Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 3, No. 2, [December 2024], pp. [70–82]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Bayesian Optimization of Temporal Features and Gradient Boosting for Short-Term Traffic Flow Prediction

Junhai Wang[1*], Putrama Alkhairi[2], Sankar Kumar Roy[3], Tutut Herawan[4]
[1] *Department of Commerce and Circulation, Zhejiang Technical Institute of Economics, China*
[2] *Informatics Management, STIKOM Tunas Bangsa, Pematangsiantar, North Sumatra, Indonesia*
[3] *Department of Applied Mathematics, Vidyasagar University, Midanpore, India*
[4] *Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia*

[1] *340017@zjtie.edu.cn;* [2] *putrama@amiktunasbangsa.ac.id;* [3] *sankroy2006@gmail.com;* [4] *tutut@um.edu.my*
** Corresponding author*

Abstract

Accurate next-hour traffic forecasting requires temporal predictors that remain valid at prediction time and robust evaluation against strong controls. This study examines Bayesian optimization of lag features and Gradient Boosting for traffic-volume prediction using the UCI Metro Interstate Traffic Volume dataset. After duplicate-hour aggregation and feature construction, 34,669 hourly observations were divided chronologically into training (20,801), development (6,934), and held-out test (6,934) partitions. The experiment compared a seasonal reference, calendar–weather model, fixed Gradient Boosting baseline, Bayesian model-only tuning, random joint tuning, and Bayesian joint lag–model tuning. On the held-out test set, the proposed method achieved RMSE 221.438, MAE 143.215, R^2 0.987247, and sMAPE 5.719%, reducing the baseline RMSE of 224.276 by 1.27%. Its 13 predictors also reduced the baseline feature count of 22 by 40.9%. Bayesian tuning of Gradient Boosting alone attained a slightly lower RMSE of 221.213 and MAE of 143.086, showing that joint optimization did not surpass the strongest control. The modest improvement and divergence between development and test rankings limit claims of general superiority. These findings support careful evaluation of feature parsimony and forecasting error without assuming that a more complex optimization strategy is necessarily better.

Keywords: traffic volume forecasting; Bayesian optimization; temporal feature selection; Gradient Boosting; chronological validation

1. Introduction

Short-term traffic-flow prediction supports congestion management, signal control, and capacity planning in intelligent transportation systems [1]. Traffic demand is influenced by recurring daily and weekly cycles as well as less predictable incidents and weather [2]. A systematic synthesis of traffic-congestion studies identifies forecasting horizon and evaluation consistency as central sources of heterogeneity [3]. Comparisons between classical and deep forecasting methods suggest that model choice depends on information availability and task design [4]. Recent reviews also emphasize that geographical transfer, prediction scale, and reproducibility should not be inferred from benchmark accuracy alone [5]. Complex traffic models capture multidimensional dependencies but can impose substantial computational and data requirements [6]. A broad survey of time-series and spatial traffic prediction distinguishes temporal forecasting from network-wide inference [7]. Recent work on congestion forecasting likewise identifies temporal structure and dataset-specific evaluation as recurring concerns [8]. International-scale traffic estimation research illustrates why transferability must be tested rather than assumed [9]. Multi-source traffic forecasting further shows that the usefulness of external variables depends on when those variables become available [10].

Temporal predictors can encode recent vehicle counts, daily periodicity, and longer seasonal dependencies without requiring a deep architecture. Chronological cross-validation is important because tuning against observations that follow the prediction origin can produce misleading results [11]. Automated hyperparameter search requires a declared search space, evaluation protocol, and comparable computational budget [12]. Gradient-boosted trees offer flexible nonlinear regression and respond differently to loss functions and tuning decisions [13].

Bayesian optimization has been combined with tree-based energy prediction, illustrating the broader relevance of sequential search [14]. A road-traffic study combined Bayesian search with XGBoost, demonstrating that boosted learners remain competitive in traffic applications [15]. Feature-selection research shows that a smaller temporal representation may improve efficiency without guaranteeing improved forecasts [16]. Bayesian tuning has also been applied to recurrent traffic models, although its results do not establish an advantage for every estimator [17] . Road-sensor inference studies further warn that sample dependence and measurement coverage can affect reported generalization [18]. Crowdsourced traffic modeling similarly highlights the sensitivity of prediction to the spatial and measurement context [19].

A useful assessment must separate gains from model-parameter tuning from gains attributable to selecting temporal feature families. Hybrid traffic forecasting has been evaluated using ensembles and decompositions, but the added components require independent comparison [20]. Adaptive time-series forecasting models illustrate the importance of resolving multiple recurring demand patterns [21]. Prediction under detector instability is also a distinct robustness problem that cannot be solved by low average error alone [22]. External-event inputs can affect traffic predictions, provided they are known by the forecast issuance time [23]. Hierarchical temporal representations can improve multi-station forecasting, although a single-site evaluation cannot establish equivalent spatial capability [24]. Multiscale feature learning suggests that temporal scales and model complexity should be assessed jointly [25]. Traffic-event detection demonstrates that feature availability may alter the apparent benefit of an otherwise strong machine-learning model [26]. This study therefore compares fixed Gradient Boosting, Bayesian model-only tuning, random joint temporal-model search, and Bayesian joint search under the same chronological holdout. The primary outcome is held-out RMSE, supplemented by MAE, R^2, sMAPE, WAPE, residual diagnostics, execution time, and predictor count.

2. Method

2.1. Related Work

Recent advances in network-scale forecasting have increased the range of available temporal and spatial representations. Pre-trained spatiotemporal networks aim to learn transferable traffic patterns from extended histories [27]. Time-window partitioning addresses changing correlations between road segments and adjacent periods [28]. Transformer-based spatial models represent heterogeneous traffic connections and long temporal dependencies [29]. Multi-channel models combine signals from different periods and road structures to support more complete forecasts [30]. Signal-aware graph approaches also embed road control conditions into movement-level traffic prediction [31]. Hierarchical multilayer perceptrons offer an alternative to graph-based temporal learning [32]. Dynamic similarity pyramid models show how changing cross-node interactions can be learned across multiple scales [16]. Dilated convolutional approaches address temporal influence and multi-scale spatial interactions in a different model family [33]. These contributions motivate temporally informed features, but their spatial claims are not directly transferable to a single-sensor boosting experiment.

Practical forecasting also requires an appropriate balance between model complexity, feature availability, and measurable error. Knowledge-augmented graph models inject temporal patterns into network-level prediction [34]. Feature-enhanced models have demonstrated the importance of representing correlations across temporal scales [25]. The relation between forecasting accuracy and sensor availability has been tested using historical traffic profiles [22]. Time-aware features should be calculated exclusively from information observed before the forecast target. The validity of chronological splits also depends on whether repeated wall-clock timestamps reflect distinct measurements rather than independent hours. The present setting consequently focuses on a single-site next-hour target instead of making unverified claims about network-wide traffic transfer. Its main methodological question is whether searching the temporal feature family adds predictive benefit beyond tuning Gradient Boosting alone.

Bayesian optimization and ordinary random search are particularly informative controls because both can explore the same candidate space. Search budgets and final evaluation conditions are essential for comparing their observed outcomes [12]. Bayesian boosted-traffic prediction has been investigated using an XGBoost formulation, although its feature and validation design differs from the present experiment [15]. Short-horizon traffic studies using filter-based and genetic feature selection show that representation reduction deserves separate measurement [16]. Temporal cross-validation studies further demonstrate that the best development model need not be best under a later evaluation period [11]. A strong fixed-feature Bayesian comparator therefore isolates the value of tuning the learner itself. The joint search is evaluated through its incremental RMSE, secondary error measures, and number of

retained raw predictors. The comparison is descriptive because a single chronological test set does not independently establish statistical superiority.

Table 1 situates the experiment among recent traffic-forecasting and optimization studies. Its entries focus on temporal representation, boosted-tree tuning, and validation principles. The selected literature demonstrates that prediction architectures vary widely across data resolutions and sensor configurations. It also shows that optimization should be judged against adequate conventional controls. The contribution of the present study is the explicit separation of temporal-feature search from Bayesian model-only tuning. The resulting comparison avoids attributing all observed accuracy changes to a hybrid architecture.

Table 1. Comparative synthesis of selected related studies

Author / year	Contribution	Research-specific limitation
Liu and Shin (2025)	Reviewed statistical, ML, and deep traffic forecasting	No matched test of lag-family Bayesian GB tuning
Lu et al. (2023)	Applied Bayesian optimization to road-traffic XGBoost	Different learner and traffic scenario
Wang et al. (2025)	Examined genetic temporal-spatial feature selection	Does not isolate Bayesian learner-only tuning
Lai et al. (2025)	Evaluated temporal cross-validation for forecasting	Different time series and optimization objective
Bischl et al. (2023)	Formalized hyperparameter-search best practices	General methodological, not traffic-specific
Matue et al. (2026)	Reviewed ML/DL congestion-forecasting methods	Does not establish single-site next-hour superiority

Table 1 differentiates methodological relevance from direct comparability. The cited studies use different traffic horizons, data modalities, or learning algorithms. Bayesian search has credible applications, but its improvement depends on what is tuned and how the search is evaluated. Feature selection may simplify the input representation without ensuring superior test accuracy. Chronological validation remains necessary to prevent future observations from influencing model choice. The experiment therefore examines fixed, model-only, random joint, and Bayesian joint alternatives under one locked evaluation boundary.

2.2. Dataset

The study uses the Metro Interstate Traffic Volume dataset from the UCI Machine Learning Repository. Its target is next-clock-hour traffic volume at the monitored interstate site. Timestamp harmonization aggregated 7,629 duplicate-hour records according to the exported source audit. The resulting modeled series contains 34,669 observations and 22 recorded predictor columns. Inputs include previous traffic counts, lag families, rolling summaries, hour and weekday indicators, holidays, and available weather descriptors. The source timestamps are local wall-clock observations without an unambiguously recorded time zone. Table 2 summarizes the dataset provenance and the constraints on future-target construction.

Table 2. Executed dataset specification and forecasting constraints

Dataset attribute	Study-specific protocol
Public source	UCI Metro Interstate Traffic Volume, dataset 492
Research endpoint	Next clock-hour traffic_volume (regression)
Analysis unit	Hourly observation after duplicate-hour aggregation
Executed partition	Chronological train/dev/test = 20,801 / 6,934 / 6,934
Processing / quality audit	7,629 duplicate-hour records aggregated; 22 predictors; lag missingness audited
Annotation / scope check	Local wall-clock timestamps; verify weather availability at issuance

Table 2 specifies an hourly forecasting task rather than a classification problem. Traffic volume is continuous, and the primary evaluation criterion is RMSE. The duplicate-hour aggregation means that observations were not treated as independently recorded identical clock hours. The feature audit shows some missingness among lag variables, requiring preprocessing within the fitted data boundary. Calendar and weather fields are exogenous descriptors, while lags and rolling summaries are derived from traffic history. Future weather values are not

presumed to be known at forecast time. Uncertainty about daylight-saving interpretation and weather issuance remains a limitation of the available evidence.

The experiment separated model development from final evaluation using chronological rather than shuffled partitions. The first 20,801 records form the training segment, followed by 6,934 development records and 6,934 test records. All baseline and search candidates were assessed before final held-out scoring. Three optimization families each recorded 32 objective evaluations. The candidate list and search scores were retained in the executed notebook outputs. Table 3 describes the main controls applied at each stage. The original external data-loader and optimizer modules were not supplied for independent source-code verification.

Table 3. Executed experimental stages and safeguards

Stage	Required handling
Stage A: audit	Hourly harmonization, source SHA-256, feature and missingness audit
Stage B: split	Chronological 60/20/20: 20,801 train, 6,934 development, 6,934 test
Stage C: fit	Fixed and candidate GB models; 32 evaluations per search method
Stage D: freeze	Development-selected settings; locked later test evaluation
Stage E: report	RMSE, MAE, R^2, sMAPE, WAPE, MedianAE, timing, predictors

Table 3 makes clear that the last 6,934 observations were reserved for the reported test comparison. Candidate selection used the preceding development segment. A fixed seed of 42 was recorded for reproducibility. The six compared methods share the same target and evaluation rows. The exported notebook identifies the protocol as a full-data run rather than a downsampled pilot. The specific internal origin-fold construction is not independently confirmed because the supporting Python modules are missing. A repeated-origin or independent-year study would provide stronger evidence of temporal robustness.

2.3. Research Framework

Figure 1 presents the relationship between the baseline and optimized forecasting pathways. The control uses calendar, weather, and fixed traffic-history variables with a nonsearched Gradient Boosting regressor. The proposed pathway searches the temporal feature family jointly with model hyperparameters. The same chronological split supplies the data for both pathways. The next-clock-hour target is defined after the forecast issuance time. The evaluation includes both regression error and representation size. A separate Bayesian model-only comparator is included to isolate the contribution of feature search.

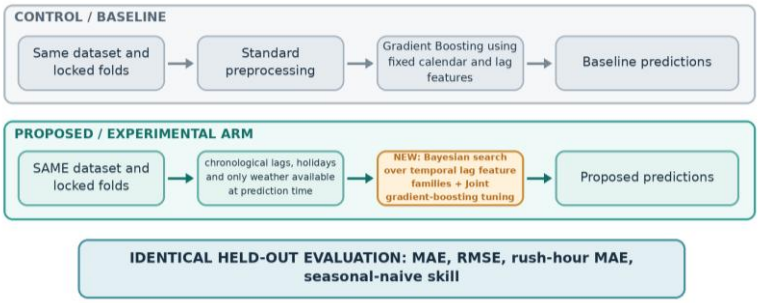


Figure 1. Controlled baseline-versus-proposed research framework

Figure 1 identifies the decision points at which candidate models differ. Ordinary preprocessing and the held-out test data remain common to all comparisons. The temporal-search branch chooses a reduced feature representation whereas the fixed baseline retains 22 columns. Using the same held-out period avoids comparing methods across different traffic regimes. The diagram is a conceptual architecture rather than proof that every proposed nested-validation step was executed. The independent evidence establishes the recorded chronological train/development/test protocol and 32 candidate evaluations per tuned method.

The empirical contrasts are summarized in Table 4. The seasonal reference measures how much a trained regressor improves upon a simple historical traffic rule. The calendar-and-weather model tests a reduced nonlag representation. The fixed Gradient Boosting baseline establishes the principal regression comparator. Bayesian GB-only tuning separates hyperparameter gains from feature-family selection. Random joint search and Bayesian joint search then compare two approaches to the combined tuning problem. All six are assessed on the same final traffic-volume target.

Table 4. Baseline, ablation, and optimization comparisons

Component	Comparative specification
Baseline	Gradient Boosting with 22 fixed temporal/weather predictors
Proposed innovations	Joint Bayesian temporal-feature selection and GB tuning
Controlled factors	Same hourly series, chronological split, next-hour target, test metrics
Required ablations	Naive; calendar/weather; fixed GB; Bayesian GB; random joint; Bayesian joint
Primary comparisons	Test RMSE, MAE, R^2, sMAPE, WAPE, predictor count, runtime

Table 4 lists six alternatives that address different aspects of model design. The naive reference has no fitted regression parameters. The calendar/weather model provides a reduced-information comparison. The fixed baseline retains all 22 recorded inputs. Both joint-search models ultimately use 13 features, while Bayesian model-only tuning retains 22. These contrasts separate predictive quality, reduced dimensionality, and optimization algorithm choice. An improvement over the fixed baseline alone would not establish superiority over the strongest tuned control.

2.4. Proposed Method

Figure 2 illustrates the proposed joint temporal-feature and Gradient Boosting architecture. The first decision block searches eligible lag and calendar feature families. The second block tunes the regressor while assessing a development criterion. The candidate features draw on observed traffic histories and contemporaneous descriptors recorded in the dataset. The trained model outputs a continuous estimate for the next clock hour. The design requires temporal ordering to prevent target information from entering earlier feature vectors. The method is evaluated against the component-only Bayesian alternative rather than solely against the default learner.

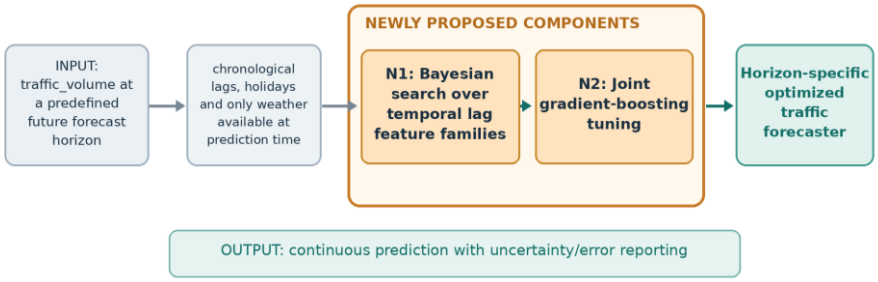


Figure 2. Proposed method architecture with explicitly highlighted novel components

Figure 2 shows the coupling between the temporal representation and model configuration. The conceptual feature search covers lag windows and summary variables. The fitted estimator is Gradient Boosting rather than a neural or probabilistic classifier. The prediction is expressed as a continuous vehicle-count estimate. The executed Bayesian joint search and its random-search comparator each evaluated 32 candidate configurations. The optimization source implementation was unavailable for independent review, limiting verification of the candidate-sampling procedure. The performance claims are consequently restricted to the exported test metrics and reported feature counts.

Table 5 defines the main implementation elements and the observable evaluation criteria. The baseline uses fixed Gradient Boosting settings and the complete predictor representation. Bayesian tuning of Gradient Boosting alone serves as the principal single-component control. Joint random and Bayesian searches explore a smaller temporal representation. Development RMSE is used to compare candidate forecast errors. The final evaluation also reports MAE, R^2, percentage error measures, and execution time. The selected model is not assumed to be superior merely because it has fewer features.

Table 5. Executed model components and mathematical objective

Design element	Operational definition
Standard input branch	Chronologically available lags, rolling means, calendar and weather
Novel block N1	Bayesian search over temporal lag-feature families
Novel block N2	Joint Gradient Boosting hyperparameter tuning
Output model	Next-hour continuous traffic-volume estimate
Optimization / decision objective	Minimize development RMSE for candidate (parameters, feature family)

Design element	Operational definition
Ablation comparison	Fixed GB; Bayesian GB only; random joint; Bayesian joint
Acceptance endpoints	Test RMSE, MAE, R^2, sMAPE, WAPE, predictors and runtime

Table 5 distinguishes the proposed optimization from the basic Gradient Boosting estimator. The model parameters and temporal feature family are selected using development error. The final test records are not used to revise candidate settings. Eight mathematical expressions below define the next-hour target, candidate inputs, the additive estimator, and the selection criterion. The remaining equations specify RMSE, MAE, R^2, sMAPE, and relative RMSE change. These definitions align the manuscript with the regression outcomes reported by the experiment. They do not imply that an unobserved probabilistic calibration step or independent nested experiment was executed.

$$y_{t+1} = v(t+1), \qquad \hat{y}_{t+1} = f_{\theta,W}(\mathbf{x}_t). \tag{1}$$

$$\mathbf{x}_t(W) = \left[\mathbf{c}_t, \mathbf{m}_t, v_t, \{v_{t-\ell}: \ell \in W\}, \overline{v}_{t,W}\right]. \tag{2}$$

$$F_M(\mathbf{x}) = F_0(\mathbf{x}) + \eta \sum_{m=1}^{M} h_m(\mathbf{x}). \tag{3}$$

$$(\theta^*, W^*) = \underset{\theta,W}{\operatorname{argmin}}\ \mathrm{RMSE}_{\mathrm{dev}}(\theta, W). \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \tag{6}$$

$$\mathrm{sMAPE}(\%) = \frac{100}{n}\sum_{i=1}^{n}\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i| + \varepsilon}. \tag{7}$$

$$\Delta\mathrm{RMSE}(\%) = 100\,\frac{\mathrm{RMSE}_{\mathrm{base}} - \mathrm{RMSE}_{\mathrm{method}}}{\mathrm{RMSE}_{\mathrm{base}}}. \tag{8}$$

2.5. Executed Experimental Protocol

The executed analysis used random seed 42 and the entire harmonized hourly series. The candidate search comprised 32 evaluations each for Bayesian GB-only, random joint, and Bayesian joint tuning, giving 96 recorded objectives. The fixed algorithm baseline achieved the best reported development objective in the exported comparison, whereas test outcomes showed small advantages for tuned models. The selected joint Bayesian model retained 13 raw predictors and had a recorded development-search fit time of 166.610 seconds. The matched random joint search retained 13 predictors and required 142.860 seconds, while Bayesian GB-only tuning retained 22 and required 256.234 seconds. Final refitting and evaluation times are reported separately from these search costs. The source modules implementing lag construction and Optuna trial evaluation were unavailable for independent inspection; consequently, a line-by-line leakage audit remains outstanding. The results establish a reproducible evidence trail for this run but do not establish robustness across other stations, daylight-saving regimes, or repeated temporal origins.

3. Results and Discussion

3.1. Data Characteristics and Held-Out Performance

The exploratory summary in Figure 3 describes the distribution and completeness of the modeled traffic series. The target spans low-demand nighttime periods and high-volume peaks. Several lag variables contain missing observations after time harmonization. The underlying distribution is multimodal, consistent with recurring daily traffic conditions. Such heterogeneity motivates time-aware predictors rather than a constant average. The figure does not establish geographic or seasonal generalization beyond this one monitoring site.

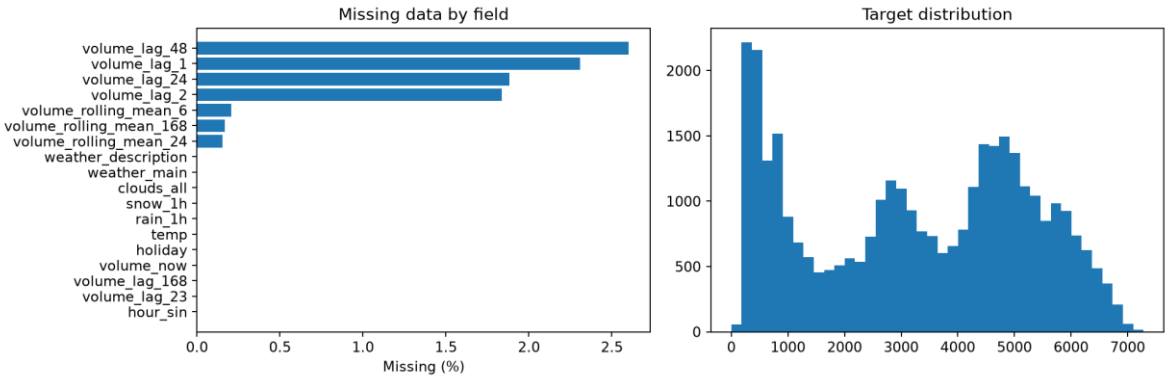


Figure 3. Feature missingness and hourly target-volume distribution

Figure 3 highlights nonuniform traffic demand and missingness among history-derived variables. The target histogram contains distinct low- and high-volume regions rather than a symmetric distribution. These regimes are expected to contribute differently to average squared error. The largest missingness proportions occur in particular lag families, not in every original field. The actual treatment of missing lag entries depends on the external preprocessing module. Evaluation uses the recorded feature matrix, without modifying the observed data distribution shown in the figure.

Table 6 compares the six recorded methods on the identical held-out segment. The seasonal reference is substantially less accurate than all trained approaches. The fixed Gradient Boosting learner establishes the main baseline. Bayesian model-only tuning produces the lowest RMSE, narrowly ahead of the proposed joint pipeline. Random joint search performs between the baseline and proposed method. The table reports predictor counts to distinguish feature compression from predictive accuracy.

Table 6. Baseline versus optimized traffic-volume forecasts on the locked test set

Method	Features	RMSE	MAE	R^2
Seasonal/persistence naive	0	1021.469	565.672	0.728635
Calendar/weather only	13	223.272	145.215	0.987035
Fixed GB baseline	22	224.276	145.960	0.986918
Bayesian GB only	22	221.213	143.086	0.987273
Random joint lag/GB	13	222.686	144.151	0.987103
Proposed Bayesian joint	13	221.438	143.215	0.987247

Table 6 indicates that the proposed Bayesian joint model reduces RMSE from 224.276 to 221.438, an improvement of 1.265%. Its MAE also declines from 145.960 to 143.215 while R^2 increases from 0.986918 to 0.987247. The stronger Bayesian GB-only control achieves RMSE 221.213, outperforming the proposed model by 0.225. Both joint approaches retain 13 predictors instead of the baseline 22, a reduction of 40.9%. The naive model remains far worse, indicating substantial predictive value in the trained approaches. These differences are observed on one held-out period without reported paired confidence intervals or significance tests.

Table 7 adds the complementary measures needed to interpret small differences in RMSE. Median absolute error describes the typical prediction deviation. sMAPE and WAPE express error relative to traffic magnitude. The test-time inference measurement provides a limited indication of computational cost. These quantities should not be interpreted as interchangeable with RMSE. No single indicator establishes that the joint pipeline dominates all alternatives.

Table 7. Secondary error and efficiency measures on the locked test set

Method	MedianAE	sMAPE %	WAPE %	Inf. ms/1k	MAE skill
Seasonal naive	262.000	21.548	17.031	0.000	0.000
Calendar/weather	96.412	5.882	4.372	6.778	0.7433
Fixed GB	96.240	5.938	4.395	9.086	0.7420
Bayesian GB	94.169	5.745	4.308	9.086	0.7471
Random joint	95.112	5.837	4.340	11.249	0.7452
Bayesian joint	94.227	5.719	4.312	9.086	0.7468

Table 7 demonstrates that the proposed model attains the lowest sMAPE at 5.719%, despite not having the lowest RMSE or MAE. Its WAPE of 4.312% is slightly higher than the Bayesian GB-only value of 4.308%. The proposed model also has a median absolute error of 94.227, close to the 94.169 recorded for the strongest control. The displayed test inference rate matches the fixed and Bayesian GB-only models in the execution log. Random joint search has the largest recorded inference time among the fitted alternatives. The ranking changes across metrics, supporting a qualified rather than universal superiority claim.

3.2. Model Predictions and Residual Behavior

Figure 4 summarizes the principal accuracy comparison and includes the naive reference. The Bayesian model-only bar is shortest for RMSE and MAE. The joint Bayesian model follows closely, while random search and fixed baselines remain less accurate. All trained approaches have R^2 near 0.987 on the held-out segment. These values do not establish statistical separation between the two strongest models. Their small differences should be considered alongside feature count and operational reproducibility.

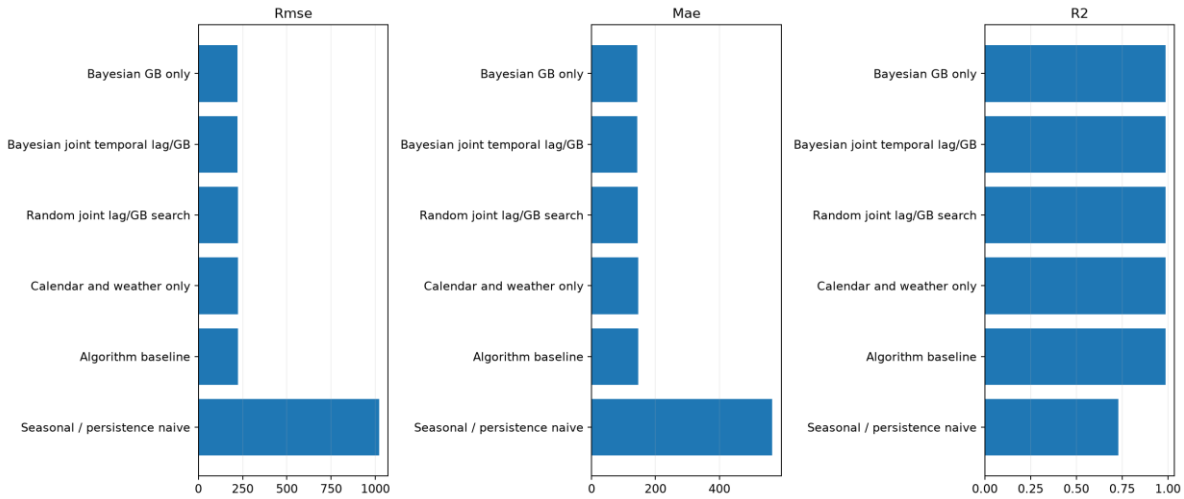


Figure 4. Held-out RMSE, MAE, and R^2 across all model variants

Figure 4 illustrates the dominance of all fitted regressors over the seasonal rule. The proposed joint model improves its primary RMSE relative to the fixed learner. Bayesian GB-only tuning retains a small accuracy advantage over joint feature search. The chart supports measured ranking but not a claim that every possible Bayesian search is superior. Relative error plots are provided separately to reveal whether rankings change under alternate losses. The finding is limited to the reported horizon and test partition.

Figure 5 compares the observed volume with model predictions across all six approaches. The diagonal reference represents perfect agreement. Seasonal predictions disperse more widely than those of the trained regressors. Predictions from the tuned and fixed boosting variants cluster much more closely around the reference line. The visual density reflects the large number of test observations. The figure provides a qualitative diagnostic rather than a substitute for the error metrics.

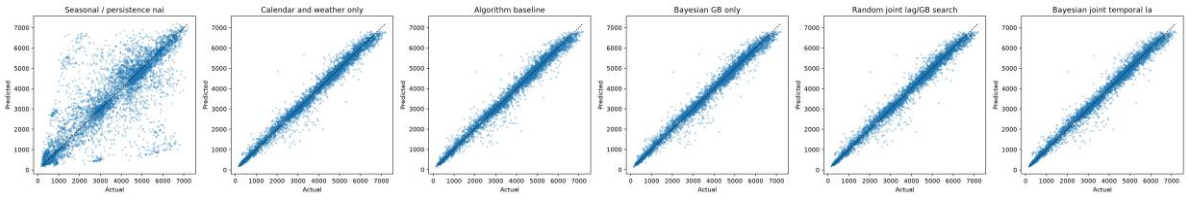


Figure 5. Actual versus predicted hourly volume for the six compared methods

Figure 5 shows strong overall alignment of boosting predictions with recorded traffic counts. Large traffic deviations remain visible around the diagonal for several hours. The seasonal reference displays the widest spread. Subtle differences among tuned models cannot be reliably ranked by visual overlap alone. Table 6 establishes their measured RMSE and MAE ordering. These observations do not determine whether the same ranking would hold at another roadway site. Figure 6 examines the distribution of residuals defined as observed minus predicted volume. The broad seasonal residual profile contrasts with the narrower boosted-model distributions. Residual spread matters

because large errors receive extra weight under RMSE. The plot also exposes tail events that can be hidden by average metrics. Both positive and negative errors remain present after optimization. A centered residual histogram alone does not guarantee calibration or temporal robustness.

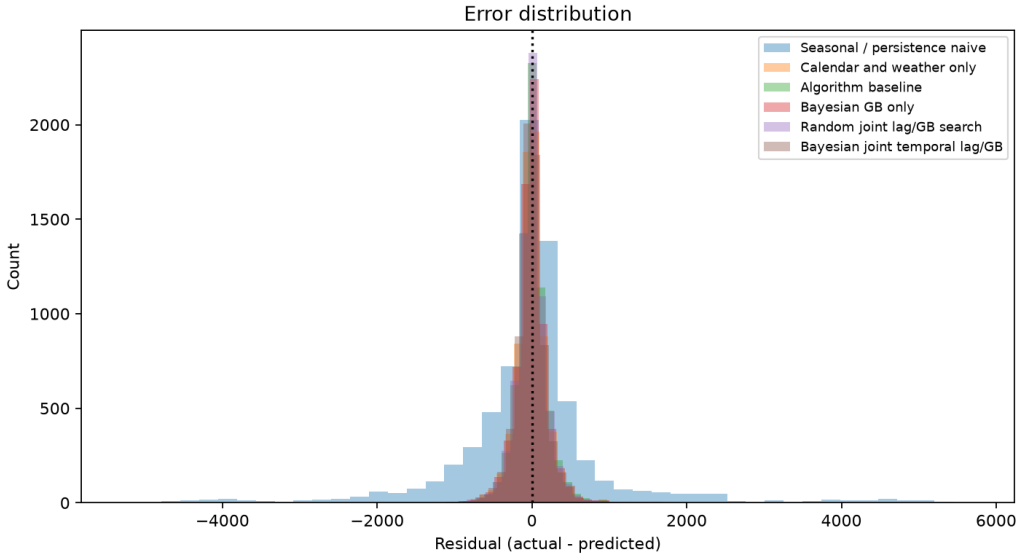


Figure 6. Held-out residual distribution across all forecasting approaches

Figure 6 indicates that trained forecasting models concentrate more residuals near zero than the naive rule. The residual tails reveal occasional substantial misses even when aggregate R^2 is high. The Bayesian joint model does not eliminate these extreme observations. Residual asymmetry may reflect changes in congestion, weather, or traffic regimes, but these causes were not individually identified. A causal interpretation would require additional event and sensor metadata. The analysis therefore supports an error-distribution description rather than an operational failure diagnosis.

Figure 7 displays the time-ordered test forecasts rather than shuffled observations. Daily and weekly peaks create repeating demand cycles across the segment. The fitted regressors generally follow the observed fluctuations. Some sharp transitions produce visible prediction errors. The plot is useful for identifying when poor forecasts occur even if their overall frequency is small. The original local wall-clock timestamps require caution near daylight-saving transitions.

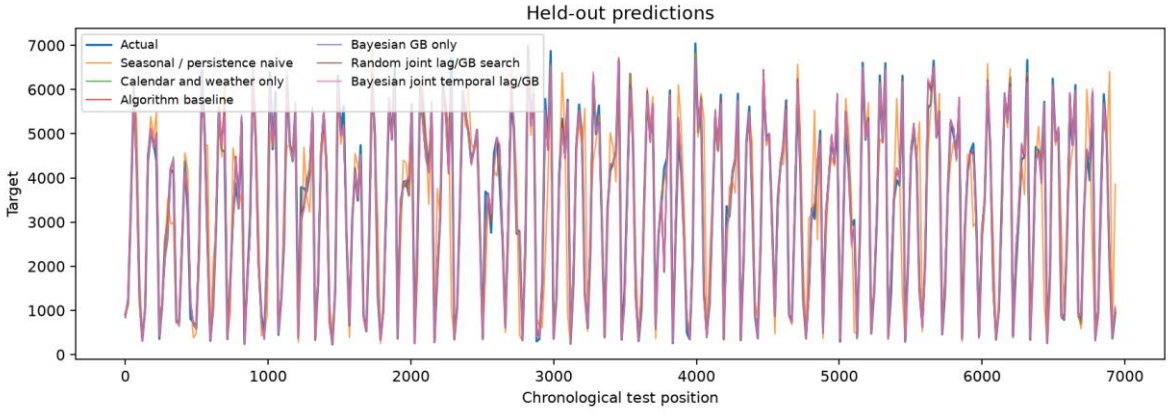


Figure 7. Chronologically ordered actual and predicted test volume

Figure 7 reveals that the trained models track much of the periodic traffic structure. Differences between competing boosting variants are visually small at this plotting scale. Demand shocks and abrupt peaks remain harder to match. The observed test trajectory does not demonstrate robustness to later years or nearby stations. Predictors derived from lagged volume are particularly influential in this task. This figure therefore complements, rather than replaces, time-blocked generalization analysis.

3.3. Optimization, Feature Contribution, and Limitations

Figure 8 groups test observations by the rank of actual traffic volume. The resulting absolute-error traces show how forecasting difficulty changes across the demand distribution. The naive reference exhibits particularly large errors for many observations. Tuned regressors have lower error magnitudes but still display isolated spikes. The figure helps distinguish performance during ordinary and extreme volume regimes. The quantile ordering is descriptive and is not a separate training or evaluation split.

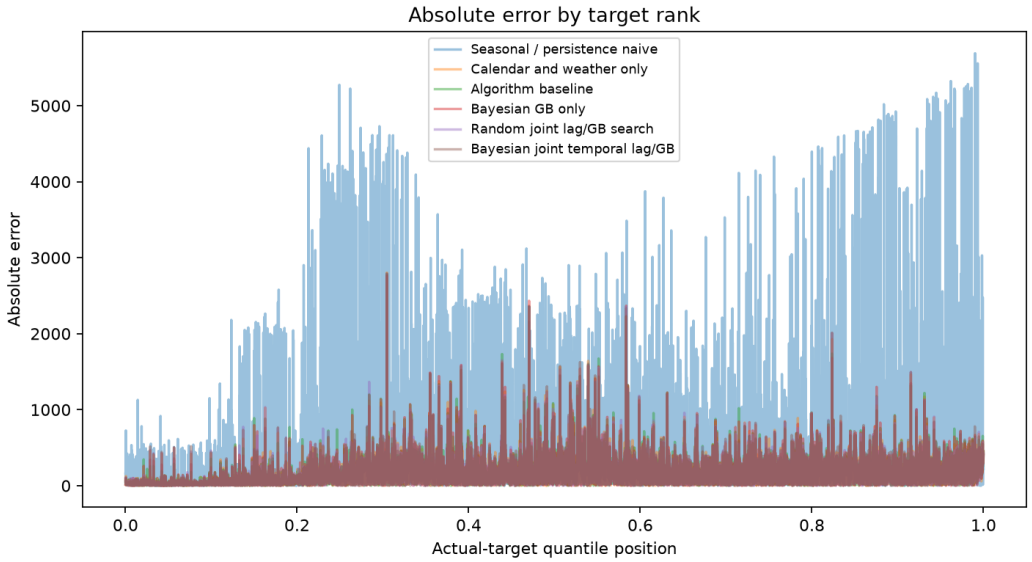


Figure 8. Absolute prediction error across observed volume quantiles

Figure 8 indicates that large errors occur in more than one traffic-volume range. The proposed joint model does not uniformly eliminate these misses. Low average MAE can coexist with several large local deviations. Errors at abrupt transitions may matter to operational control but their specific causes were not measured. No independent incident labels were supplied to attribute those peaks to accidents or closures. Evaluation during identified high-demand or incident intervals remains a useful future test.

Figure 9 displays development objective scores across 96 recorded candidate evaluations. The optimizer trace combines the three 32-evaluation search arms in the displayed run history. Higher negative-objective values indicate lower reported development error. The best-to-date line reflects the selected development candidate at each evaluation index. Search traces reveal candidate exploration but do not independently establish generalization. The final comparison uses later held-out test observations rather than this trace alone.

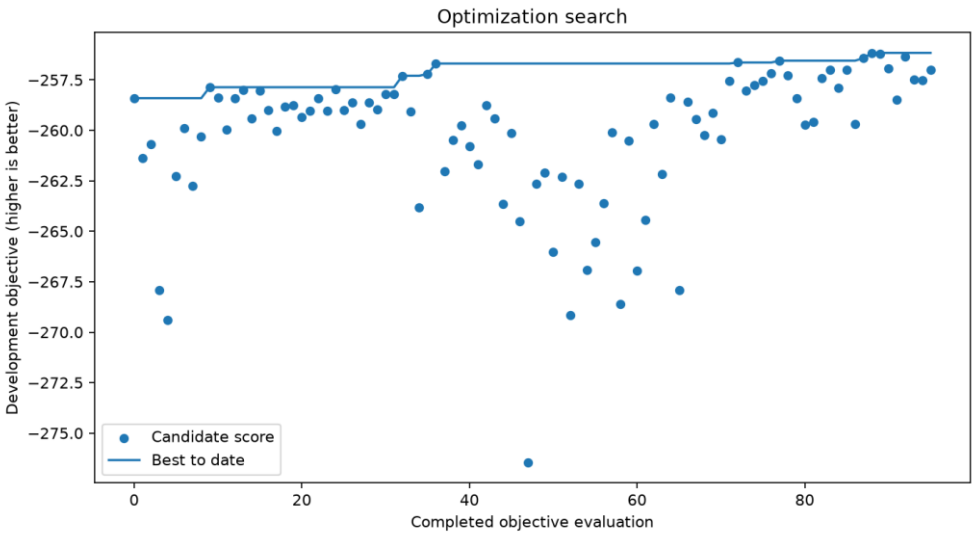


Figure 9. Development objective trace across the recorded optimizer evaluations

Figure 9 shows that search results vary appreciably across candidate configurations. The joint Bayesian branch obtained a development objective of −256.162, while random joint search reached −256.690. The fixed baseline had a better reported development objective of −241.916. This development ranking contrasts with the modest test improvements obtained by tuned models. The discrepancy is consistent with possible sensitivity to temporal segment and optimization protocol, but it is not a formal drift test. The exact search module should be independently examined before attributing this pattern to a specific optimizer mechanism.

Figure 10 summarizes permutation-importance estimates from the fitted traffic model. The current volume field dominates the displayed decrease in predictive quality when permuted. Hourly periodicity variables are the next most influential predictors. Other weather and calendar variables make smaller marginal contributions in this specific diagnostic. The importance computation was performed on a sampled held-out set for plotting. Permutation importance reflects predictive dependence rather than causal influence.

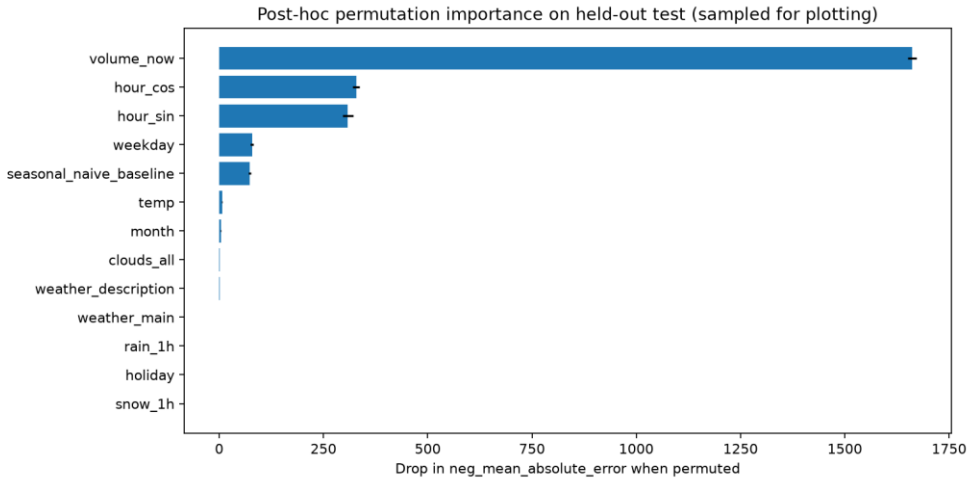


Figure 10. Post hoc permutation-importance diagnostic for temporal predictors

Figure 10 emphasizes the central role of recent observed traffic volume. The strong contribution of the current-volume predictor is consistent with one-hour temporal persistence. Smaller importances of many weather fields do not imply that weather never affects congestion. Correlations among lags and calendar variables may redistribute the apparent contributions. The diagnostic does not prove which variables remain available under real-time deployment constraints. Future studies should verify data availability and stable feature rankings across independent time periods.

The comparative results identify model-only Bayesian tuning as the strongest observed accuracy control. Its RMSE advantage over the joint method is small but directionally unfavorable to the central incremental-optimization hypothesis. The joint approach nevertheless compresses the raw predictor representation by 40.9% while improving upon the fixed baseline. Development and test rankings differ, so optimization stability cannot be inferred from the single recorded split. Local wall-clock ambiguity, weather issuance timing, and unavailable supporting source modules restrict independent leakage assessment. The experiment does not include paired uncertainty intervals or external validation across roads, years, and forecasting horizons. Accordingly, the findings support a limited performance–parsimony trade-off, not a proven advantage of joint Bayesian temporal optimization.

4. Conclusion

Bayesian joint optimization of temporal features and Gradient Boosting was evaluated for next-hour traffic-volume prediction on 34,669 hourly observations. The approach achieved held-out RMSE 221.438 and MAE 143.215, improving upon the fixed Gradient Boosting baseline of 224.276 RMSE and 145.960 MAE. The joint model used 13 predictors instead of 22, reducing its raw input dimensionality by 40.9%. Bayesian tuning of Gradient Boosting alone nevertheless achieved the lowest RMSE of 221.213 and MAE of 143.086. The primary hypothesis of superiority over the strongest tuned control was therefore not supported by this evaluation. The modest gains, development–test ranking discrepancy, and single-site temporal data constrain generalization. Future work should verify the complete search implementation, predictor issuance times, repeated chronological origins, and out-of-site performance before considering operational deployment.

References

[1] R. Liu and S.-Y. Shin, "A Review of Traffic Flow Prediction Methods in Intelligent Transportation System Construction," *Applied Sciences*, vol. 15, no. 7, 2025, doi: 10.3390/app15073866.

[2] A. Mystakidis, P. Koukaras, and C. Tjortjis, "Advances in Traffic Congestion Prediction: An Overview of Emerging Techniques and Methods," *Smart Cities*, vol. 8, no. 1, 2025, doi: 10.3390/smartcities8010025.

[3] M. Attioui and M. Lahby, "A Systematic Literature Review of Traffic Congestion Forecasting: From Machine Learning Techniques to Large Language Models," *Vehicles*, vol. 7, no. 4, 2025, doi: 10.3390/vehicles7040142.

[4] S. Afandizadeh, S. Abdolahi, and H. Mirzahossein, "Deep Learning Algorithms for Traffic Forecasting: A Comprehensive Review and Comparison with Classical Ones," *J. Adv. Transp.*, vol. 2024, 2024, doi: 10.1155/2024/9981657.

[5] Y. García-Ramírez, "Artificial Intelligence and Machine Learning for Road Traffic Congestion Prediction and Forecasting: A Systematic Review of Methods, Validation, Explainability, and Reproducibility," *Encyclopedia*, vol. 6, no. 9, 2026, doi: 10.3390/encyclopedia6090205.

[6] A. Ali, A. Sharafian, H. M. Y. Naeem, M. Zakarya, Z. Wu, and X. Bai, "Advanced computational models for urban traffic flow prediction: A comprehensive review and future directions," *Comput. Sci. Rev.*, vol. 60, 2026, doi: 10.1016/j.cosrev.2025.100886.

[7] K. Du *et al.*, "Traffic prediction in time series, spatialtemporal, and OD data: A systematic survey," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 12, no. 3, pp. 666–700, 2025, doi: 10.1016/j.jtte.2025.03.001.

[8] T. Matue, A. A. Akinyelu, and M. Mokotsolane, "Advances in intelligent transportation systems: ML and DL techniques for traffic congestion forecasting," *Int. J. Data Sci. Anal.*, vol. 22, 2026, doi: 10.1007/s41060-026-01203-9.

[9] A. Sheehan, C. Jephcote, and J. Gulliver, "Development and evaluation of statistical models for international-scale road traffic flow estimation," *Environ. Int.*, vol. 210, 2026, doi: 10.1016/j.envint.2026.110231.

[10] M. Leinonen, A. Al-Tachmeesschi, B. Turkmen, N. Atashi, and L. Ruotsalainen, "Long Short Term Memory Based Traffic Prediction Using Multi-Source Data," *International Journal of Intelligent Transportation Systems Research*, vol. 23, pp. 354–371, 2025, doi: 10.1007/s13177-024-00451-y.

[11] M. Lai, S. S. Wulff, Y. Cao, T. J. Robinson, and R. Rajapaksha, "Temporal cross-validation in forecasting: A case study of COVID-19 incidence using wastewater data," *Qual. Reliab. Eng. Int.*, 2025, doi: 10.1002/qre.3686.

[12] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, 2023, doi: 10.1002/widm.1484.

[13] A. Hjort, J. Pensar, I. Scheel, and D. E. Sommervoll, "House price prediction with gradient boosted trees under different loss functions," *Journal of Property Research*, vol. 39, no. 4, pp. 338–364, 2022, doi: 10.1080/09599916.2022.2070525.

[14] J. Pan, X. Guo, H. Fu, Y. Zhang, and J. Zhang, "Machine learning prediction of building energy consumption by combining Bayesian optimization and the SHAP model," *Int. J. Green Energy*, vol. 23, no. 8, pp. 1530–1545, 2026, doi: 10.1080/15435075.2026.2619448.

[15] X. Lu, C. Chen, R. Gao, and Z. Xing, "Prediction of High-Speed Traffic Flow around City Based on BO-XGBoost Model," *Symmetry (Basel).*, vol. 15, no. 7, 2023, doi: 10.3390/sym15071453.

[16] B. Wang, P. Qian, Y. Cheng, Y. Qian, and J. Zhang, "A short-term traffic flow prediction network using filter-genetic feature selection method," *Urban Lifeline*, vol. 3, 2025, doi: 10.1007/s44285-025-00049-0.

[17] H. M. Khan, A. Khan, M. I. Majid, S. U. Haq, and M. H. Zafar, "A Bayesian-Optimized Long Short-Term Memory Deep Learning Model for Traffic Prediction in Intelligent Transportation," *Arab. J. Sci. Eng.*, 2026, doi: 10.1007/s13369-026-11437-1.

[18] J. Klingwort and J. Burger, "A framework for population inference: Combining machine learning, network analysis, and non-probability road sensor data," *Comput. Environ. Urban Syst.*, vol. 103, 2023, doi: 10.1016/j.compenvurbsys.2023.101976.

[19] L. Tay, J. M.-Y. Lim, S.-N. Liang, K. K. Chua, and Y. H. Tay, "Urban traffic volume estimation using intelligent transportation system crowdsourced data," *Eng. Appl. Artif. Intell.*, vol. 126, 2023, doi: 10.1016/j.engappai.2023.107064.

[20] G. Dai, J. Tang, and W. Luo, "Short-term traffic flow prediction: An ensemble machine learning approach," *Alexandria Engineering Journal*, vol. 74, pp. 467–480, 2023, doi: 10.1016/j.aej.2023.05.015.

[21] Q. Shao *et al.*, "An adaptive composite time series forecasting model for short-term traffic flow," *J. Big Data*, vol. 11, 2024, doi: 10.1186/s40537-024-00967-w.

[22] X. Zou, E. Chung, H. Ye, and H. Zhang, "Deep Learning for Traffic Prediction and Trend Deviation Identification: A Case Study in Hong Kong," *Data Science for Transportation*, vol. 6, 2024, doi: 10.1007/s42421-024-00112-2.

[23] Z. Qiu, Z. Xie, Z. Ji, X. Liu, and G. Wang, "Integrating query data for enhanced traffic forecasting: A Spatio-Temporal Graph Attention Convolution Network approach with delay modeling," *Knowl. Based. Syst.*, vol. 301, 2024, doi: 10.1016/j.knosys.2024.112315.

[24] J. Lou, X. Zhang, R. Wang, Z. Liu, K. Zhao, and Q. Shen, "HSFE: A hierarchical spatial-temporal feature enhanced framework for traffic flow forecasting," *Inf. Sci. (N. Y).*, vol. 679, 2024, doi: 10.1016/j.ins.2024.121070.

[25] S. Du, T. Yang, F. Teng, J. Zhang, T. Li, and Y. Zheng, "Multi-scale feature enhanced spatio-temporal learning for traffic flow forecasting," *Knowl. Based. Syst.*, vol. 294, 2024, doi: 10.1016/j.knosys.2024.111787.

[26] P. Moriano, A. Berres, H. Xu, and J. Sanyal, "Spatiotemporal features of traffic help reduce automatic accident detection time," *Expert Syst. Appl.*, vol. 244, 2024, doi: 10.1016/j.eswa.2023.122813.

[27] S. Fang, W. Ji, S. Xiang, and W. Hua, "PreSTNet: Pre-trained Spatio-Temporal Network for traffic forecasting," *Information Fusion*, vol. 106, 2024, doi: 10.1016/j.inffus.2024.102241.

[28] D. Yu *et al.*, "Modeling dynamic spatio-temporal correlations and transitions with time window partitioning for traffic flow prediction," *Expert Syst. Appl.*, vol. 252, 2024, doi: 10.1016/j.eswa.2024.124187.

[29] G. Shi, L. Luo, Y. Song, J. Li, and S. Pan, "Deep transformer-based heterogeneous spatiotemporal graph learning for geographical traffic forecasting," *iScience*, vol. 27, no. 7, 2024, doi: 10.1016/j.isci.2024.110175.

[30] J. Xiao and B. Long, "A multi-channel spatial-temporal transformer model for traffic flow forecasting," *Inf. Sci. (N. Y).*, vol. 671, 2024, doi: 10.1016/j.ins.2024.120648.

[31] M. Xu, T. Z. Qiu, J. Fang, H. He, and H. Chen, "Signal-control refined dynamic traffic graph model for movement-based arterial network traffic volume prediction," *Expert Syst. Appl.*, vol. 228, 2023, doi: 10.1016/j.eswa.2023.120393.

[32] Y. Qin, H. Luo, F. Zhao, Y. Fang, X. Tao, and C. Wang, "Spatio-temporal hierarchical MLP network for traffic forecasting," *Inf. Sci. (N. Y).*, vol. 632, pp. 543–554, 2023, doi: 10.1016/j.ins.2023.03.063.

[33] W. Li *et al.*, "Enhancing traffic flow prediction through multi-view attention mechanism and dilated convolutional networks," *Complex & Intelligent Systems*, vol. 12, 2026, doi: 10.1007/s40747-025-02146-7.

[34] Y. Wang, J. Hu, F. Teng, L. Peng, S. Du, and T. Li, "KaTaGCN: Knowledge-Augmented and Time-Aware Graph Convolutional Network for efficient traffic forecasting," *Information Fusion*, vol. 111, 2024, doi: 10.1016/j.inffus.2024.102542.