HIGHWAY, MACHINE LEARNING

Assessing the Predictive Contribution of Socioeconomic, Environmental and Highway-Specific Factors to Traffic Demand Using Machine Learning

Miracle Chukwuebuka Ekwueme[1], Henry Kene Ugwuanyi[2], Gloria Nkechi Ugwuanyi[3]

[1]Copcan Civil LP, Nanaimo, British Columbia, Canada
[2]Department of Civil Engineering, Michael Okpara University of Agriculture, Umudike, Nigeria
[3]Department of Education Foundation, University of Nigeria, Nsukka, Nigeria

Published November 2026

Abstract

Reliable highway traffic-demand estimation supports capacity assessment, infrastructure investment, and transportation planning. This study assesses the predictive contribution of socioeconomic, environmental, temporal, and highway-specific factors to annual average daily traffic (AADT) on five interstate highway corridors in Southeastern Nigeria. A longitudinal dataset of 100 corridor-year observations (2001–2020) was analysed using Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbours (KNN), with a temporal trend benchmark. Chronological validation used 2001–2016 for training and 2017–2020 for holdout testing. RF impurity importance, GB holdout permutation importance, and sensitivity analysis interpreted predictive contribution. GB achieved the best holdout performance (R^2=0.5338; RMSE=396,916 PCU/day; MAE=249,437 PCU/day; MAPE=28.45%). RF importance was dominated by the Enugu–Port Harcourt corridor indicator (0.6150), population density (0.1443), and Abakaliki–Ogoja (0.1247). GB permutation importance likewise ranked corridor identity (133,679.86) and density (74,081.95) highest. Across observed ranges, mean predictions changed by +522,318 PCU/day for density and +342,891 for GDP, compared with smaller changes for rainfall (+2,984) and pothole volume (−3,103). Prediction error increased sharply in 2020, demonstrating vulnerability to structural disruptions. Findings identify corridor context and population density as leading predictive information sources and stress chronological validation and noncausal interpretation of importance.

Keywords: AADT; machine learning; traffic demand; feature importance; highway corridors; chronological validation

1 Introduction

Annual average daily traffic (AADT) is a fundamental highway-planning indicator used in capacity analysis, pavement design, traffic operations, safety assessment, maintenance programming, and investment appraisal. Traffic demand emerges from interacting economic, demographic, spatial, environmental, and roadway factors, making accurate estimation a persistent engineering challenge.

Machine-learning methods increasingly support traffic forecasting by capturing nonlinear patterns and interactions. Reviews discuss artificial intelligence in traffic prediction and highlight data quality, interpretability, and generalisation challenges (Shaygan et al., 2022; Du et al., 2025). AADT studies have incorporated land use, road density, demographic, and spatial context (Pulugurtha & Mathew, 2021; Sfyridis & Agnolucci, 2020, 2023).

This study addresses temporal leakage risks in model evaluation, limited interpretation of predictor contribution, and confusion between predictive importance and causal influence. It compares RF, GB, and KNN using chronological validation and combines feature importance with sensitivity analysis. Objectives are to compare performance, identify leading predictors, evaluate model response across observed ranges, and derive highway-planning implications.

2 Study area

Five interstate corridors were analysed: Umuahia–Ikot Ekpene (Abia), Onitsha–Owerri (Anambra), Abakaliki–Ogoja (Ebonyi), Enugu–Port Harcourt (Enugu), and Owerri–Port Harcourt (Imo). The panel comprises annual records from 2001–2020 (five corridors × 20 years = 100 observations). The response is AADT in passenger-car units per day (PCU/day). Predictors include GDP, passenger population density, annual rainfall, pothole volume, year, and corridor/state indicators (Table 1).

Table 1. Predictor groups and outcome variable.

Group	Variables	Interpretation
Socioeconomic	GDP; population density	Economic and demographic demand context
Environmental/condition	Rainfall; pothole volume	Climate and road-condition proxies
Temporal	Year	Long-run temporal pattern
Highway-specific	Corridor indicators	Persistent corridor heterogeneity
Outcome	AADT (PCU/day)	Annual traffic demand

3 Methods

3.1 Chronological validation

The chronological split used 2001–2016 (80 records) for model development and 2017–2020 (20 records) for holdout evaluation. This preserves temporal ordering and better approximates future forecasting than random splitting. The small holdout warrants caution in interpreting rankings.

3.2 Models and metrics

Random Forest aggregates predictions from randomized decision trees; Gradient Boosting sequentially fits trees to reduce preceding errors; K-Nearest Neighbours predicts from nearby observations in predictor space. A temporal trend benchmark was included. Evaluation metrics were R^2, RMSE, MAE, and MAPE. RMSE penalizes large errors more heavily, MAE reports average absolute error, and MAPE expresses error relative to observed traffic.

3.3 Predictive contribution and sensitivity

RF impurity importance quantifies predictor-associated reductions in tree impurity. GB permutation importance measures prediction deterioration after randomly permuting a predictor. Scores are model- and sample-dependent; correlated predictors can share or obscure importance, and negative scores may reflect noise. Sensitivity analysis reports change in mean predicted AADT across each variable's observed minimum–maximum range. These diagnostics describe model behaviour and do not estimate causal effects.

4 Results

4.1 Model performance

Gradient Boosting produced the highest holdout R^2 (Table 2; Figures 1 and 2) and lowest reported errors among the machine-learning models. KNN ranked second by R^2 and RF was close behind. The trend benchmark was weaker (R^2=0.4712), indicating added predictive information from nonlinear and corridor-related predictors.

Table 2. Chronological holdout performance of the models (2017–2020).

Model	R^2	RMSE (PCU/day)	MAE (PCU/day)	MAPE
Gradient Boosting	0.5338	396,916.31	249,437.28	28.45%
KNN	0.5198	402,801.03	298,293.81	32.33%
Random Forest	0.5126	405,847.31	275,226.98	29.97%
Temporal trend	0.4712	—	—	—

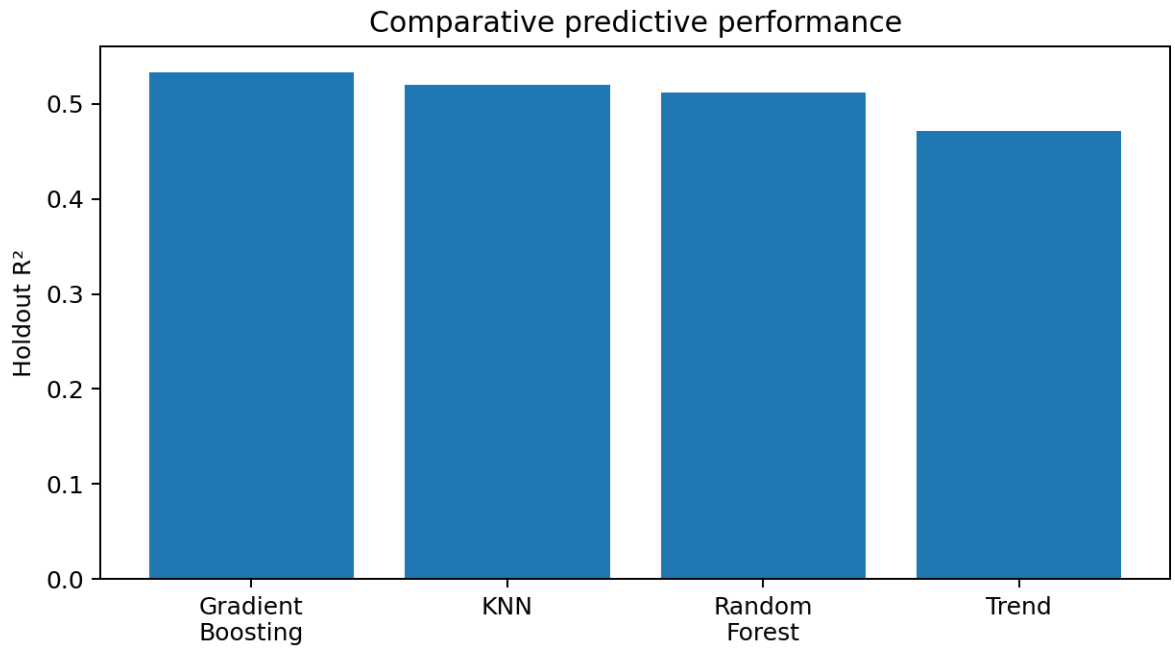


Figure 1. Comparative model R^2 on the 2017–2020 chronological holdout.

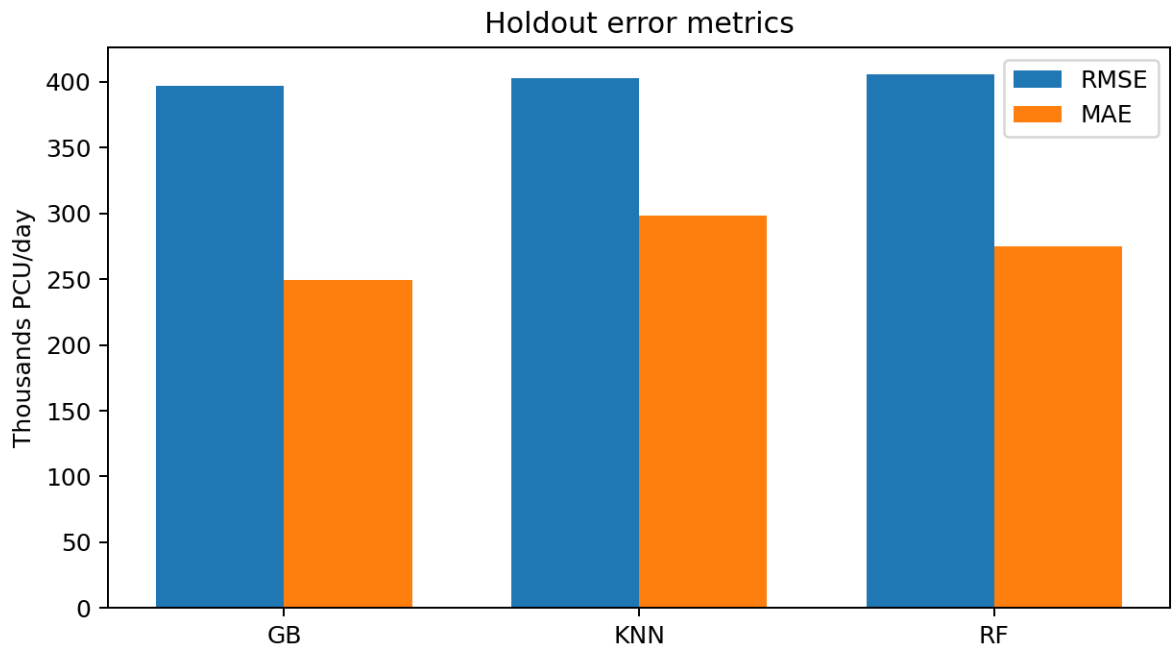


Figure 2. Holdout RMSE and MAE for machine-learning models.

4.2 Random Forest importance

Enugu–Port Harcourt had the largest impurity importance (0.6150; Table 3 and Figure 3), followed by population density (0.1443) and Abakaliki–Ogoja (0.1247). GDP contributed 0.0721, whereas year, pothole volume, and rainfall were comparatively small. Corridor indicators likely encode unmeasured differences in geometry, network position, connectivity, and commercial function.

Table 3. Random Forest impurity-based predictor importance.

Predictor	Importance
Enugu–PHC	0.6150
Density	0.1443
Abakaliki–Ogoja	0.1247
GDP	0.0721
Year	0.0212
Pothole	0.0149
Rainfall	0.0055
Anambra	0.0015
Imo	0.0009

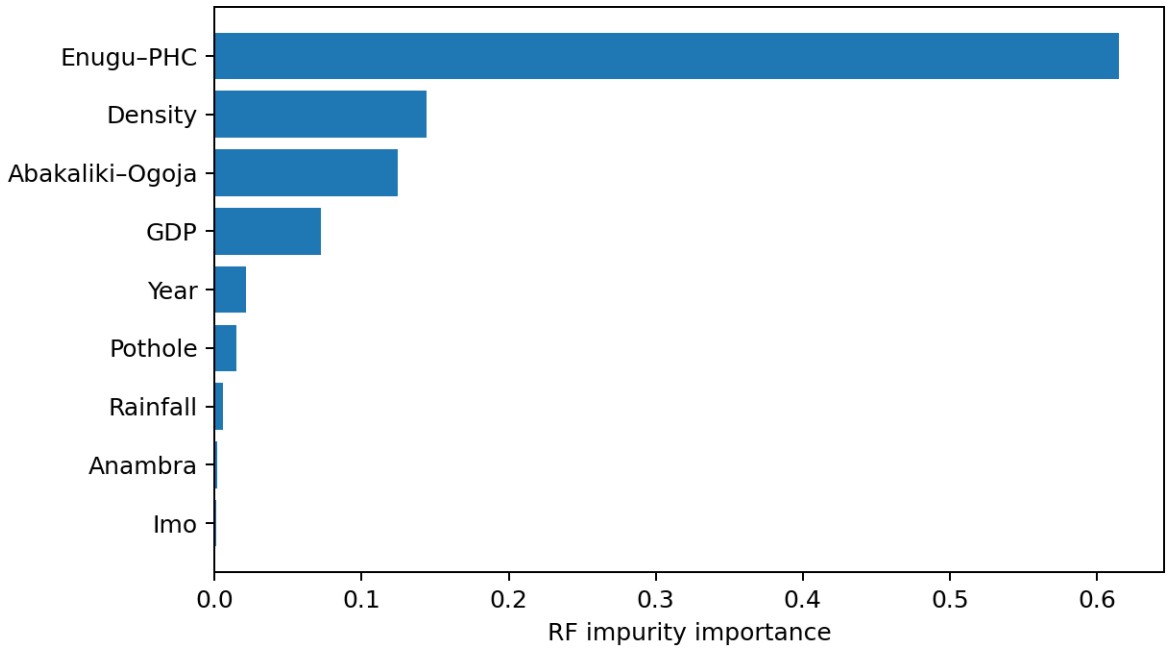


Figure 3. RF impurity-based predictor importance.

4.3 Gradient Boosting permutation importance

Holdout permutation results (Table 4; Figure 4) corroborated corridor identity and population density as leading sources of predictive information. GDP and rainfall scores were small relative to variability; year was zero and pothole volume slightly negative. These results indicate limited unique contribution in this fitted model and test sample, not that these factors are unimportant to highway engineering.

Table 4. Gradient Boosting holdout permutation importance.

Predictor	Mean importance	SD
Corridor/state	133,679.86	71,716.13
Population density	74,081.95	29,443.14
GDP	3,122.10	31,350.49
Rainfall	2,492.21	1,639.21
Year	0	0
Pothole volume	−406.09	2,857.58

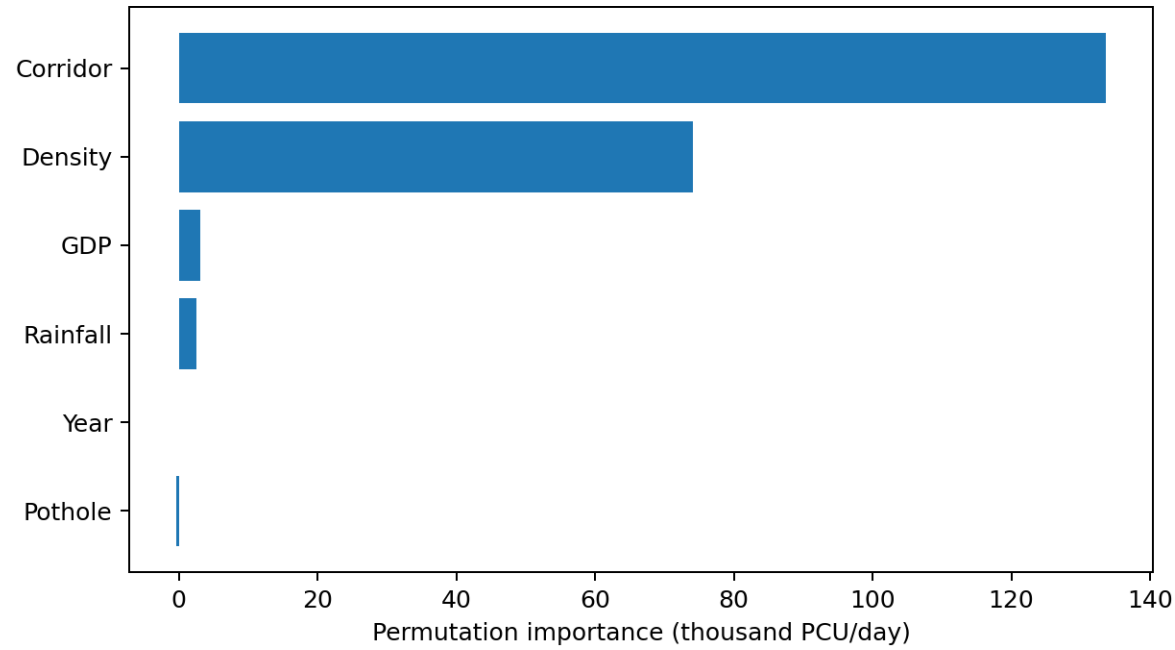


Figure 4. GB holdout permutation importance (mean score).

4.4 Sensitivity analysis

Predictions changed most (Table 5; Figure 5) across the observed population-density range, followed by GDP. Rainfall and pothole volume produced much smaller responses. These are descriptive model sensitivities and not causal effects or intervention elasticities.

Table 5. Sensitivity of mean predicted AADT across observed predictor ranges.

Predictor	Observed range	Change in mean predicted AADT
Population density	1,489,500–21,361,955.9	+522,318 PCU/day
GDP	18,500,000–92,765,000	+342,891 PCU/day
Rainfall	1,820.45–3,308	+2,984 PCU/day
Pothole volume	0.390–2.871	−3,103 PCU/day

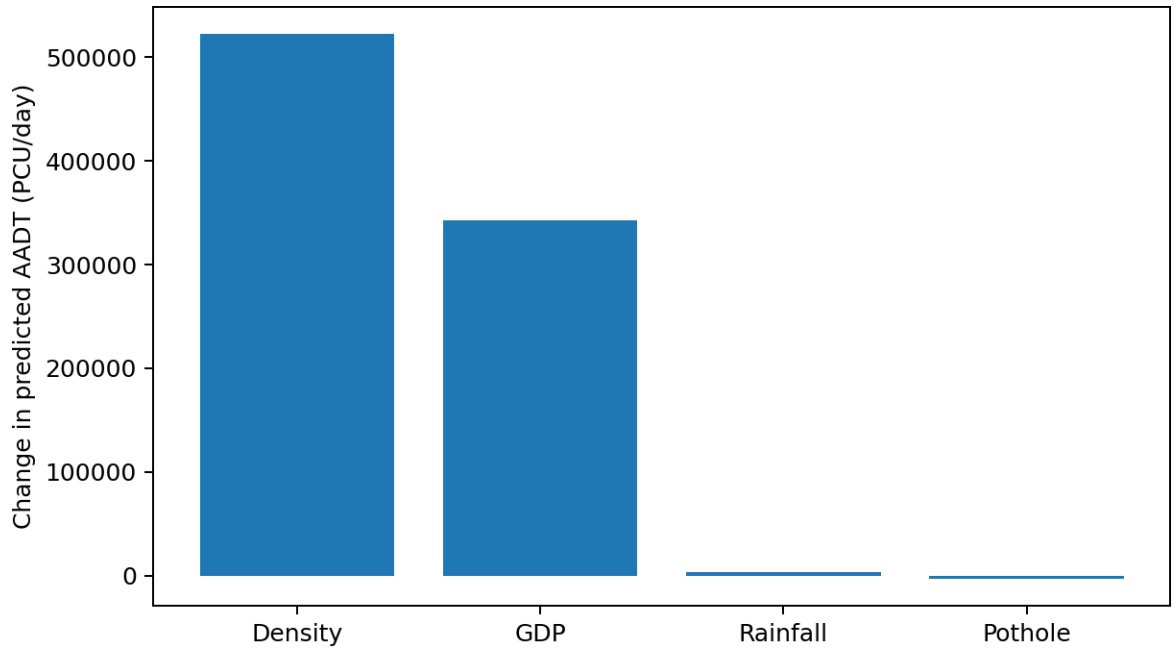


Figure 5. Sensitivity across observed predictor ranges.

4.5 Annual holdout error

Errors remained relatively low during 2017–2019 but increased sharply in 2020 (MAPE = 89.0%; Table 6). This pattern is consistent with an abrupt disruption or structural break that historical relationships could not anticipate, underscoring the need for shock-aware modelling.

Table 6. Gradient Boosting MAPE by holdout year.

Year	GB MAPE
2017	11.2%
2018	7.9%
2019	5.7%
2020	89.0%

5 Discussion

The convergence of RF and GB diagnostics suggests corridor context and population density contain the strongest predictive information. Corridor indicators may absorb persistent omitted attributes such as route connectivity, network position, commercial activity, geometry, and local travel patterns. Their importance reflects predictive representation of corridor heterogeneity, not a causal effect of corridor labels. Population density aligns with AADT studies using demographic and spatial context (Pulugurtha & Mathew, 2021; Ma et al., 2026). GDP showed a positive sensitivity response but modest permutation importance, potentially because its information overlaps with density, corridor identity, or time.

Environmental predictors had comparatively low model importance in this annual dataset. Rainfall and pothole volume may influence travel speed, route choice, vehicle operating costs, and road deterioration, but effects may be indirect, nonlinear, lagged, or obscured by annual aggregation. Low

importance is not proof of no engineering relevance. GB's modest advantage over KNN and RF is indicative given only 20 holdout records. The 2020 failure illustrates how stable-period performance can break under regime change; traffic prediction reviews likewise stress spatial-temporal generalisation and robustness (Shaygan et al., 2022; Du et al., 2025).

6 Engineering and planning implications

Highway agencies should preserve corridor identifiers and collect demographic/economic context alongside traffic counts. Density and corridor characteristics can guide monitoring prioritisation and provide useful inputs where direct counts are sparse. Forecasts used for capacity or investment decisions should include uncertainty and scenarios. The 2020 error supports incorporating disruption indicators, scenario analysis, and periodic recalibration. Future operational models could add vehicle registration, fuel price, rehabilitation/roadworks, accessibility, and higher-frequency traffic observations.

7 Limitations

The five-corridor, 100-record dataset limits power and generalisation. A 20-record holdout yields uncertain performance and importance estimates. Annual aggregation may hide short-term weather and road-condition effects. Corridor indicators proxy for unobserved characteristics; importance and sensitivity are not causal. The 2020 shock is poorly represented by training history. Further work should expand corridors, use rolling-origin validation, add highway-supply/accessibility data, and test uncertainty-aware or regime-switching methods.

8 Conclusion

This study evaluated predictive contributions to AADT across five Southeastern Nigerian highway corridors. GB led chronological holdout performance (R^2 = 0.5338; RMSE = 396,916 PCU/day; MAE = 249,437 PCU/day; MAPE = 28.45%), but performance deteriorated markedly in 2020. RF and GB diagnostics consistently ranked corridor identity and population density highest. Sensitivity analysis showed substantial prediction changes across observed population-density and GDP ranges and comparatively small responses for rainfall and pothole volume. The contribution is a temporally validated and interpretable assessment of predictor contribution in a multicorridor, data-constrained context. Findings can support traffic monitoring and planning, provided predictive importance is not confused with causal influence.

References

[1] Du K, Guo X, Li L, Song J, Shi Q, Hu M, Fang J (2025). Traffic prediction in time series, spatial-temporal, and OD data: A systematic survey. *Journal of Traffic and Transportation Engineering (English Edition)*, 12(3), 666–700. doi:10.1016/j.jtte.2025.03.001

[2] Ma L, Al-Shukairi R, Stettler MEJ, Graham DJ (2026). Estimating annual average daily traffic on local roads: Integrating spatial insights with machine learning. *Journal of Transport Geography*, 134, 104715. doi:10.1016/j.jtrangeo.2026.104715

[3] Pulugurtha SS, Mathew S (2021). Modeling AADT on local functionally classified roads using land use, road density, and nearest nonlocal road data. *Journal of Transport Geography*, 93, 103071. doi:10.1016/j.jtrangeo.2021.103071

[4] Sfyridis A, Agnolucci P (2020). Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *Journal of Transport Geography*, 83, 102658. doi:10.1016/j.jtrangeo.2020.102658

[5] Sfyridis A, Agnolucci P (2023). Factors affecting road traffic: Identifying drivers of annual average daily traffic using least absolute shrinkage and selection operator regression. *Transportation Research Record*, 2677(5), 373–388. doi:10.1177/03611981221141435

[6] Shaygan M, Meese C, Li W, Zhao XG, Nejad M (2022). Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities. *Transportation Research Part C: Emerging Technologies*, 145, 103921. doi:10.1016/j.trc.2022.103921