge and S. Qin, "Digital twin intelligent transportation system (DT-ITS): A systematic review," *IET Intelligent Transport Systems*, vol. 18, no. 12, pp. 2325-2358, 2024.

[11] J. Guo, M. Bilal, Y. Qiu, C. Qian, X. Xu, and K.-K. R. Choo, "Survey on digital twins for Internet of Vehicles: Fundamentals, challenges, and opportunities," *Digital Communications and Networks*, vol. 10, no. 2, pp. 237-247, 2024.

[12] 3rd Generation Partnership Project (3GPP), "Study on LTE-based V2X services," 3GPP TR 36.885, V14.0.0, Jun. 2016. [Online]. Available: https://www.3gpp.org/dynareport/36885.htm

[13] Z. Hasirci Tugcu, K. Kuzulugil, and I. H. Cavdar, "Measurement-based V2V propagation modeling in highway, suburban, and urban environments," *Vehicular Communications*, vol. 48, p. 100791, 2024.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785-794.

[16] U. Ahmed, R. Tu, J. Xu, G. Amirjamshidi, M. Hatzopoulou, and M. J. Roorda, "GPS-based traffic conditions classification using machine learning approaches," *Transportation Research Record*, vol. 2677, no. 2, pp. 1445-1454, 2023.